

Pattern Recognition (Elective VI)

CS 745

Professional Elective Course

4 Credits : 4:0:0

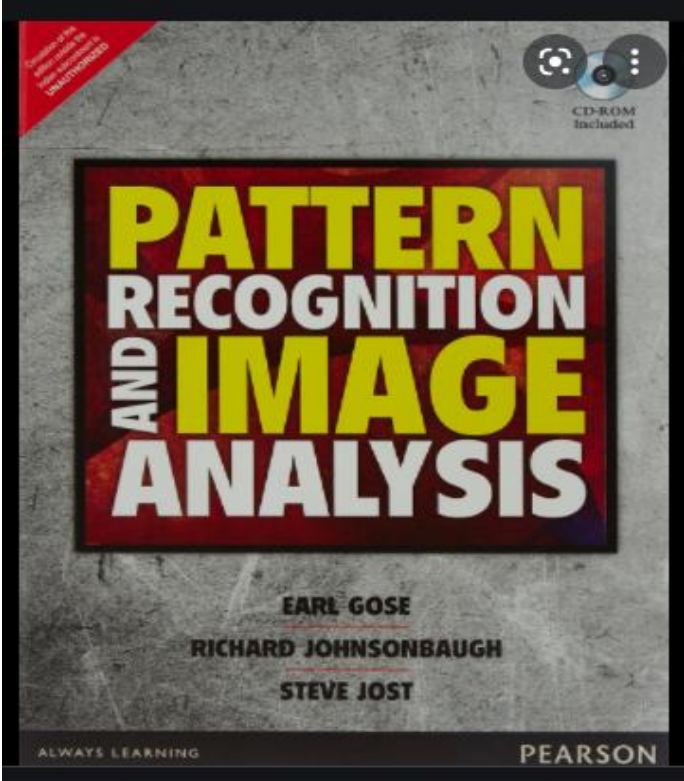
Course Syllabus for 5 Units

Unit No.	Course Content	No. of Hours
1.	Introduction: Applications of Pattern Recognition, Statistical Decision Theory and Analysis. Probability: Introduction to probability, Probabilities of Events, Random Variables, Joint Distributions and Densities, Moments of Random Variables, Estimation of Parameters from samples	11
2.	Statistical Decision Making: Introduction, Bayes' Theorem, Conditionally Independent Features, Decision Boundaries.	10
3	Nonparametric Decision Making: Introduction, Histograms, kernel and Window Estimators, Nearest Neighbour Classification Techniques: K Nearest neighbour algorithm, Adaptive Decision Boundaries, Minimum Squared Error Discriminant Functions, Choosing a decision-making technique.	11
4.	Clustering: Introduction, Hierarchical Clustering, Agglomerative clustering algorithm, The single linkage algorithm, The complete linkage algorithm, Partitional Clustering: Forgy's algorithm, The K-Means algorithm	10
5.	Dimensionality Reduction: Singular Value Decomposition, Principal Component Analysis, Linear Discriminated Analysis.	10

Course Outcomes

CO-1	Estimating Parameters from Samples.
CO-2	Classify Patterns using Parametric and Non-Parametric Techniques.
CO-3	Clustering of Samples using different Clustering Algorithms.
CO-4	Apply various Dimensionality Reduction Techniques to reduce the Dimension.

Text Books:



+ Text Books:

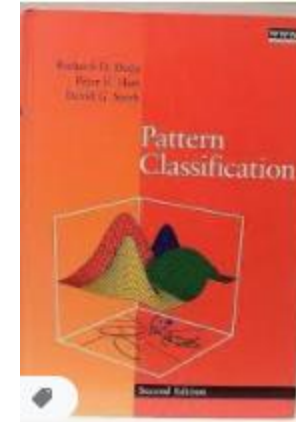
Sl. No.	Author/s	Title	Publisher Details
1	Earl Gose, Richard Johnsonbaugh, Steve Jost.	Pattern recognition and Image analysis	Pearson 2015



Reference Books and Web sources

Reference Books:

Sl. No.	Author/s	Title	Publisher Details
1	<u>Richard O.Duda</u> , <u>Peter E.Hart</u> , <u>David G. Stork</u>	Pattern Classification	John Wiley publication, 2nd edition, 2001.
2	<u>A.K.Jain</u> , <u>R.Bolle</u> , <u>S.Pankanti</u>	Biometric: Personal Identification in network society	Kluwer academic publishers, 1999.
3	<u>Robert Schalkoff</u>	Pattern Recognition: Statistical, Structural and Neural Approaches	John Wiley & Sons, Inc.1992.
4	Christopher M. Bishop	Pattern Recognition and Machine Learning	Springer publication, 2006



Web Resources:

Sl. No.	Web link
1	https://nptel.ac.in/courses/106/108/106108057/
2	https://nptel.ac.in/courses/106/106/106106046/

Course Tutor Details:

Dr. Srinath. S

Associate Professor

Dept. of Computer Science and Engineering

SJCE, JSS S&TU

Mysuru- 6

Email: srinath@sjce.ac.in

Mobile: 9844823201

CO vs PO and PSO mapping

Course Outcomes	Program Outcomes												PSO's			
	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3	PSO4
CO-1	3	3	2	3	1	0	0	0	1	0	0	2	3	3	2	3
CO-2	3	3	3	3	3	0	0	0	1	0	1	2	3	3	2	3
CO-3	3	3	3	3	3	1	0	0	0	0	1	2	3	3	2	3
CO-4	3	3	3	3	3	1	0	0	0	0	1	2	3	3	2	3

0 -- No association 1---Low association, 2--- Moderate association, 3---High association

UNIT - 1

- Introduction: Applications of Pattern Recognition, Statistical Decision Theory and Analysis. Probability: Introduction to probability, Probabilities of Events, Random Variables, Joint Distributions and Densities, Moments of Random Variables, Estimation of Parameters from samples

Introduction: Definition

- **Pattern recognition** is the theory or algorithm concerned with the automatic detection (recognition) and later classification of objects or events using a machine/computer.

- **Applications of Pattern Recognition**

- Some examples of the problems to which pattern recognition techniques have been applied are:

- Automatic inspection of parts of an assembly line
- Human speech recognition
- Character recognition
- Automatic grading of plywood, steel, and other sheet material
- Identification of people from
 - finger prints,
 - hand shape and size,
 - Retinal scans
 - voice characteristics,
 - Typing patterns and
 - handwriting
- Automatic inspection of printed circuits and printed characters
- Automatic analysis of satellite picture to determine the type and condition of agricultural crops, weather conditions, snow and water reserves and mineral prospects.
- Classification and analysis in medical images. : to detect a disease

Features and classes

- **Properties** or **attributes** used to classify the objects are called **features**.
- A collection of “**similar**” (**not necessarily same**) objects are grouped together as **one “class”**.

- For example:



- All the above are classified as character T
- Classes are identified by a **label**.
- Most of the pattern recognition tasks are first done by humans and automated later.
- Automating the classification of objects using the same features as those used by the people can be difficult.
- Some times features that would be impossible or difficult for humans to estimate are useful in automated system. For example satellite images use wavelengths of light that are invisible to humans.

Two broad types of classification

- **Supervised classification**

- **Guided by the humans**

- It is called supervised learning because **the process of learning from the training dataset can be thought of as a teacher supervising the learning process.**

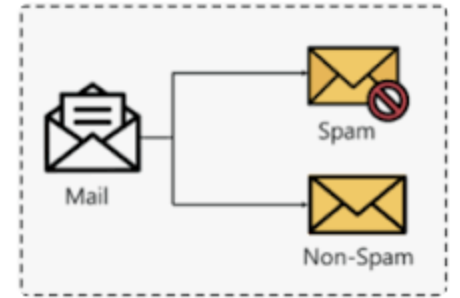
- We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher.

- Classify the mails as span or non span based on redecided parameters.

- **Unsupervised classification**

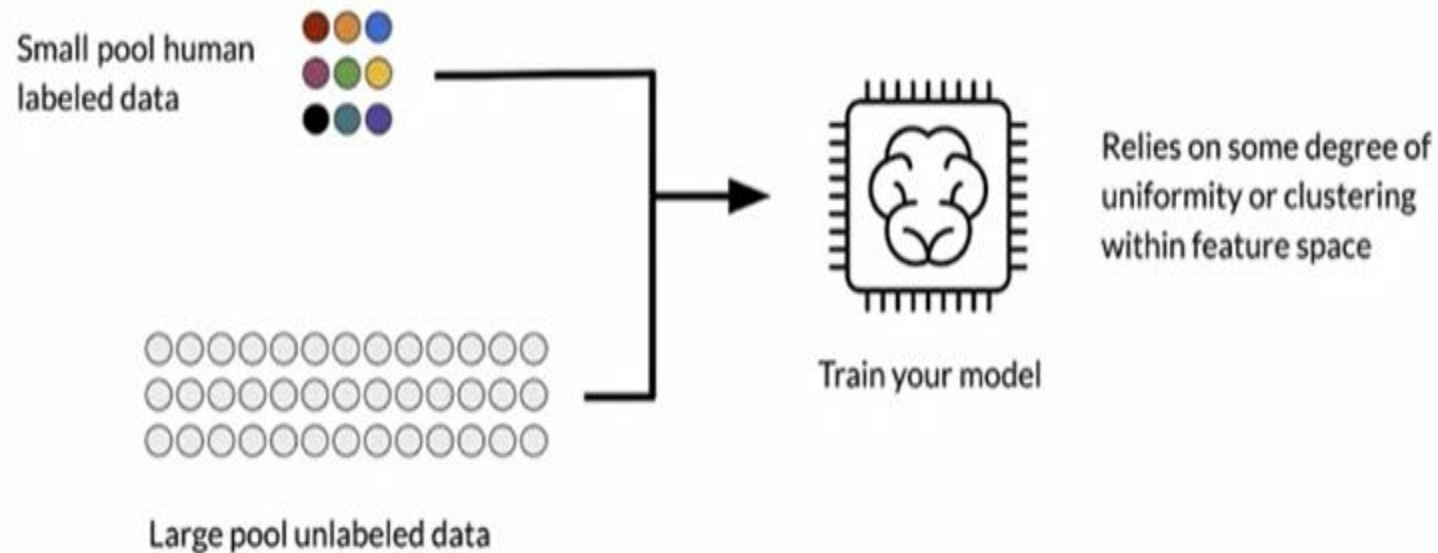
- Not guided by the humans.

- Unsupervised Classification is called **clustering.**



Another classifier : Semi supervised learning

It makes use of a small number of labeled data and a large number of unlabeled data to learn



Samples or patterns

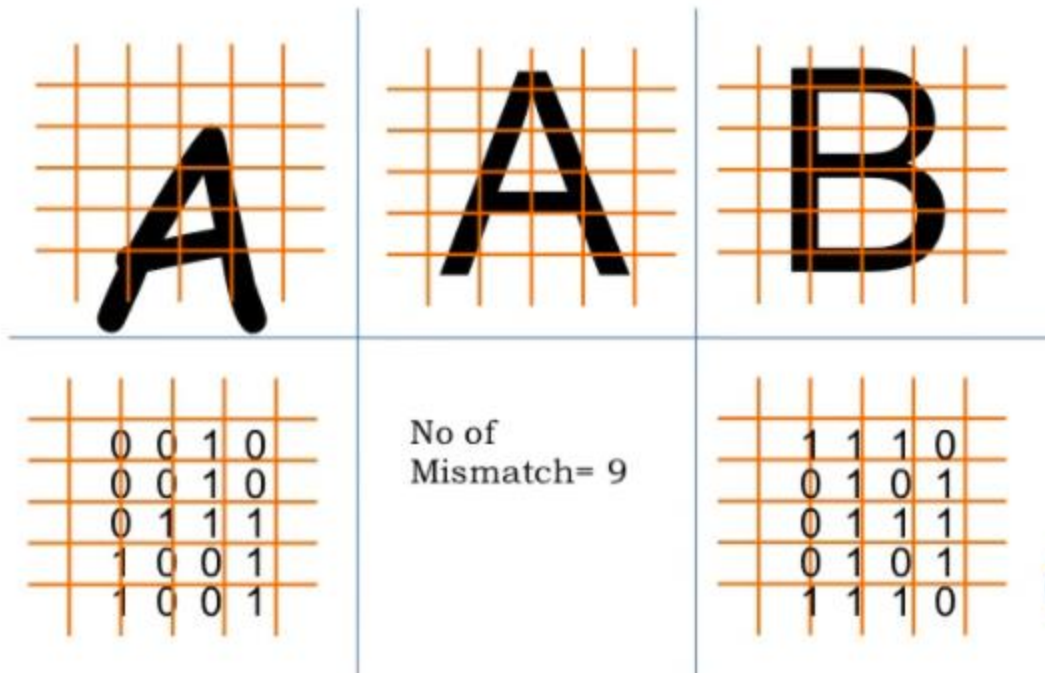
- The individual items or objects or situations to be classified will be referred as **samples or patterns or data**.
- **The set of data is called “Data Set”.**

Training and Testing data

- Two types of data set **in supervised classifier**.
 - **Training set** : 70 to 80% of the available data will be used for training the system.
 - In Supervised classification Training data is **the data you use to train an algorithm or machine learning model to predict the outcome you design your model to predict**.
 - **Testing set** : around 20-30% will be used for testing the system. Test data is used to measure the performance, such as accuracy or efficiency, of the algorithm you are using to train the machine.
 - Testing is the measure of quality of your algorithm.
 - Many a times even after 80% testing, failures can be see during testing, reason being not good representation of the test data in the training set.
- **Unsupervised classifier does not use training data**

Statistical Decision Theory

- Decision theory, in statistics, **a set of quantitative methods for reaching optimal decisions.**



Example for Statistical Decision Theory

- Consider Hypothetical Basket ball Association:
- The prediction could be based on the difference between the home team's average number of points per game (apg) and the visiting team's 'apg' for previous games.
- The training set consists of scores of previously played games, with each home team is classified as **winner or loser**
- Now the prediction problem is : given a game to be played, predict the home team to be a winner or loser using the feature 'dapg',
- Where **$dapg = \text{Home team apg} - \text{Visiting team apg}$**

Game	<i>dapg</i>	Home Team	Game	<i>dapg</i>	Home Team
1	1.3	Won	16	-3.1	Won
2	-2.7	Lost	17	1.7	Won
3	-0.5	Won	18	2.8	Won
4	-3.2	Lost	19	4.6	Won
5	2.3	Won	20	3.0	Won
6	5.1	Won	21	0.7	Lost
7	-5.4	Lost	22	10.1	Won
8	8.2	Won	23	2.5	Won
9	-10.8	Lost	24	0.8	Won
10	-0.4	Won	25	-5.0	Lost
11	10.5	Won	26	8.1	Won
12	-1.1	Lost	27	-7.1	Lost
13	2.5	Won	28	2.7	Won
14	-4.2	Won	29	-10.0	Lost
15	-3.4	Lost	30	-6.5	Won

Data set of games showing outcomes, differences between average numbers of points scored and differences between winning percentages for the participating teams in previous games

- The figure shown in the previous slide, lists 30 games and gives the value of d_{apg} for each game and tells whether the home team won or lost.
- Notice that in this data set the team with the larger apg usually wins.
- For example in the 9th game the home team on average, scored 10.8 fewer points in previous games than the visiting team, on average and also the home team lost.
- When the teams have about the same apg 's the outcome is less certain. For example, in the 10th game, the home team on average scored 0.4 fewer points than the visiting team, on average, but the home team won the match.
- Similarly 12th game, the home team had an apg 1.1. less than the visiting team on average and the team lost.

Histogram of dapg

- Histogram is a convenient way to describe the data.
- To form a histogram, the data from a single class are grouped into intervals.
- Over each interval rectangle is drawn, with height proportional to number of data points falling in that interval. In the example interval is chosen to have width of two units.
- General observation is that, the prediction is not accurate with single feature 'dgpa'

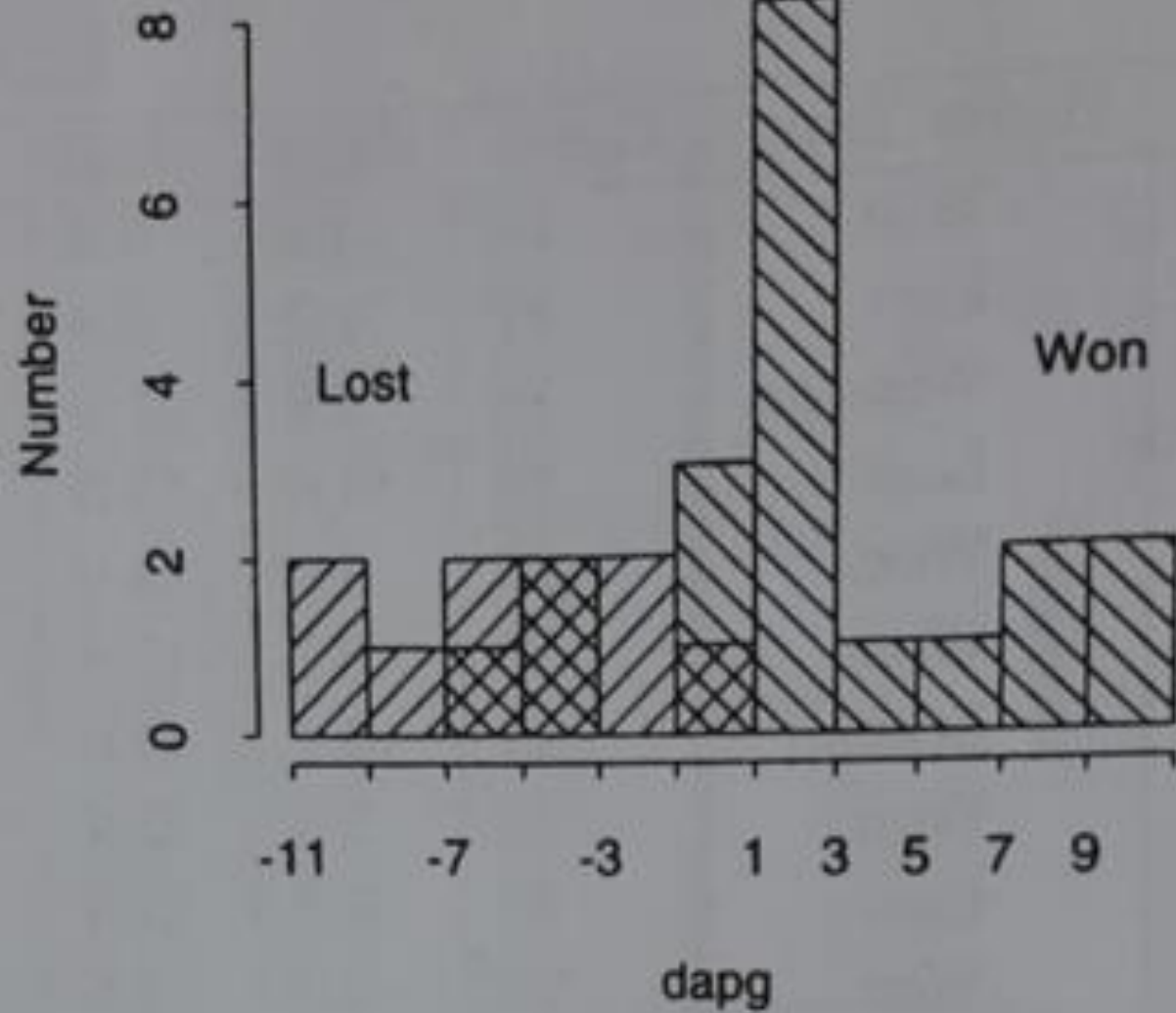


Figure 1.2: Histogram of *dapg*.

Prediction

- To predict normally a threshold value T is used.
- ' $dgpa$ ' $> T$ consider to be won
- ' $dgpa$ ' $< T$ consider to be lost

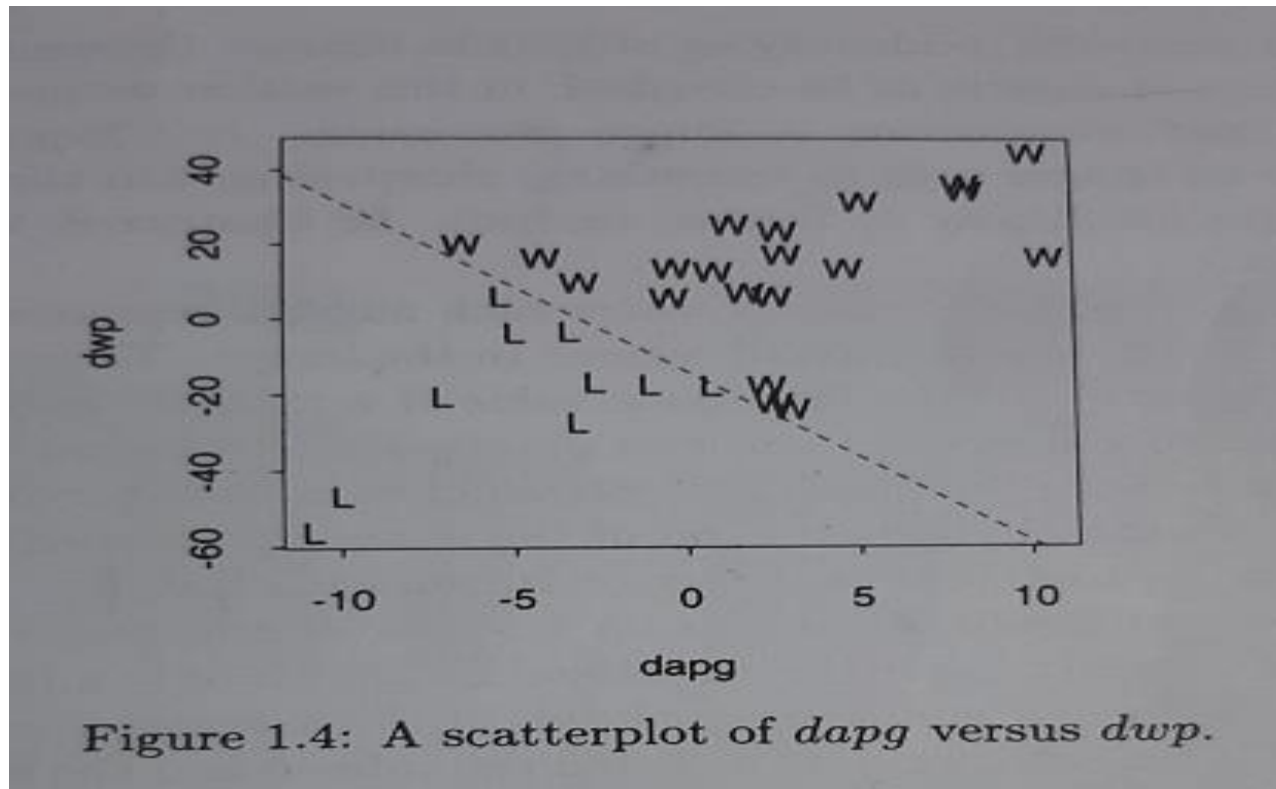
- T is called decision boundary or threshold.
- If $T=-1$, four samples in the original data are misclassified.
 - Here 3 winners are called losers and one loser is called winner.
- If $T=0.8$, results in no samples from the loser class being misclassified as winner, but 5 samples from the winner class would be misclassified as loser.
- If $T=-6.5$, results no samples from the winner class being misclassified as losers, but 7 samples from the loser would be misclassified as winners.
- By inspection, we see that when a decision boundary is used to classify the samples the minimum number of samples that are misclassified is four.
- In the above observations, the minimum number of samples misclassified is 4 when $T=-1$

- To make it more accurate let us consider two features.
 - Additional features often increases the accuracy of classification.
 - Along with 'dapg' another feature 'dwp' is considered.
-
- wp = winning percentage of a team in previous games
 - dwp = difference in winning percentage between teams
 - dwp = Home team wp – visiting team wp

Game	<i>dapg</i>	<i>dwp</i>	Home Team	Game	<i>dapg</i>	<i>dwp</i>	Home Team
1	1.3	25.0	Won	16	-3.1	9.4	Won
2	-2.7	-16.9	Lost	17	-1.7	6.8	Won
3	-0.5	5.3	Won	18	2.8	17.0	Won
4	-3.2	-27.5	Lost	19	4.6	13.3	Won
5	2.3	-18.0	Won	20	3.0	-24.0	Won
6	5.1	31.2	Won	21	0.7	-17.8	Lost
7	-5.4	5.8	Lost	22	10.1	44.6	Won
8	8.2	34.3	Won	23	2.5	-22.4	Won
9	-10.8	-56.3	Lost	24	0.8	12.3	Won
10	-0.4	13.3	Won	25	-5.0	-3.8	Lost
11	10.5	16.3	Won	26	8.1	36.0	Won
12	-1.1	-17.6	Lost	27	-7.1	-20.6	Lost
13	2.5	5.7	Won	28	2.7	23.2	Won
14	-4.2	16.0	Won	29	-10.0	-46.9	Lost
15	-3.4	-3.4	Lost	30	-6.5	19.7	Won

Data set of games showing outcomes, differences between average number of points scored and differences between winning percentages for the participating teams in previous games

- Now observe the results on a **scatterplot**



- Each sample has a corresponding feature vector ($dapg$, dwp), which determines its position in the plot.
- Note that the feature space can be classified into two decision regions by a straight line, called a **linear decision boundary**. (refer line equation). **Prediction of this line is logistic regression.**
- If the sample lies above the decision boundary, the home team would be classified as the winner and if it is below the decision boundary it is classified as loser.

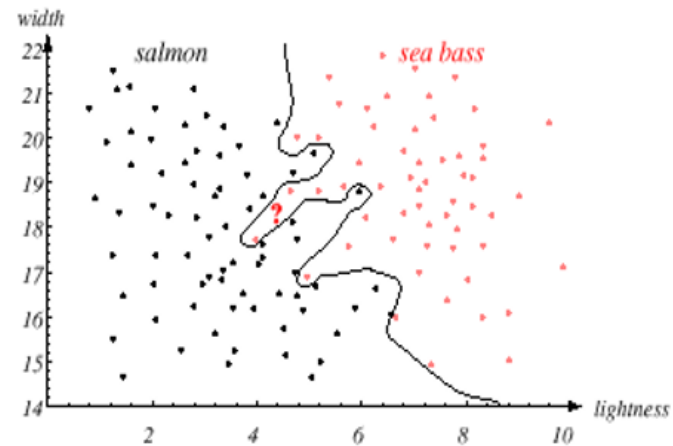
Prediction with two parameters.

- Consider the following : springfield (Home team)

Springfield's *apg* = 98.3
Centerville's *apg* = 102.9
Springfield's *wp* = 21.4
Centerville's *wp* = 58.1.

- $d_{apg} = \text{home team } apg - \text{visiting team } apg = 98.3 - 102.9 = -4.6$
- $d_{wp} = \text{Home team } wp - \text{visiting team } wp = 21.4 - 58.1 = -36.7$
- Since the point $(d_{apg}, d_{wp}) = (-4.6, -36.7)$ lies below the decision boundary, we predict that the home team will lose the game.

- If the feature space cannot be perfectly separated by a straight line, a **more complex boundary might be used**. (non-linear)

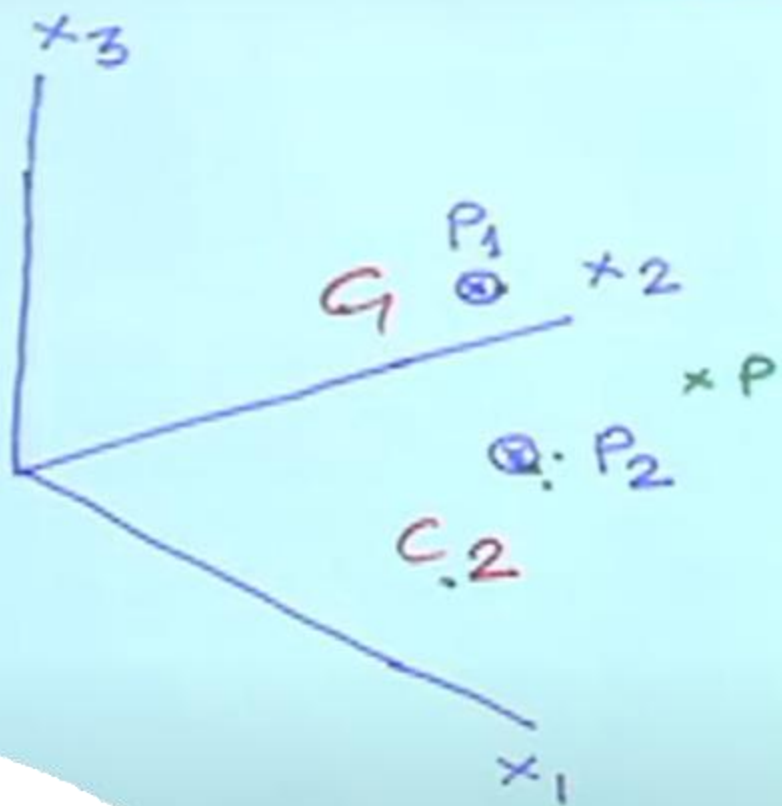


- Alternatively a simple decision boundary such as straight line might be used even if it did not perfectly separate the classes, provided that the error rates were acceptably low.

Simple illustration of Pattern Classification

- A pattern/object can be identified by set of features.
- Collection of features for a pattern forms feature vector.
- Example : (in next slide)
- P1 and P2 are two patterns with 3 features, so 3 Dimensional feature vector.
- There are two classes C1 and C2.
- P1 belongs to C1 and P2 belongs to C2
- Given P, a new pattern with feature vector, it has to be classified into one of the class based on the similarity value.
- If d_1 is the distance between (p and p1) and d_2 is the distance between (p and p2) then p will be classified into the class having least difference.

$M = 3.$



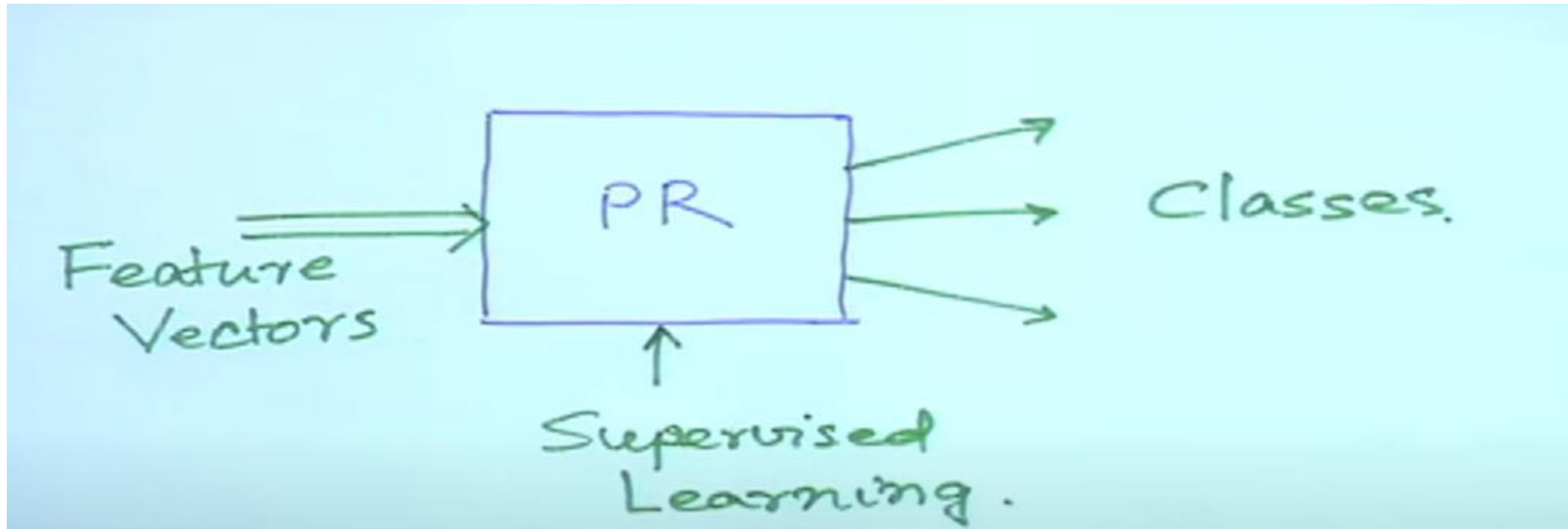
$$P_1 \approx \langle 3, 5, 1 \rangle$$

$$P_2 \approx \langle f_1, f_2, f_3 \rangle$$

$$P \approx \langle x_1, x_2, x_3 \rangle$$

$$d(P_1, P) > d(P_2, P)$$

Block diagram of Pattern recognition and classification



Input to our pattern recognition system will be feature vectors and output will be decision about selecting the classes

- Having the model shown in previous slide, we can use it for any type of recognition and classification.
- It can be
 - speaker recognition
 - Speech recognition
 - Image classification
 - Video recognition and so on...

- It is now very important to learn:
 - Different techniques to extract the features
 - Then in the second stage, different methods to recognize the pattern and classify
 - Some of them use statistical approach
 - Few uses probabilistic model using mean and variance etc.
 - Other methods are - neural network, deep neural networks
 - Hyper box classifier
 - Fuzzy measure
 - And mixture of some of the above

Examples for pattern recognition and classification



Handwriting Recognition

From
Jim Elder
829 Loop Street, Apt 300
Allentown, New York 14707

To
Dr. Bob Grant
602 Queensberry Parkway
Omara, West Virginia 26028

We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Zack.

It all started around six months ago while attending the "Rubeq" Jazz Concert. Organizing such an event is no picnic, and as President of the Alumni Association, a co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm.

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

Thank you!
Jim



From
Jim Elder
829 Loop Street, Apt 300
Allentown, New York 14707

To
Dr. Bob Grant
602 Queensberry Parkway
Omara, West Virginia 26028

We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Zack.

It all started around six months ago while attending the "Rubeq" Jazz Concert. Organizing such an event is no picnic, and as President of the Alumni Association, a co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm.

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

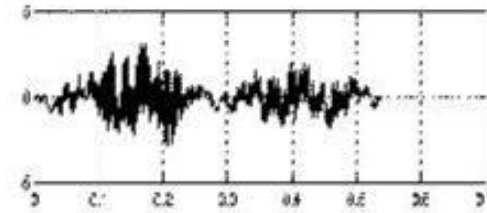
Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

Thank you!
Jim

License Plate Recognition



Biometric Recognition



Face Detection/Recognition

Detection



↓ Matching



→ Recognition

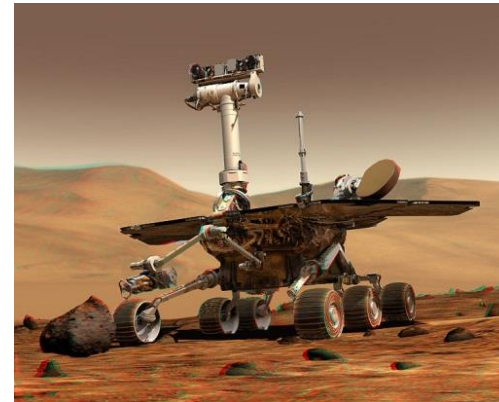
Fingerprint Classification

Important step for speeding up identification



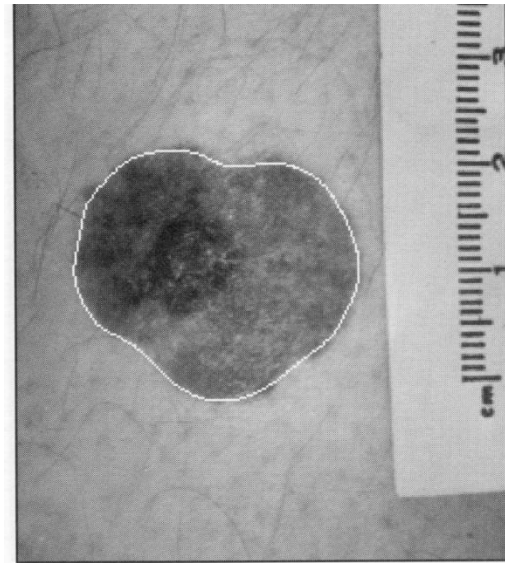
Autonomous Systems

Obstacle detection and avoidance
Object recognition



Medical Applications

Skin Cancer Detection



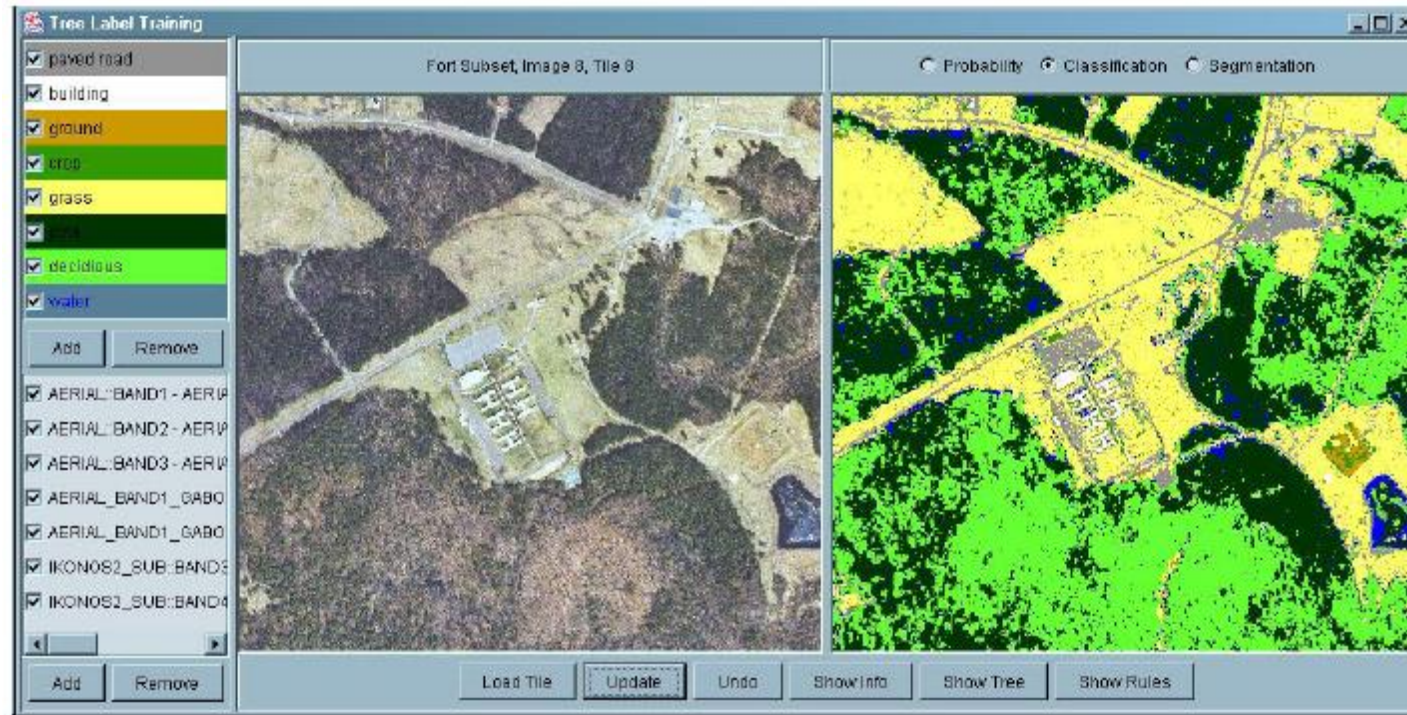
Breast Cancer Detection



Land Cover Classification

(using aerial or satellite images)

Many applications including “precision” agriculture.



Probability:

Introduction to probability

Probabilities of Events

What is covered?

- Basics of Probability
- Combination
- Permutation
- Examples for the above
- Union
- Intersection
- Complement

What is a probability

- **Probability** is the branch of mathematics concerning numerical descriptions of how likely an event is to occur
- The probability of an event is a number between 0 and 1, where, roughly speaking, 0 indicates that the event is not going to happen and 1 indicates event happens all the time.

$$\text{Probability of an event} = \frac{\text{Chance favouring the event}}{\text{Total possible events}}$$

Experiment

- The **term experiment** is used in probability theory to describe a process for which the outcome is not known with certainty.

Example of experiments are:

Rolling a fair six sided die.

Randomly choosing 5 apples from a lot of 100 apples.

Event

- An event is an outcome of an experiment. It is denoted by capital letter. Say E_1, E_2, \dots or A, B, \dots and so on
- For example toss a coin, H and T are two events.
- The event consisting of all possible outcomes of a statistical experiment is called the “Sample Space”. Ex: $\{ E_1, E_2, \dots \}$

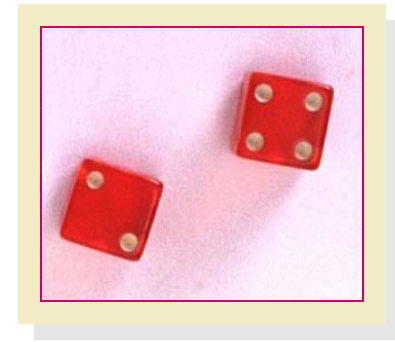
Examples

Sample Space of Tossing a coin = $\{H,T\}$

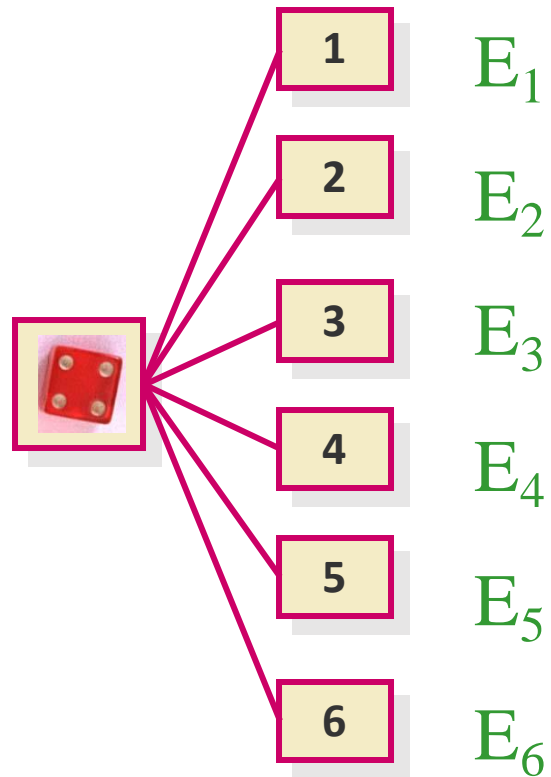
Tossing 2 Coins = $\{HH,HT,TH,TT\}$

Example

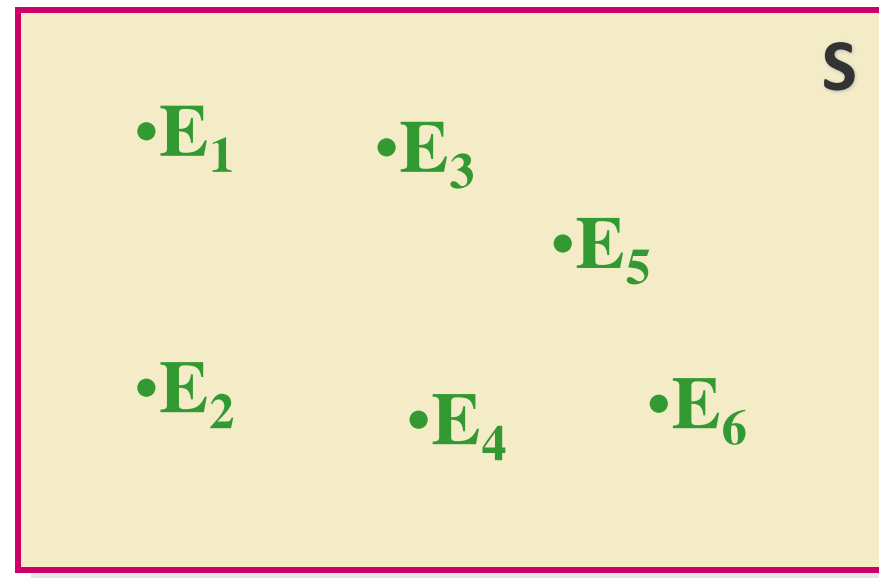
- The die toss:
- Simple events:



Sample space:



$$S = \{E_1, E_2, E_3, E_4, E_5, E_6\}$$



The Probability of an Event $P(A)$



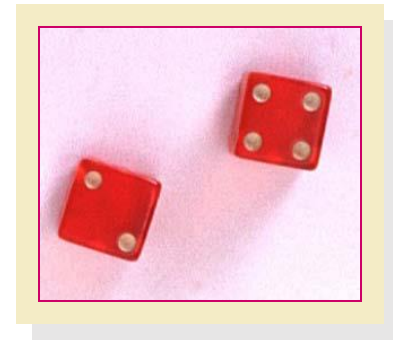
- The probability of an event A measures “how often” A will occur. We write $P(A)$.
- Suppose that an experiment is performed n times. The relative frequency for an event A is

$$\frac{\text{Number of times } A \text{ occurs}}{n} = \frac{f}{n}$$

- If we let n get infinitely large,

$$P(A) = \lim_{n \rightarrow \infty} \frac{f}{n}$$

The Probability of an Event



- $P(A)$ must be between 0 and 1.
 - If event A can never occur, $P(A) = 0$. If event A always occurs when the experiment is performed, $P(A) = 1$.
 - **Then $P(A) + P(\text{not } A) = 1$.**
 - So $P(\text{not } A) = 1 - P(A)$
- The sum of the probabilities for all simple events in S equals 1.

Example 1












Toss a fair coin twice. What is the probability of observing at least one head?

1st Coin	2nd Coin	E_i	$P(E_i)$
H	H	HH	$1/4$
	T	HT	$1/4$
T	H	TH	$1/4$
	T	TT	$1/4$

$$\begin{aligned} &P(\text{at least 1 head}) \\ &= P(E_1) + P(E_2) + P(E_3) \\ &= 1/4 + 1/4 + 1/4 = 3/4 \end{aligned}$$

Example 2

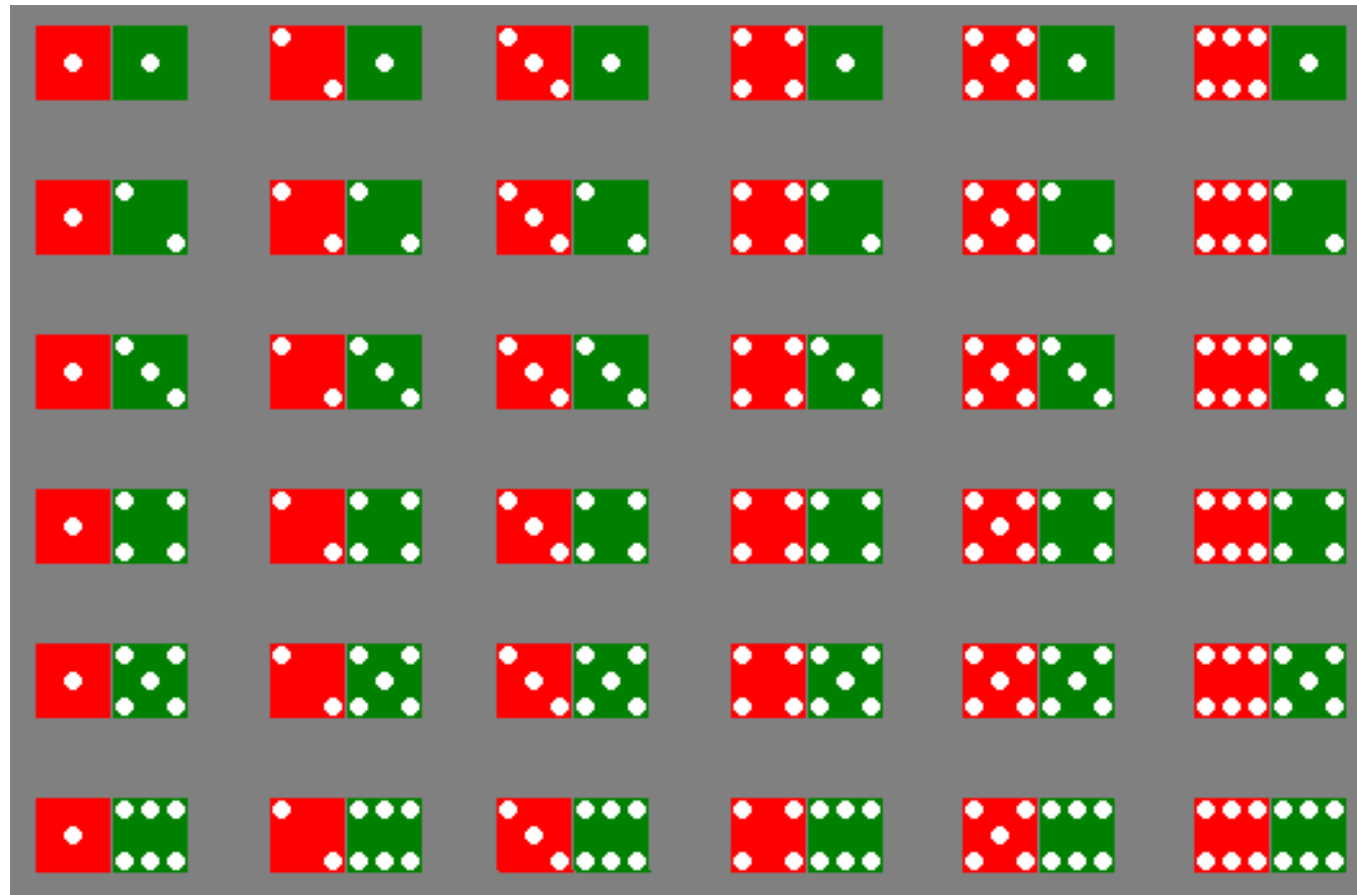
A bowl contains three colour Ms[®], one red, one blue and one green. A child selects two M&Ms at random. What is the probability that at least one is red?

1st M&M	2nd M&M	E_i	$P(E_i)$
		RB	1/6
		RG	1/6
		BR	1/6
		BG	1/6
		GB	1/6
		GR	1/6

$$\begin{aligned} &P(\text{at least 1 red}) \\ &= P(\text{RB}) + P(\text{BR}) + P(\text{RG}) + P(\text{GR}) \\ &= 4/6 = 2/3 \end{aligned}$$

Example 3

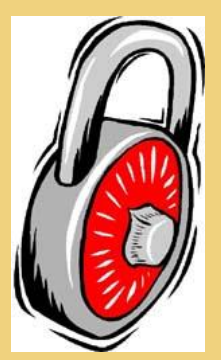
The sample space of throwing a pair of dice is



Example 3

Event	Simple events	Probability
Dice add to 3	(1,2),(2,1)	2/36
Dice add to 6	(1,5),(2,4),(3,3), (4,2),(5,1)	5/36
Red die show 1	(1,1),(1,2),(1,3), (1,4),(1,5),(1,6)	6/36
Green die show 1	(1,1),(2,1),(3,1), (4,1),(5,1),(6,1)	6/36

Permutations



- The number of ways you can arrange n distinct objects, taking them r at a time is

$$P_r^n = \frac{n!}{(n-r)!}$$

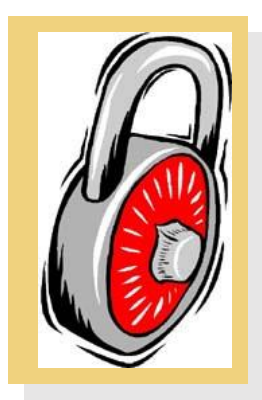
where $n! = n(n-1)(n-2)\dots(2)(1)$ and $0! \equiv 1$.

Example: How many 3-digit lock combinations can we make from the numbers 1, 2, 3, and 4?

The order of the choice is important!

$$P_3^4 = \frac{4!}{1!} = 4(3)(2) = 24$$

Examples



Example: A lock consists of five parts and can be assembled in any order. A quality control engineer wants to test each order for efficiency of assembly. How many orders are there?

The order of the choice is important!

$$P_5^5 = \frac{5!}{0!} = 5(4)(3)(2)(1) = 120$$

Combinations

- The number of distinct combinations of n distinct objects that can be formed, taking them r at a time is

$$C_r^n = \frac{n!}{r!(n-r)!}$$

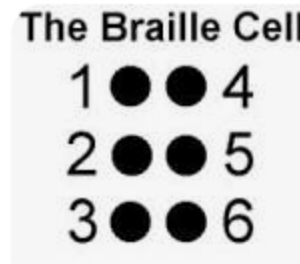
Example: Three members of a 5-person committee must be chosen to form a subcommittee. How many different subcommittees could be formed?

The order of the choice is not important!

$$C_3^5 = \frac{5!}{3!(5-3)!} = \frac{5(4)(3)(2)1}{3(2)(1)(2)1} = \frac{5(4)}{(2)1} = 10$$

Is it combination or permutation?

- Having 6 dots in a braille cell, how many different character can be made?



- It is a problem of combination
- $C_{6,0} + C_{6,1} + C_{6,2} + C_{6,3} + C_{6,4} + C_{6,5} + C_{6,6} = 1 + 6 + 15 + 20 + 15 + 6 + 1 = 64$
- (Why combination is used not permutation? : reason each dots is of same nature)
- 64 different characters can be made.
- Where N is from 0 to 6. (It is the summation of combinations..)

ways of counting the

$$\sum_i C(N, i) = 2^N$$

Having 4 characters, how many 2 character words can be formed:

Permutation : $P_{6,2} = 12$

Combination: $C_{6,2} = 6$

Remember Permutation is larger than combination

The diagram illustrates the difference between permutations and combinations using the characters A, B, C, and D. At the top, the characters are listed: A B C D. Below this, a 3x4 grid of two-letter pairs is shown. The first column contains AB, AC, and AD, each enclosed in a colored box (red, blue, and green respectively). The second column contains BA, BC, and BD, with BA crossed out and BC and BD boxed in red and yellow. The third column contains CA, CB, and CD, with CA and CB crossed out and CD boxed in orange. The fourth column contains DA, DB, and DC, with DA and DB crossed out and DC boxed in orange. To the right of the grid, the permutation value is written as $P = 12$ in blue, and the combination value is written as $C = 6$ in red.

Summary:

- So formula for Permutation is : (order is relevant)

$$P_r^n = \frac{n!}{(n-r)!}$$

- Formula for Combination is: (Order is not relevant)

$$C_r^n = \frac{n!}{r!(n-r)!}$$

Event Relations

Special Events

**The Null Event, is also called as empty event
represented by - ϕ**

$\phi = \{ \} =$ the event that contains no outcomes

The Entire Event, The Sample Space - S

$S =$ the event that contains all outcomes

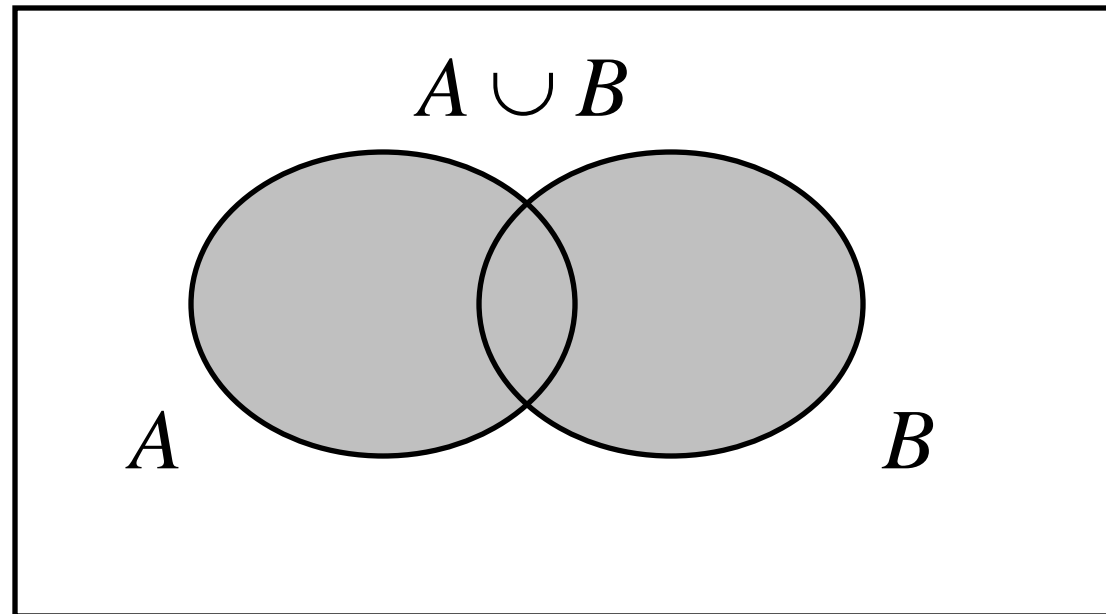
3 Basic Event relations

1. **Union** if you see the word **or**,
2. **Intersection** if you see the word **and**,
3. **Complement** if you see the word **not**.

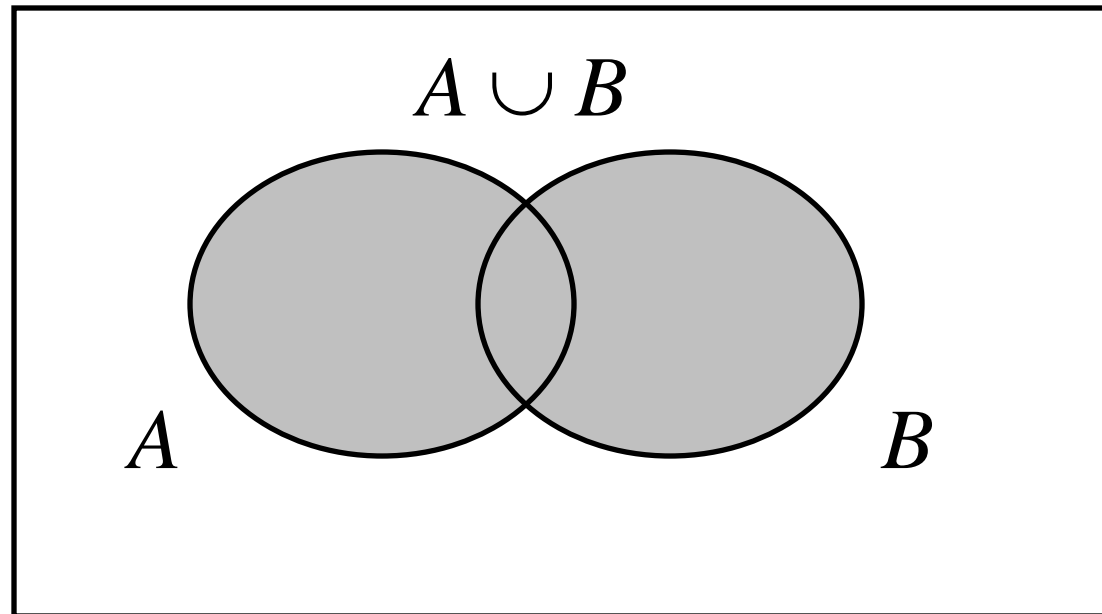
Union

Let A and B be two events, then the **union** of A and B is the event (denoted by $A \cup B$) defined by:

$$A \cup B = \{e \mid e \text{ belongs to } A \text{ or } e \text{ belongs to } B\}$$



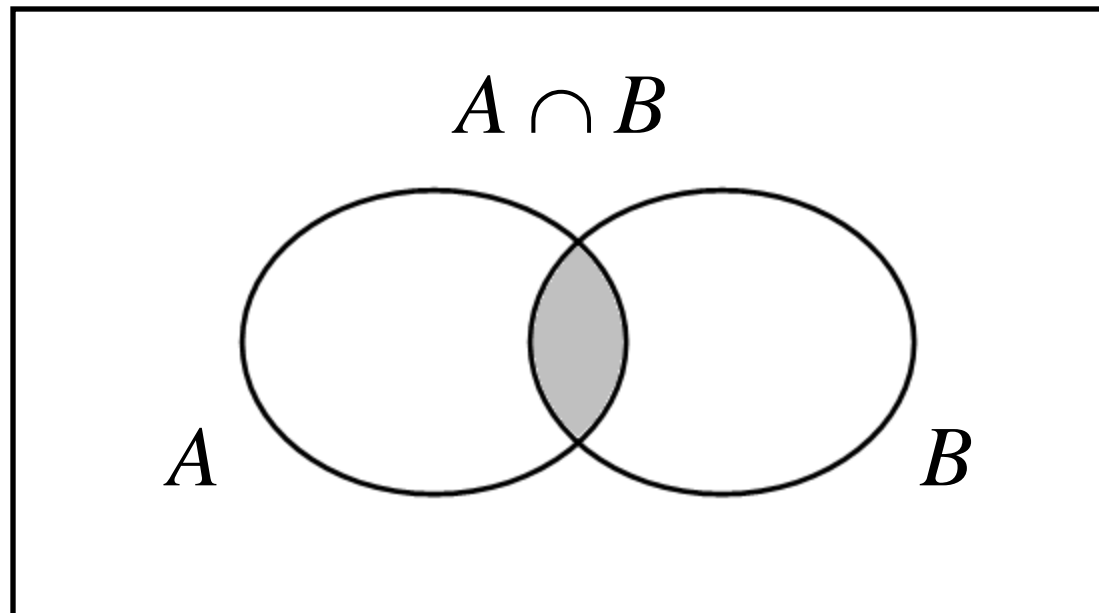
The event $A \cup B$ **occurs** if the event A **occurs** or the event and B **occurs** or **both** occurs.



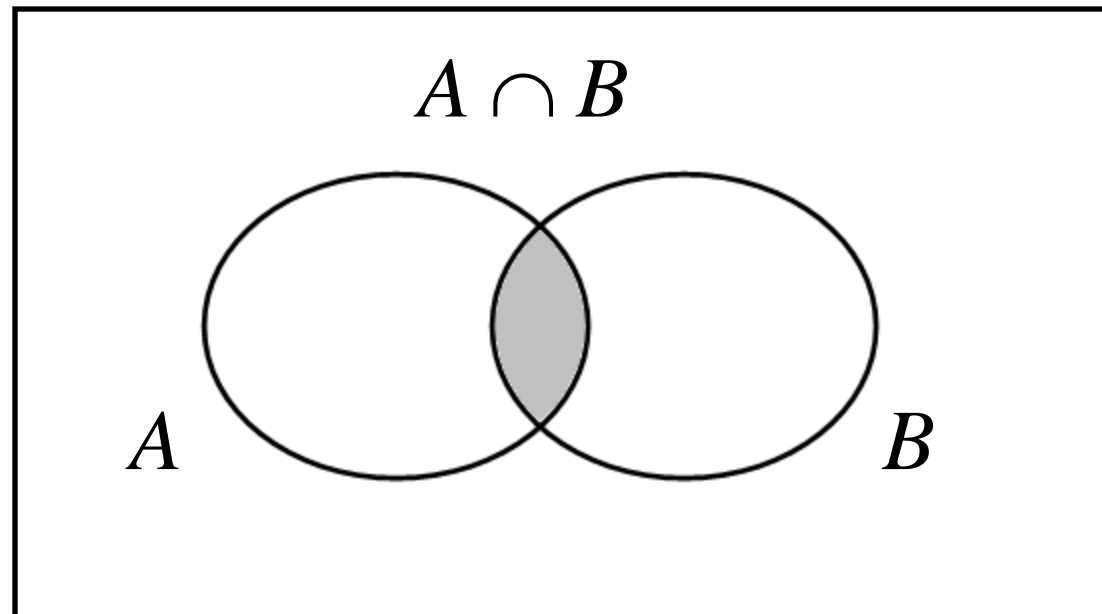
Intersection

Let A and B be two events, then the **intersection** of A and B is the event (denoted by $A \cap B$) defined by:

$$A \cap B = \{e \mid e \text{ belongs to } A \text{ and } e \text{ belongs to } B\}$$



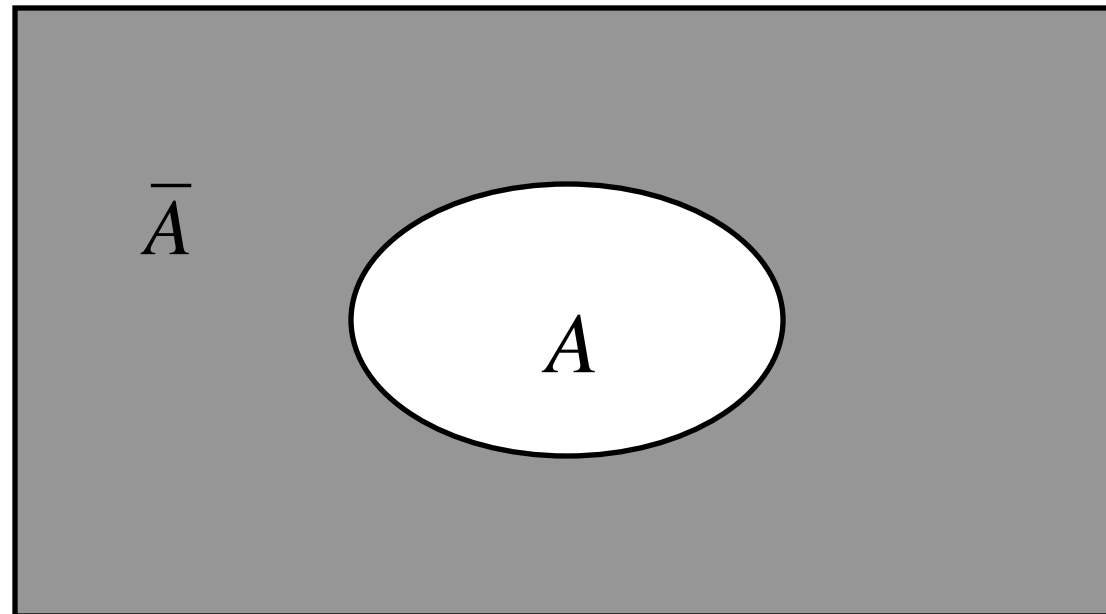
The event $A \cap B$ **occurs** if the event A **occurs** and the event and B **occurs** .



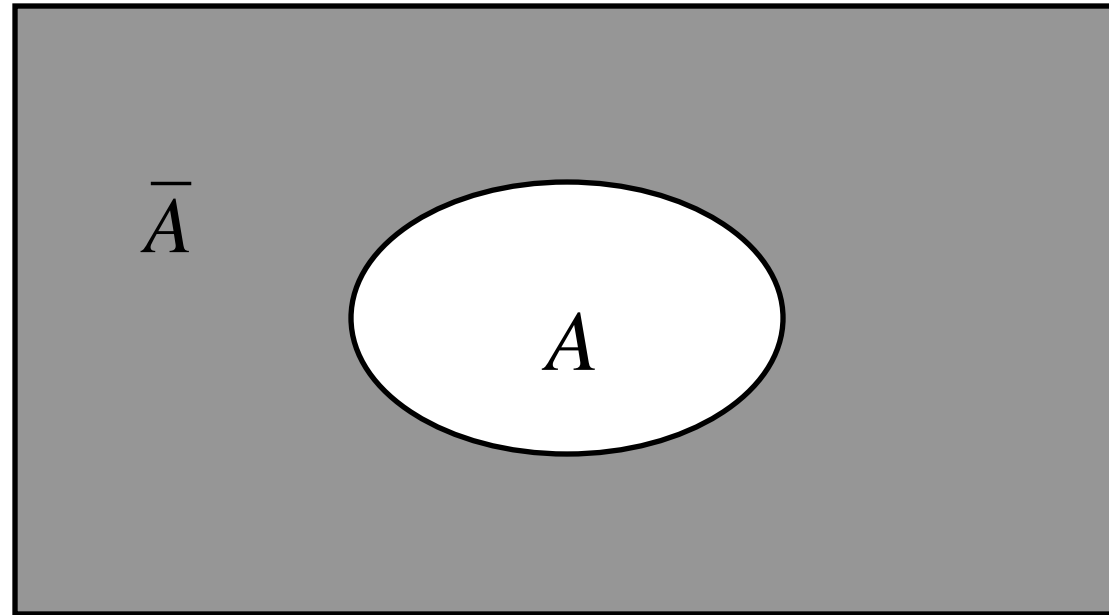
Complement

Let A be any event, then the **complement** of A (denoted by \bar{A}) defined by:

$$\bar{A} = \{e \mid e \text{ does not belongs to } A\}$$



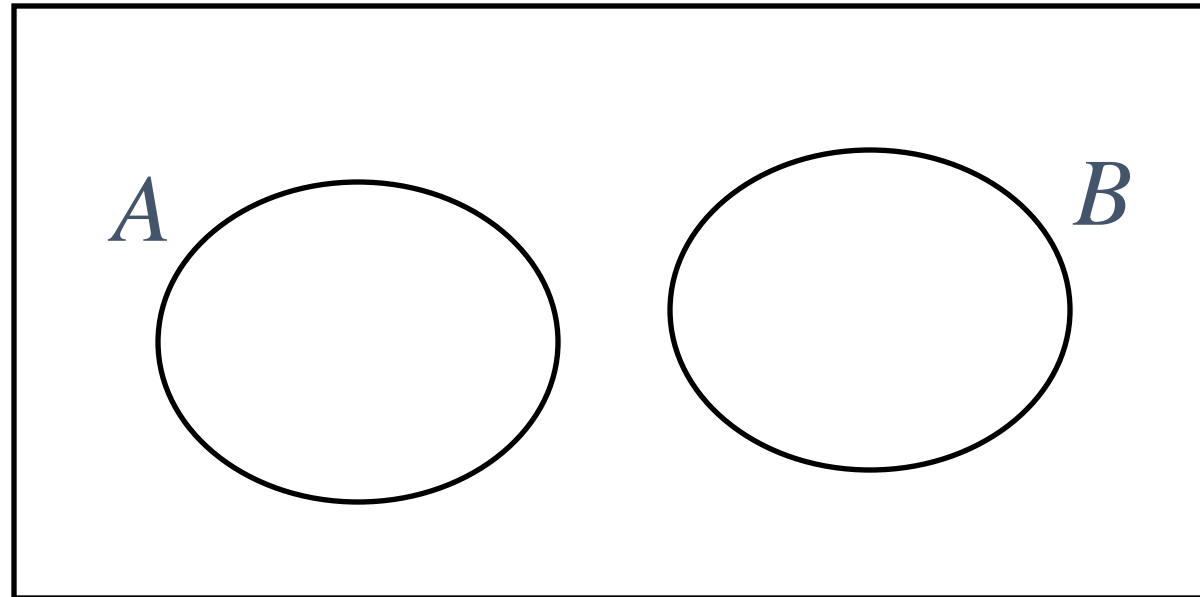
The event \bar{A} **occurs** if the event A **does not occur**



Mutually Exclusive

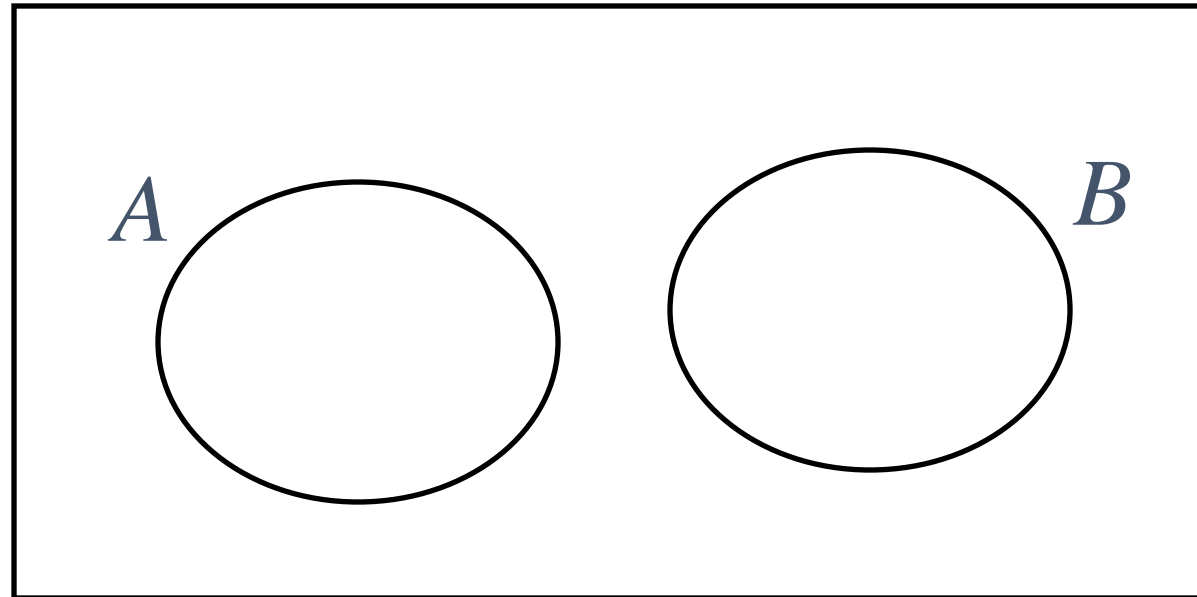
Two events A and B are called **mutually exclusive** if:

$$A \cap B = \phi$$



If two events A and B are **mutually exclusive** then:

1. They have no outcomes in common. They can't occur at the same time. The outcome of the random experiment can not belong to both A and B .



Rules of Probability

Additive Rule

Rule for complements

Probability of an Event E .

(revisiting ... discussed in earlier slides)

Suppose that the sample space $S = \{o_1, o_2, o_3, \dots, o_N\}$ has a finite number, N , of outcomes.

Also each of the outcomes is equally likely (because of symmetry).

Then for any event E

$$P[E] = \frac{n(E)}{n(S)} = \frac{n(E)}{N} = \frac{\text{no. of outcomes in } E}{\text{total no. of outcomes}}$$

Note: the symbol $n(A) =$ no. of elements of A

Additive rule (In general)

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

or

$$P[A \text{ or } B] = P[A] + P[B] - P[A \text{ and } B]$$

The additive rule (Mutually exclusive events) if $A \cap B = \phi$

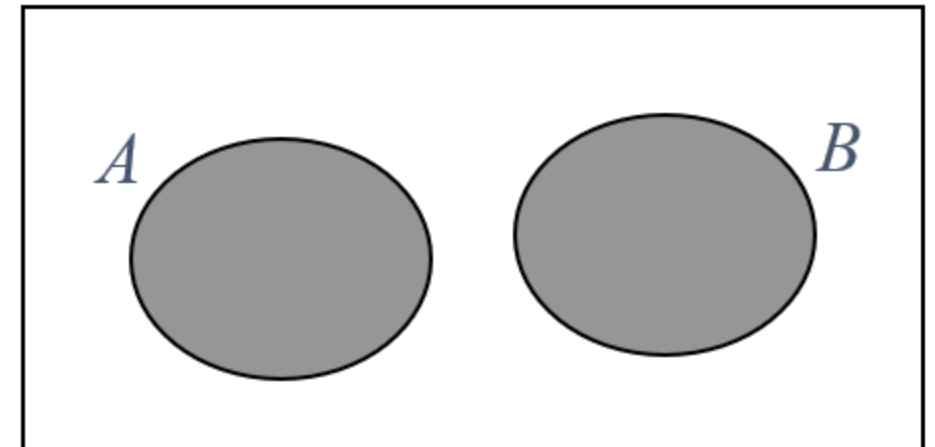
$$P[A \cup B] = P[A] + P[B]$$

i.e.

$$P[A \text{ or } B] = P[A] + P[B]$$

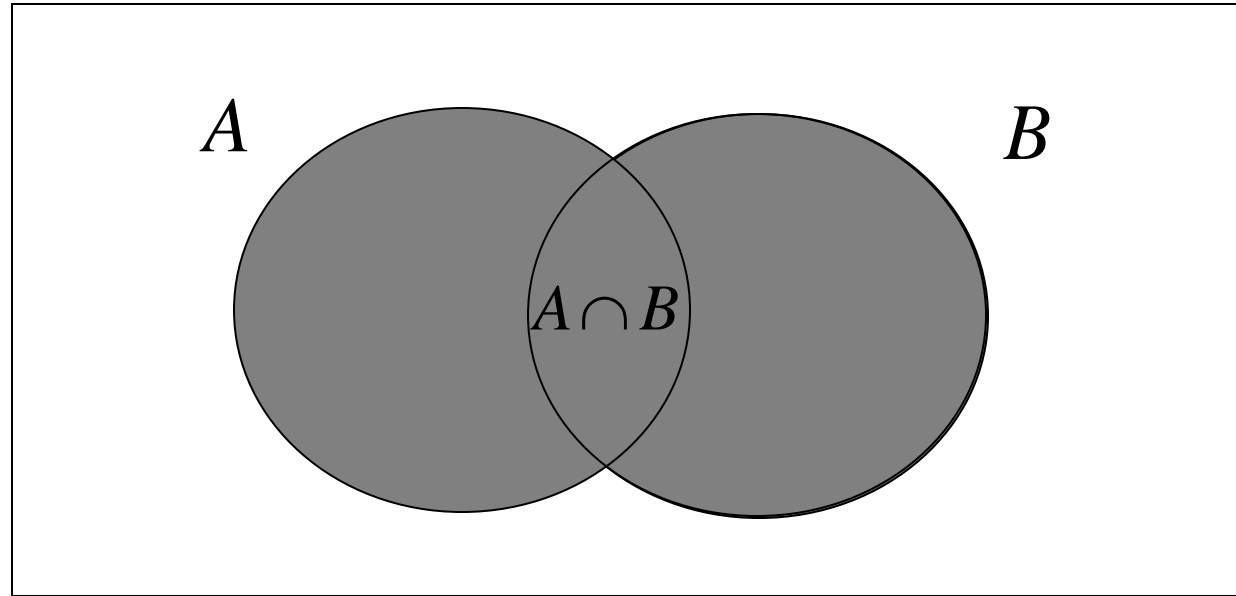
if $A \cap B = \phi$

(A and B mutually exclusive)



Logic

$$A \cup B$$



When $P[A]$ is added to $P[B]$ the outcome in $A \cap B$ are counted twice

hence

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

Example:

Bangalore and Mohali are two of the cities competing for the National university games. (There are also many others).

The organizers are narrowing the competition to the final 5 cities.

There is a 20% chance that Bangalore will be amongst the final 5.

There is a 35% chance that Mohali will be amongst the final 5 and

an 8% chance that both Bangalore and Mohali will be amongst the final 5.

What is the probability that **Bangalore or Mohali** will be amongst the final 5.

Solution:

Let A = the event that Bangalore is amongst the **final 5**.

Let B = the event that Mohali is amongst the **final 5**.

Given $P[A] = 0.20$, $P[B] = 0.35$, and $P[A \cap B] = 0.08$

What is $P[A \cup B]$?

Note: “and” $\equiv \cap$, “or” $\equiv \cup$.

$$\begin{aligned} P[A \cup B] &= P[A] + P[B] - P[A \cap B] \\ &= 0.20 + 0.35 - 0.08 = 0.47 \end{aligned}$$

Find the probability of drawing an ace or a spade from a deck of cards.

There are 52 cards in a deck; 13 are spades, 4 are aces.

Probability of a single card being spade is: $13/52 = 1/4$.

Probability of drawing an Ace is : $4/52 = 1/13$.

Probability of a single card being both Spade and Ace = $1/52$.

Let A = Event of drawing a spade .

Let B = Event drawing Ace.

Given $P[A] = 1/4$, $P[B] = 1/13$, and $P[A \cap B] = 1/52$

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

$$P[A \cup B] = 1/4 + 1/13 - 1/52$$

Rule for complements

Rule for complements

The Complement Rule states that the **sum of the probabilities of an event and its complement must equal 1**, or for the event A , $P(A) + P(A') = 1$.

$$P[\bar{A}] = 1 - P[A]$$

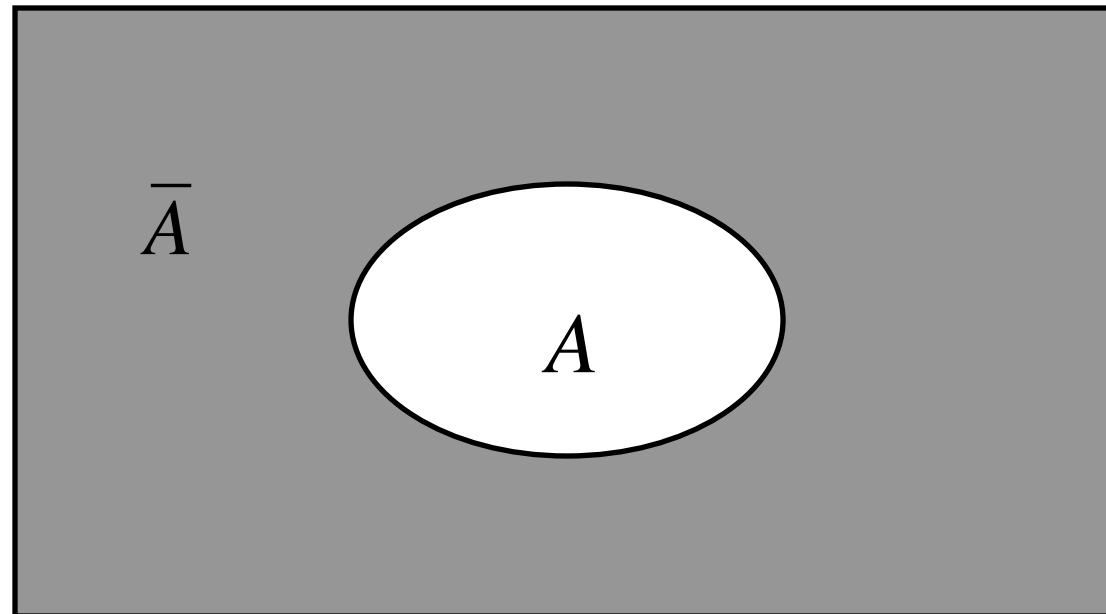
or

$$P[\text{not } A] = 1 - P[A]$$

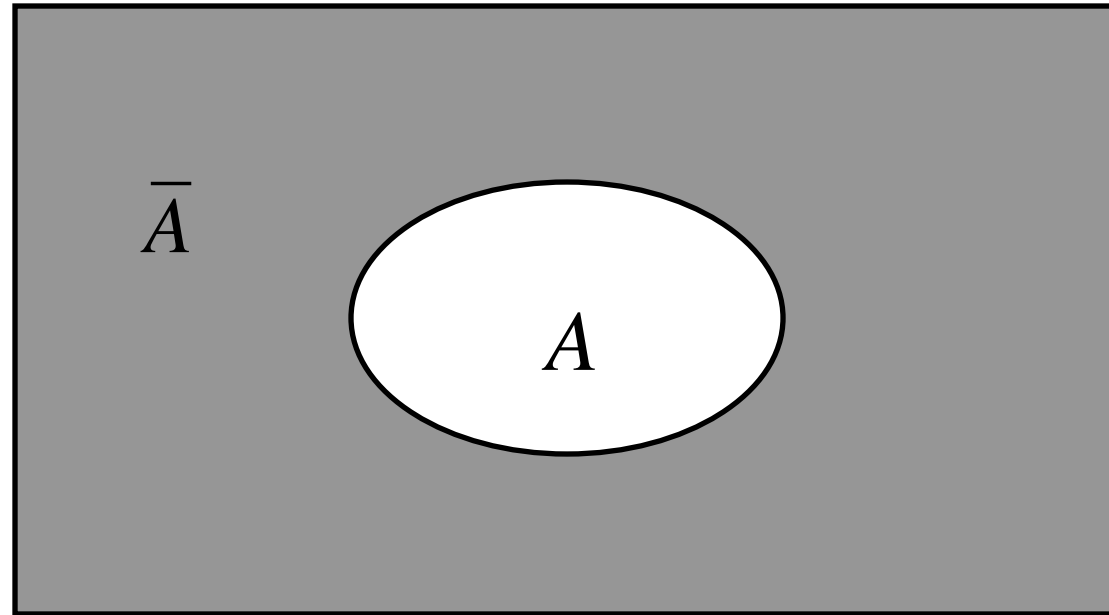
Complement

Let A be any event, then the **complement** of A (denoted by \bar{A}) defined by:

$$\bar{A} = \{e \mid e \text{ does not belongs to } A\}$$



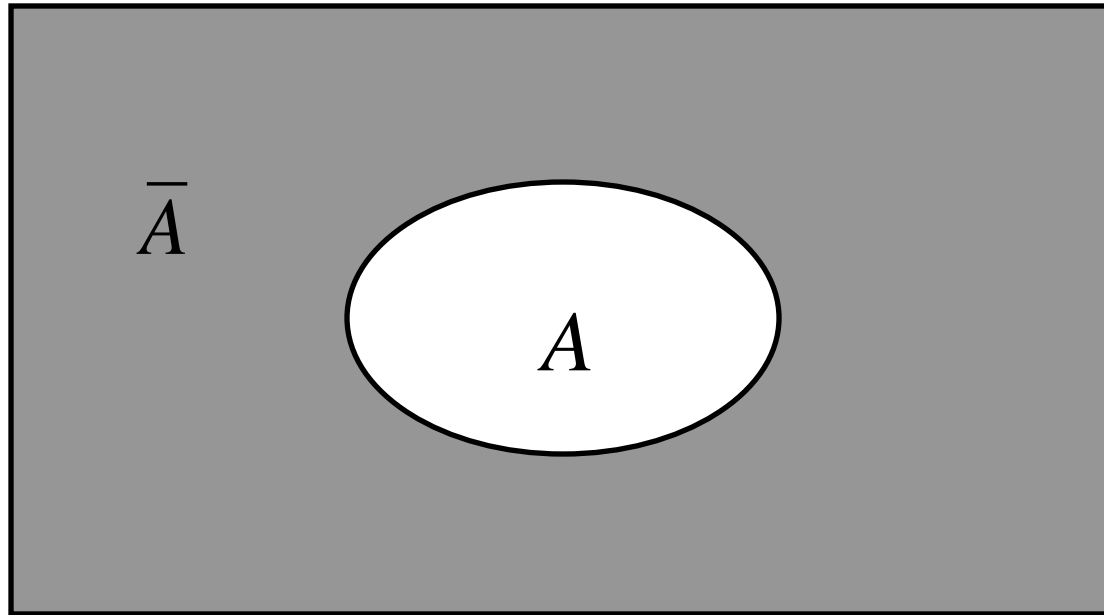
The event \bar{A} **occurs** if the event A **does not occur**



Logic:

\bar{A} and A are **mutually exclusive**.

and $S = A \cup \bar{A}$



thus $1 = P[S] = P[A] + P[\bar{A}]$

and $P[\bar{A}] = 1 - P[A]$

What Is Conditional Probability?

- Conditional probability is defined as the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome.
- Conditional probability is calculated by multiplying the probability of the preceding event by the updated probability of the succeeding, or conditional, event.
- Bayes' theorem is a mathematical formula used in calculating conditional probability.

Definition

Suppose that we are interested in computing the probability of event A and we have been told event B has occurred.

Then the conditional probability of A given B is defined to be:

$$P[A|B] = \frac{P[A \cap B]}{P[B]} \quad \text{if } P[B] \neq 0$$

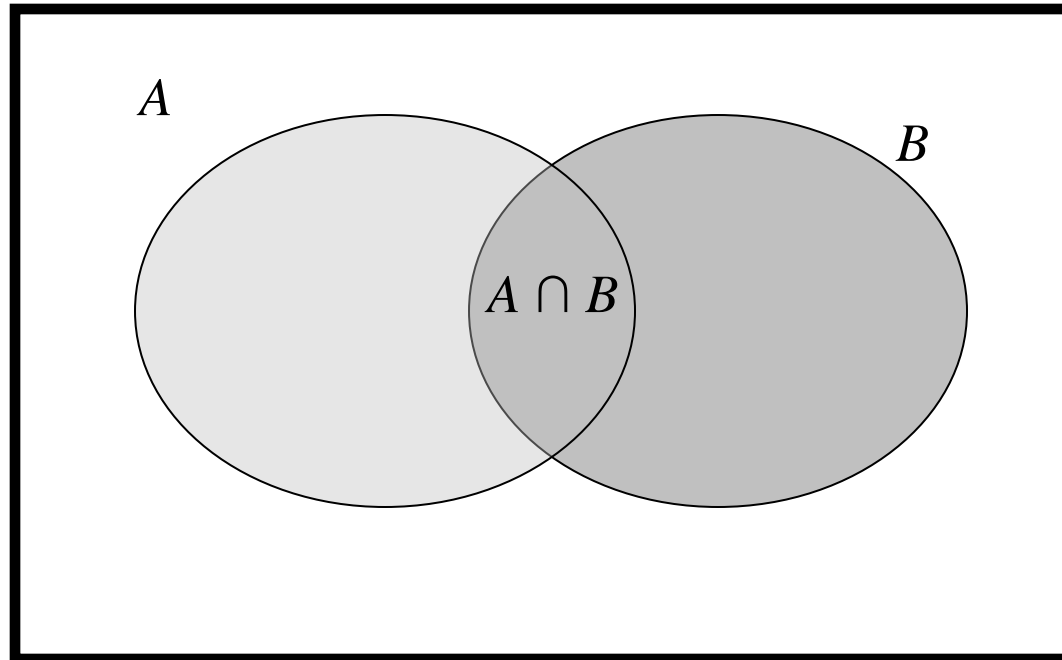
Illustrates that probability of A , given($|$) probability of B occurring

Rationale:

If we're told that event B has occurred then the sample space is restricted to B .

The event A can now only occur if the outcome is in of $A \cap B$. Hence the new probability of A *in B* is:

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$



An Example

Twenty – 20 World cup started:

For a specific married couple the probability that the husband watches the match is 80%,

the probability that his wife watches the match is 65%,

while the probability that they both watch the match is 60%.

If the husband is watching the match, what is the probability that his wife is also watching the match

Solution:

Let B = the event that the husband watches the match

$$P[B] = 0.80$$

Let A = the event that his wife watches the match

$$P[A] = 0.65 \text{ and}$$

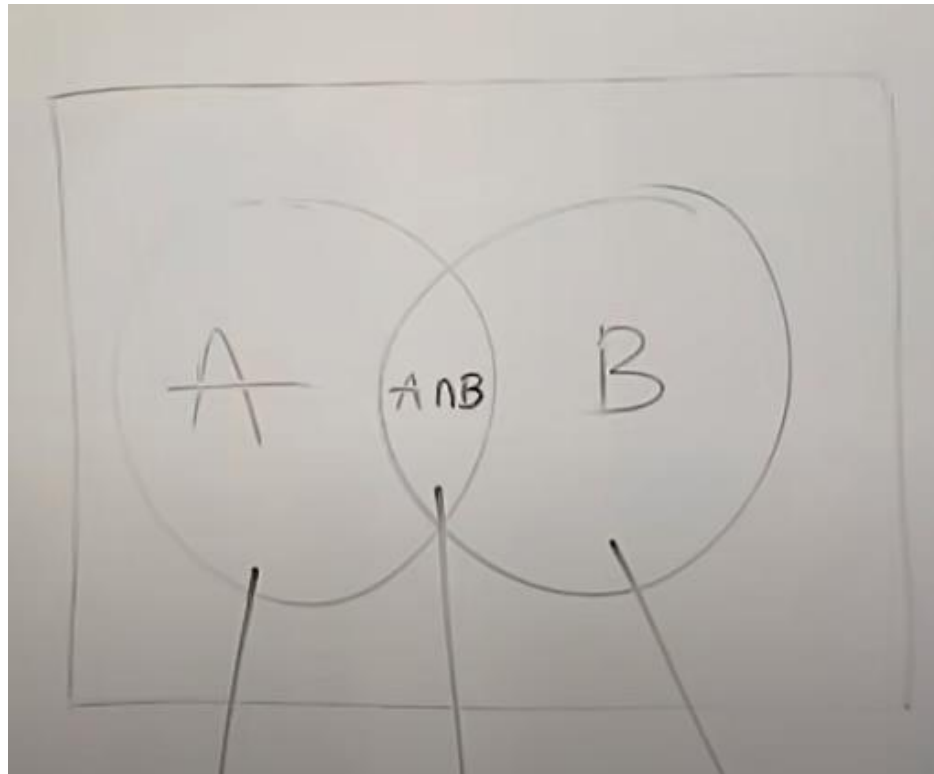
$$P[A \cap B] = 0.60$$

$$P[A|B] = \frac{P[A \cap B]}{P[B]} = \frac{0.60}{0.80} = 0.75$$

Another example

- There are 100 Students in a class.
- 40 Students likes Apple
 - Consider this event as A, So probability of occurrence of A is $40/100 = 0.4$
- 30 Students likes Orange.
 - Consider this event as B, So probability of occurrence of B is $30/100=0.3$
- 20 Students likes Both Apple and Orange, So probability of Both A and B occurring is = A intersect B = $20/100 = 0.2$
- Remaining Students does not like either Apple nor Orange
- What is the probability of A in B, means what is the probability that A is occurring given B :

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



40

20

30

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A|B) = 0.2/0.3 = 0.67$$

$P(A|B)$ indicates that A occurring in the sample space of B.

Here we are not considering the entire sample space of 100 students, but only 30 students.

More Example Problem for Conditional Probability

Example : Calculating the conditional probability of rain given that the biometric pressure is high.

Weather record shows that high barometric pressure (defined as being over 760 mm of mercury) occurred on 160 of the 200 days in a data set, and it rained on 20 of the 160 days with high barometric pressure. If we let R denote the event “rain occurred” and H the event “ High barometric pressure occurred” and use the frequentist approach to define probabilities.

$$P(H) = 160/200 = 0.8$$

and $P(R \text{ and } H) = 20/200 = 0.10$ (rain and high barometric pressure intersection)

We can obtain the probability of rain given high pressure, directly from the data.

$$P(R|H) = 20/160 = 0.10/0.80 = 0.125$$

Representing in conditional probability

$$P(R|H) = P(R \text{ and } H)/P(H) = 0.10/0.8 = 0.125.$$

In my town, it's rainy one third ($1/3$) of the days.

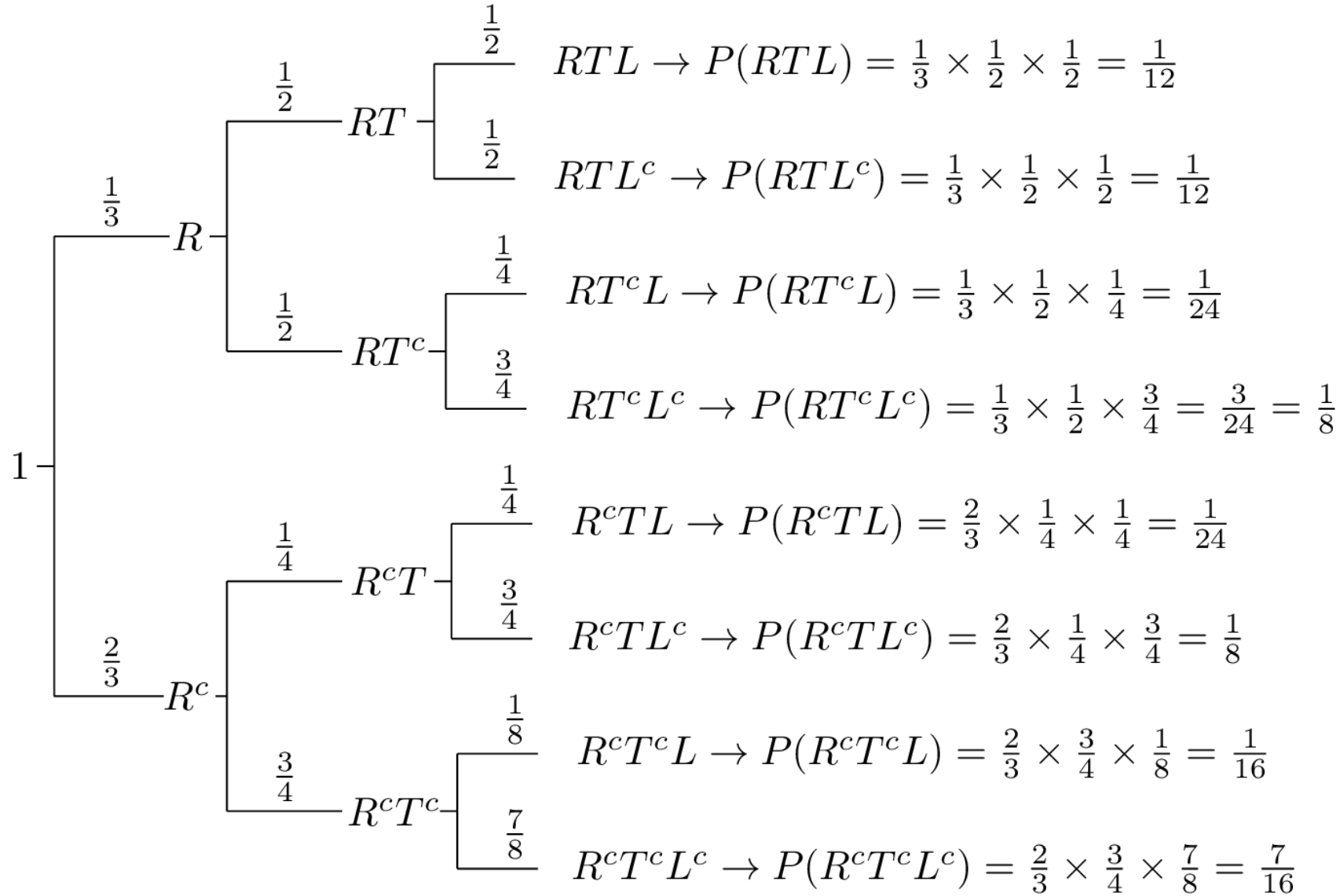
Given that it is rainy, there will be heavy traffic with probability $1/2$, and given that it is not rainy, there will be heavy traffic with probability $1/4$.

If it's rainy and there is heavy traffic, I arrive late for work with probability $1/2$.

On the other hand, the probability of being late is reduced to $1/8$ if it is not rainy and there is no heavy traffic.

In other situations (rainy and no traffic, not rainy and traffic) the probability of being late is 0.25. You pick a random day.

- What is the probability that it's not raining and there is heavy traffic and I am not late?
- What is the probability that I am late?
- Given that I arrived late at work, what is the probability that it rained that day?



Let **R** be the event that it's rainy, **T** be the event that there is heavy traffic, and **L** be the event that I am late for work. As it is seen from the problem statement, we are given conditional probabilities in a chain format. Thus, it is useful to draw a tree diagram for this problem. In this figure, each leaf in the tree corresponds to a single outcome in the sample space. We can calculate the probabilities of each outcome in the sample space by multiplying the probabilities on the edges of the tree that lead to the corresponding outcome.

- a. The probability that it's not raining and there is heavy traffic and I am not late can be found using the tree diagram which is in fact applying the chain rule:

$$\begin{aligned}P(R^c \cap T \cap L^c) &= P(R^c)P(T|R^c)P(L^c|R^c \cap T) \\ &= 2/3 \cdot 1/4 \cdot 3/4 \\ &= 1/8.\end{aligned}$$

- b. The probability that I am late can be found from the tree. All we need to do is sum the probabilities of the outcomes that correspond to me being late. In fact, we are using the law of total probability here.

$$\begin{aligned} P(L) &= P(R \text{ and } T \text{ and } L) + P(R \text{ and } T^c \text{ and } L) + P(R^c \text{ and } T \text{ and } L) + P(R^c \text{ and } T^c \text{ and } L) \\ &= 1/12 + 1/24 + 1/24 + 1/16 \\ &= 11/48. \end{aligned}$$

- c. We can find $P(R|L)$ using

$$P(R|L) = \frac{P(R \cap L)}{P(L)}$$

We have already found $P(L) = 11/48$ and we can find $P(R \cap L)$ similarly by adding the probabilities of the outcomes that belong to $R \cap L$.

Random Variables

Random variable takes a random value, which is real and can be finite or infinite and it is generated out of random experiment.

The random value is generated out of a function.

Example: Let us consider an experiment of tossing two coins.

Then sample space is $S = \{ HH, HT, TH, TT \}$

Given X as random variable with condition: **number of heads.**

$$X(HH) = 2$$

$$X(HT) = 1$$

$$X(TH) = 1$$

$$X(TT) = 0$$

$X = \text{no. of heads}$

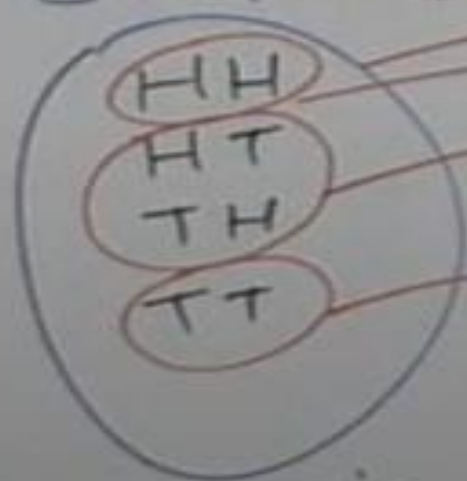
$$X(HH) = 2$$

$$X(HT) = 1$$

$$X(TH) = 1$$

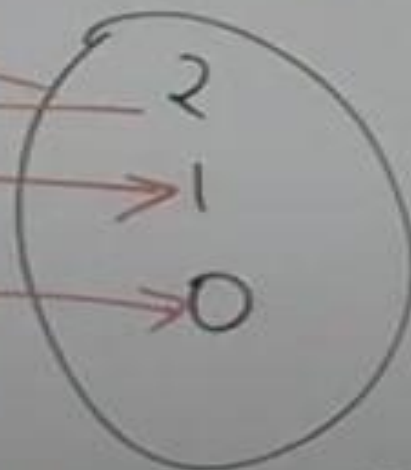
$$X(TT) = 0$$

Sample space Ω



Domain

Real value



Range $\{0, 1, 2\}$
subset of
Real no.

- Two types of random variables
 - **Discrete random variables**
 - **Continuous random variable**

Discrete random variables

- If the variable value is finite or infinite but countable, then it is called discrete random variable.
- Example of tossing two coins and to get the count of number of heads is an example for discrete random variable.
- Sample space of real values is fixed.

Continuous Random Variable

- If the random variable values lies between two certain fixed numbers then it is called continuous random variable. The result can be finite or infinite.
- Sample space of real values is not fixed, but it is in a range.
- If X is the random value and it's values lies between a and b then,

It is represented by : $a \leq X \leq b$

Example: Temperature, age, weight, height...etc. ranges between specific range.
Here the values for the sample space will be infinite

Probability distribution

- Frequency distribution is a listing of the observed frequencies of all the output of an experiment that actually occurred when experiment was done.
- Where as a probability distribution is a listing of the probabilities of all possible outcomes that could result if the experiment were done. (distribution with expectations).

Broad classification of Probability distribution

- Discrete probability distribution
 - Binomial distribution
 - Poisson distribution
- Continuous Probability distribution
 - Normal distribution

Discrete Probability Distribution: Binomial Distribution

- A binomial distribution can be thought of as simply **the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times.** (When we have only two possible outcomes)
- Example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.

Assumptions of Binomial distribution

(It is also called as Bernoulli's Distribution)

- Assumptions:
 - Random experiment is performed repeatedly with a fixed and finite number of trials. The number is denoted by 'n'
 - There are two mutually exclusive possible outcome on each trial, which are know as "Success" and "Failure". Success is denoted by 'p' and failure is denoted by 'q'. and $p+q=1$ or $q=1-p$.
 - The outcome of any give trail does not affect the outcomes of the subsequent trail. That means all trials are independent.
 - The probability of success and failure (p&q) remains constant for all trials. If it does not remain constant then it is not binomial distribution. For example tossing a coin the probability of getting head or getting a red ball from a pool of colored balls, here every time after the ball is taken out it is again replaced to the pool.
 - With this assumption let see the formula

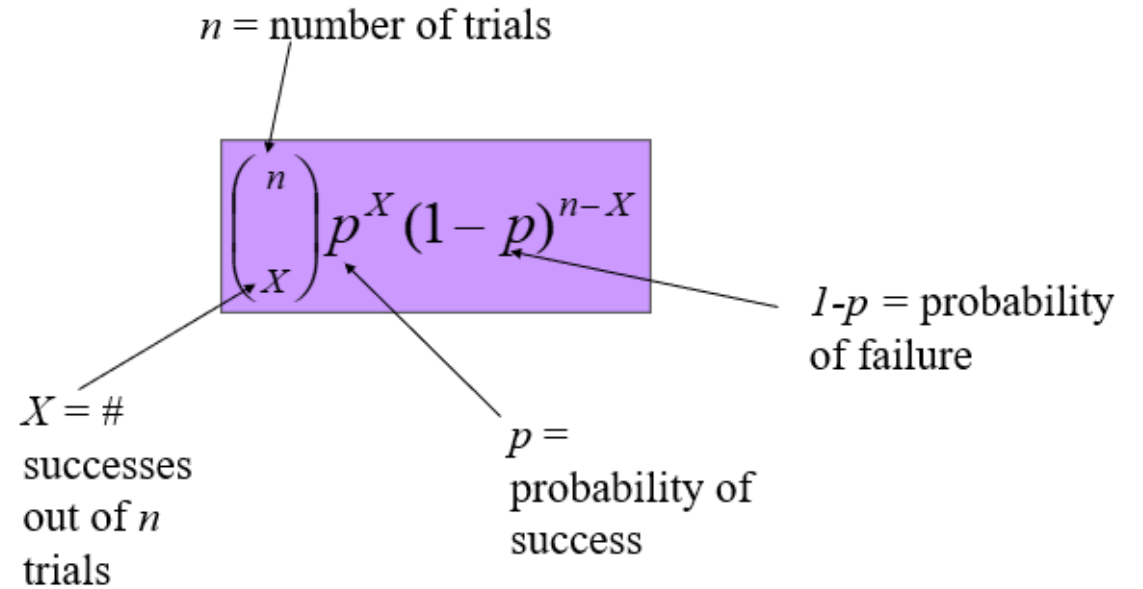
Formula for Binomial Distribution

$$P_x = \binom{n}{x} p^x q^{n-x}$$

OR

$$P(X=r) = {}^n C_r p^r q^{n-r}$$

Where P is success and
q is failure



Binomial Distribution: Illustration with example

- Consider a pen manufacturing company
- 10% of the pens are defective
- (i) Find the probability that **exactly 2 pens** are defective in a box of 12
- So $n=12$,
- $p=10\% = 10/100 = 1/10$
- $q = (1-p) = 90/100 = 9/10$
- $X=2$

$$P(X=r) = {}^n C_r p^r q^{n-r}$$

$$\begin{aligned} P[X=2] &= {}^n C_x p^x q^{n-x} \\ &= {}^{12} C_2 \left(\frac{1}{10}\right)^2 \left(\frac{9}{10}\right)^{10} \end{aligned}$$

- Consider a pen manufacturing company
- 10% of the pens are defective

- (i) Find the probability that **at least 2 pens** are defective in a box of 12
- So $n=12$,
- $p=10\% = 10/100 = 1/10$
- $q = (1-p) = 90/100 = 9/10$
- $X \geq 2$
- $P(X \geq 2) = 1 - [P(X < 2)]$
- $\quad = 1 - [P(X=0) + P(X=1)]$

Binomial distribution: Another example

- If I toss a coin 20 times, what's the probability of getting exactly 10 heads?

$$\binom{20}{10} (.5)^{10} (.5)^{10} = .176$$

The Binomial Distribution: another example

- Say 40% of the class is female.
- What is the probability that 6 of the first 10 students walking in will be female?

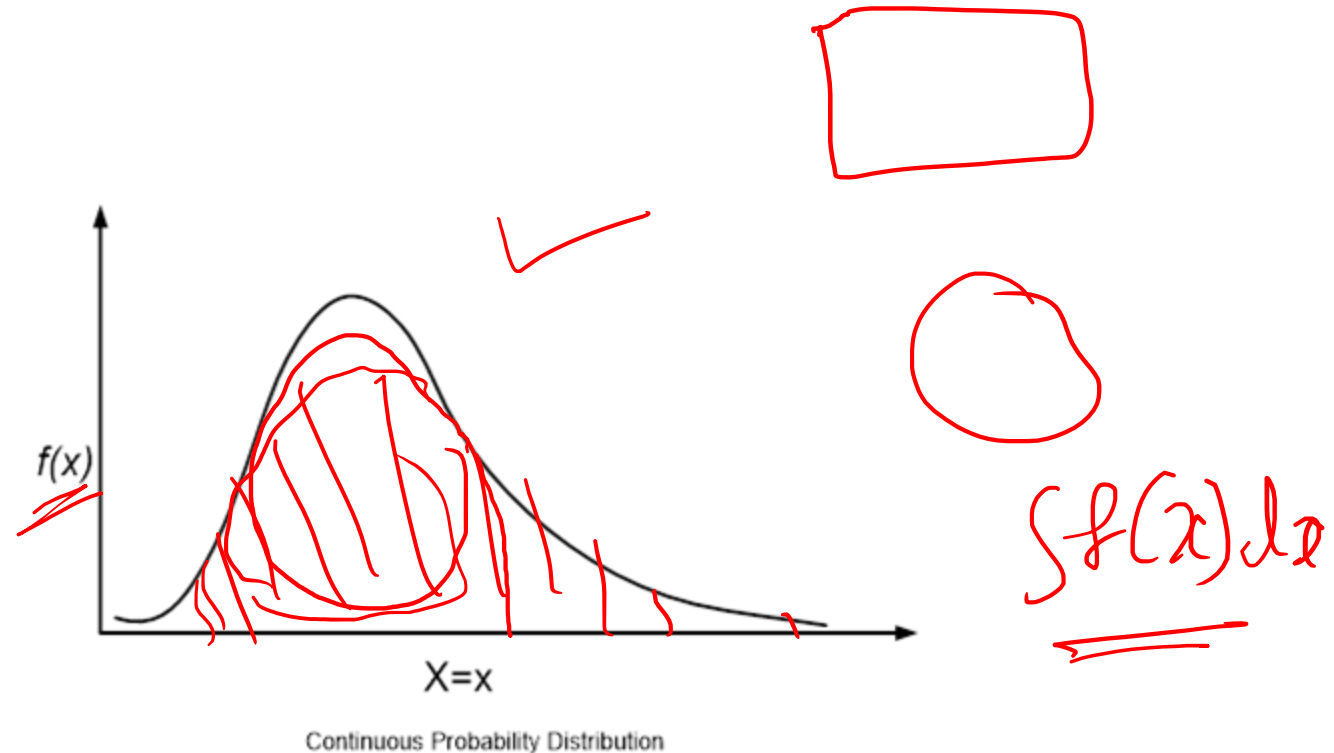
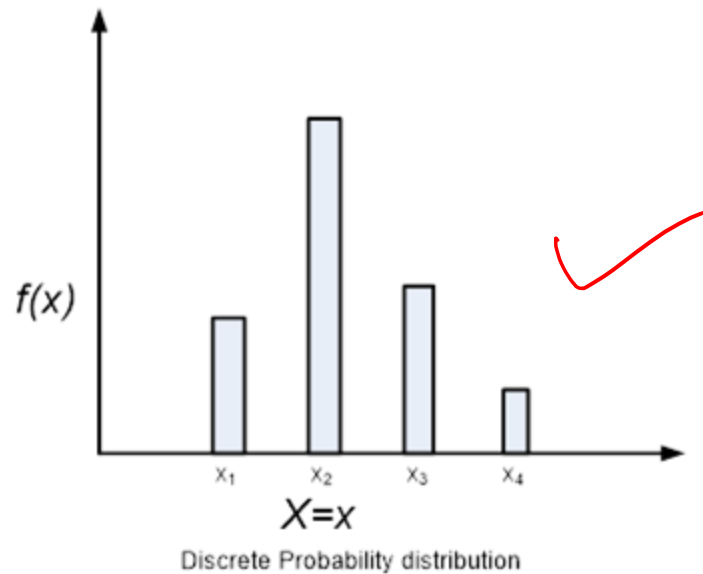
$$\begin{aligned}P(x) &= \binom{n}{x} p^x q^{n-x} \\&= \binom{10}{6} (.4^6) (.6^{10-6}) \\&= 210(.004096)(.1296) \\&= .1115\end{aligned}$$

Continuous Probability Distributions

- When the random variable of interest can take any value in an interval, it is called **continuous random variable**.
 - Every continuous random variable has **an infinite, uncountable number of possible values** (i.e., any value in an interval).
- **Examples** Temperature on a given day, Length, height, intensity of light falling on a given region.
- The length of time it takes a truck driver to go from New York City to Miami.
- The depth of drilling to find oil.
- The weight of a truck in a truck-weighing station.
- The amount of water in a 12-ounce bottle.

For each of these, if the variable is X , then $x > 0$ and less than some maximum value possible, but it can take on any value within this range

- Continuous random variable differs from discrete random variable. **Discrete random variables can take on only a finite number of values or at most a countable infinity of values.**
- A continuous random variable is described by **Probability density function.** This function is used to obtain the probability that the value of a continuous random variable is in the given interval.



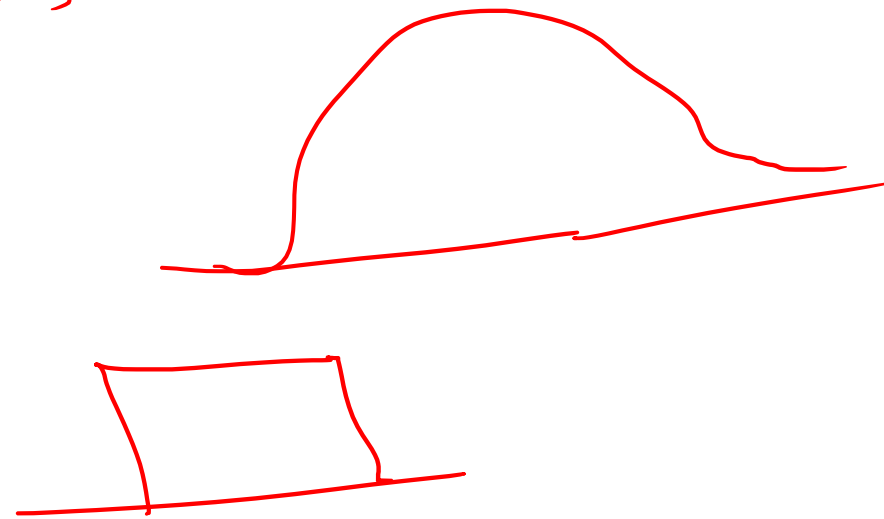
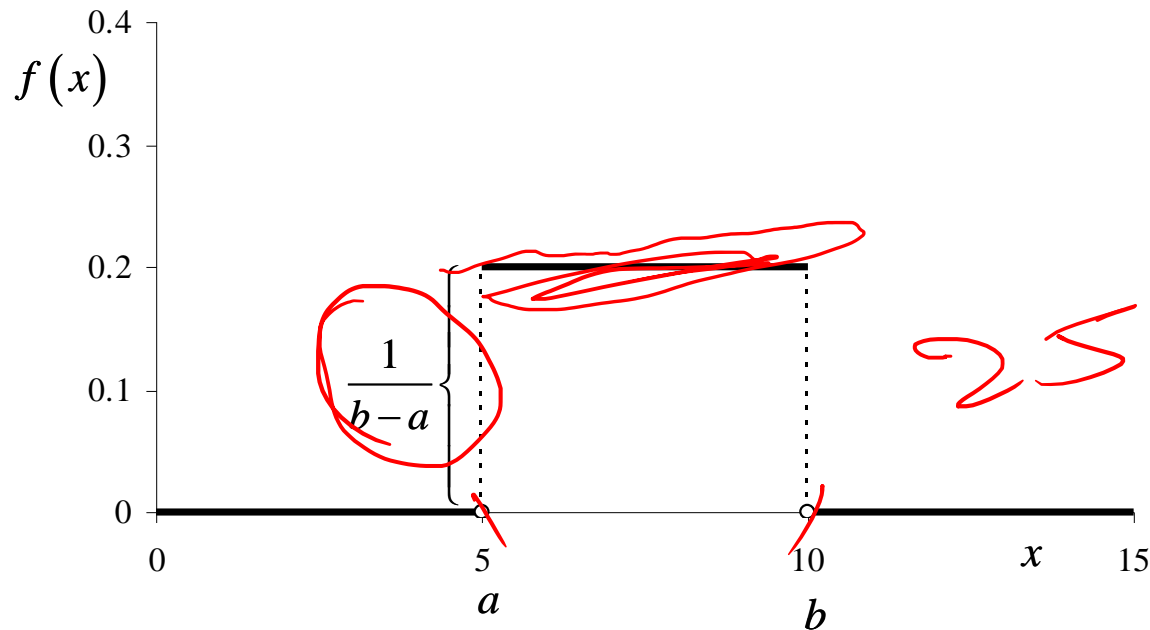
Continuous Uniform Distribution

- For Uniform distribution, $f(x)$ is constant over the possible value of x .
- Area looks like a rectangle.
- For the area in continuous distribution we need to do integration of the function.
- However in this case it is the area of rectangle.
- Example to time taken to wash the cloths in a washing machine. (for a standard condition)

Continuous Distributions

The Uniform distribution from a to b

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$



25

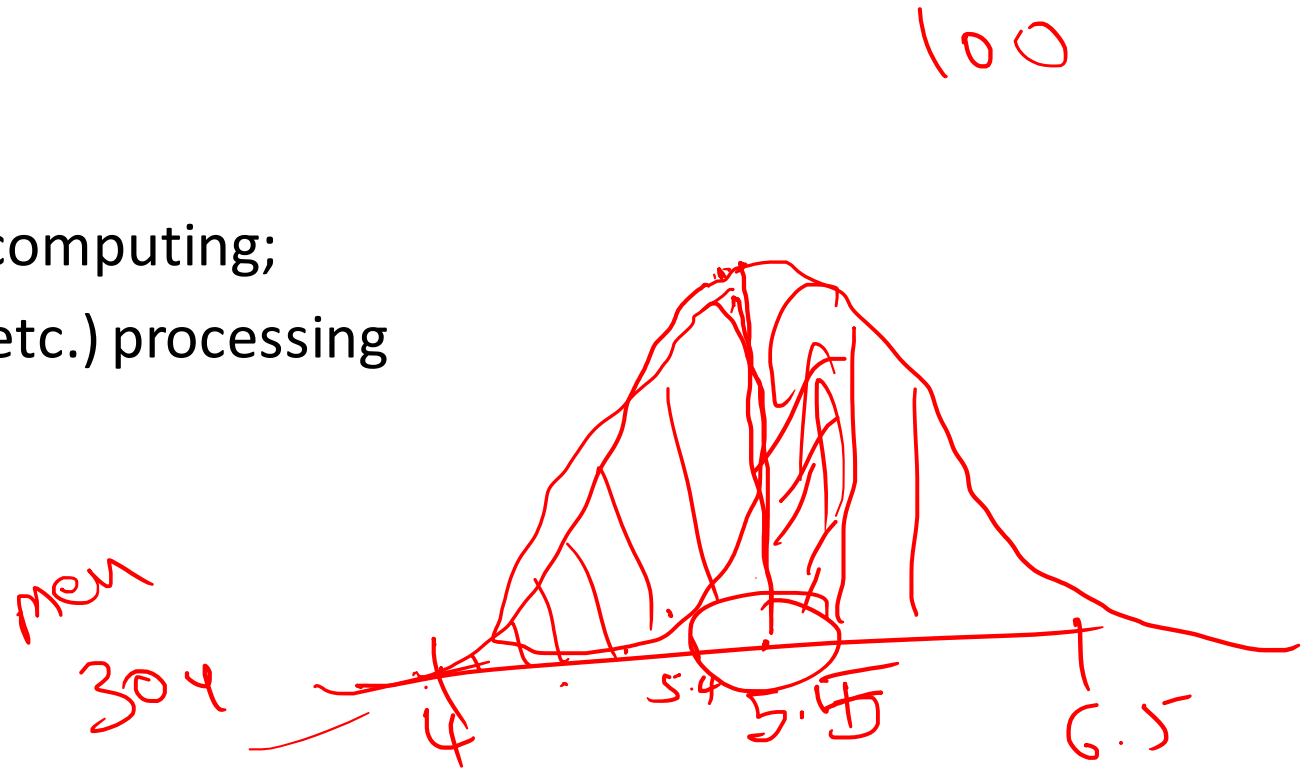
NORMAL DISTRIBUTION

- The most often used continuous probability distribution is the normal distribution; it is also known as Gaussian distribution.
- Its graph called the normal curve is the bell-shaped curve.
- Such a curve approximately describes many phenomenon occur in nature, industry and research.
 - Physical measurement in areas such as meteorological experiments, rainfall studies and measurement of manufacturing parts are often more than adequately explained with normal distribution.

NORMAL DISTRIBUTION Applications:

The normal (or Gaussian) distribution, is a very commonly used (occurring) function in the fields of probability theory, and has wide applications in the fields of:

- Pattern Recognition; ✓
- Machine Learning;
- Artificial Neural Networks and Soft computing;
- Digital Signal (image, sound , video etc.) processing
- Vibrations, Graphics etc.



The probability distribution of the normal variable depends upon the two parameters μ and σ

- **The parameter μ is called the mean or expectation of the distribution.**
- **The parameter σ is the standard deviation; and variance is thus σ^2 .**

- **Few terms:**

- **Mode: Repeated terms**
- **Median: middle data (if there are 9 data, the 5th one is the median)**
- **Mean: is the average of all the data points**
- **SD- standard Deviation, indicates how much the data is deviated from the mean.**

- Low SD indicates that all data points are placed close by
- High SD indicates that the data points are distributed and are not close by.

- **SD given by the formula (S)**
- **Where S is sample SD**

- **If you want population SD, represented by**

σ

and then divide by N not N-1

12

$$\sigma^2 = \sum_{x=1}^{x=n} (x_i - \bar{x})^2$$

10 11 12 12

\bar{x}

14, 15, 14

$$\sigma^2 = \frac{\sum_{x=1}^{x=n} (x_i - \bar{x})^2}{n}$$

S²
S²
S

$S_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$

- n = The number of data points
- x_i = Each of the values of the data
- \bar{x} = The mean of x_i

Normal Distribution Curve

99.7%
95%
68%

-3 -2 -1 1 2 3

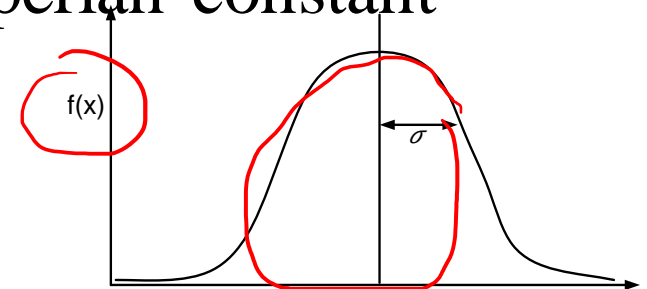
1000 (95D)

The probability distribution of the normal variable depends upon the two parameters μ and σ

- **The parameter μ is called the mean or expectation of the distribution.**
- **The parameter σ is the standard deviation; and variance is thus σ^2 .**
- **standard deviation** is a measure of the amount of variation or dispersion of a set of values.
- A low standard deviation indicates that the values tend to be close to the mean (expected value) of the set,
- a high standard deviation indicates that the values are spread out over a wider range.
- The density of the normal variable x with mean μ and variance σ^2 is

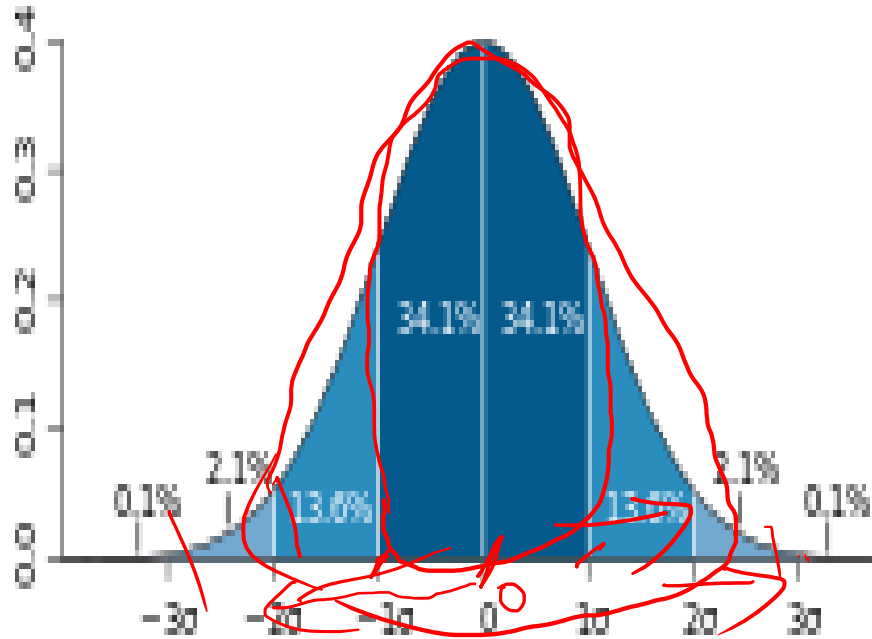
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty$$

where $\pi = 3.14159 \dots$ and $e = 2.71828 \dots$, the Naperian constant

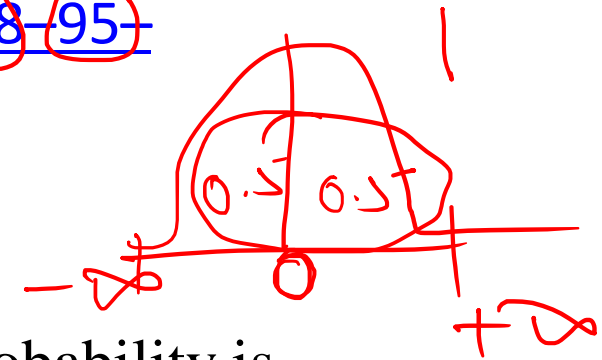


The Normal distribution
(mean μ , standard deviation σ)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



A plot of normal distribution (or bell-shaped curve) where each band has a width of 1 standard deviation – See also: 68 95 99.7 rule.



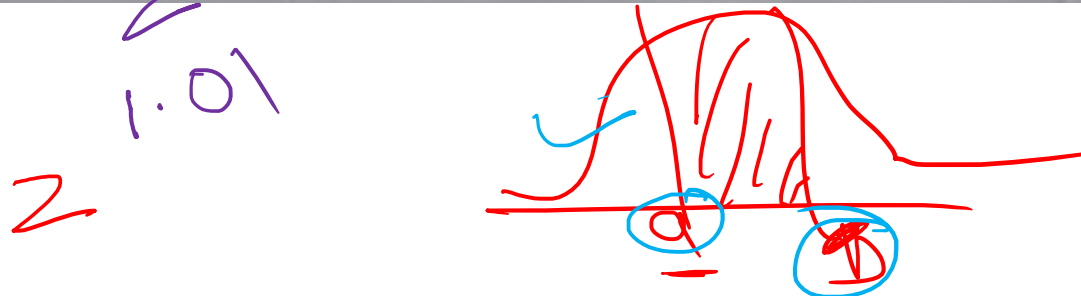
Standard Normal Distribution : In the above equation probability is computed for particular value of x . If you want a range then it has to be integrated.

For Standard Normal distribution:

- For standard normal distribution, the area under the given range is given by:

$$\int_a^b p(x) dx = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$
$$= C\left(\frac{b-\mu}{\sigma}\right) - C\left(\frac{a-\mu}{\sigma}\right)$$

Handwritten annotations: A red checkmark is next to the first integral. A red arrow points from the first integral to the second. A blue circle is drawn around $\frac{b-\mu}{\sigma}$ in the second equation, with a purple arrow pointing to it. A purple 'Z' is written below the circle. A purple 'SD' is written to the right of the second equation. A red arrow points from the second equation to the diagram below.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986
3.0	0.998650								
3.5	0.9998674								
4.0	0.99996833								
4.5	0.999996602								
5.0	0.9999997133								
5.5	0.99999998101								
6.0	0.999999999013								
6.5	0.9999999999588								
7.0	0.99999999999872								

$$C(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$



Figure 2.11: Areas under the standard normal curve

z	0.00	-0.01	-0.02	-0.03	-0.04	-0.05	-0.06	-0.07	-0.08	-0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	1.350×10^{-3}									
3.5	2.326×10^{-4}									
4.0	3.167×10^{-5}									
4.5	3.398×10^{-6}									
5.0	2.867×10^{-7}									
5.5	1.800×10^{-8}									
6.0	9.866×10^{-10}									
6.5	4.016×10^{-11}									
7.0	1.280×10^{-12}									

$$C(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$



Figure 2.12: Areas under the standard normal curve from $-\infty$ to z , where $z < 0$. For

Problem: Normal distribution

- Consider an electrical circuit in which the voltage is normally distributed with mean 120 and standard deviation of 3. What is the probability that the next reading will be between 119 and 121 volts? ✓

between 119 and 121

The z -transformation is

$$z = (x - 120)/3,$$

so we obtain from Figure 2.11 or 2.12

$$\begin{aligned} P(119 \leq x \leq 121) &= C\left(\frac{121 - 120}{3}\right) - C\left(\frac{119 - 120}{3}\right) \\ &= C(0.333) - C(-0.333) \\ &= 0.631 - (1 - 0.631) = 0.631 - 0.369 = 0.262 \end{aligned}$$

The value 0.631 was obtained by linear interpolation between $C(0.33)$ and $C(0.34)$

Handwritten notes and diagrams:

- 180
- 2300
- 270
- $a = 121$
- $b = 119$
- 121
- 120

1. Most graduate schools of business require applicants for admission to take the Graduate Management Admission Council's GMAT examination. Scores on the GMAT are roughly normally distributed with a mean of 527 and a standard deviation of 112. What is the probability of an individual scoring above 500 on the GMAT?

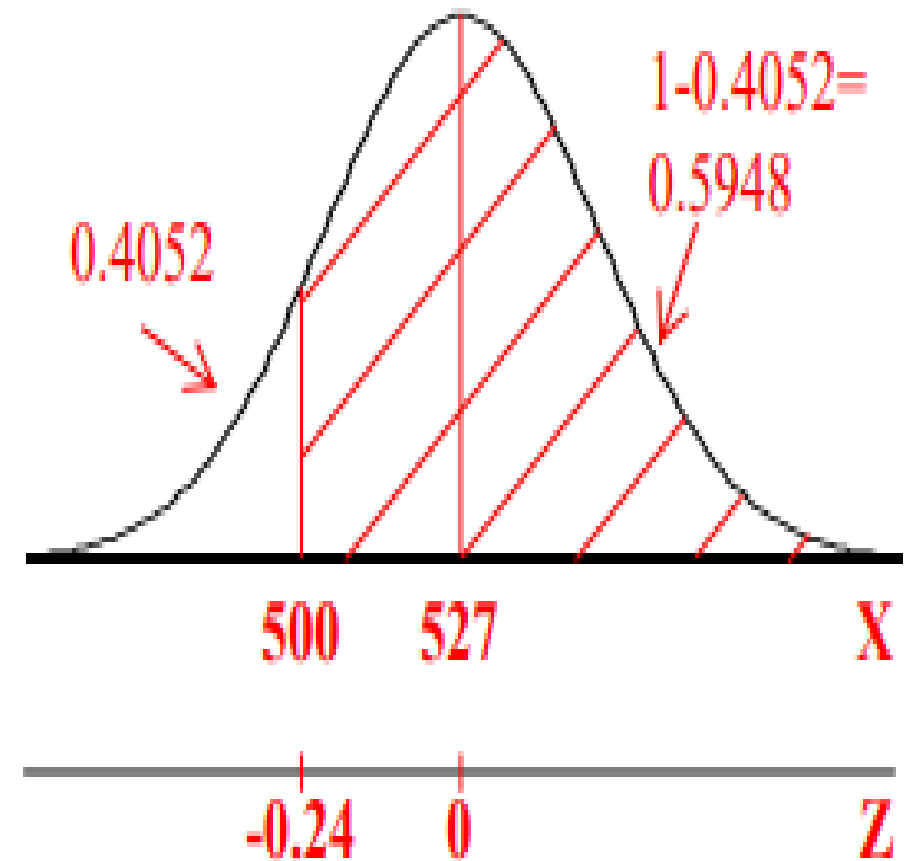
Normal Distribution

$$Z = \frac{500 - 527}{112} = -0.24107$$

$$\mu = 527$$

$$\sigma = 112$$

$$\Pr\{X > 500\} = \Pr\{Z > -0.24\} = 1 - 0.4052 = \boxed{0.5948}$$



Another problem

5. The average number of acres burned by forest and range fires in a large New Mexico county is 4,300 acres per year, with a standard deviation of 750 acres. The distribution of the number of acres burned is normal. What is the probability that between 2,500 and 4,200 acres will be burned in any given year?

Normal Distribution $Z = \frac{2500 - 4300}{750} = -2.40$

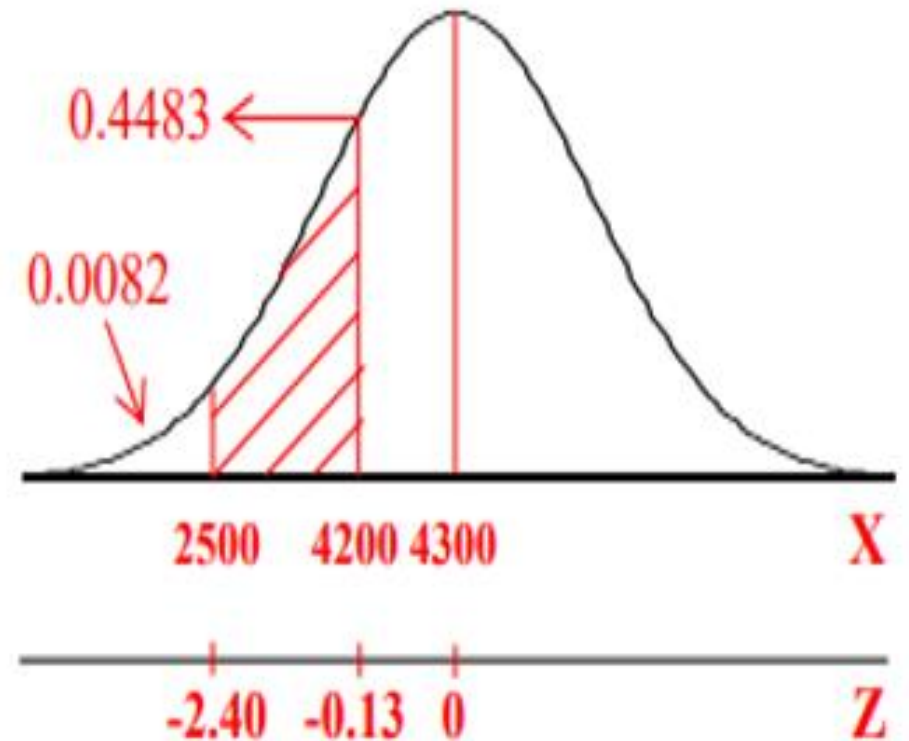
$\mu = 4300$ $Z = \frac{4200 - 4300}{750} = -0.13333$

$\sigma = 750$

$P(2500 < X < 4200) = P(-2.40 < Z < -0.13)$

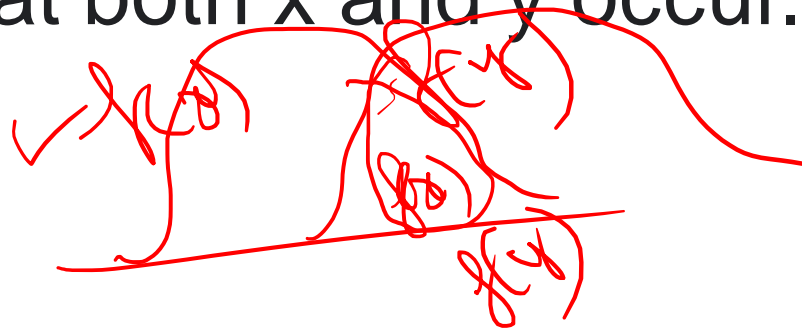
$P(-2.40 < Z < -0.13) = P(Z < -0.13) - P(Z < -2.40)$

$P(-2.40 < Z < -0.13) = 0.4483 - 0.0082 = \boxed{0.4401}$



Joint Distributions and Densities

- The joint random variables (x,y) signifies that , simultaneously, the first feature has the value x and the second feature has the value y .
- If the random variables **x and y are discrete**, the joint distribution function of the joint random variable (x,y) is the probability of $P(x,y)$ that both x and y occur.



Joint distribution in continuous random variable

- If x and y are continuous, then the probability density function is used over the region R , where x and y is applied is used.
- It is given by:

$$P((x, y) \text{ is in } R) = \iint_R p(x, y) dx dy,$$

- Where the integral is taken over the region R . This integral represents a volume in the xyp plane.

Moments of Random Variables

Moments are very useful in statistics because they tell us much about our data.

- In mathematics, the **moments** of a function are quantitative measures related to the shape of the function's graph.
- It gives information about the spread of data, skewedness and kurtosis.

- If the function is a probability distribution, then there are four commonly used moments in statistics

- ✓ The first moment is the expected value - measure of center of the data
- ✓ The second central moment is the variance - spread of our data about the mean
- ✓ The third standardized moment is the skewness - the shape of the distribution
- ✓ The fourth standardized moment is the kurtosis - measures the peakedness or flatness of the distribution.

Computing Moments for population

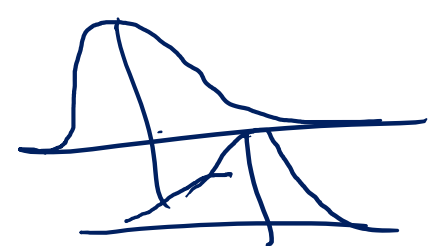
Higher Order Moments

✓ 1	$\frac{\sum x}{n}$	CENTRED
✓ 2	$\frac{\sum x^2}{n}$	$\frac{\sum (x - \mu)^2}{n}$
✓ 3	$\frac{\sum x^3}{n}$	$\frac{\sum (x - \mu)^3}{n}$
✓ 4	$\frac{\sum x^4}{n}$	$\frac{\sum (x - \mu)^4}{n}$

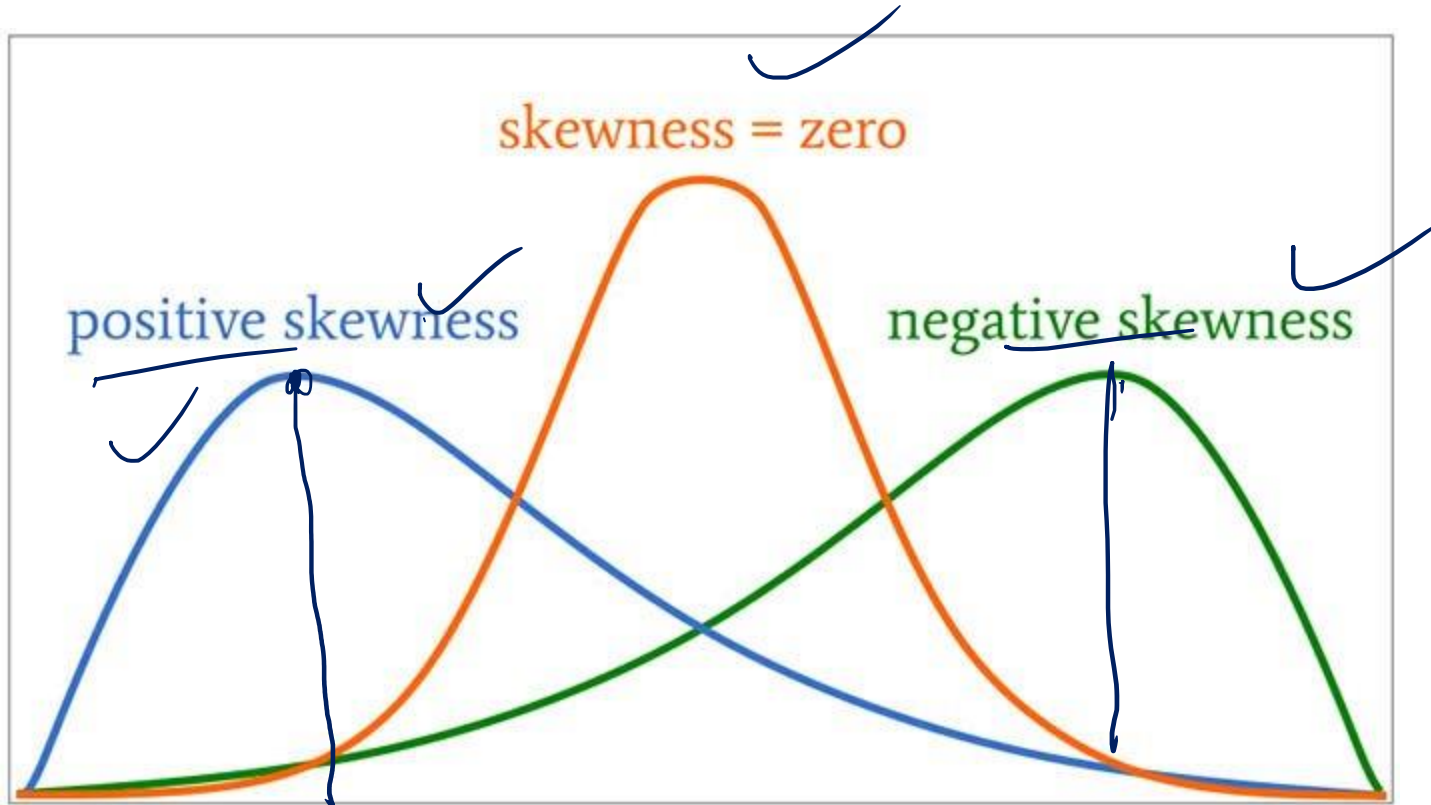
The image shows a slide titled "Higher Order Moments" with a list of formulas for population moments. The first row shows the first moment $\frac{\sum x}{n}$ and the word "CENTRED" with a downward arrow. The second, third, and fourth rows show the second, third, and fourth moments, respectively, with their uncentred and centred forms. Handwritten blue annotations include checkmarks next to each row, a large bracket grouping the centred formulas, and a circle around the number "2" in the second row with the word "variance" written next to it. To the right of the slide, there are two hand-drawn diagrams: the top one shows a bell curve with a mean line and a shaded area under the curve, and the bottom one shows a skewed distribution curve with a mean line.



variance
1 2 3 4 14
1 1 1 1 14
1 1 1 1 14



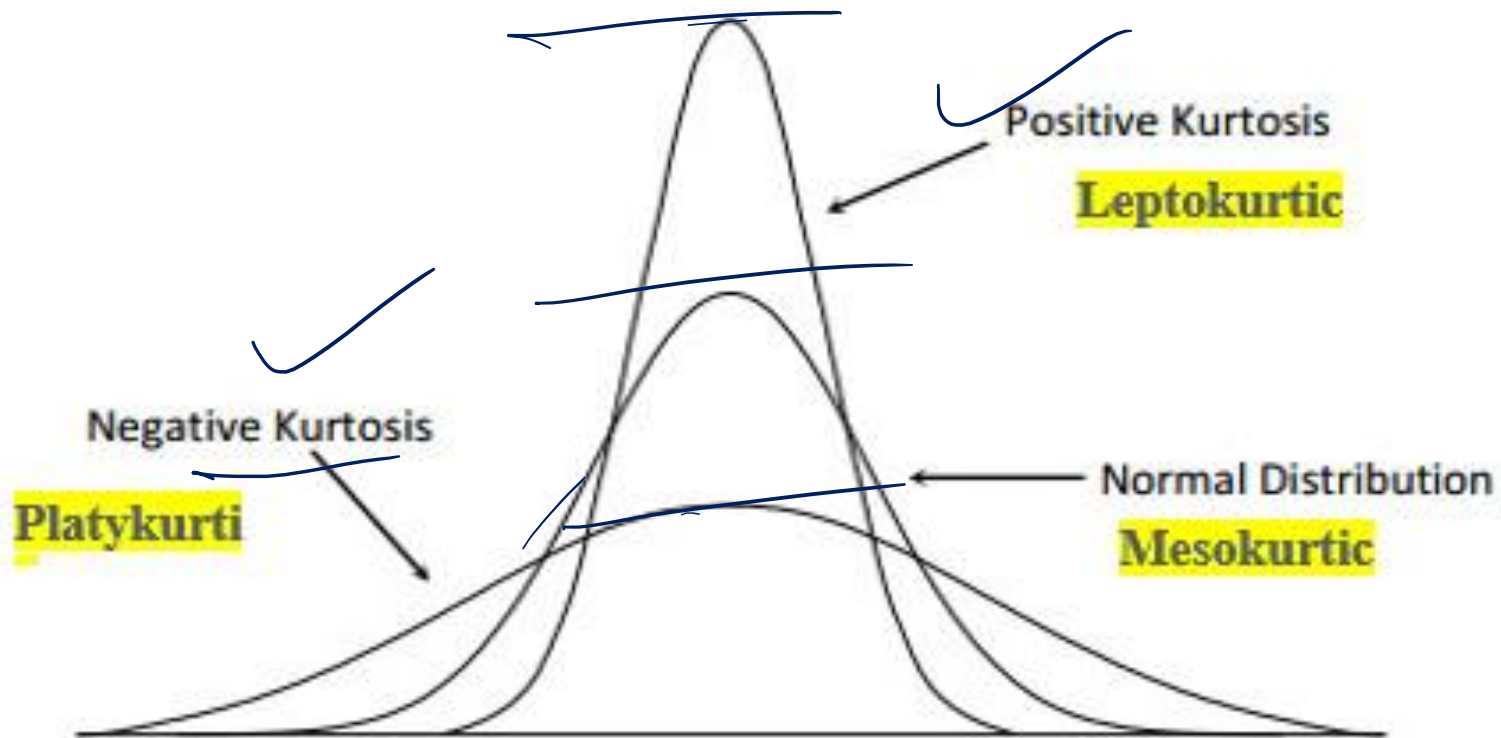
Moment 3: To know the Skewness



In positive Skewness,
Mean is $>$ median
and
Median $>$ mode

And it is reverse in case
of $-ve$ skewness

Moment 4 : To know the Kurtosis

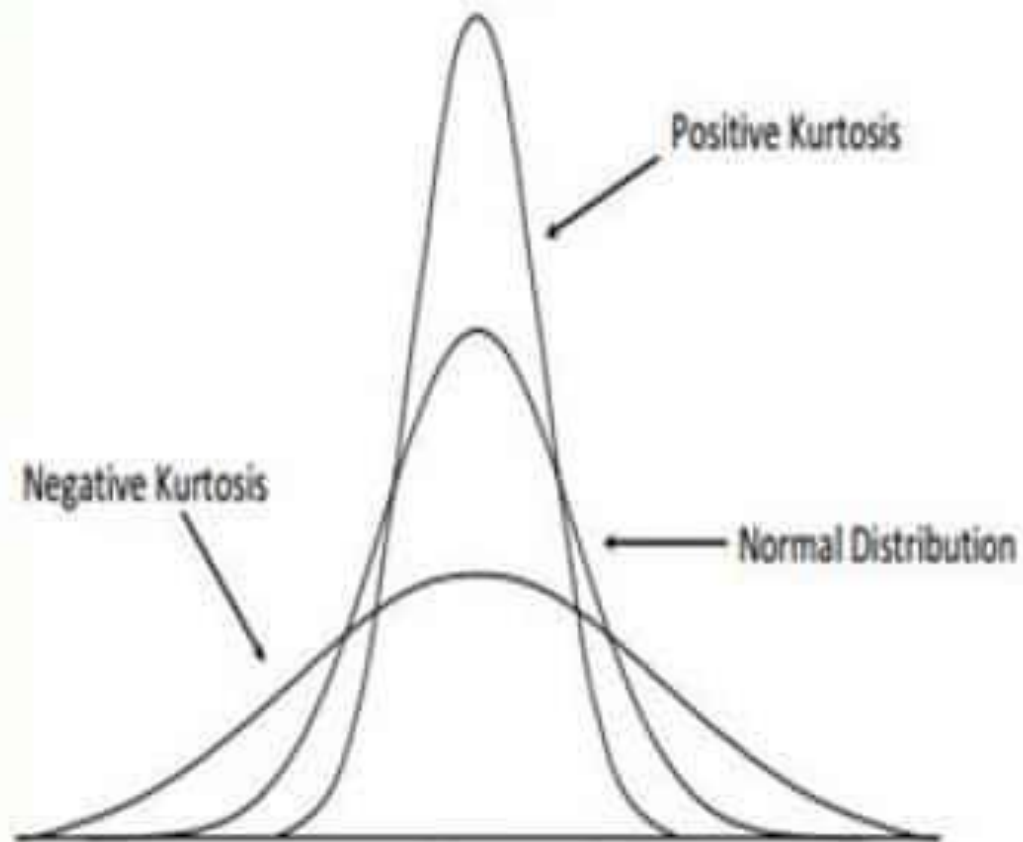
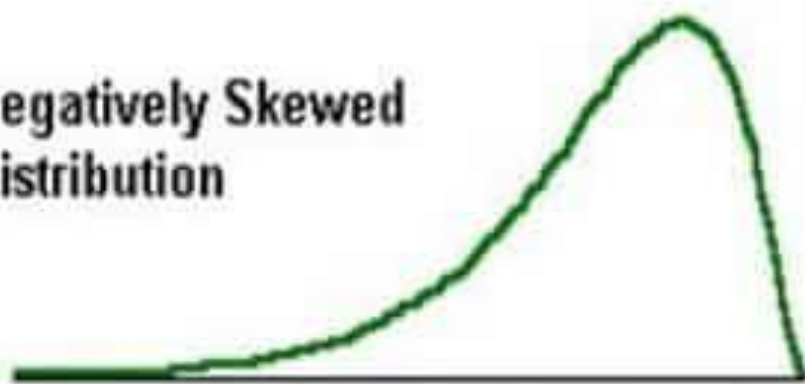


Skewness

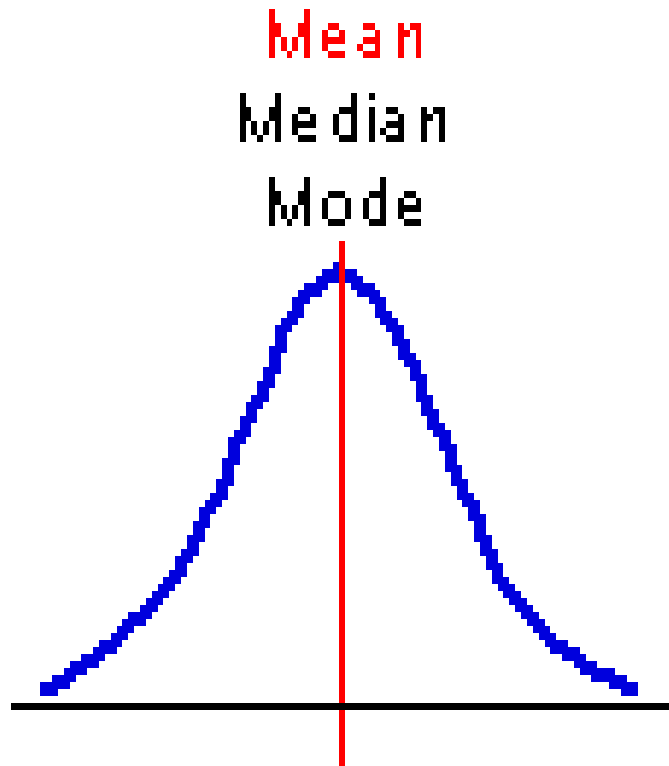
Kurtosis



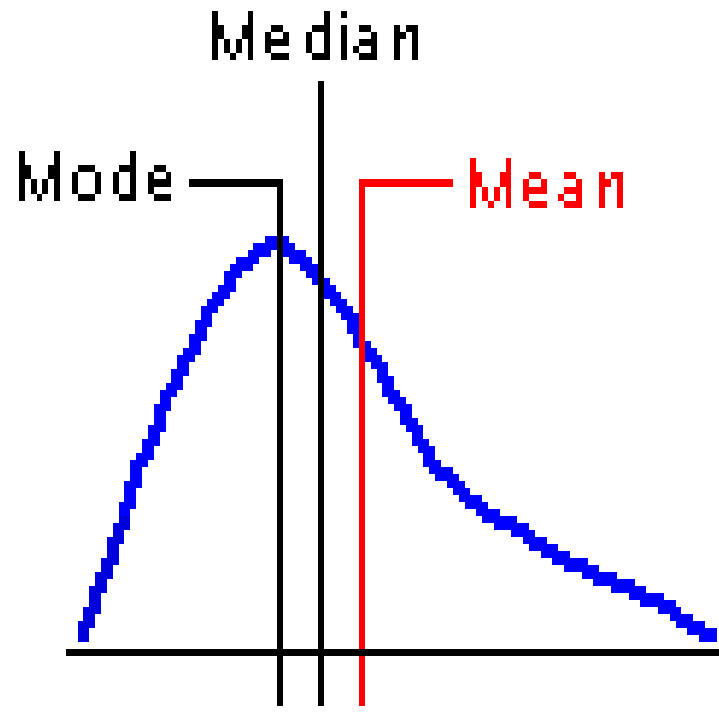
(-) Negatively Skewed Distribution



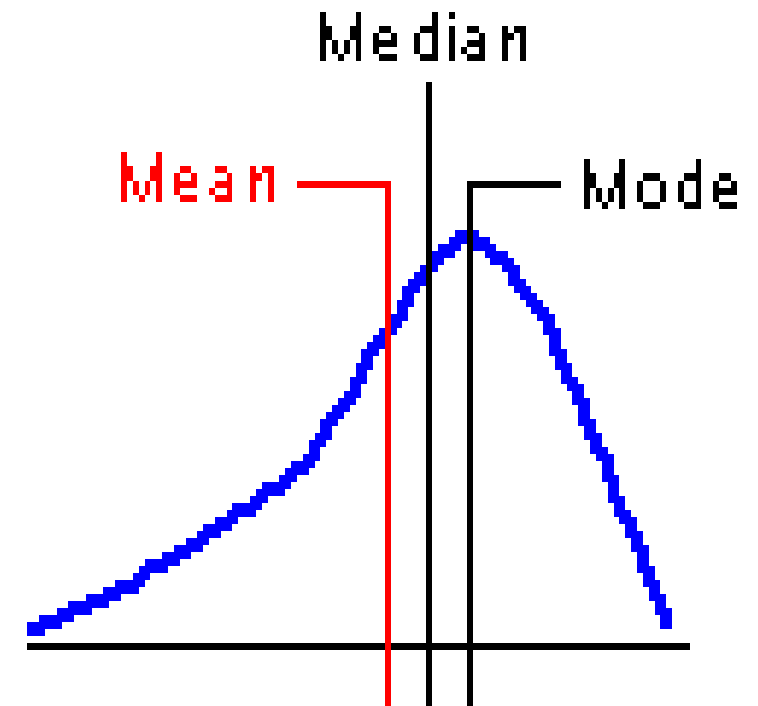
D



Symmetrical
Distribution



Positive
Skew

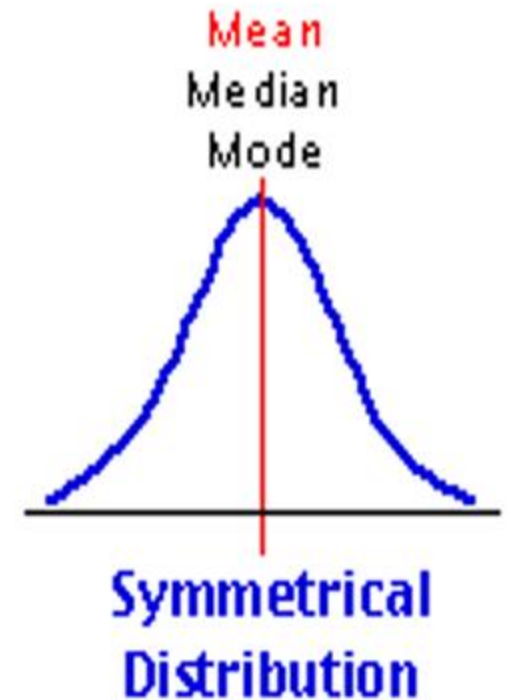


Negative
Skew

Normal Distribution

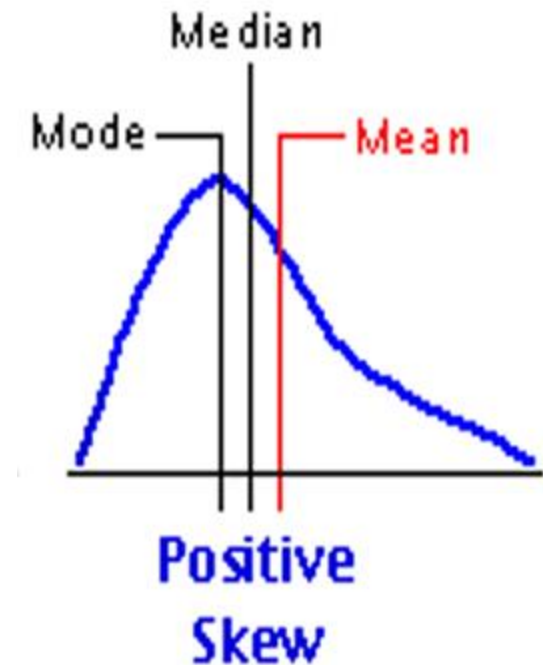
- Consider an example of x values:
- 4,5,5,6,6,6,7,7,8
- Mode, Median and mean all will be equal

- = Mode is 6
- = Median is 6
- = Mean is also 6



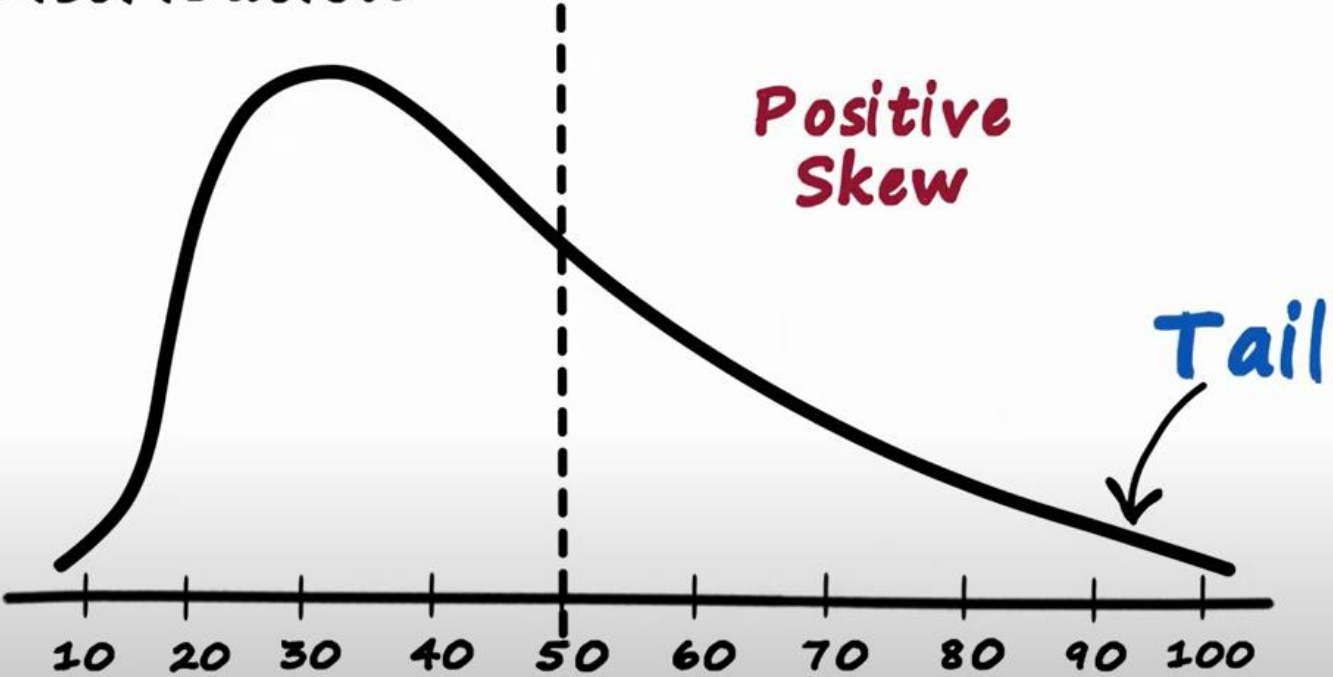
Positive Skew

- Consider an example of x values:
- 5,5,5,6,6,7,8,9,10
- (It is an example for Normal Distribution)
- = Mode is 5
- = Median is 6
- = Mean is also 6.8



+ve skew

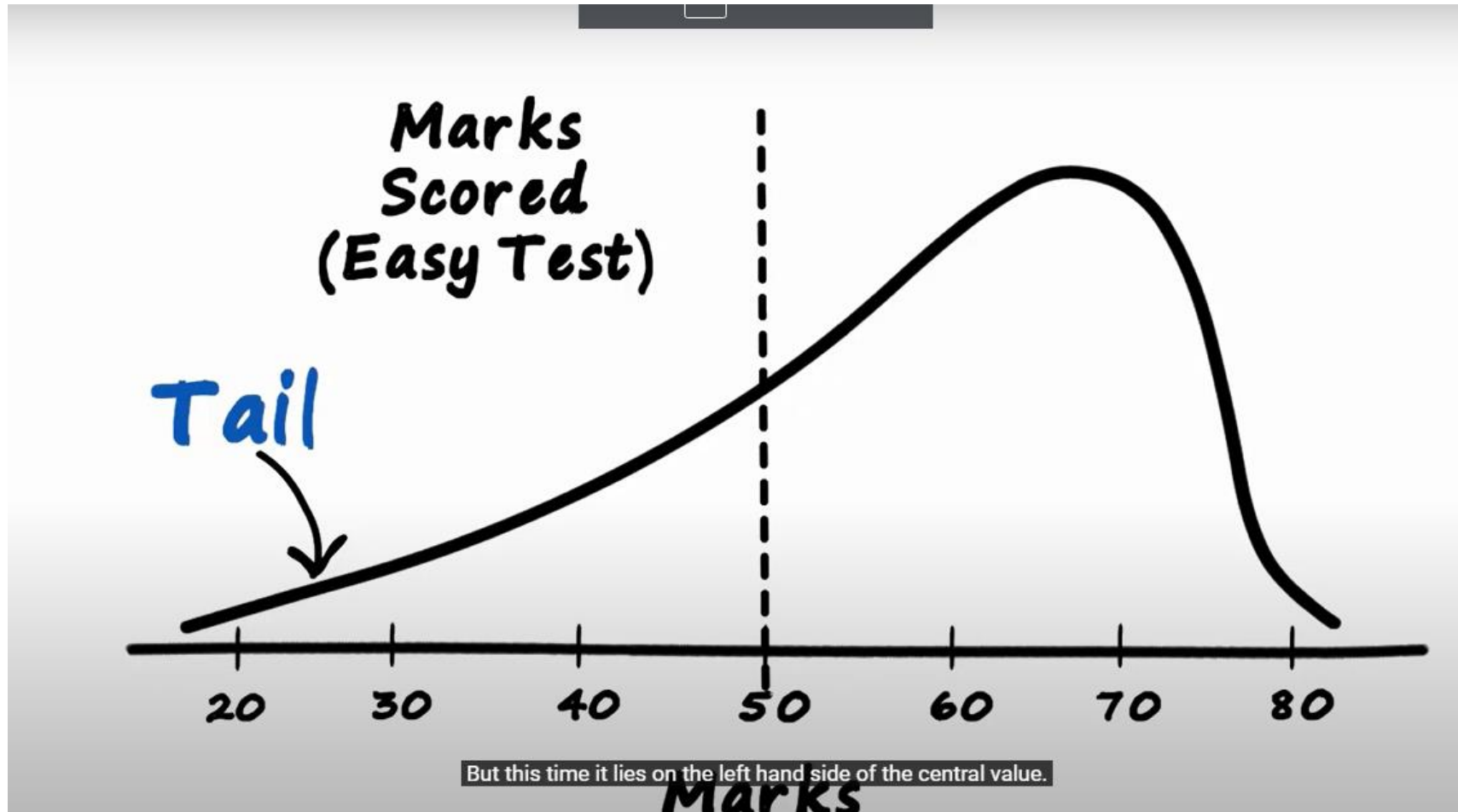
Income Distribution



It is not equally distributed on both sides of the Central value.

income (in 100's)

-ve skew



Difference between PDF and PMF

1. The full form of PDF is Probability Density Function whereas the full form of PMF is Probability Mass Function
 2. PMF is used when there is a need to find a solution in a range of discrete random variables whereas PDF is used when there is a need to find a solution in a range of continuous random variables.
 3. PDF uses continuous random variables whereas PMF uses discrete random variables.
 4. Pdf formula is $F(x) = P(a < x < b) = \int_a^b f(x)dx > 0$ whereas pmf formula is $p(x) = P(X=x)$
 5. The solutions of PDF falls in the radius of continuous random variables whereas the solutions of PMF falls in the radius
-

Moments for random variable:

- The “moments” of a random variable (or of its distribution) are expected values of powers or related functions of the random variable.

Formula for Computing Kth Central moment of Random variable

The k^{th} **central moment** of X

$$\mu_k^0 = E \left[(X - \mu)^k \right]$$

$$= \begin{cases} \sum_x (x - \mu)^k p(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x - \mu)^k f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

Let X be a discrete random variable having support $x = \langle 1, 2 \rangle$ and the pmf is

$$p_X(x) = \begin{cases} 3/4 & \text{if } x = 1 \\ 1/4 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases} \quad \text{Using this compute mean (first order moment)}$$


First order moment is the mean.

Solution:

$$\begin{aligned} \mu_X(1) &= E[X] \\ &= \sum_{x \in \mathcal{R}_X} p_X(x)x \\ &= \frac{3}{4} \cdot 1 + \frac{1}{4} \cdot 2 \\ &= \frac{5}{4} \end{aligned}$$

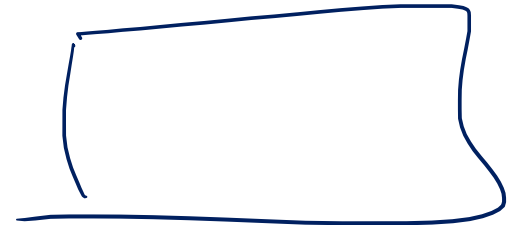
For example computation of 3rd Order moment

- The third central moment of can be computed as follows:
- Here X value is 1 and 2 and Probability is $\frac{3}{4}$ and $\frac{1}{4}$ respectively. Consider Mean is $\frac{5}{4}$

$$\begin{aligned}\mu_X(3) &= E[(X - E[X])^3] \\ &= \sum_{x \in R_X} p_X(x) \left(x - \frac{5}{4}\right)^3 \\ &= \frac{3}{4} \cdot \left(1 - \frac{5}{4}\right)^3 + \frac{1}{4} \cdot \left(2 - \frac{5}{4}\right)^3 \\ &= \frac{3}{4} \cdot \left(-\frac{1}{4}\right)^3 + \frac{1}{4} \cdot \left(\frac{3}{4}\right)^3 \\ &= -\frac{3}{4^4} + \frac{27}{4^4} \\ &= \frac{24}{256} = \frac{3}{32}\end{aligned}$$


✓ Estimation of Parameters from Samples

- There are 3 kinds of estimates for these parameters:
 - Method of moments estimates ✓
 - Maximum likelihood estimates ✓
 - Unbiased estimates. ✓



End of Unit 1

Unit - 2
Pattern Recognition
Statistical Decision Making
Dr. Srinath. S

Syllabus for Unit - 2

- Statistical Decision Making:
- Introduction, Bayes' Theorem
- Conditionally Independent Features
- Decision Boundaries

Classification (Revision)

It is the task of assigning a class label to an input pattern. The class label indicates one of a given set of classes. The classification is carried out with the help of a model obtained using a learning procedure. There are two categories of classification. **supervised learning** and **unsupervised learning**.

- **Supervised learning** makes use of a set of examples which already have the class labels assigned to them.
- **Unsupervised learning** attempts to find inherent structures in the data.
- **Semi-supervised learning** makes use of a small number of labeled data and a large number of unlabeled data to learn the classifier.

Learning - Continued

- The classifier to be designed is built using input samples which is a mixture of all the classes.
- The classifier learns how to discriminate between samples of different classes.
- If the Learning is offline i.e. Supervised method then, the classifier is first given a set of training samples and the optimal decision boundary found, and then the classification is done.
- Supervised Learning refers to the process of designing a pattern classifier by using a Training set of patterns to assign class labels.
- If the learning involves no teacher and no training samples (Unsupervised). The input samples are the test samples itself. The classifier learns from the samples and classifies them at the same time.

Statistical / Parametric decision making

This refers to the situation in which we assume the general form of probability distribution function or density function for each class.

- Statistical/Parametric Methods **uses a fixed number of parameters to build the model.**
- Parametric methods are assumed to be a normal distribution.
- Parameters for using the normal distribution is –
 - Mean
 - Standard Deviation
- For each feature, we first estimate the mean and standard deviation of the feature for each class.

Statistical / Parametric decision making (Continued)

- If a group of features – multivariate normally distributed, estimate mean and standard deviation and covariance.
- Covariance is a measure of the relationship between two random variables, in statistics.
- The covariance indicates the relation between the two variables and helps to know if the two variables vary together. (To find the relationship between two numerical variable)
- In the covariance formula, the covariance between two random variables X and Y can be denoted as $\text{Cov}(X, Y)$.
- x_i is the values of the X-variable
- y_j is the values of the Y-variable
- \bar{x} is the mean of the X-variable
- \bar{y} is the mean of the Y-variable
- N is the number of data points

$$\text{Cov}(X, Y) = \frac{\sum (X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

Positive and negative covariance

- Positive Co variance: If temperature goes high sale of ice cream also goes high. This is positive covariance. Relation is very close.



- On the other hand cold related disease is less as the temperature increases. This is negative covariance.



- No co variance : Temperature and stock market links



Example: Two set of data X and Y

Day	x	y
1	30	5
2	35	8
3	40	8
4	25	4
5	35	5
Mean	33	6

Compute $x - \bar{x}$ and $y - \bar{y}$

Day	x	y	$x - \bar{x}$	$y - \bar{y}$
1	30	5	-3	-1
2	35	8	+2	+2
3	40	8	+7	+2
4	25	4	-8	-2
5	35	5	+2	-1
Mean	33	6		

Apply Covariance formula

$$\text{Cov}(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

Day	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
1	30	5	-3	-1	3
2	35	8	+2	+2	4
3	40	8	+7	+2	14
4	25	4	-8	-2	16
5	35	5	+2	-1	-2
Mean	33	6			Sum = 35

- Final result will be $35/5 = 7 =$ is a positive covariance

Statistical / Parametric Decision making - continued

- Parametric Methods can perform well in many situations but its performance is at peak (top) when the spread of each group is different.
- Goal of most classification procedures is to estimate the probabilities that a pattern to be classified belongs to various possible classes, based on the values of some feature or set of features.

Ex1: To classify the fish on conveyor belt as salmon or sea bass

Ex2: To estimate the probabilities that a patient has various diseases given some symptoms or lab tests. (Use laboratory parameters).

Ex3: Identify a person as Indian/Japanese based on statistical parameters like height, face and nose structure.

- In most cases, we decide which is the **most likely class**.
- We need a **mathematical decision making algorithm**, to obtain classification or decision.

Bayes Theorem

When the joint probability, $P(A \cap B)$, is hard to calculate or if the inverse or Bayes probability, $P(B|A)$, is easier to calculate then **Bayes theorem** can be applied.

Revisiting conditional probability

Suppose that we are interested in computing the probability of event A and we have been told event B has occurred.

Then the conditional probability of A given B is defined to be:

$$P[A|B] = \frac{P[A \cap B]}{P[B]} \quad \text{if } P[B] \neq 0$$

Similarly,

$$P[B|A] = \frac{P[A \cap B]}{P[A]} \quad \text{if } P[A] \text{ is not equal to } 0$$



- Original Sample space is the red coloured rectangular box.
- What is the probability of A occurring given sample space as B.
- Hence $P(B)$ is in the denominator.
- And area in question is the intersection of A and B

$$P[A|B] = \frac{P[A \cap B]}{P[B]} \quad \text{and} \quad P[B|A] = \frac{P[A \cap B]}{P[A]}$$

From the above expressions, we can rewrite

$$P[A \cap B] = P[B].P[A|B]$$

$$\text{and} \quad P[A \cap B] = P[A].P[B|A]$$

This can also be used to calculate $P[A \cap B]$

So

$$P[A \cap B] = P[B].P[A|B] = P[A].P[B|A]$$

or

$$P[B].P[A|B] = P[A].P[B|A]$$

$$P[A|B] = P[A].P[B|A] / P[B] \quad - \text{ Bayes Rule}$$

Bayes Theorem

Intersection or probability
of both events occurring

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

Conditional Probability

Likelihood

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

Posterior Probability

Predictor Prior

Class Prior

Bayes Theorem

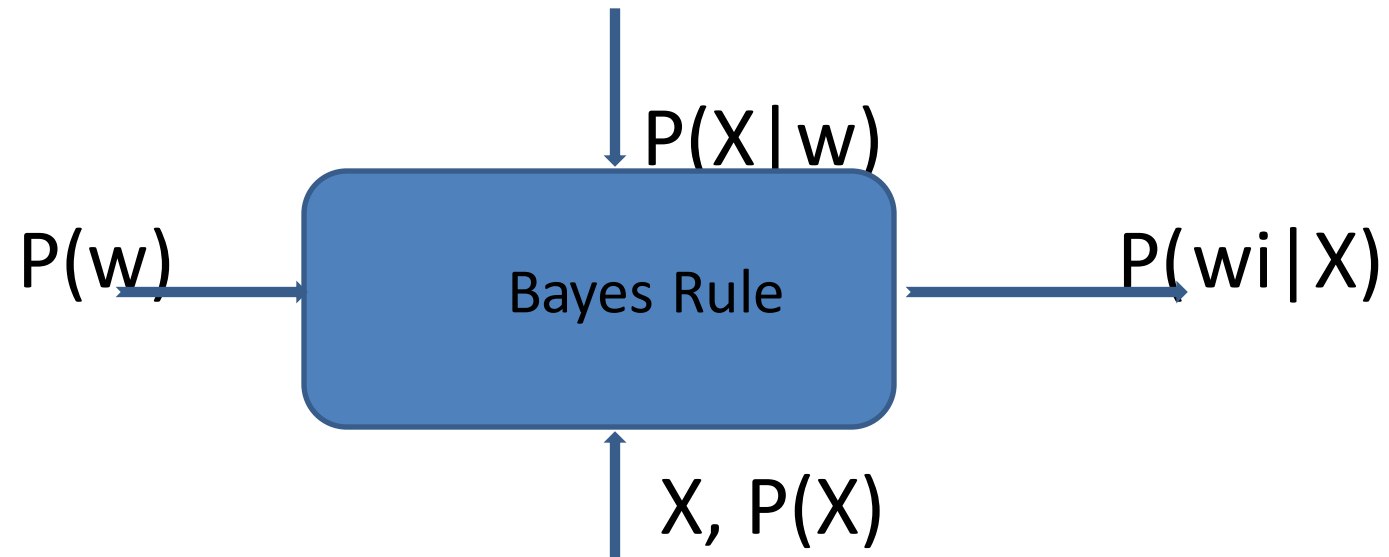
- **Class prior or prior probability:** probability of event A occurring before knowing anything about event B.
- **Predictor prior or evidence:** same as class prior but for event B.
- **Posterior probability:** probability of event A after learning about event B.
- **Likelihood:** reverse of the posterior probability.

Bayes Theorem:

$$P(\omega_j / x) = \frac{p(x / \omega_j)P(\omega_j)}{p(x)} = \frac{\textit{likelihood} \times \textit{prior}}{\textit{evidence}}$$

The goal is to measure: $P(w_i | X)$

Measured-conditioned or posteriori probability, from the above three values.



This is the Prob. of any vector X being assigned to class w_i .

Example for Bayes Rule/ Theorem

- Given Bayes' Rule :

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

The diagram illustrates the components of Bayes' Theorem. The equation $P(A | B) = \frac{P(B | A) P(A)}{P(B)}$ is centered. Four arrows point from the terms in the equation to their respective labels: 'Likelihood' points to $P(B | A)$, 'Class Prior' points to $P(A)$, 'Posterior Probability' points to $P(A | B)$, and 'Predictor Prior' points to $P(B)$.

Bayes Theorem

Example1:

- Compute : Probability in the deck of ca

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

Likelihood ← $P(B | A)$ $P(A)$ → Class Prior
↓ $P(A | B)$ ↓ $P(B)$ → Predictor Prior
Posterior Probability

Bayes Theorem

- Probability of (King/Face)
- It is given by $P(\text{King/Face}) = P(\text{Face/King}) * P(\text{King}) / P(\text{Face})$
 $= 1 * (4/52) / (12/52)$
 $= 1/3$

Example2:

Cold (C) and not-cold (C'). Feature is fever (f).

Prior probability of a person having a cold, $P(C) = 0.01$.

Prob. of having a fever, given that a person has a cold is, $P(f|C) = 0.4$.

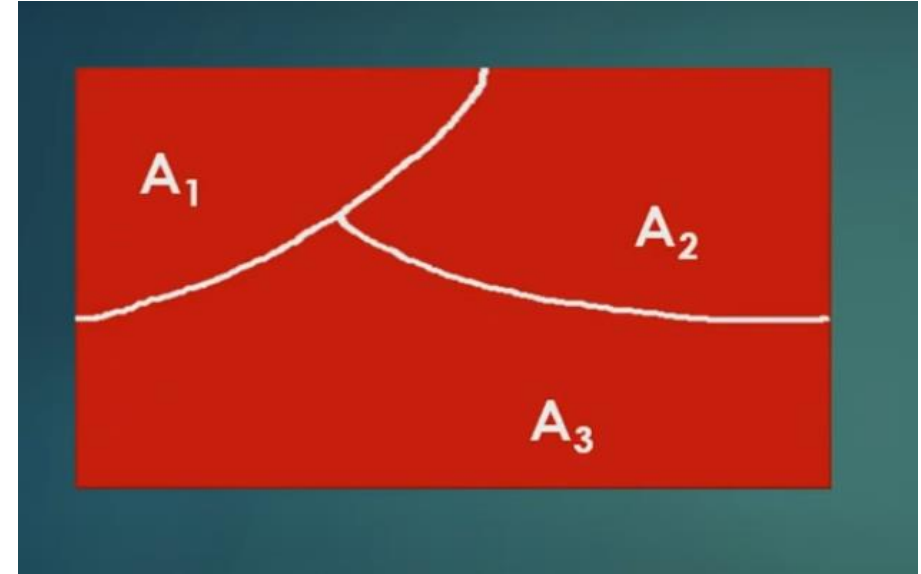
Overall prob. of fever $P(f) = 0.02$.

Then using Bayes Th., the Prob. that a person has a cold, given that she (or he) has a fever is:

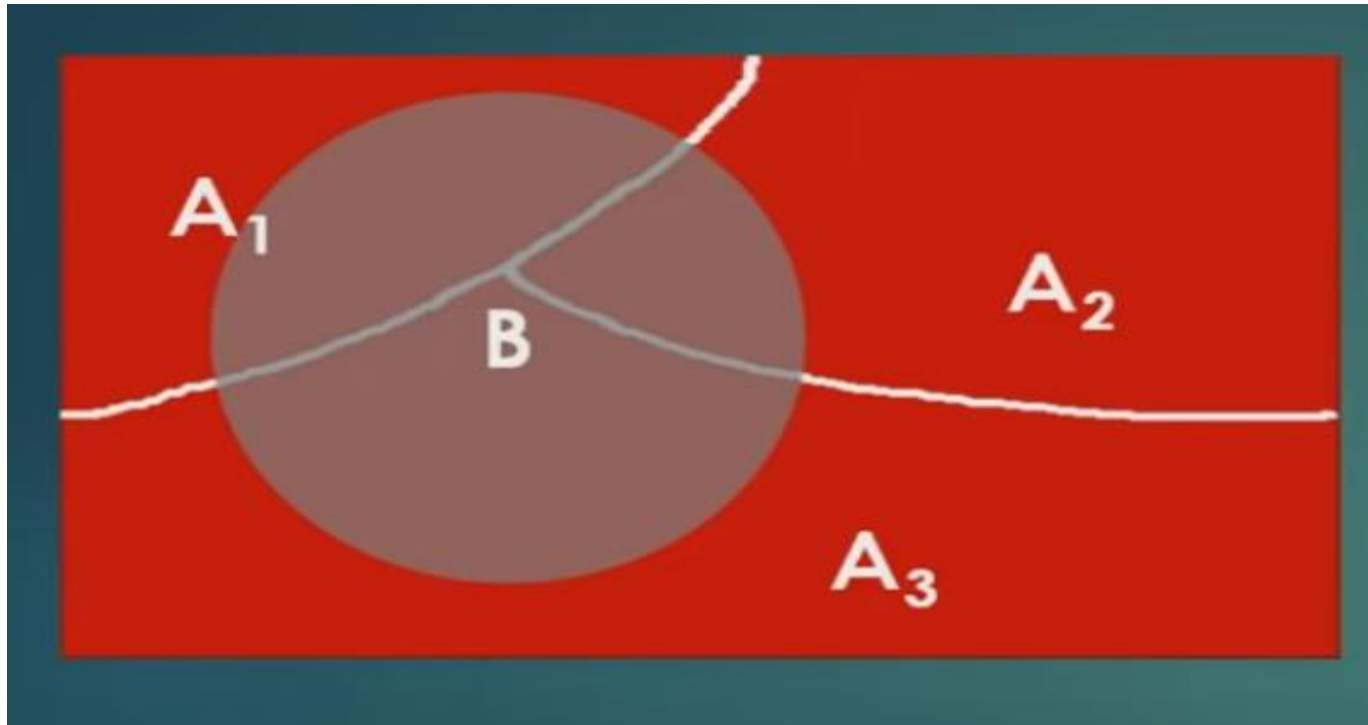
$$P(C|f) = \frac{P(f|C) P(C)}{P(f)} = \frac{0.4 * 0.01}{0.02} = 0.2$$

Generalized Bayes Theorem

- Consider we have 3 classes A_1 , A_2 and A_3 .
- Area under Red box is the sample space
- Consider they are mutually exclusive and collectively exhaustive.
- Mutually exclusive means, if one event occurs then another event cannot happen.
- Collectively exhaustive means, if we combine all the probabilities, i.e $P(A_1)$, $P(A_2)$ and $P(A_3)$, it gives the sample space, i.e the total rectangular red coloured space.



- Consider now another event B occurs over A1,A2 and A3.
- Some area of B is common with A1, and A2 and A3.
- It is as shown in the figure below:



• Portion common with A1 and B is shown by:

$$P(A_1 \cap B)$$

• Portion common with A2 and B is given by :

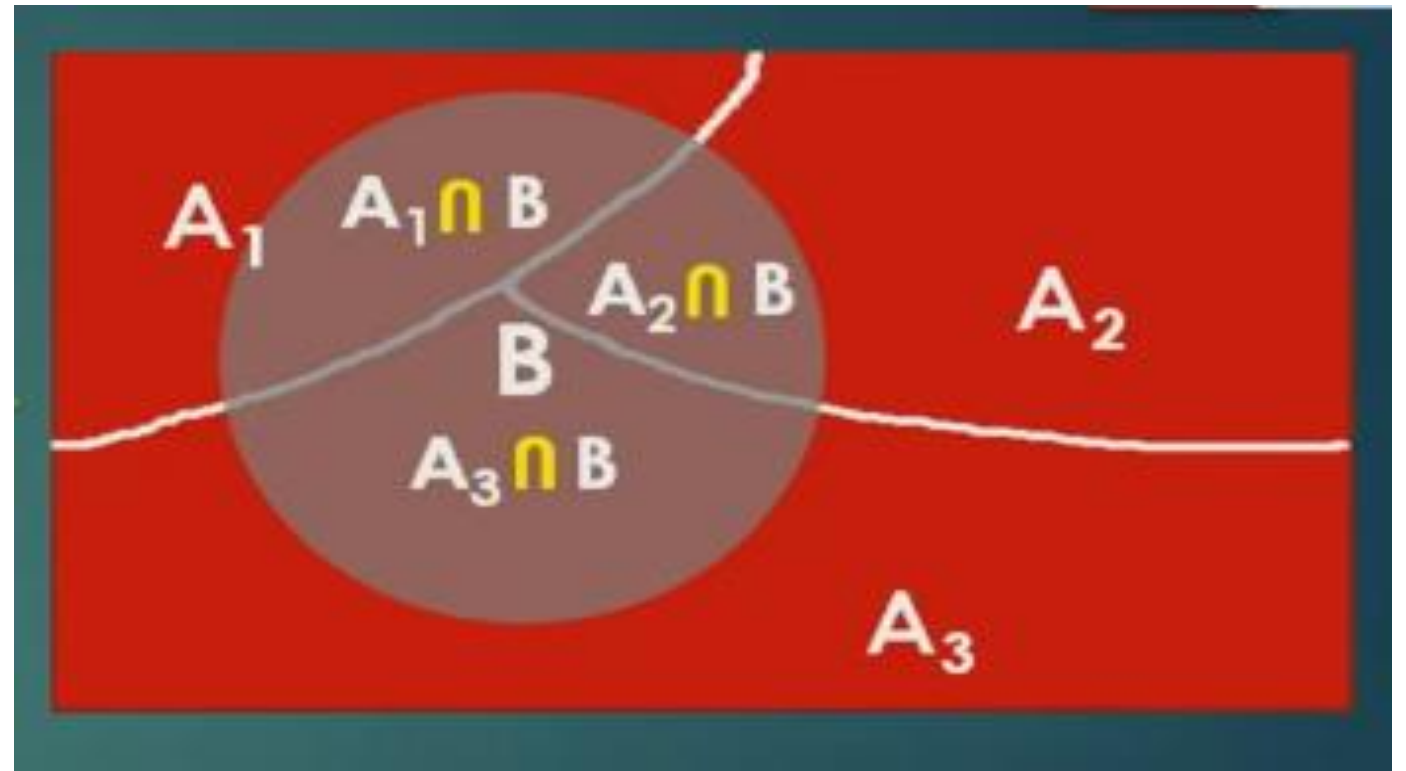
$$P(A_2 \cap B)$$

• Portion common with A3 and B is given by:

$$P(A_3 \cap B)$$

• Probability of B in total can be given by

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$



- Remember :

$$P(A \cap B) = P(A | B) * P(B) = P(B | A) * P(A)$$

- Equation from the previous slide:

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$

- Replacing first in the second equation in this slide, we will get:

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$

$$P(B) = P(B | A_1) * P(A_1) + P(B | A_2) * P(A_2) + P(B | A_3) * P(A_3)$$

Further simplified P(B)

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$

$$P(B) = P(B | A_1) * P(A_1) + P(B | A_2) * P(A_2) + P(B | A_3) * P(A_3)$$

$$P(B) = \sum_{i=1}^n P(B | A_i) * P(A_i)$$

Arriving at Generalized version of Bayes theorem

$$P(A_i | B) = \frac{P(B | A_i) * P(A_i)}{P(B)}$$

$$P(B) = \sum_{i=1}^n P(B | A_i) * P(A_i)$$

$$P(A_i | B) = \frac{P(B | A_i) * P(A_i)}{\sum_{i=1}^n P(B | A_i) * P(A_i)}$$

Example 3: Problem on Bayes theorem with 3 class case

In order to manage the Credit Risk, a bank regularly rates each of its borrowers as A_1 or A_2 or A_3 , based on their Credit history. A_1 implies lowest risk and A_3 implies highest risk. Risk means the chance that a borrower might fail to payback the loan amount.

Based on historical data, on an average, 30% customers are rated A_1 , 60% are rated A_2 , and 10% are rated A_3 . It was found that 1% of the customers who were rated A_1 , 10% of the customers who were rated A_2 , and 18% of the customers who were rated A_3 , eventually became defaulters (failed to payback).

If you randomly pick up a customer from defaulter's pool, what is the probability that he had received an A_1 rating?

What is being asked

- While solving problem based on Bayes theorem, we need to split the given information carefully:
- Asked is:

*“If you randomly pickup a customer from **defaulter’s pool**, what is the probability that he had received an **A₁ rating**?”*

$$P(\text{Rating } A_1 \mid \text{Defautler}) = ?$$

- Note, the flip of what is asked will be always given:

Note: Flip of what is being asked i.e. $P(\text{Defautler} \mid \text{Rating } A_1)$ will always be given in such problems.

- It is found in the following statement :

*“It was found that **1% of the customers who were rated A_1** , 10% of the customers who were rated A_2 , and 18% of the customers who were rated A_3 , eventually became defaulters (failed to payback).”*

$$P(\text{Defautler} \mid \text{Rating } A_1) = 1\% \text{ or } 0.01$$

$$P(\text{Defautler} \mid \text{Rating } A_2) = 10\% \text{ or } 0.10$$

$$P(\text{Defautler} \mid \text{Rating } A_3) = 18\% \text{ or } 0.18$$

- What else is given:

“Based on historical data, on an average, 30% customers are rated A_1 , 60% are rated A_2 , and 10% are rated A_3 .”

- Represented by:

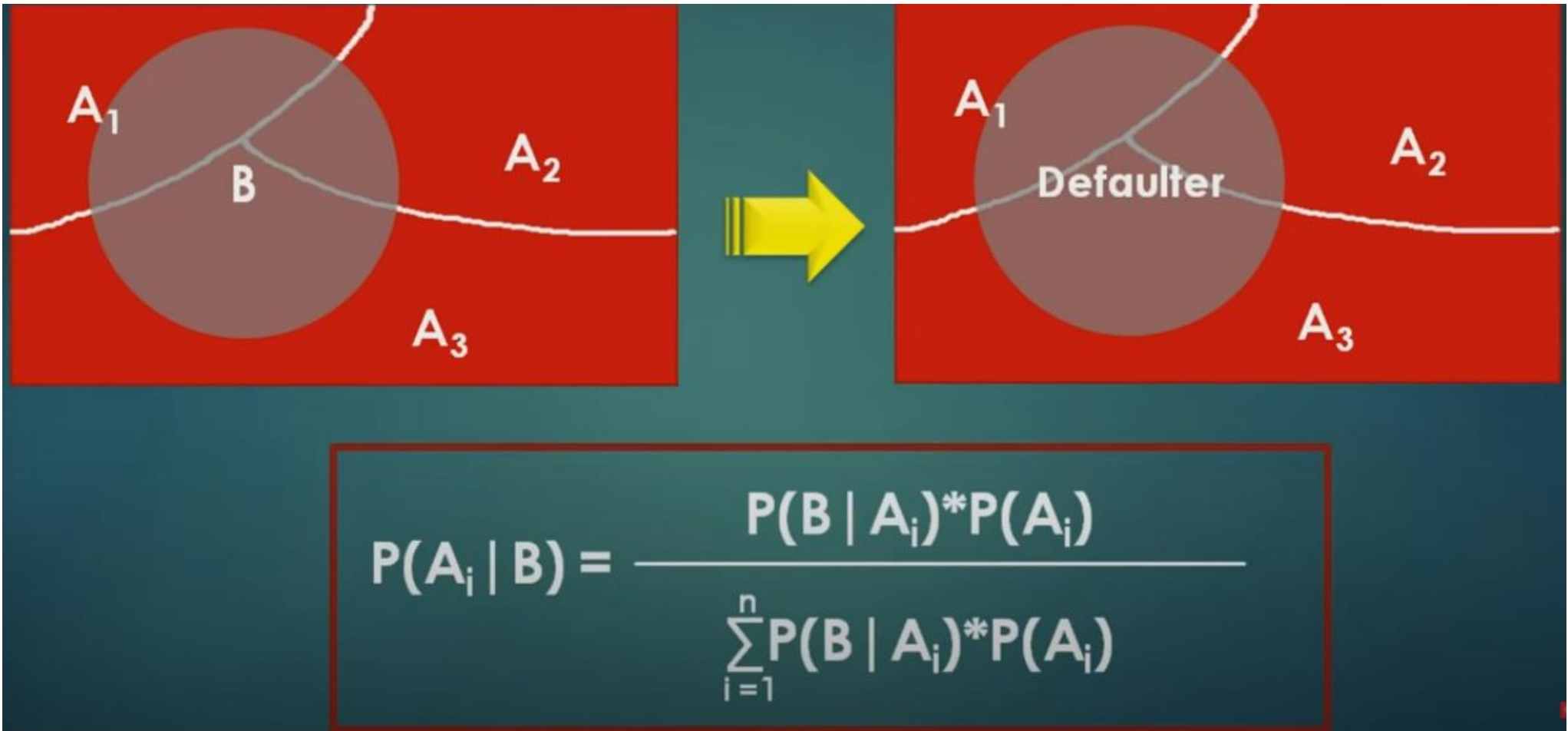
$$P(\text{Rating } A_1) = 30\% \text{ or } 0.30$$

$$P(\text{Rating } A_2) = 10\% \text{ or } 0.60$$

$$P(\text{Rating } A_3) = 18\% \text{ or } 0.10$$

Note: $P(\text{Rating } A_1) + P(\text{Rating } A_2) + P(\text{Rating } A_3) = 0.30 + 0.60 + 0.10 = 1$

So.. Given Problem can be represented as:



$$P(A_i | B) = \frac{P(B | A_i) * P(A_i)}{\sum_{i=1}^n P(B | A_i) * P(A_i)}$$

Numerator:

$$P(\text{Defautler} | \text{Rating } A_1) * P(\text{Rating } A_1) = 0.01 * 0.30 = 0.003$$

Denominator:

$$\begin{aligned} \sum_{i=1}^n P(B | A_i) * P(A_i) &= P(B | A_1) * P(A_1) + P(B | A_2) * P(A_2) + P(B | A_3) * P(A_3) \\ &= 0.01 * 0.30 + 0.10 * 0.60 + 0.18 * 0.10 \\ &= 0.081 \end{aligned}$$

$$P(\text{Rating } A_1 | \text{Defautler}) = \frac{0.003}{0.081} = 0.0370 \text{ or } 3.7\%$$

Example-4.

Given 1% of people have a certain genetic defect. (It means 99% don't have genetic defect)
90% of tests on the genetic defected people, the defect/disease is found positive(true positives).
9.6% of the tests (on non diseased people) are false positives

**If a person gets a positive test result,
what are the Probability that they actually have the genetic defect?**

A = chance of having the genetic defect. That was given in the question as 1%. ($P(A) = 0.01$)

That also means the probability of *not* having the gene ($\sim A$) is 99%. ($P(\sim A) = 0.99$)

X = A positive test result.

$P(A|X)$ = Probability of having the genetic defect given a positive test result. (To be computed)

$P(X|A)$ = Chance of a positive test result given that the person actually has the genetic defect = 90%. (0.90)

$p(X|\sim A)$ = Chance of a positive test if the person *doesn't* have the genetic defect. That was given in the question as 9.6% (0.096)

Now we have all of the information, we need to put into the equation:

$$P(A|X) = (.9 * .01) / (.9 * .01 + .096 * .99) = 0.0865 (8.65%).$$

The probability of having the faulty gene on the test is 8.65%.

Example - 5

Given the following statistics, what is the probability that a woman has cancer if she has a positive mammogram result?

One percent of women over 50 have breast cancer.

Ninety percent of women who have breast cancer test positive on mammograms.

Eight percent of women will have false positives.

Let women having cancer is W and $\sim W$ is women not having cancer.

Positive test result is PT .

Solution for Example 5

What is asked: what is the probability that a woman has cancer if she has a positive mammogram result?

- $P(W)=0.01$
- $P(\sim W)=0.99$
- $P(PT|W)=0.9$
- $P(PT|\sim W)=0.08$ Compute $P(\text{testing positive})$
$$(0.9 * 0.01) / ((0.9 * 0.01) + (0.08 * 0.99)) = 0.10.$$

Example-6

A disease occurs in 0.5% of the population
(5% is 5/10% removing % $(5/10)/100=0.005$)

A diagnostic test gives a positive result in:

- 99% of people with the disease
- 5% of people without the disease (false positive)

A person receives a positive result

What is the probability of them having the disease, given a positive result?

- $$P(\text{disease}|\text{positive test}) = \frac{P(PT|D) \times P(D)}{P(PT|D) \times P(D) + P(PT|\sim D) \times P(\sim D)}$$

$$= \frac{(0.99 \times 0.005)}{(0.99 \times 0.005) + (0.05 \times 0.995)}$$

Therefore:

$$P(\text{disease}|\text{positive test}) = \frac{0.99 \times 0.005}{0.0547} = 0.09$$

i. e. 9%

- We know:

- $P(D)$ = chance of having the disease

- $P(\sim D)$ = chance of not having the disease

- $P(\text{positive test}|\text{disease}) = 0.99$

- $P(\text{disease}) = 0.005$

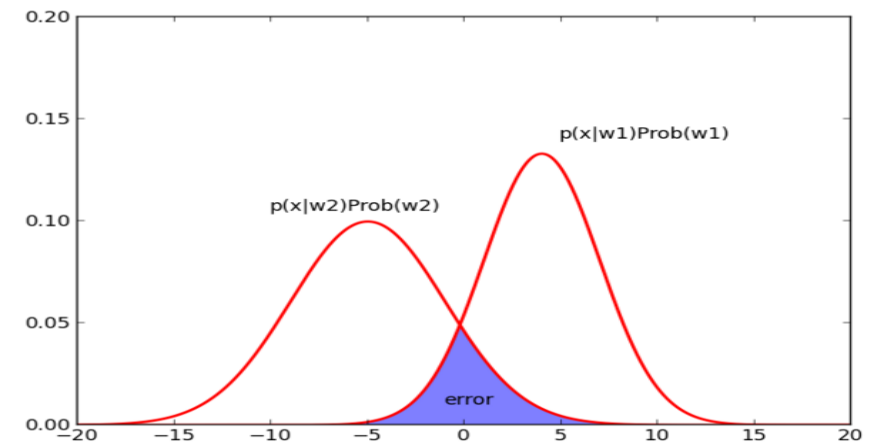
Decision Regions

- Likelihood ratio R between two classes can be computed by dividing posterior probability of two classes.
- So $P(C_i | x)$ (posterior probability of class C_i) and $P(C_j | x)$ (posterior probability of class C_j) are to be divided to understand the likelihood.
- If there are only two classes, then C_i and C_j can be replaced by A and B and the equation becomes: (the equation obtained is so because, the denominator gets cancelled)

$$R = \frac{P(A|x)}{P(B|x)} = \frac{P(A)p(x|A)}{P(B)p(x|B)}$$

- If the likelihood ratio R is greater than 1, we should select class A as the most likely class of the sample, otherwise it is class B
- A boundary between the **decision regions is called decision boundary**
- **Optimal decision boundaries separate the feature space into decision regions R_1, R_2, \dots, R_n such that class C_i is the most probable for values of x in R_i than any other region**

- For feature values exactly on the decision boundary between two classes, the two classes are equally probable.
- Thus to compute the optimal decision boundary between two classes A and B, we can equate their posterior probabilities if the densities are continuous and overlapping.



– $P(A|x) = P(B|x)$.

- Substituting Bayes Theorem and cancelling $p(x)$ term:

– $P(A)p(x|A) = P(B)p(x|B)$

- If the feature x in both the classes are normally distributed

- $$P(A) \frac{1}{\sigma_A \sqrt{2\pi}} e^{-\frac{(x-\mu_A)^2}{2\sigma_A^2}} = P(B) \frac{1}{\sigma_B \sqrt{2\pi}} e^{-\frac{(x-\mu_B)^2}{2\sigma_B^2}}$$

-

- Cancelling $\sqrt{2\pi}$ and taking natural logarithm

- $$-2\ln\left(\frac{P(A)}{\sigma_A}\right) + \left(\frac{x-\mu_A}{\sigma_A}\right)^2 = -2\ln\left(\frac{P(B)}{\sigma_B}\right) + \left(\frac{x-\mu_B}{\sigma_B}\right)^2$$

- $D = -2\ln(P(A)/\sigma_A) + (\frac{x-\mu_A}{\sigma_A})^2 + 2\ln(P(B)/\sigma_B) + (\frac{x-\mu_B}{\sigma_B})^2$
- D equals 0 then : on the decision boundary;
- D is positive in the decision region in which B is most likely the class;
- and D is negative in the decision region in which A is most likely.

- Example problem can be seen in the next slide

Example 3.4 Computing optimal one-dimensional decision boundaries.

To compute the decision boundaries for Example 3.3, we substitute the following data into (3.15): $\mu_G = 26$, $\sigma_G = 2$, $\mu_{\bar{G}} = 22$, $\sigma_{\bar{G}} = 3$, $P(G) = 0.8$, and $P(\bar{G}) = 0.2$. This produces

$$\begin{aligned} -2 \ln(0.8/2) + \left(\frac{x-26}{2}\right)^2 &= -2 \ln(0.2/3) + \left(\frac{x-22}{3}\right)^2 & (3.15) \\ 2 \cdot 36 \ln\left(\frac{0.2 \cdot 2}{0.8 \cdot 3}\right) &= 4(x-22)^2 - 9(x-26)^2 \\ 5x^2 - 292x + 4018.99 &= 0, \end{aligned}$$

so

$$x = \frac{292 \pm \sqrt{292^2 - 4 \cdot 5 \cdot 4018.99}}{2 \cdot 5} = 22.2 \text{ and } 36.2.$$

These two decision boundaries partition the feature space into three decision regions and produce the following decision rule:

1. Classify the sample as class G if $22.2 < x < 36.2$.
2. Classify the sample as class \bar{G} if $x < 22.2$ or $36.2 < x$.

Independence

- Independent random variables: Two random variables X and Y are said to be statistically independent if and only if :
- $p(x,y) = p(x).p(y)$
- Ex: Tossing two coins... are independent.
- Then the joint probability of these two will be product of their probability
- Another Example: X – Throw of dice, Y Toss of a coin
- (Event X and Y are joint probabilities and are independent)
- X=height and Y=Weight are joint probabilities are not independent... usually they are dependent.

- Independence is equivalent to saying
- $P(y|x) = P(y)$ or
- $P(x|y) = P(x)$

Conditional Independence

- Two random variables X and Y are said to be independent given Z if and only if
- $P(x,y|z)=P(x|z).P(y|z)$: indicates that X and Y are independent given Z .
- Example: X : Throw a dice
 Y : Toss a coin
 Z : Card from deck

So X and Y are conditionally independent and also conditionally independent.

Joint probabilities are dependent but conditionally independent

- Let us consider:
 - X: height
 - Y: Vocabulary
 - Z: Age
- Height is less indicates age is less and hence vocabulary might vary.
- So Vocabulary is dependent on height.
- Further let us add a condition Z.
- If Age is fixed say 30, then consider samples of people with age 30, but now the vocabulary of people with age 30 ..as the height increases vocabulary does not changes.
- So it is conditionally independent but joint probabilities are dependent without condition.

Reverse:

- **Two events are independent, but conditionally they are becoming dependent.**
- Let us say X : Dice throw 1
- Y : Dice throw 2
-
- Basically they are independent.
- Let us add $Z = \text{sum of the dice}$
- Given Z and X value is fixed then Y value depends on X value.
- It is Denoted by $x \perp y \mid z$
- X is said to be orthogonal or perpendicular to y , given z .

Multiple Features

- A single feature may not discriminate well between classes.
- Recall the example of just considering the 'dapg' or 'dwp' we can not discriminate well between the two classes. (Example for hypothetical basket ball games – unit 1).
- If the joint conditional density of multiple features is known for each class, Bayesian classification is very similar to classification with one feature.
- Replace the value of single feature x by feature vector X which has single feature as the component.

- $$P(w_i | X) = \frac{P(w_i)P(X | w_i)}{\sum_{j=1}^k P(w_j)P(x | w_j)}$$
 for single feature

- $$P(w_i | X) = \frac{P(w_i)p(X | w_i)}{\sum_{j=1}^k P(w_j)p(x | w_j)}$$

- For multiple features with Vector X replaces the conditional probabilities $P(X | W_i)$ by the conditional densities $p(x | w_i)$

Example of Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

Solution

A: attributes

M: mammals

N: non-mammals

$$P(A | M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A | N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A | M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A | N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

$$P(A|M)P(M) > P(A|N)P(N)$$

=> Mammals

Example. 'Play Tennis' data

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
<i>Day1</i>	Sunny	Hot	High	Weak	<i>No</i>
<i>Day2</i>	Sunny	Hot	High	Strong	<i>No</i>
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
<i>Day6</i>	Rain	Cool	Normal	Strong	<i>No</i>
Day7	Overcast	Cool	Normal	Strong	Yes
<i>Day8</i>	Sunny	Mild	High	Weak	<i>No</i>
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
<i>Day14</i>	Rain	Mild	High	Strong	<i>No</i>

Question: For the day <sunny, cool, high, strong>, what's the play prediction?

- Naïve based classifier is very popular for document classifier
- (naïve means: all are equal and independent: all the attributes will have equal weightage and are independent)

Based on the examples in the table, classify the following datum \mathbf{x} :

$\mathbf{x}=(\text{Outl}=\text{Sunny}, \text{Temp}=\text{Cool}, \text{Hum}=\text{High}, \text{Wind}=\text{strong})$

- That means: Play tennis or not?

$$h_{NB} = \arg \max_{h \in [\text{yes}, \text{no}]} P(h)P(\mathbf{x} | h) = \arg \max_{h \in [\text{yes}, \text{no}]} P(h) \prod_t P(a_t | h)$$

$$= \arg \max_{h \in [\text{yes}, \text{no}]} P(h)P(\text{Outlook} = \text{sunny} | h)P(\text{Temp} = \text{cool} | h)P(\text{Humidity} = \text{high} | h)P(\text{Wind} = \text{strong} | h)$$

- Working:

$$P(\text{PlayTennis} = \text{yes}) = 9 / 14 = 0.64$$

$$P(\text{PlayTennis} = \text{no}) = 5 / 14 = 0.36$$

$$P(\text{Wind} = \text{strong} | \text{PlayTennis} = \text{yes}) = 3 / 9 = 0.33$$

$$P(\text{Wind} = \text{strong} | \text{PlayTennis} = \text{no}) = 3 / 5 = 0.60$$

etc.

$$P(\text{yes})P(\text{sunny} | \text{yes})P(\text{cool} | \text{yes})P(\text{high} | \text{yes})P(\text{strong} | \text{yes}) = 0.0053$$

$$P(\text{no})P(\text{sunny} | \text{no})P(\text{cool} | \text{no})P(\text{high} | \text{no})P(\text{strong} | \text{no}) = \mathbf{0.0206}$$

$\Rightarrow \text{answer: } \text{PlayTennis}(x) = \text{no}$

What is our probability of error?

- For the two class situation, we have
- $P(\text{error} | x) = \begin{cases} P(\omega_1 | x) & \text{if we decide } \omega_2 \\ P(\omega_2 | x) & \text{if we decide } \omega_1 \end{cases}$
- We can minimize the probability of error by following the posterior:

Decide ω_1 if $P(\omega_1 | x) > P(\omega_2 | x)$

Probability of error becomes $P(\text{error} | x) = \min [P(\omega_1 | x), P(\omega_2 | x)]$

Equivalently, Decide ω_1 if $p(x | \omega_1)P(\omega_1) > p(x | \omega_2)P(\omega_2)$;

otherwise decide ω_2 i.e., the evidence term is not used in decision making.

Conversely, if we have uniform priors, then the decision will rely exclusively on the likelihoods.

Take Home Message: Decision making relies on both the priors and the likelihoods and Bayes Decision Rule combines them to achieve the minimum probability of error.

Application of Naïve Bayes Classifier for NLP

- Consider the following sentences:
 - S1 : The food is Delicious : Liked
 - S2 : The food is Bad : Not Liked
 - S3 : Bad food : Not Liked
- Given a new sentence, whether it can be classified as **liked sentence** or **not liked**.
- **Given Sentence: Delicious Food**

- Remove stop words, then perform stemming

	F1	F2	F3	Output
	Food	Delicious	Bad	
• S1	1	1	0	1
• S2	1	0	1	0
• S3	1	0	1	0

- $P(\text{Liked} \mid \text{attributes}) = P(\text{Delicious} \mid \text{Liked}) * P(\text{Food} \mid \text{Liked}) * P(\text{Liked})$
- $= (1/1) * (1/1) * (1/3) = 0.33$

- $P(\text{Not Liked} \mid \text{attributes}) = P(\text{Delicious} \mid \text{Not Liked}) * P(\text{Food} \mid \text{Not Liked}) * P(\text{Not Liked})$
- $= (0) * (2/2) * (2/3) = 0$
- Hence the given sentence belongs to Liked class

End of Unit 2

Unit-3
Non-Parametric Decision Making
Dr. Srinath.S

Syllabus

- Nonparametric Decision Making:
- Introduction, Histograms,
- kernel and Window Estimators,
- Nearest Neighbour Classification Techniques: Nearest neighbour algorithm, Adaptive Decision Boundaries, Minimum Squared Error Discriminant Functions, Choosing a decision-making technique

NON-PARAMETRIC DECISION MAKING

In parametric decision making, Only the parameters of the densities, such as their MEAN or VARIANCE had to be estimated from the data before using them to estimate probabilities of class membership.

In Nonparametric approach, distribution of data **is not defined by a finite set of parameters**

Nonparametric model does not take a predetermined form but the model is constructed according to information derived from the data.

It does not uses MEAN or VARIANCE.

Non-Parametric Decision making is considered as more robust.

Some of the popular Non – Parametric Decision making includes:

Histogram, Scatterplots or Tables of data

Kernel Density Estimation

KNN

Support Vector Machine (SVM)

HISTOGRAM

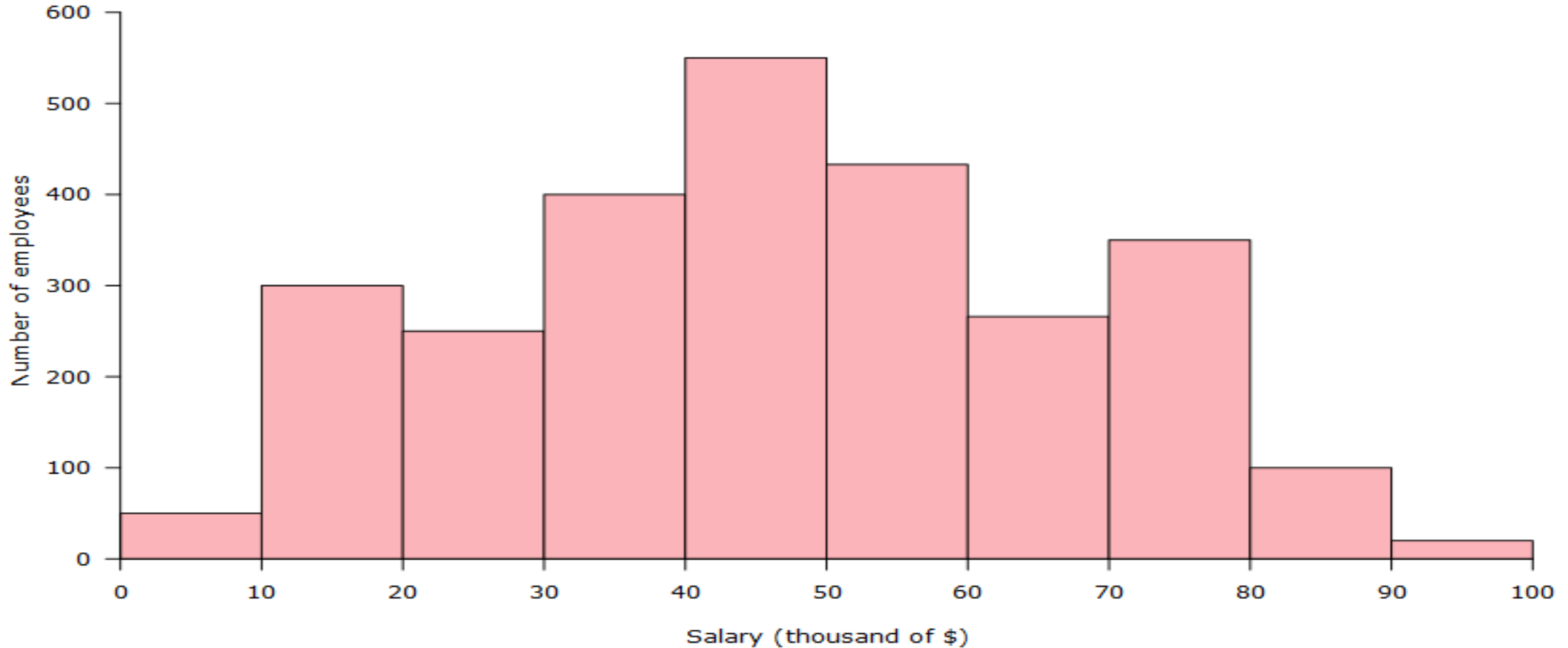
- Histogram is one of the easiest ways of obtaining the approximate density functions $\hat{p}(x)$ from the sampled data.
- Histogram is a way to estimate the distribution of data without assuming any particular shape for distribution (Gaussian, beta, etc.).
- Histogram shows the proportion of cases that fall into each of several categories.
- The total area of a histogram is always normalized to 1, to display a valid probability.(thus, it is a frequentist approach)
- Histogram plots provide a fast and reliable way to visualize the probability density of a data sample.
- A histogram is a plot that involves first grouping the observations into bins and counting the number of events that fall into each bin.

HISTOGRAM Continued

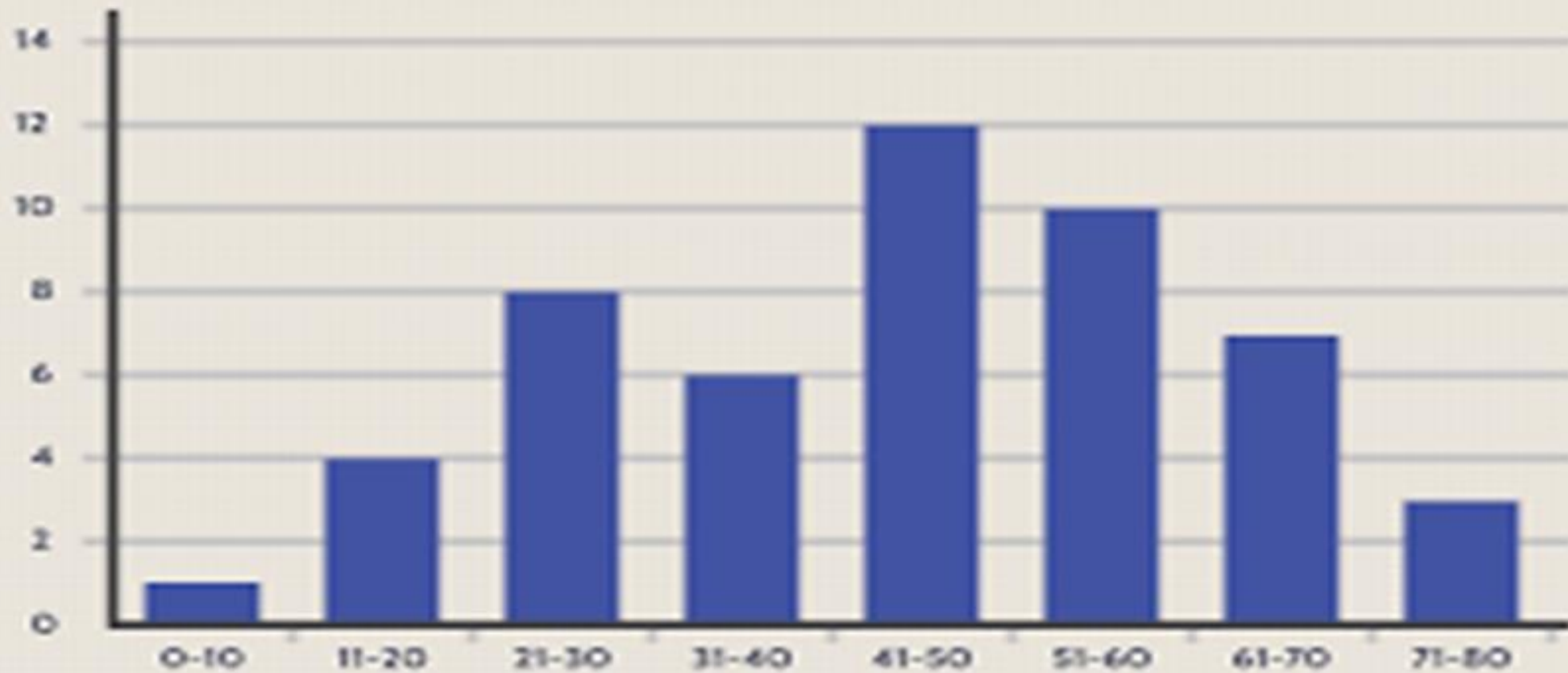
- The counts, or frequencies of observations, in each bin are then plotted as a bar graph with the bins on the x-axis and the frequency on the y-axis.
- One of the thumb rule to choose the number of intervals to be equal to the square root of the number of samples

Histogram Example

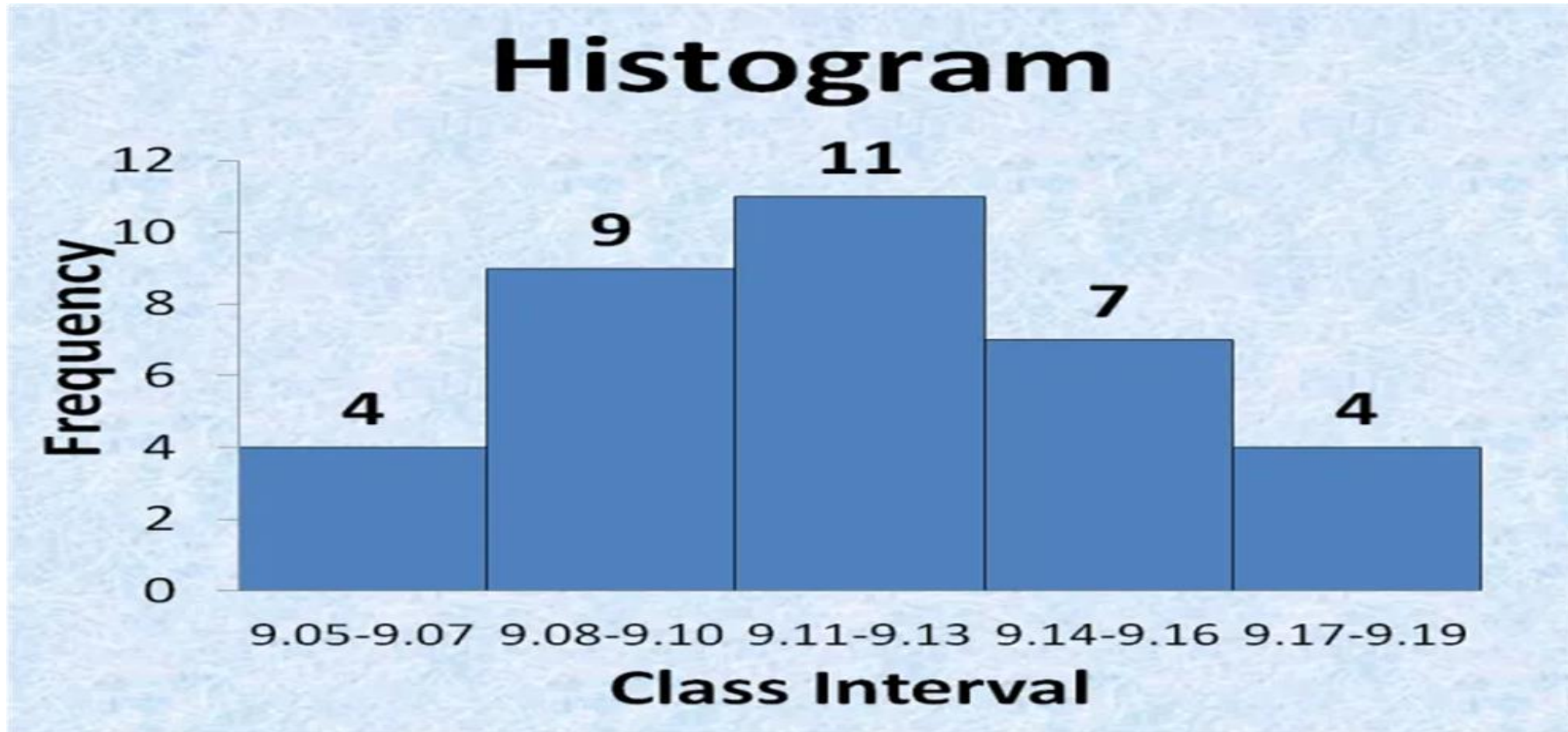
Distribution of salaries of the employees of ABC Corporation



Histogram Example



Histogram Example



HISTOGRAM Continued

Example 4.1 *Constructing a density histogram with unequal interval widths.*

Produce a histogram approximation to the density function for grapefruit volumes. A sample of 100 grapefruit for which the volume x was measured produced the following data:

Interval of x	Length of interval	Number of samples	Proportion of samples (=area)	Height of rectangle $\hat{p}(x)$
$[0, 4)$	4	10	0.1	0.025
$[4, 6)$	2	30	0.3	0.150
$[6, 7)$	1	30	0.3	0.300
$[7, 8)$	1	20	0.2	0.200
$[8, 10]$	2	10	0.1	0.050

The height of each rectangle is equal to the fraction of samples falling within its interval divided by the length of the interval (the base of the rectangle). For example, the height

For Example : the height in the first row is $0.1/4 = 0.025$

HISTOGRAM Continued

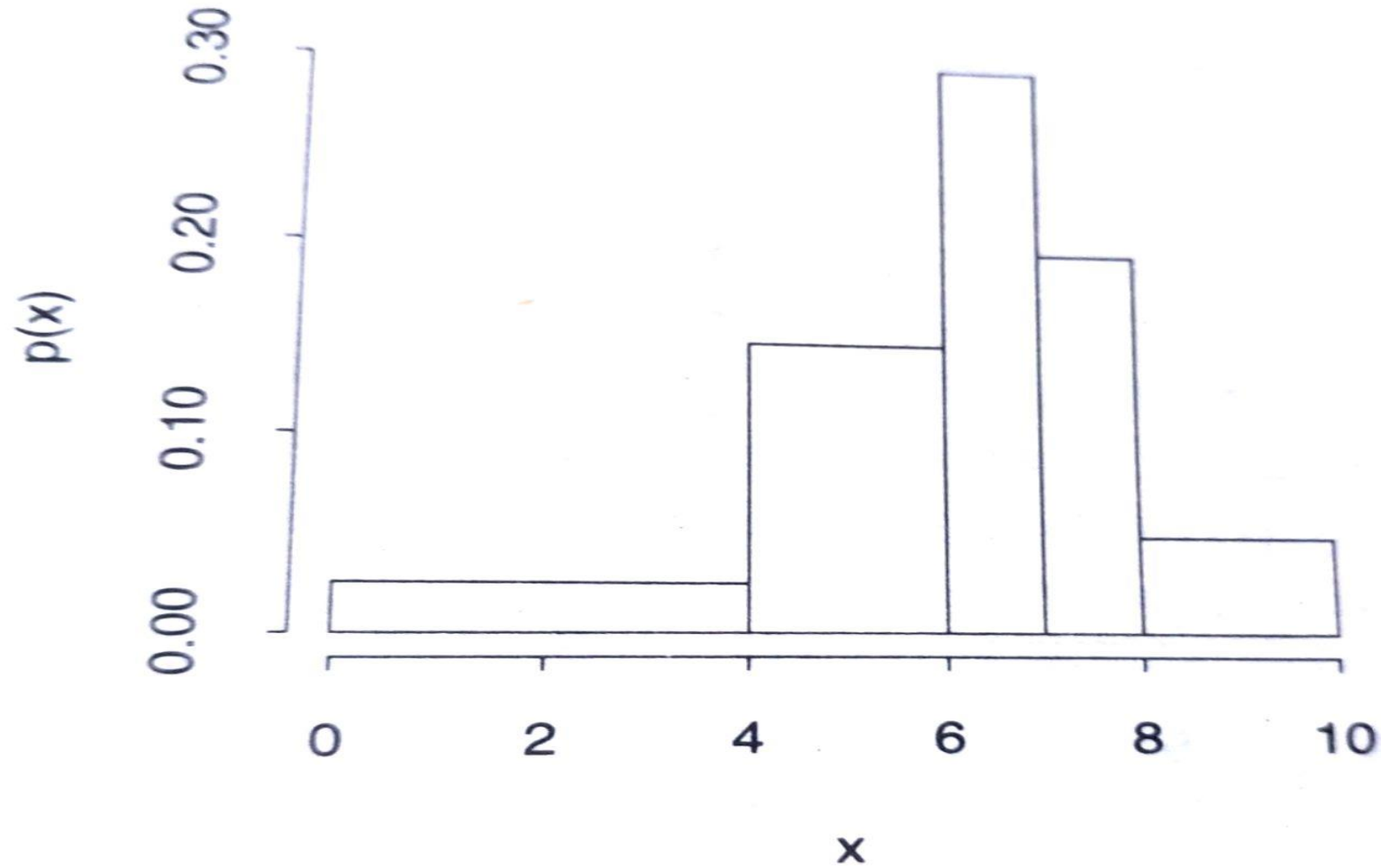


Figure 4.2: The histogram of Example 4.1.

HISTOGRAM Continued

Example 4.2 *Classification of samples using histograms and Bayes' Theorem.*

Use the following data to classify a sample with $x = 7.5$, given that $P(A) = P(B) = 0.5$. The following data are the values of feature x for 60 randomly chosen samples from class A :

0.80	0.91	0.93	0.95	1.32	1.53	1.57	1.63	1.67	1.74
2.01	2.18	2.27	2.31	2.40	2.61	2.64	2.64	2.67	2.85
2.96	2.97	3.17	3.17	3.38	3.67	3.73	3.83	3.99	4.06
4.10	4.12	4.18	4.20	4.23	4.27	4.27	4.39	4.40	4.46
4.47	4.61	4.64	4.89	4.96	5.12	5.15	5.33	5.33	5.47
5.64	5.85	5.99	6.29	6.42	6.53	6.70	6.78	7.18	7.22

And the following measurements are 60 values of x for some random samples from class B :

3.54	3.88	4.24	4.30	4.30	4.70	4.75	4.97	5.21	5.42
5.60	5.77	5.87	5.94	5.95	6.04	6.05	6.15	6.19	6.21
6.33	6.41	6.43	6.49	6.52	6.58	6.60	6.63	6.65	6.75
6.90	6.92	7.03	7.08	7.18	7.29	7.33	7.41	7.41	7.46
7.61	7.67	7.68	7.68	7.78	7.96	8.03	8.12	8.20	8.22
8.33	8.36	8.44	8.45	8.49	8.75	8.76	9.14	9.20	9.86

HISTOGRAM Continued

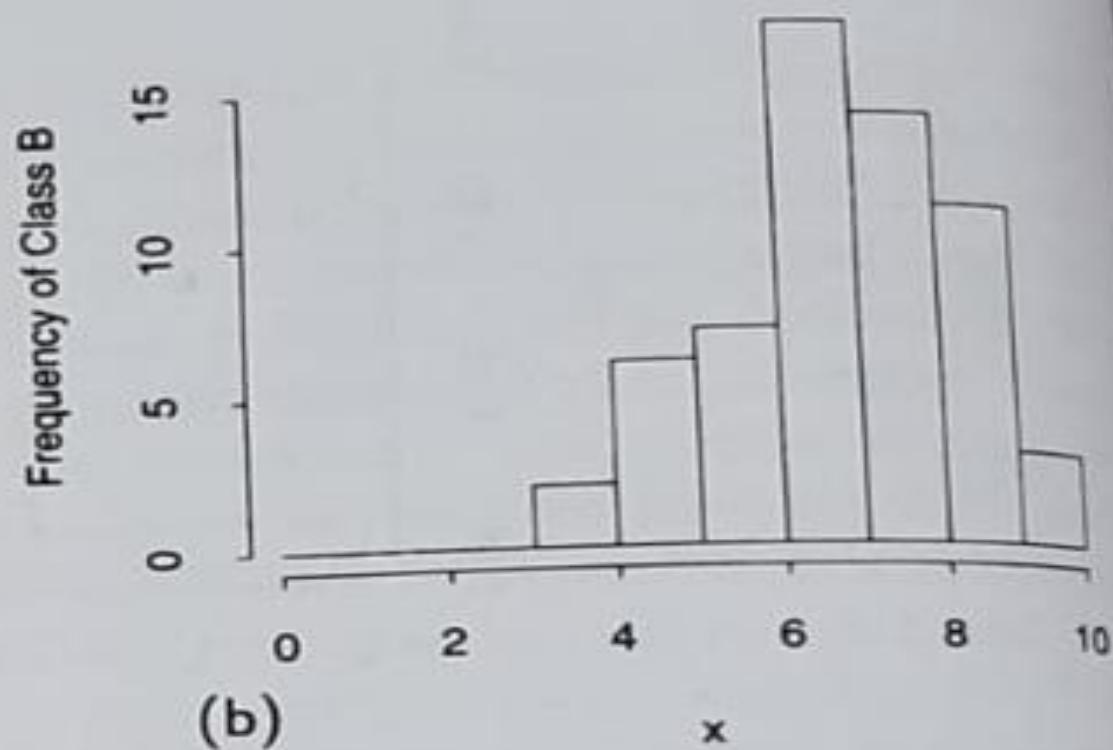
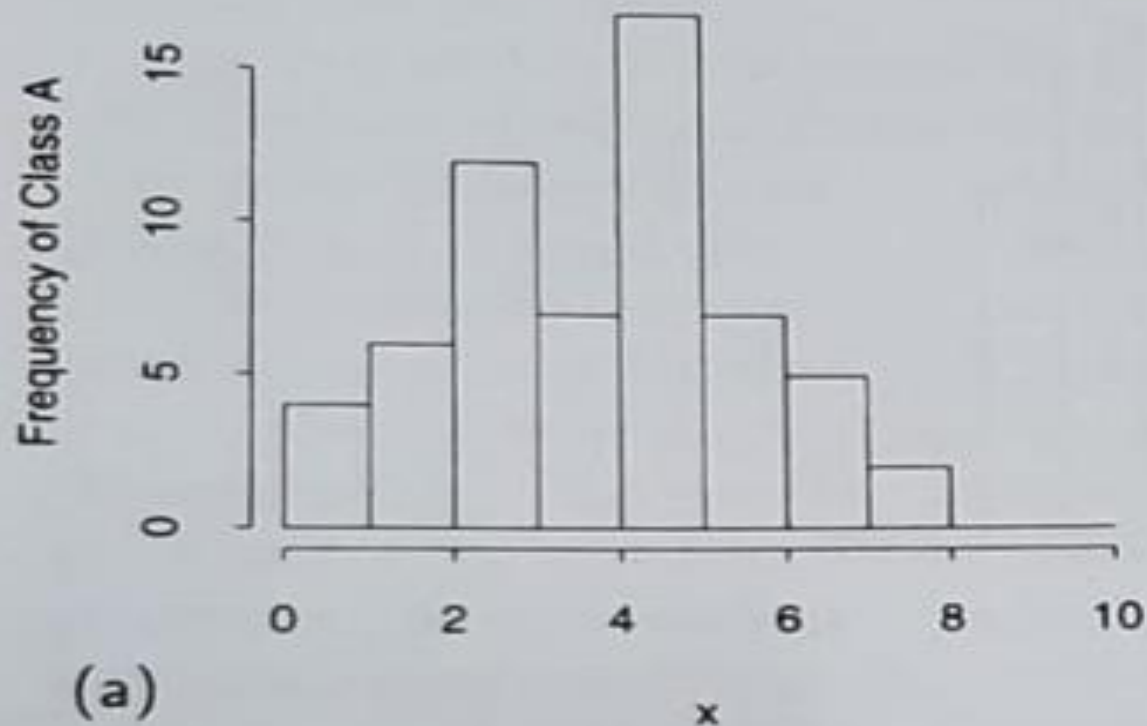


Figure 4.3: Histograms of the feature x for (a) class A and (b) class B .

Figure 4.3 shows histograms of the samples in each unit interval of x for classes A and B . To convert these to density functions, these numbers must be divided by the total number of samples (60) and the interval width (1).

Continued ...

To classify a sample with $x = 7.5$, compare the heights of the two histograms at 7.5. Because the class interval containing 7.5 is $[7, 8]$ for both classes A and B, $\hat{p}(7.5|A) = 2/[60(8 - 7)]$, and $\hat{p}(7.5|B) = 14/[60(8 - 7)]$. Using Bayes' Theorem,

$$\begin{aligned} P(A|7.5) &= \frac{p(7.5|A)P(A)}{p(7.5|A)P(A) + p(7.5|B)P(B)} \\ &= \frac{(2/60)(0.5)}{(2/60)(0.5) + (14/60)(0.5)} = 1/8 = 0.125. \end{aligned}$$

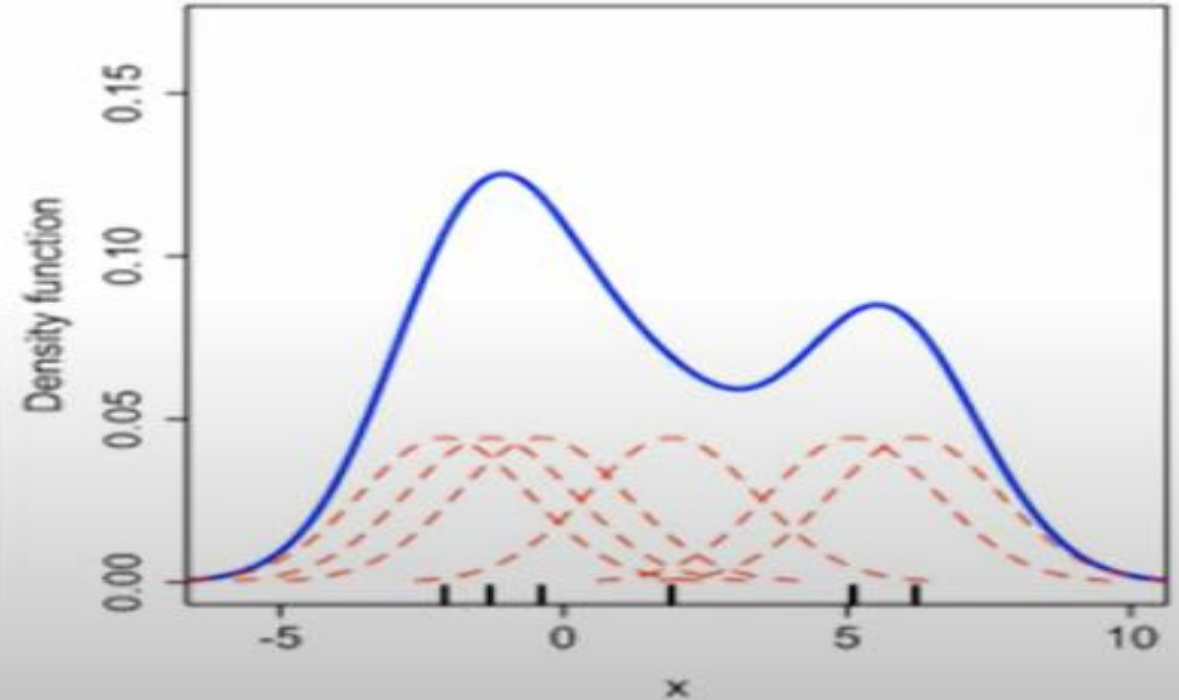
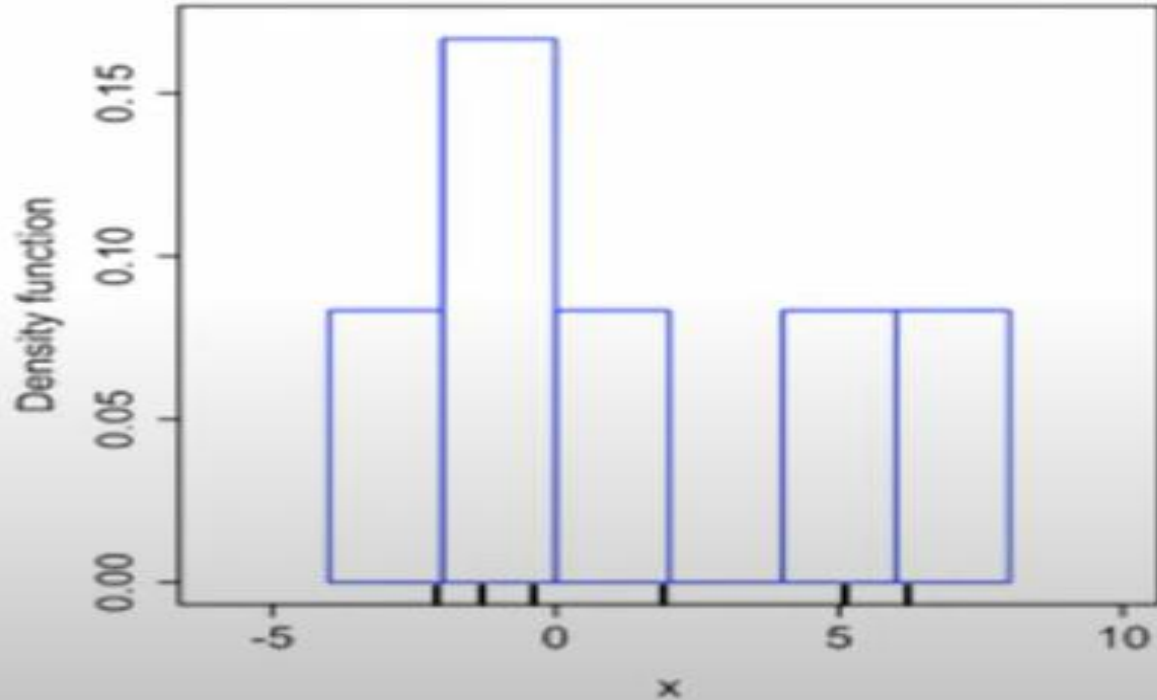
Also $P(B|7.5) = 1 - P(A|7.5) = 0.875$. Therefore, $P(A|7.5) < P(B|7.5)$, so the sample should be classified into class B.

Kernel and Window Estimators

- Histogram is a good representation for discrete data. It will show the spikes for each bin.
- But may not suite for continuous data. Then we will be using Kernel (function) for each of the data points. And the total density is estimated by the kernel density function.
- It is useful for applications like audio density estimation.
- **This approximation to a continuous density estimation is not useful in decision making.**
- Each Delta function is replaced by Kernel Functions such as rectangles, triangles or normal density functions which have been scaled so that their combined area should be equal to one.

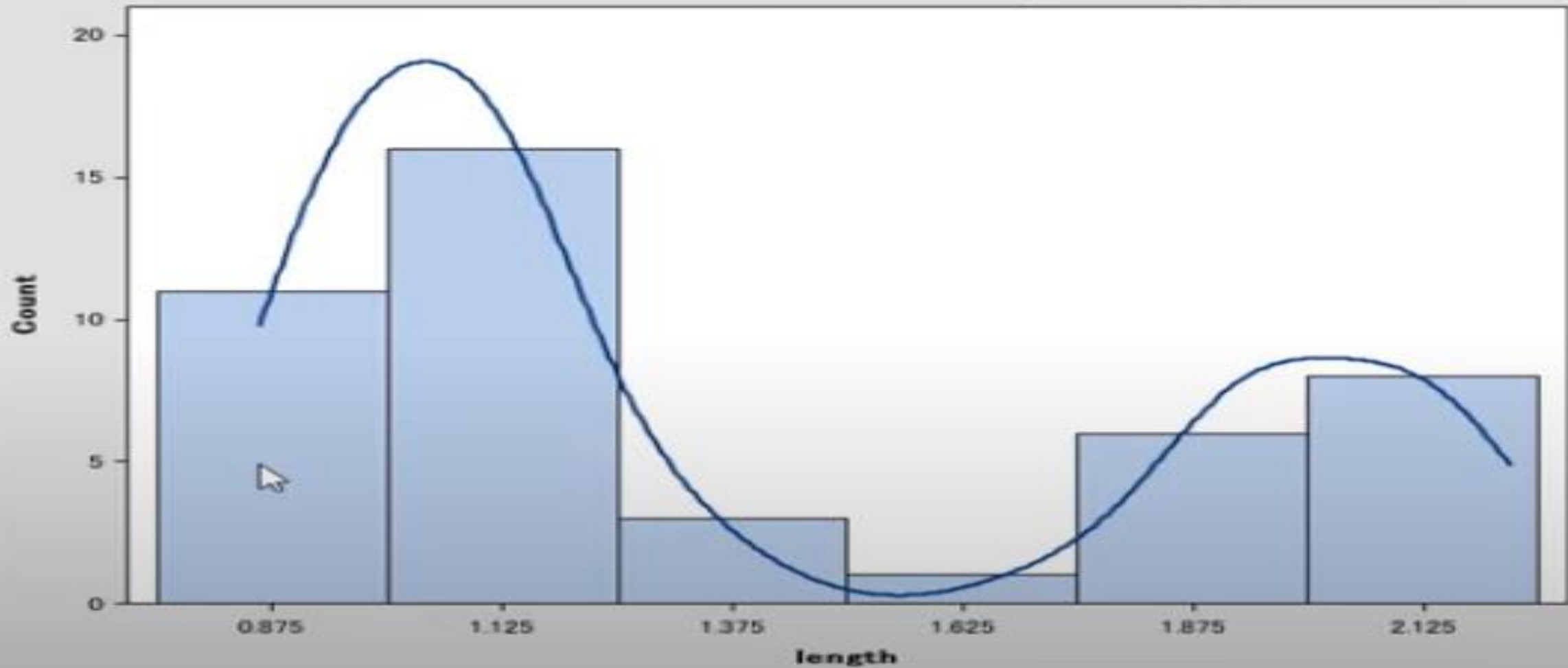
Kernel Density function

$x_1 = -2.1, x_2 = -1.3, x_3 = -0.4, x_4 = 1.9, x_5 = 5.1, x_6 = 6.2.$



-4 to -2 = 1, -2 to 0 = 2, 0 to -2 = 1 -2 to -4 = 0, -4 to -6 = 1 and -6 to -8 = 1
Height = $1/6 * 2 = 0.08$ (first case) and so on

Distribution and Kernel Density for length



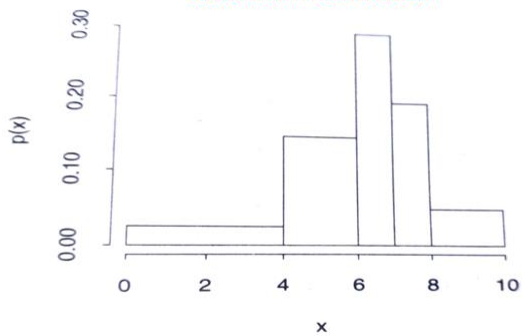
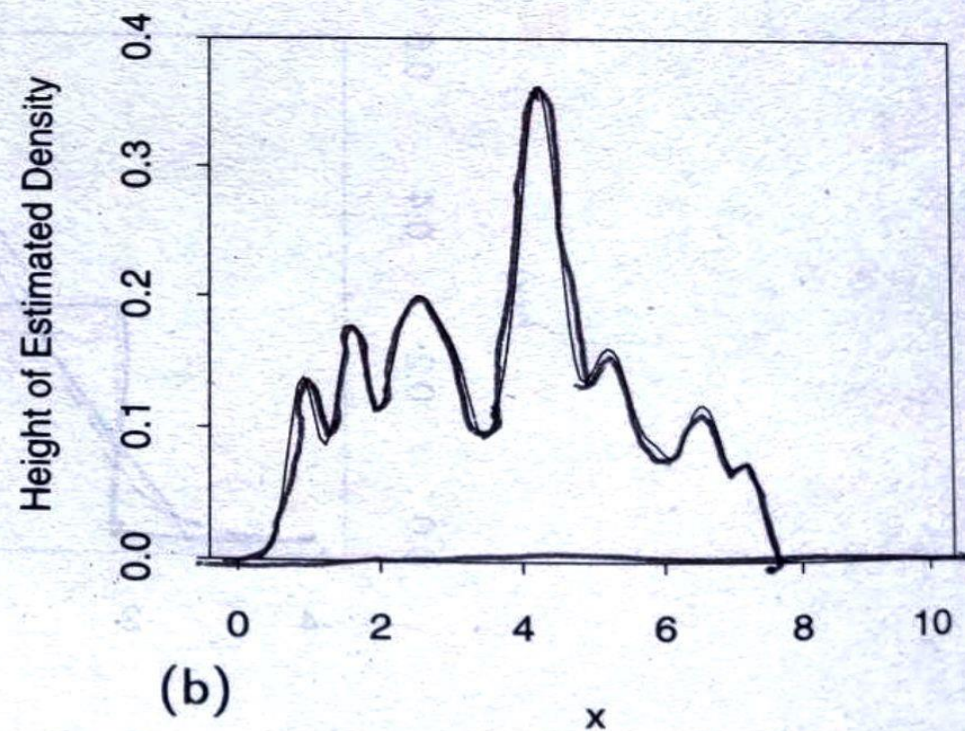
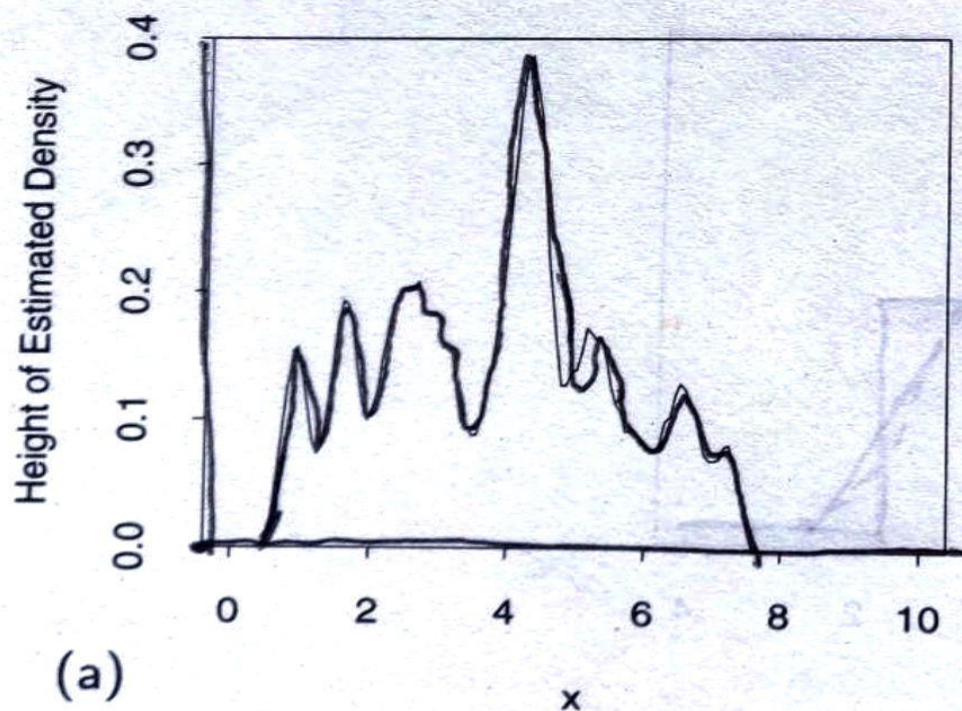


Figure 4.2: The histogram of Example 4.1.

CHAPTER 4. NONPARAMETRIC DECISION MAKING

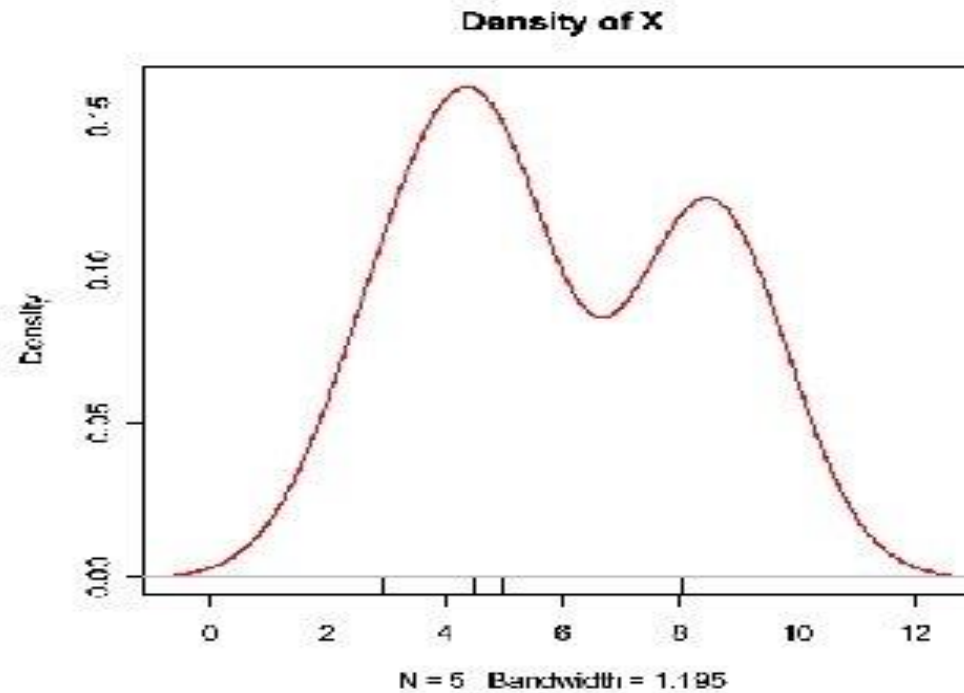
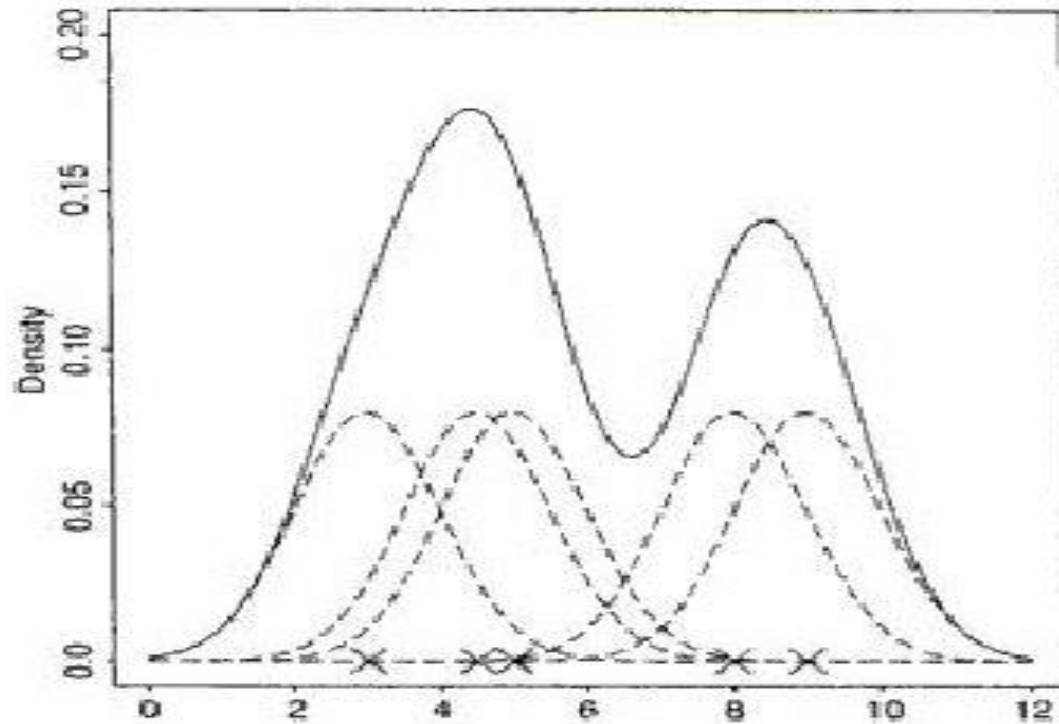
Figure 4.7: The estimated density functions $\hat{p}(x|A)$ for Example 4.2 using (a) a triangular kernel and (b) a normal kernel.

KERNEL DENSITY ESTIMATION

KDE – Based on Five Observations

Kernel density estimate constructed using five observations with the kernel chosen to be the $N(0,1)$ density.

$$x=c(3, 4.5, 5.0, 8, 9)$$



Similarity and Dissimilarity

- Distance are used to measure similarity
- There are many ways to measure the distance **S** between two instances

Distance or similarity measures are essential in solving many pattern recognition problems such as classification and clustering. Various distance/similarity measures are available in the literature to compare two data distributions.

As the names suggest, a similarity measures how close two distributions are.

For algorithms like the k-nearest neighbor and [k-means](#), it is essential to measure the distance between the data points.

- In KNN we calculate the distance between points to find the nearest neighbor.
- In K-Means we find the distance between points to group data points into clusters based on similarity.
- It is vital to choose the right distance measure as it impacts the results of our algorithm.

Euclidean Distance

- We are most likely to use Euclidean distance when calculating the distance between two rows of data that have numerical values, such as floating point or integer values.
- If columns have values with differing scales, it is common to normalize or standardize the numerical values across all columns prior to calculating the Euclidean distance. Otherwise, columns that have large values will dominate the distance measure.

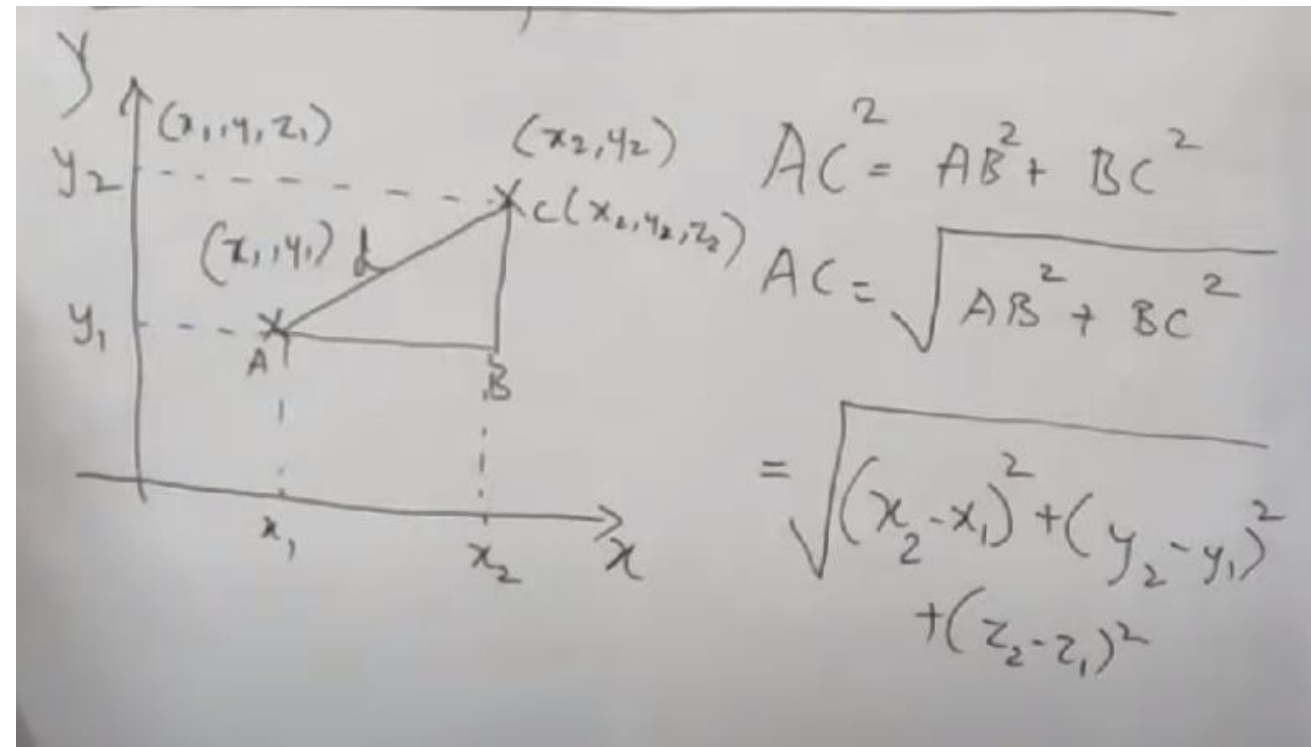
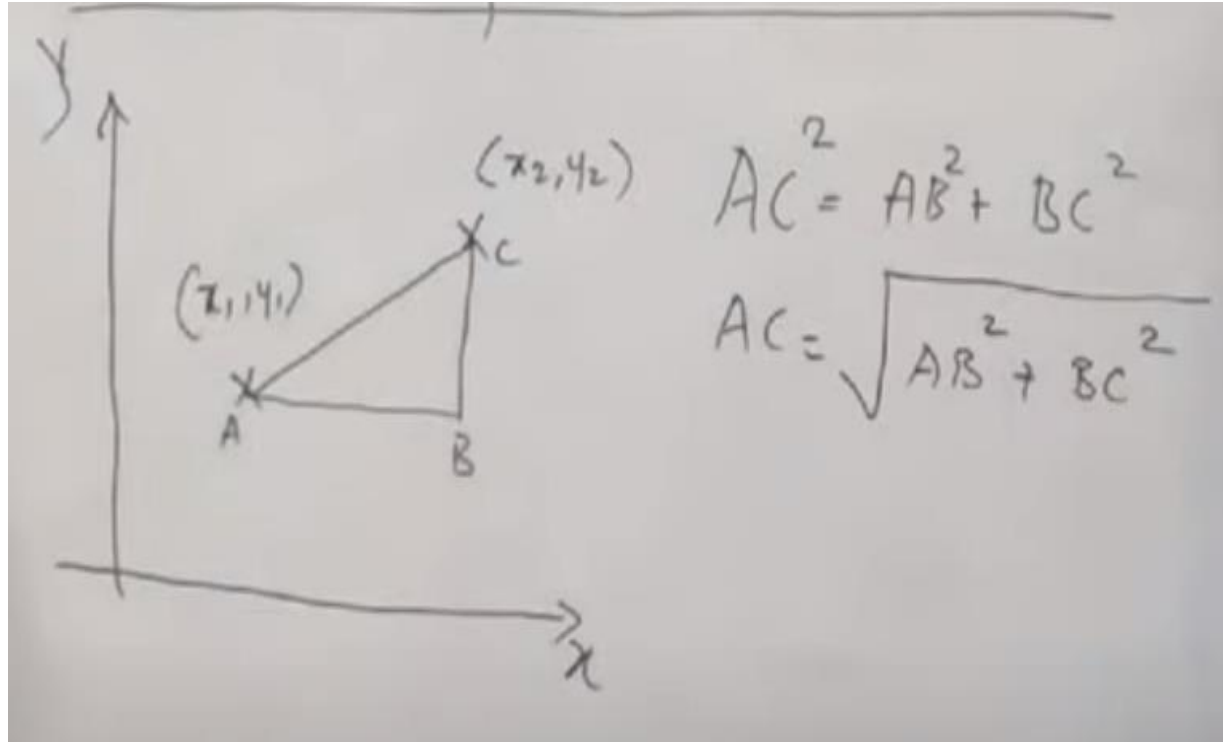
$$\mathit{dist} = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

- Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^{th} attributes (components) or data objects p and q .
- **Euclidean distance is also known as the L2 norm of a vector.**

Compute the Euclidean Distance between the following data set

- **D1= [10, 20, 15, 10, 5]**
- **D2= [12, 24, 18, 8, 7]**

Apply Pythagoras theorem for Euclidean distance



Manhattan distance:

Manhattan distance is a metric in which the distance between two points is the sum of the absolute differences of their Cartesian coordinates. In a simple way of saying it is the total sum of the difference between the x-coordinates and y-coordinates.

Formula: In a plane with p1 at (x1, y1) and p2 at (x2, y2)

$$|x_1 - x_2| + |y_1 - y_2|$$

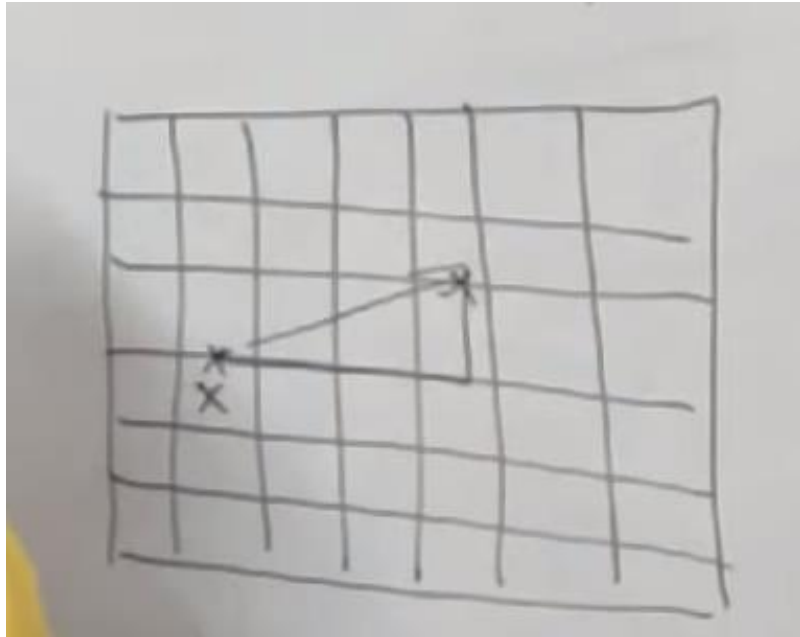
- **The Manhattan distance is related to the L1 vector norm**
- In general ManhattanDistance = sum for i to N sum of $|v1[i] - v2[i]|$

Compute the Manhattan distance for the following

- **D1 = [10, 20, 15, 10, 5]**
- **D2 = [12, 24, 18, 8, 7]**

Manhattan distance:

is also popularly called city block distance



Euclidean distance is like flying distance

Manhattan distance is like travelling by car

Minkowski Distance

- It calculates the distance between two real-valued vectors.
- It is a generalization of the Euclidean and Manhattan distance measures and adds a parameter, called the “*order*” or “*r*”, that allows different distance measures to be calculated.
- The Minkowski distance measure is calculated as follows:

$$\mathit{dist} = \left(\sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k th attributes (components) or data objects p and q .

Minkowski is called generalization of Manhattan and Euclidean:

Manhattan Distance is called L1 Norm and
Euclidean distance is called L2

Minkowski is called L_p where p can be 1 or 2

Cosine Similarity

(widely used in recommendation system and NLP)

- If A and B are two document vectors.
- Cosine similarity ranges between (-1 to +1)
 - -1 indicates not at all close and +1 indicates it is very close in similarity
- In cosine similarity data objects are treated as vectors.
- It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. It is often used to measure document similarity in text analysis.

–**Cosine Distance = 1- Cosine Similarity**

$\cos(A, B) =$ 1: exactly the same

0: orthogonal

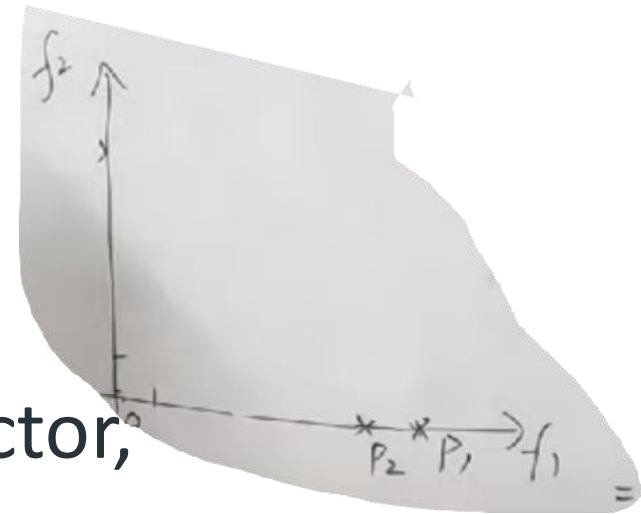
-1: exactly opposite

Formula for Cosine Similarity

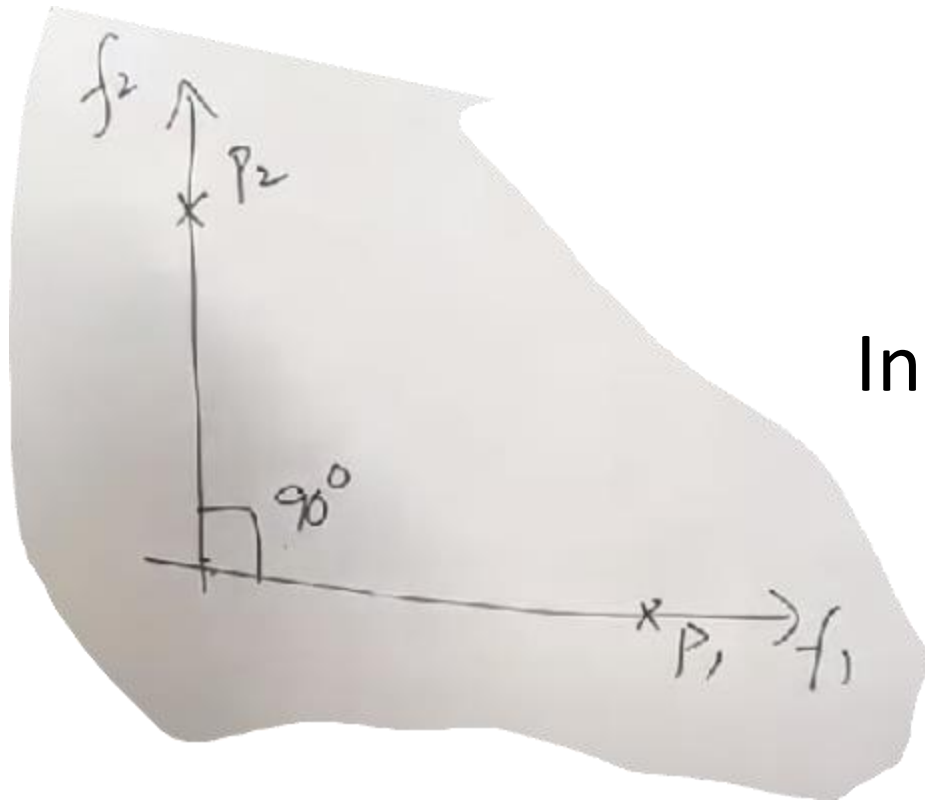
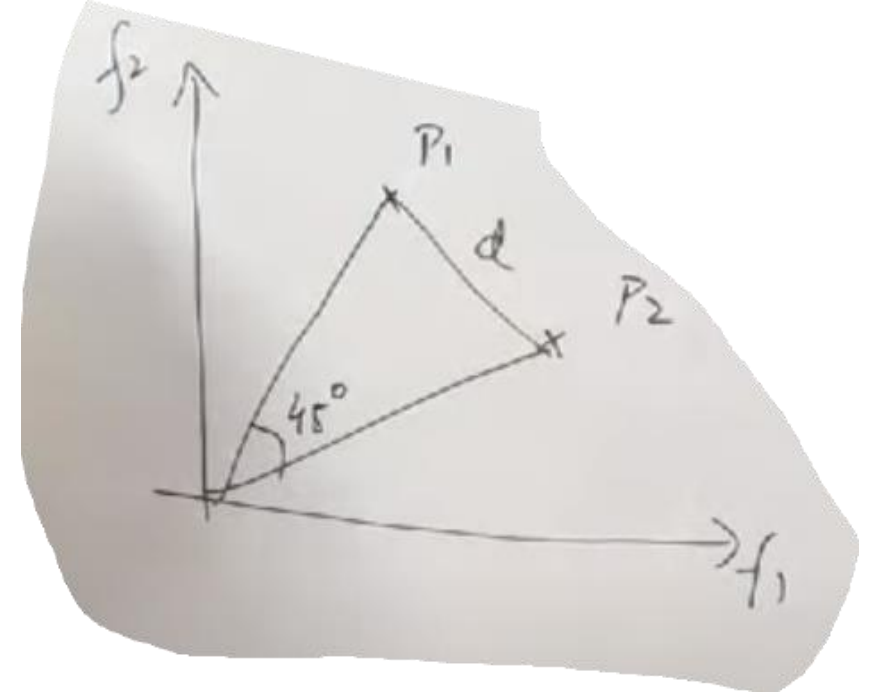
$$\text{cosine similarity} = S_C(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

- The cosine similarity between two vectors is measured in ' θ '.
- If $\theta = 0^\circ$, the 'x' and 'y' vectors overlap, thus proving they are similar.
- If $\theta = 90^\circ$, the 'x' and 'y' vectors are dissimilar.
- If two points are on the same plane or same vector
- In the example given P1 and P2 are on the same vector, and hence the angle between them is 0, so $\text{COS}(0) = 1$ indicates they are of high similarity



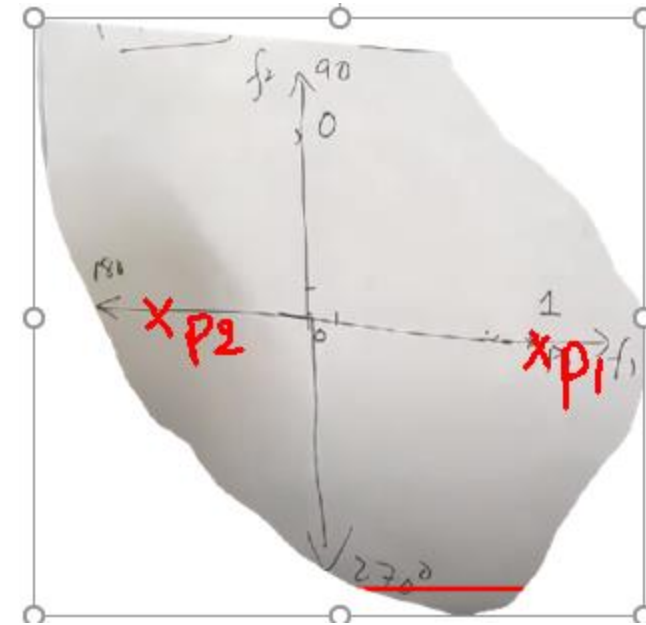
- In this example two points P1 and P2 are separated by 45 degree, and hence Cosine similarity is $\text{COS}(45) = 0.53$.



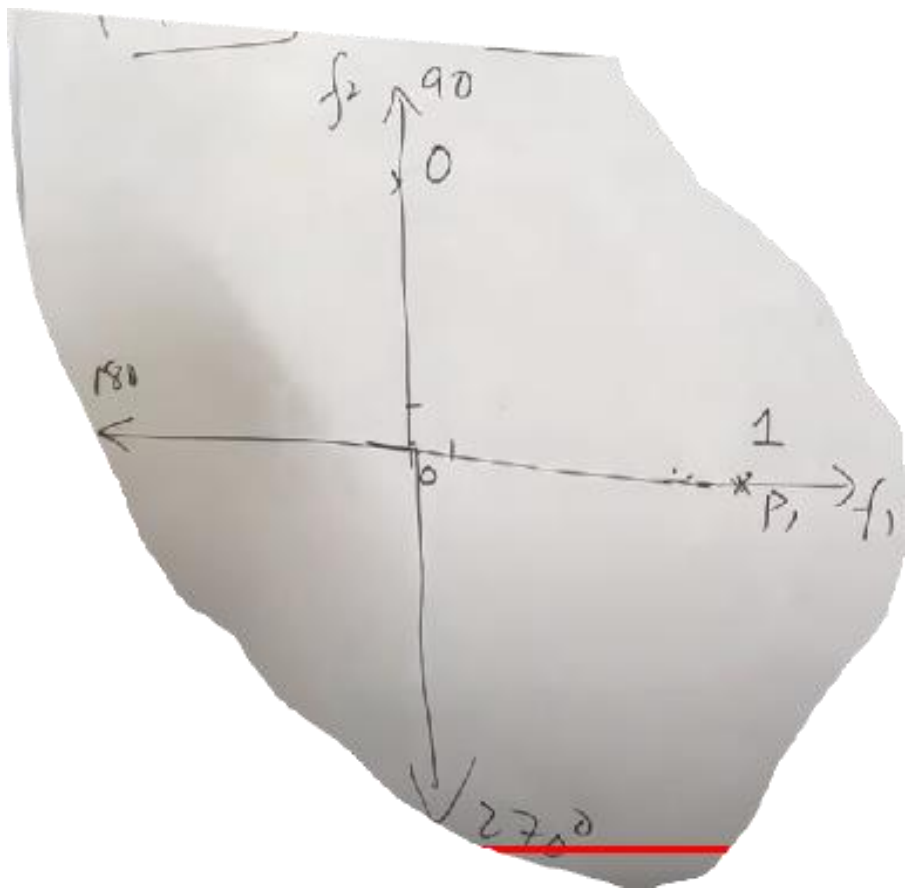
In this example P1 and P2 are separated by 90 degree, and hence the Cosine similarity is $\text{COS}(90) = 0$

If P1 and P2 are on the opposite side

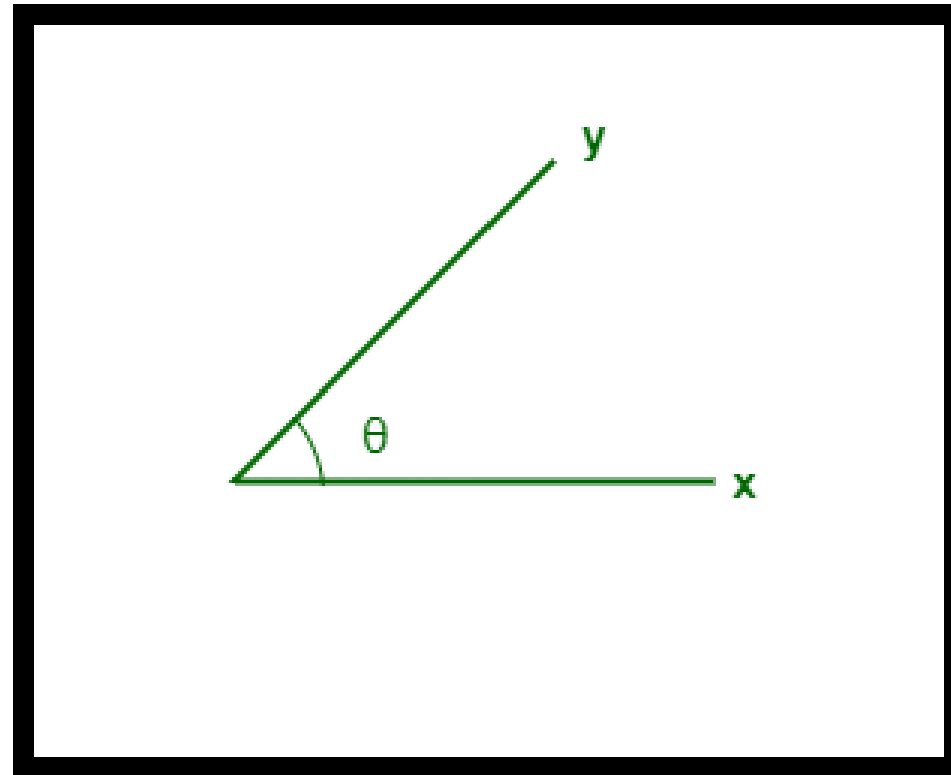
- If P1 and P2 are on the opposite side then the angle between them is 180 degree and hence the $\text{COS}(180) = -1$



- If it is 270, then again it will be 0, and 360 or 0 it will be 1.



Cosine Similarity



Advantages of Cosine Similarity

- The cosine similarity is beneficial because **even if the two similar data objects are far apart by the Euclidean distance because of the size, they could still have a smaller angle between them. Smaller the angle, higher the similarity.**
- When plotted on a multi-dimensional space, the cosine similarity **captures the orientation (the angle) of the data objects** and not the magnitude.

Example 1 for computing cosine distance

Consider an example to find the similarity between two vectors – ‘**x**’ and ‘**y**’, using Cosine Similarity. (if angle can not be estimated directly)

The ‘**x**’ vector has values, **x** = { 3, 2, 0, 5 }

The ‘**y**’ vector has values, **y** = { 1, 0, 0, 0 }

The formula for calculating the cosine similarity is : **Cos(x, y) = x . y / ||x|| * ||y||**

$$x \cdot y = 3*1 + 2*0 + 0*0 + 5*0 = 3$$

$$||x|| = \sqrt{(3)^2 + (2)^2 + (0)^2 + (5)^2} = 6.16$$

$$||y|| = \sqrt{(1)^2 + (0)^2 + (0)^2 + (0)^2} = 1$$

$$\therefore \text{Cos}(x, y) = 3 / (6.16 * 1) = 0.49$$

Example2 for computing cosine distance

$$d_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0; \quad d_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$$

$$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$||d_1|| = (3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} = 6.481$$

(square root of sum of squares of all the elements)

$$||d_2|| = (1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2)^{0.5} = (6)^{0.5} = 2.245$$

$$\begin{aligned} \text{So cosine similarity} &= \cos(d_1, d_2) = (d_1 \bullet d_2) / (||d_1|| * ||d_2||) \\ &= (5 / (6.481 * 2.245)) = 0.3150 \end{aligned}$$

Cosine distance (or it can be called dis-similarity)

$$= 1 - \cos(d_1, d_2) = 1 - 0.3436 = 0.6564$$

Find Cosine distance between

D1 = [5 3 8 1 9 6 0 4 2 1] D2 = [1 0 3 6 4 5 2 0 0 1]

When to use Cosine Similarity

- Cosine similarity looks at the angle between two vectors, euclidian similarity at the distance between two points. Hence it is very popular for NLP applications.
- Let's say you are in an e-commerce setting and you want to compare users for product recommendations:
 - User 1 bought 1x eggs, 1x flour and 1x sugar.
 - User 2 bought 100x eggs, 100x flour and 100x sugar
 - User 3 bought 1x eggs, 1x Vodka and 1x Red Bull
- By cosine similarity, user 1 and user 2 are more similar. By euclidean similarity, user 3 is more similar to user 1.

JACCARD SIMILARITY AND DISTANCE:

In Jaccard similarity instead of vectors, we will be using sets.

It is used to find the similarity between two sets.

Jaccard similarity is defined as the intersection of sets divided by their union. (count)

Jaccard similarity between two sets A and B is $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$

A simple example using set notation: How similar are these two sets?

$$A = \{0, 1, 2, 5, 6\}$$

$$B = \{0, 2, 3, 4, 5, 7, 9\}$$

$$J(A, B) = \{0, 2, 5\} / \{0, 1, 2, 3, 4, 5, 6, 7, 9\} = 3/9 = 0.33$$

Jaccard Similarity is given by : Overlapping vs Total items.

- Jaccard Similarity value ranges between 0 to 1
- 1 indicates highest similarity
- 0 indicates no similarity

Jaccard Similarity

overlapping

total items

$$0 \leq \text{Jaccard Similarity} \leq 1$$

Application of Jaccard Similarity

- Language processing is one example where jaccard similarity is used.

"Today I need to learn about Jaccard similarity."

"Later I will need other things to learn."

- In this example it is $4/12 = 0.33$

Jaccard Similarity is popularly used for ML model performance analysis

- In this example, table is designed against Actual vs predicted.

This gives an idea how our algorithm is working

- In the example is shows the overlapping +ve vs Total positives including actual and predicted

actual	predicted
0	0
1	0
0	0
1	1
0	1
0	0
1	1
0	1
0	0
1	1

Jaccard Similarity

$$\frac{\text{overlapping positives}}{\text{total positives}} = \frac{3}{6} = 0.5$$

		predicted	
		negative	positive
actual	negative	4	2
	positive	1	3

Jaccard Similarity

$$\frac{TP}{TP + FP + FN} = \frac{3}{6}$$

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

1. $d(p, q) \geq 0$ for all p and q and $d(p, q) = 0$ only if $p = q$. (Positive definiteness)
2. $d(p, q) = d(q, p)$ for all p and q . (Symmetry)
3. $d(p, r) \leq d(p, q) + d(q, r)$ for all points $p, q,$ and r . (Triangle Inequality)

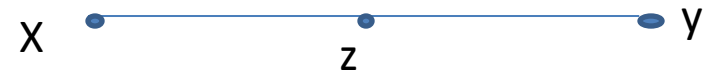
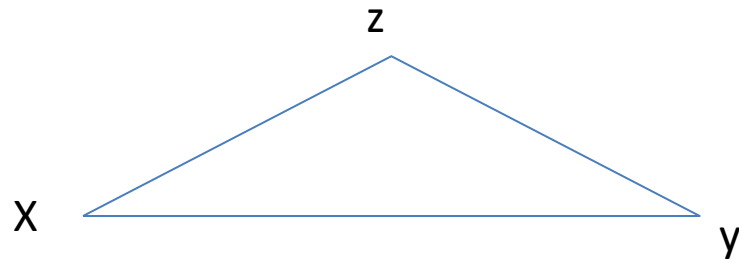
where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q .

- A distance that satisfies these properties is a **metric**, and a space is called a **metric space**

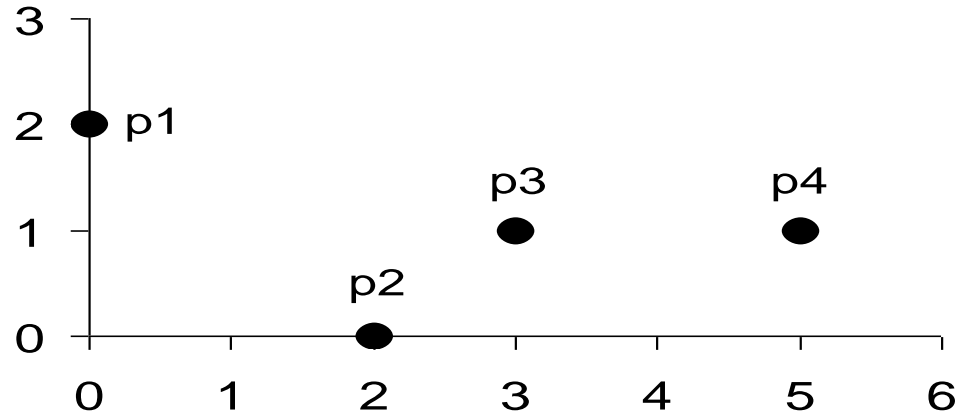
Distance Metrics Continued

- $\text{Dist}(x,y) \geq 0$
- $\text{Dist}(x,y) = \text{Dist}(y,x)$ are Symmetric
- Detours can not Shorten Distance

$$\text{Dist}(x,z) \leq \text{Dist}(x,y) + \text{Dist}(y,z)$$



Euclidean Distance

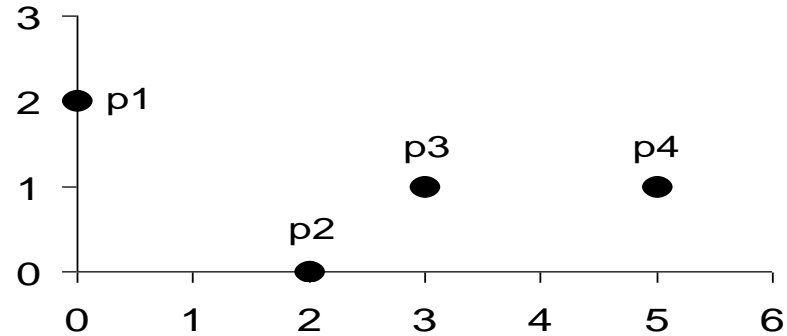


point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance



L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

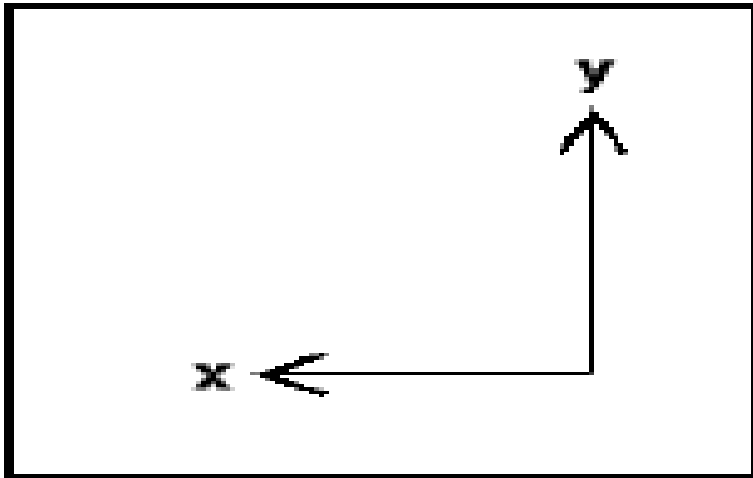
point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Distance Matrix

Summary of Distance Metrics

- Manhattan Distance

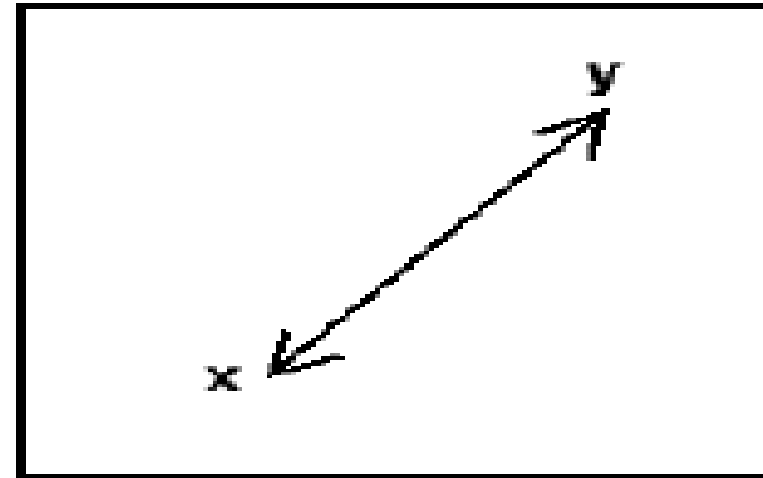
$$|X1-X2| + |Y1-Y2|$$



Manhattan

- Euclidean Distance

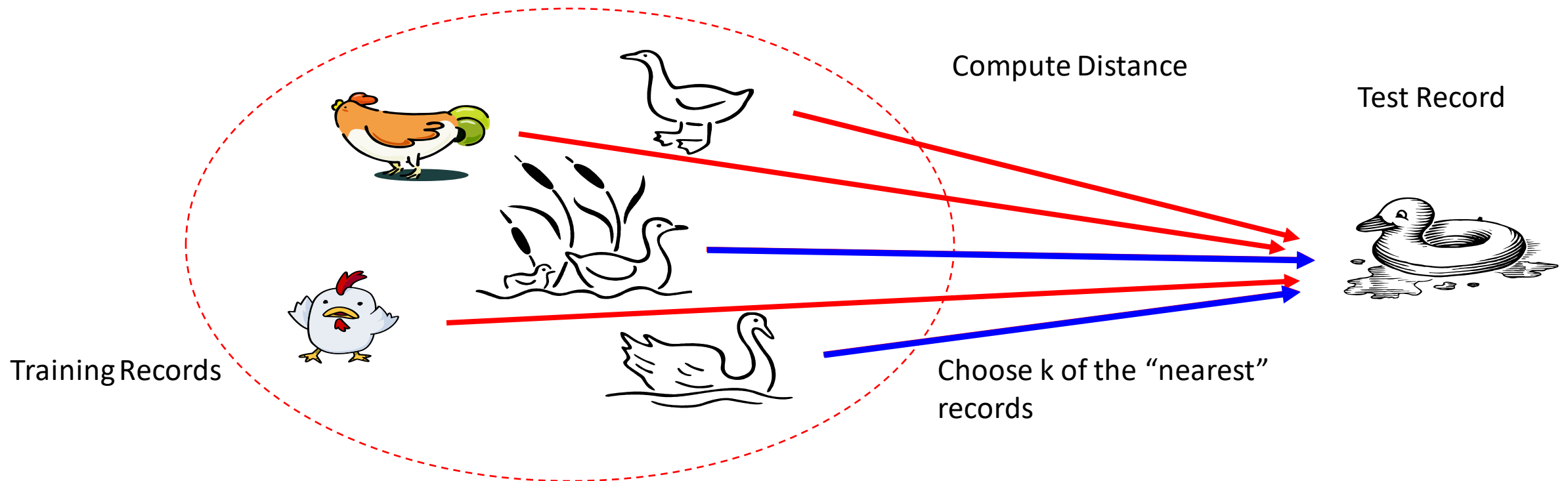
$$\sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$



Euclidean

Nearest Neighbors Classifiers

- Basic idea:
 - If it walks like a duck, quacks like a duck, then it's probably a duck



K-Nearest Neighbors (KNN) : ML algorithm

- Simple, but a very powerful classification algorithm
- Classifies based on a similarity measure
- This algorithm does not build a model
- Does not “learn” until the test example is submitted for classification
- Whenever we have a new data to classify, we find its K-nearest neighbors from the training data
- Classified by “MAJORITY VOTES” for its neighbor classes
- Assigned to the most common class amongst its K-Nearest Neighbors
(by measuring “distant” between data)
- In practice, k is usually chosen to be odd, so as to avoid ties
- The $k = 1$ rule is generally called the “nearest-neighbor classification” rule

K-Nearest Neighbors (KNN)

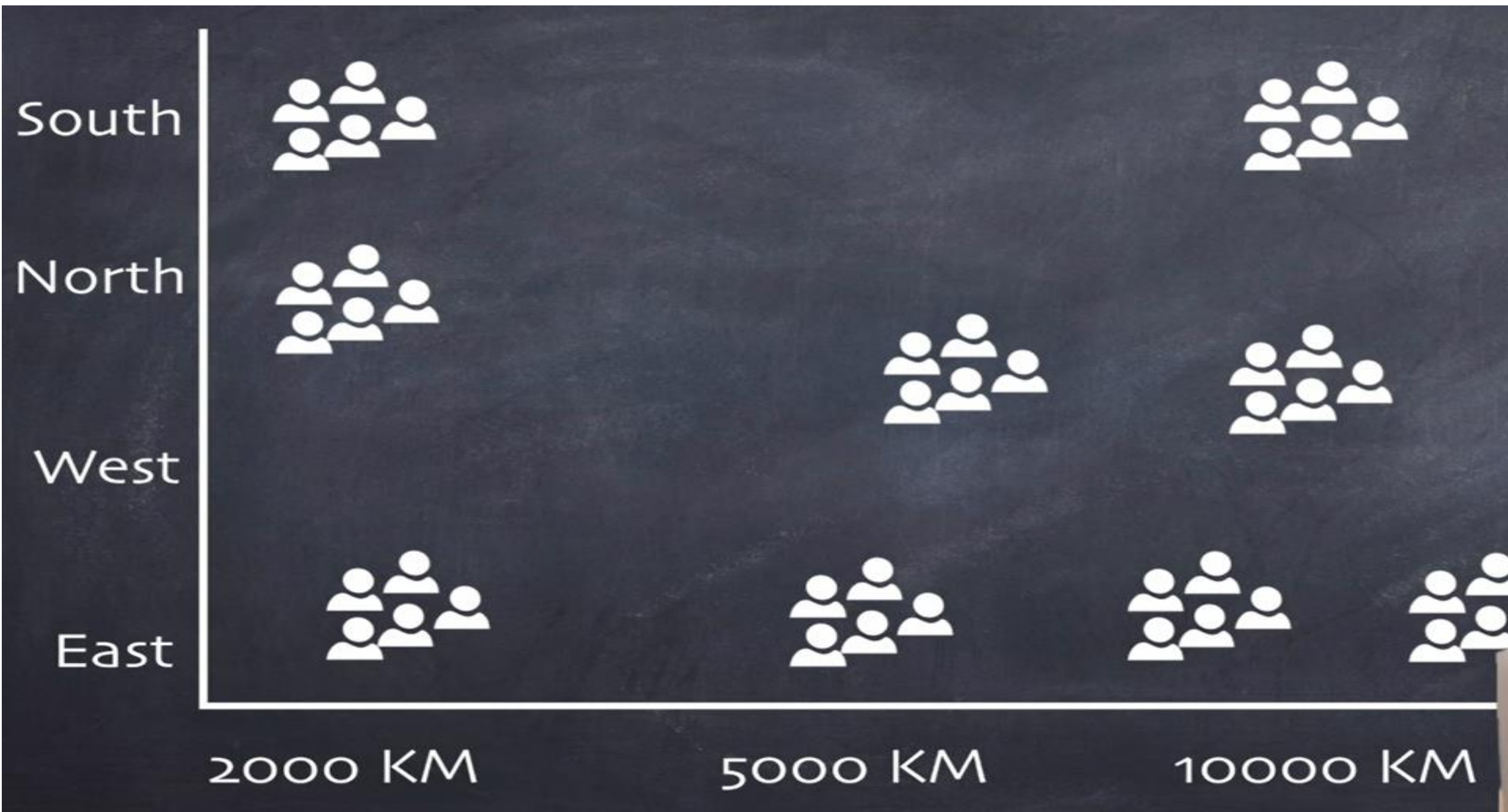
- K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data/Pattern and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm can be used for Regression as well as for Classification but **mostly it is used for the Classification problems.**
- K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.

- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Illustrative Example for KNN

Let's assume that a group of people are travelling to LONDON for a marathon

Based on the distance they travelled, guess the continent they are arriving from



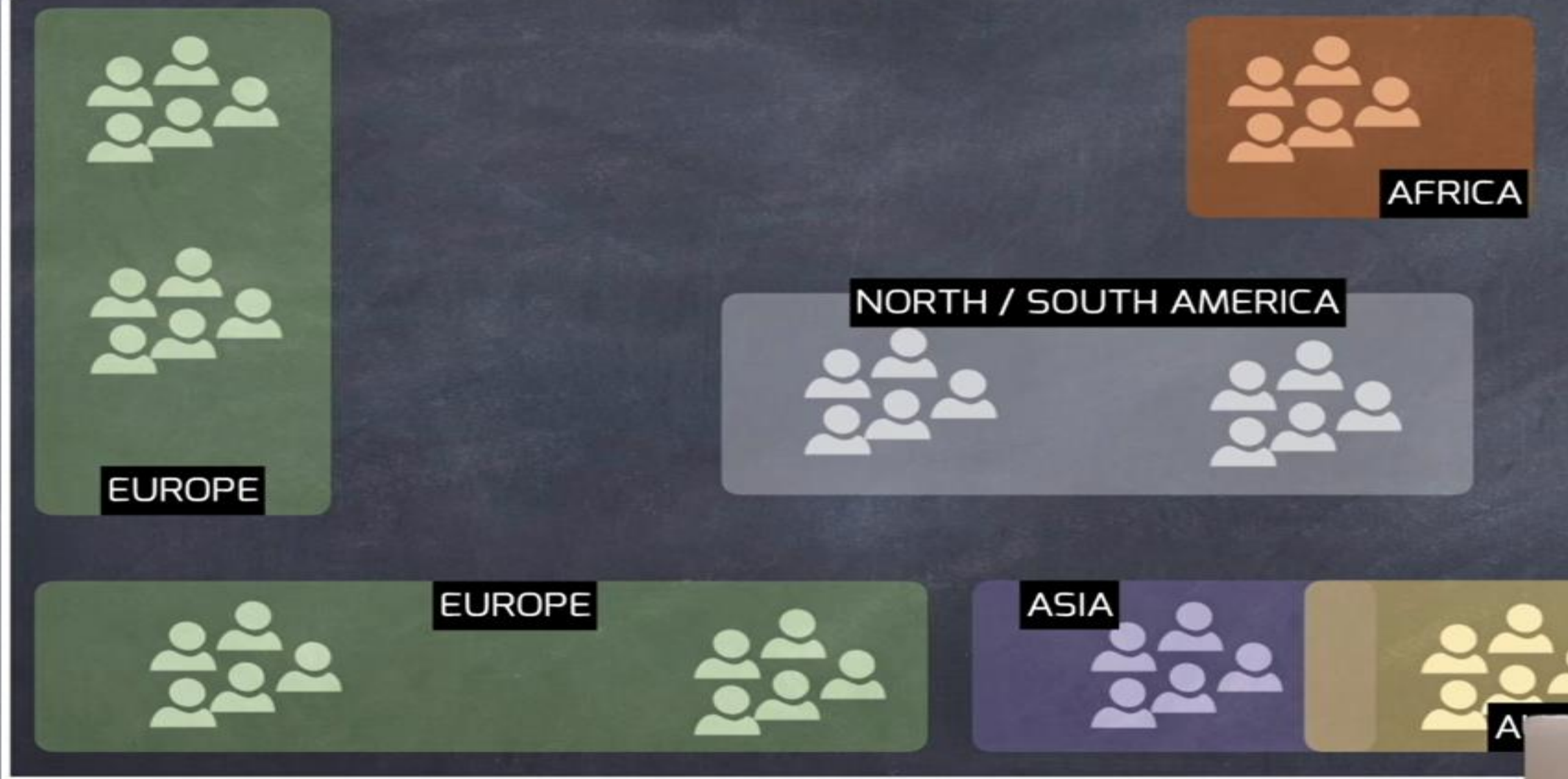


South

North

West

East



2000 KM

5000 KM

10000 KM

AFRICA

NORTH / SOUTH AMERICA

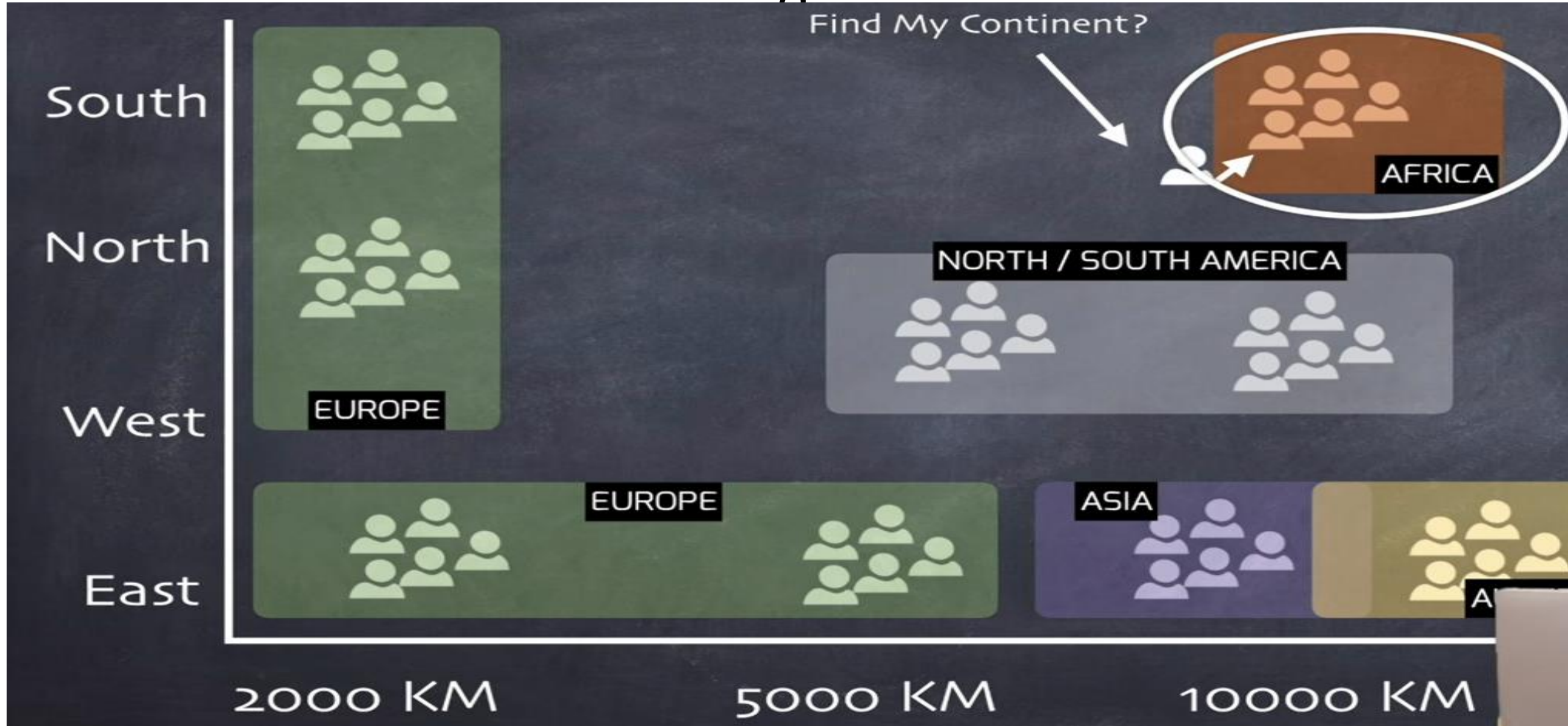
EUROPE

EUROPE

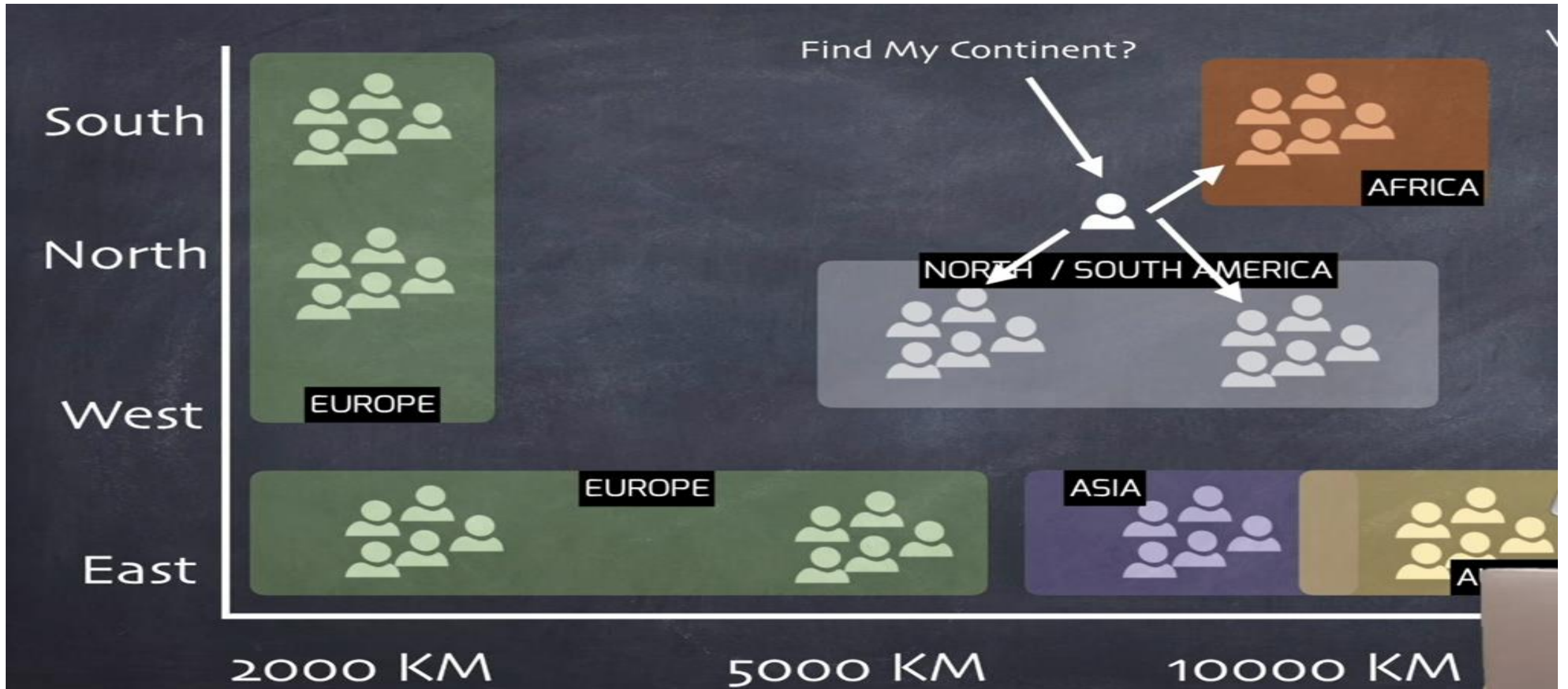
ASIA

A

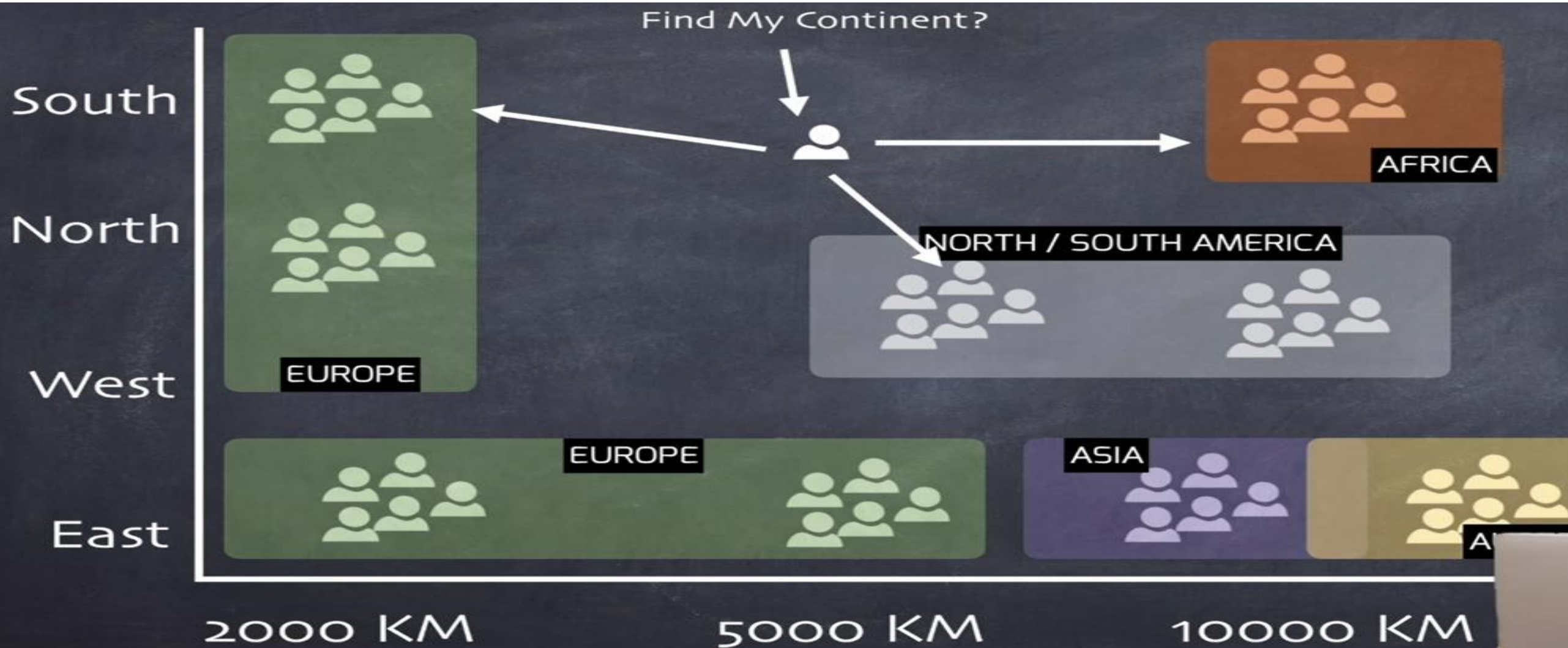
Considering $K=1$, based on nearest neighbor find the test data class- It belongs to class of africa



Now we have used $K=3$, and 2 are showing it is close to North/South America and hence the new data or data under testing belongs to that class.



In this case $K=3$... but still not a correct value to classify...Hence select a new value of K



Algorithm

- Step-1: Select the number K of the neighbors
- Step-2: Calculate the Euclidean distance to all the data points in training.
- Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
- Step-4: Among these k neighbors, apply voting algorithm
- Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.
- Step-6: Our model is ready.

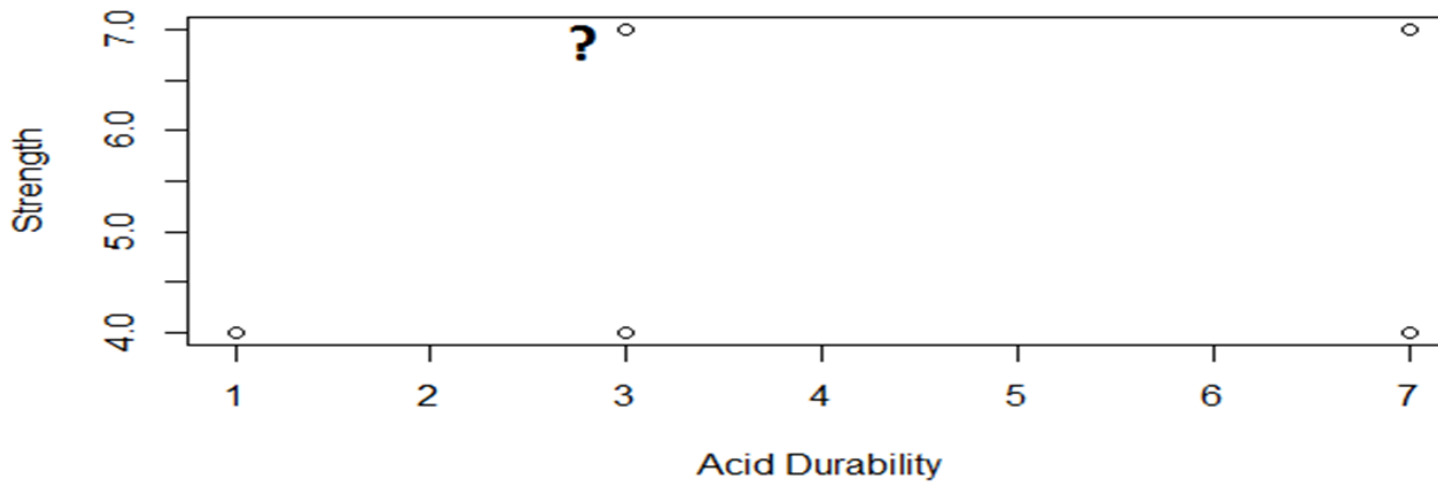
Consider the following data set of a pharmaceutical company with assigned class labels, using K nearest neighbour method classify a new unknown sample using $k = 3$ and $k = 2$.

Points	X1 (Acid Durability)	X2(strength)	Y=Classification
P1	7	7	BAD
P2	7	4	BAD
P3	3	4	GOOD
P4	1	4	GOOD

New pattern with $X1=3$, and $X2=7$ Identify the Class?

Points	X1(Acid Durability)	X2(Strength)	Y(Classification)
P1	7	7	BAD
P2	7	4	BAD
P3	3	4	GOOD
P4	1	4	GOOD
P5	3	7	?

Scatter plot



KNN








	P1	P2	P3	P4
Euclidean Distance of P5(3,7) from	(7,7)	(7,4)	(3,4)	(1,4)
	$\text{Sqrt}((7-3)^2 + (7-7)^2) = \sqrt{16}$ $= 4$	$\text{Sqrt}((7-3)^2 + (4-7)^2) = \sqrt{25}$ $= 5$	$\text{Sqrt}((3-3)^2 + (4-7)^2)$ $= \sqrt{9}$ $= 3$	$\text{Sqrt}((1-3)^2 + (4-7)^2)$ $= \sqrt{13}$ $= 3.60$

	P1	P2	P3	P4
Euclidean Distance of P5(3,7) from	(7,7)	(7,4)	(3,4)	(1,4)
	$\text{Sqrt}((7-3)^2 + (7-7)^2) = \sqrt{16} = 4$	$\text{Sqrt}((7-3)^2 + (4-7)^2) = \sqrt{25} = 5$	$\text{Sqrt}((3-3)^2 + (4-7)^2) = \sqrt{9} = 3$	$\text{Sqrt}((1-3)^2 + (4-7)^2) = \sqrt{13} = 3.60$
	BAD	BAD	GOOD	GOOD

Points	X1(Durability)	X2(Strength)	Y(Classification)
P1	7	7	BAD
P2	7	4	BAD
P3	3	4	GOOD
P4	1	4	GOOD
P5	3	7	GOOD







k-Nearest Neighbor Classifier

Example ($k=3$)

Customer	Age	Income (K)	No. of cards	Response	Distance from David
John 	35	35	3	Yes	$\sqrt{[(35-37)^2+(35-50)^2+(3-2)^2]}=15.16$
Rachel 	22	50	2	No	$\sqrt{[(22-37)^2+(50-50)^2+(2-2)^2]}=15$
Ruth 	63	200	1	No	$\sqrt{[(63-37)^2+(200-50)^2+(1-2)^2]}=152.23$
Tom 	59	170	1	No	$\sqrt{[(59-37)^2+(170-50)^2+(1-2)^2]}=122$
Neil 	25	40	4	Yes	$\sqrt{[(25-37)^2+(40-50)^2+(4-2)^2]}=15.74$
David 	37	50	2		

k-Nearest Neighbor Classifier

Example (k=3)

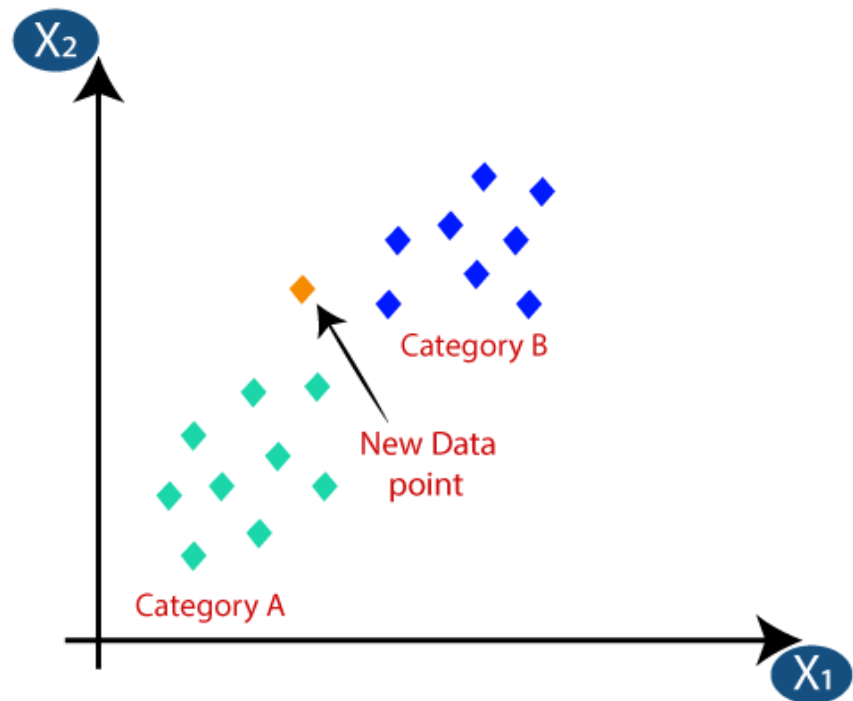
Customer	Age	Income (K)	No. of cards	Response	Distance from David
John 	35	35	3	Yes	$\text{sqrt} [(35-37)^2+(35-50)^2+(3-2)^2]=15.16$
Rachel 	22	50	2	No	$\text{sqrt} [(22-37)^2+(50-50)^2+(2-2)^2]=15$
Ruth 	63	200	1	No	$\text{sqrt} [(63-37)^2+(200-50)^2+(1-2)^2]=152.23$
Tom 	59	170	1	No	$\text{sqrt} [(59-37)^2+(170-50)^2+(1-2)^2]=122$
Neil 	25	40	4	Yes	$\text{sqrt} [(25-37)^2+(40-50)^2+(4-2)^2]=15.74$
David 	37	50	2	Yes	

Height (in cms)	Weight (in kgs)	T Shirt Size
158	58	M
158	59	M
158	63	M
160	59	M
160	60	M
163	60	M
163	61	M
160	64	L
163	64	L
165	61	L
165	62	L
165	65	L
168	62	L
168	63	L
168	66	L
170	63	L
170	64	L
170	68	L

New customer named 'Mary' has height 161cm and weight 61kg.

Suggest the T shirt Size with $K=3,5$ using Euclidean Distance and also Manhattan Distance

There is a Car manufacturer company that has manufactured a new SUV car. The company wants to give the ads to the users who are interested in buying that SUV. So for this problem, we have a dataset that contains multiple user's information through the social network. The dataset contains lots of information but the Estimated Salary and Age we will consider for the independent variable and the Purchased variable is for the dependent variable. Dataset is as shown in the table. Using $K = 5$ classify the new sample



User ID	Gender	Age	EstimatedSalary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0
15570769	Female	26	80000	0
15606274	Female	26	52000	0
15746139	Male	20	86000	0
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1
15617482	Male	45	26000	1
15704583	Male	46	28000	1
15621083	Female	48	29000	1
15649487	Male	45	22000	1
15736760	Female	47	49000	1

- There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
- A very low value for K such as $K=1$ or $K=2$, can be noisy and lead to the effects of outliers in the model.
- Large values for K are good, but it may find some difficulties.

- Advantages of KNN Algorithm:
- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

- Disadvantages of KNN Algorithm:
- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

Consider the training examples shown in the following table for a binary classification. The table shows a training set for a problem of predicting whether a loan applicant will repay his/her loan obligation or defaulting on his/her loan.

<i>Tid</i>	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

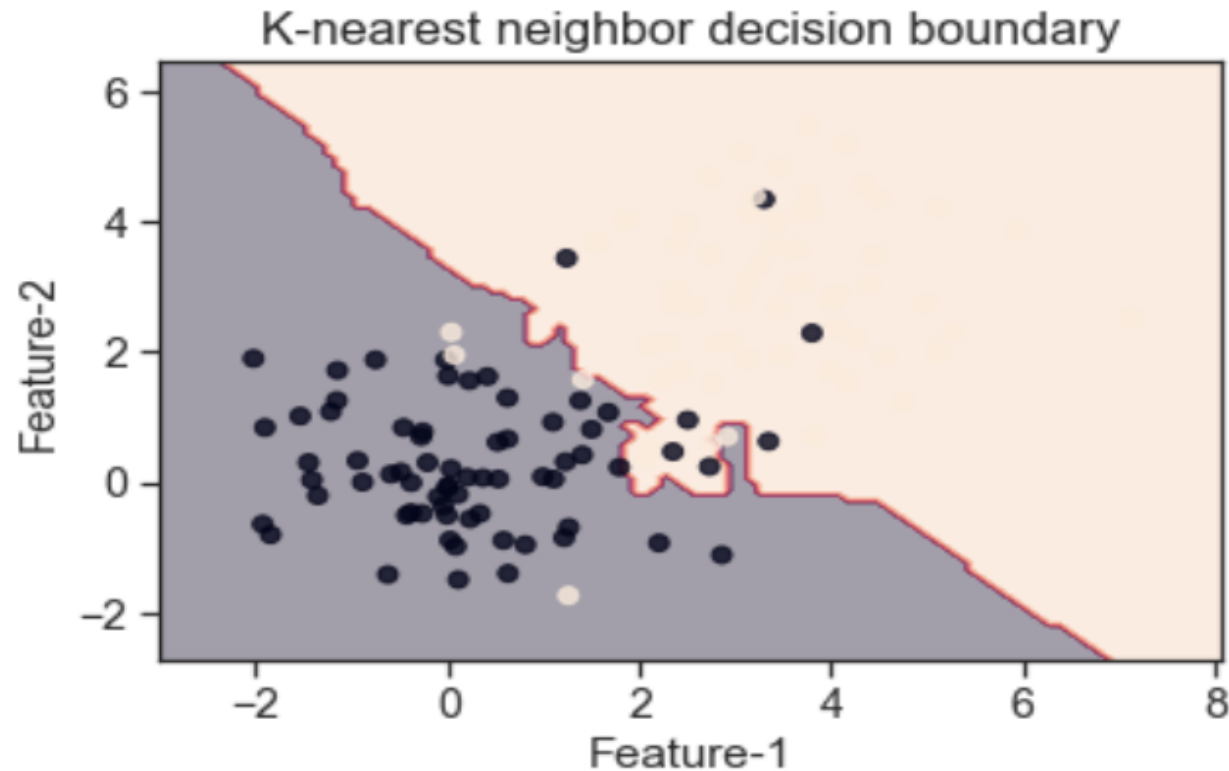
Using the kNN approach that we discussed in the class, predict the class label for this test example, $X = (\text{Home Owner} = \text{No}, \text{Marital Status} = \text{Married}, \text{Income} = \$120\text{K})$. Assume that $k = 3$ and distance is L2 norm.

Another example: solve

ID	Height	Age	Weight
1	5	45	77
2	5.11	26	47
3	5.6	30	55
4	5.9	34	59
5	4.8	40	72
6	5.8	36	60
7	5.3	19	40
8	5.8	28	60
9	5.5	23	45
10	5.6	32	58
11	5.5	38	?

- Because the distance function used to find the k nearest neighbors is not linear, so **it usually won't lead to a linear decision boundary.**

```
plt.figure()  
plt.title("K-nearest neighbor decision boundary",fontsize=16)  
plot_decision_boundaries(X_train,y_train,KNeighborsClassifier)  
plt.show()
```



Adaptive decision Boundaries

- Nearest neighbour techniques can approximate arbitrarily complicated decision regions, but their error rates may be larger than Bayesian rates.
- Experimentation may be required to choose K and to edit the reference samples.
- Classification may be time consuming if the number of reference samples is large.
- An alternate solution is to assume that the functional form of the decision boundary between each pair of classes is given, and to find the decision boundary of that form which best separates the classes in some sense.

Adaptive Decision Boundaries. Continued

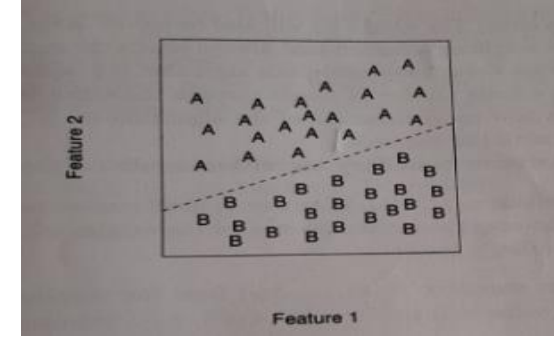
- For example, assume that a linear decision boundary will be used to classify samples into two classes and each sample has M features.
- Then the discriminant function has the form

$$D = w_0 + w_1x_1 + \dots + w_Mx_M$$

- If $D = 0$ is the equation of the decision boundary between the two classes.
- The weights w_0, w_1, \dots, w_M are to be chosen to provide good performance on the set.
- A sample with vector (x_1, x_2, \dots, x_M) is classified into one class, say class 1 if $D > 0$ and into another class say -1 if $D < 0$
- **w_0 is the intercept and w_1, w_2, \dots, w_M are all the weights related to slopes.**

It is of the form $Y = Mx + C$ or $Y = C + Mx$

Adaptive decision boundaries ...continued



- Geometrically with $D=0$ is the equation of a hyperplane decision boundary that divides the M -dimensional feature space into two regions
- Two classes are said to be linearly separable if there exists a hyperplane decision boundary such that $D>0$ for all the samples in class 1 and $D<0$ for all the samples in the class -1.
- Figure shows two classes which are separated by a hyperplane.
- Weights w_1, w_2, \dots, w_M can be varied. Boundary will be adapted based on the weights.
- During the **adaptive or training phase**, samples are presented to the current form of the classifier. Whenever a sample is correctly classified no change is made in the weights.
- When a sample is incorrectly classified, each weight is changed to correct the output.

Adaptive decision boundary algorithm

1. Initialize the weights w_0, w_1, \dots, w_M to zero or to small random values to some initial guesses.
2. Choose the next sample $x=(x_1, x_2, \dots, x_M)$ from the training set. Let the 'true' class or desired value of D be d , so that $d=1$ or -1 represents the true class of x .
3. Compute $D=w_0+w_1x_1+\dots+w_Mx_M$.
4. If D not equal to d , replace w_i by (w_i+cdx_i) (small change).
5. Repeat the steps 2 to 4 with each samples in the training set. When finished run through the entire training data set again.
6. Stop and report perfect classification when all the samples are classified properly.

- If there are N classes and M features the set of linear discriminant function is
- $D_1 = w_{10} + w_{11}x_1 + \dots + w_{1M}x_M$
- $D_2 = w_{20} + w_{21}x_1 + \dots + w_{2M}x_M$
-
- $D_n = w_{n0} + w_{n1}x_1 + \dots + w_{nM}x_M$

Minimum Squared Error Discriminant Functions

- Although the adaptive decision boundary and adaptive discriminate function techniques have considerable appeal, it requires lot of iterations.
- Alternate solution is to have the “ Minimum Squared Error (MSE)” classification procedure.
- MSE does not require iteration.
- MSE uses single discriminant function regardless of the number of classes.

MSE

- If there are V samples and M features for each sample, there will be V feature vectors
 $x_i = (x_{i1}, x_{i2}, \dots, x_{iM}), i=1$ to V
- Let the true class of x_i be represented by d_i , which can have any numerical value. We want to find a set of weights $w_j, j=0, \dots, M$ for single linear discriminant function.
 - $D(x_i) = w_0 + w_1 x_{i1} + \dots + w_M x_{iM}$
 - Such that $D(x_i) = d_i$ for all the samples i . In general it will not be possible
- But by properly choosing the weights w_0, w_1, \dots, w_M , the sum of the squared differences between the set of desired values d_i and the actual values $D(x_i)$ can be minimized. The sum of the squared errors E is
- $$E = \sum_{i=1}^V ((D(x_i) - d_i)^2).$$
- The values of the weights that minimize E may be found by computing the partial derivatives of E with respect to each of the w_j , setting each derivative to zero and solving for the weights w_0, \dots, w_M

End of Unit 3

Unit4
Clustering
Dr. Srinath.S

Unit – 4 Syllabus

- **Clustering:** Introduction
- Hierarchical Clustering:
 - Agglomerative Clustering Algorithm
 - The single Linkage Algorithm
 - The Complete Linkage Algorithm
 - The Average – Linkage Algorithm
- Partitional Clustering:
 - Forgy's Algorithm
 - The K-Means Algorithm

Introduction

- In the earlier chapters, we saw that how samples may be classified if a training set is available to use in the design of a classifier.
- However in many situations classes are themselves are initially undefined.
- Given a set of feature vectors sampled from some population, we would like to know if the data set consists of a number of relatively distinct subsets, then we can define them to be classes.
- This is sometimes called as class discovery or unsupervised classification
- Clustering refers to the process of grouping samples so that the samples are similar within each group. The groups are called clusters.

What is Clustering?

- Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups.
- In simple words, the aim is to segregate groups with similar traits and assign them into clusters.
- A good clustering will have high intra-class similarity and low inter-class similarity

Applications of Clustering

- Recommendation engines
- Market segmentation
- Social network analysis
- Search result grouping
- Medical imaging
- Image segmentation
- Anomaly detection

Types of clustering:

- Hierarchical Clustering:
 - Agglomerative Clustering Algorithm
 - The single Linkage Algorithm
 - The Complete Linkage Algorithm
 - The Average – Linkage Algorithm
 - Divisive approach
 - Polythetic The division is based on more than one feature.
 - Monothetic Only one feature is considered at a time.
- Partitional Clustering:
 - Forgy's Algorithm
 - The K-Means Algorithm
 - The Isodata Algorithm.

Example: Agglomerative

- 100 students from India join MS program in some particular university in USA.
- Initially each one of them looks like single cluster.
- After some times, 2 students from SJCE, Mysuru makes a cluster.
- Similarly another cluster of 3 students(patterns / Samples) from RVCE meets SJCE students.
- Now these two clusters makes another bigger cluster of Karnataka students.
- Later ... south Indian student cluster and so on...

Example : Divisive approach

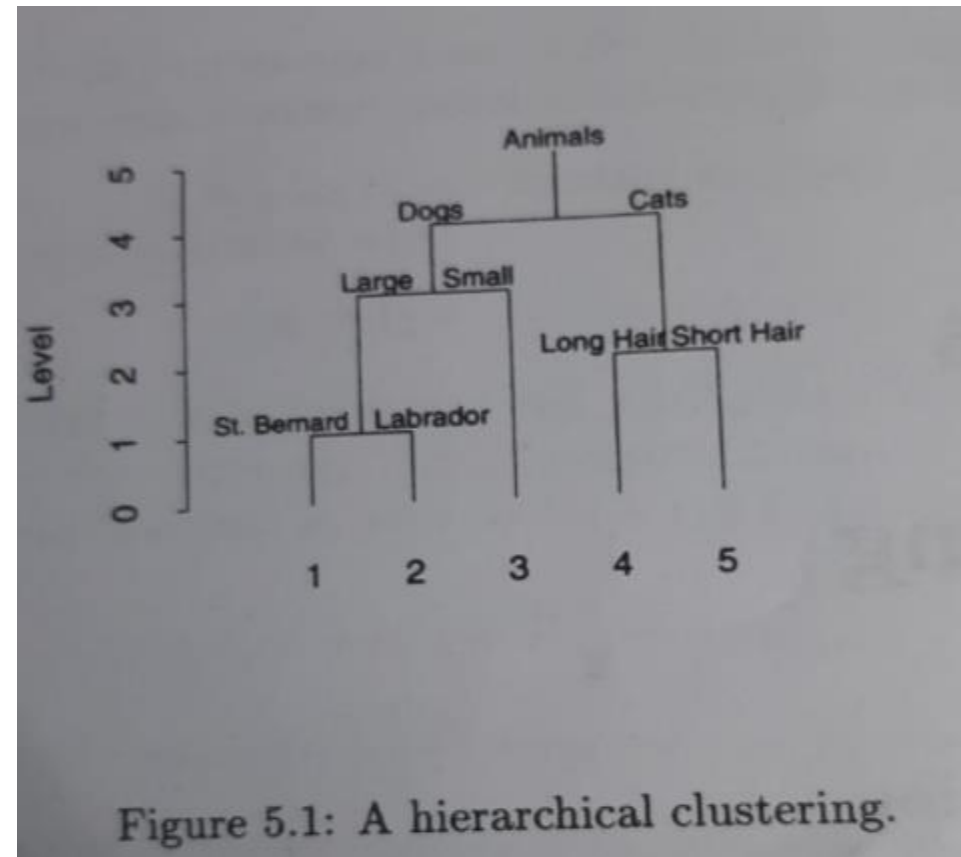
- In a large gathering of engineering students..
 - Separate JSS S&TU students
 - Further computer science students
 - Again ..7th sem students
 - » In sub group and divisive cluster is C section students.

Hierarchical clustering

- Hierarchical clustering refers to a clustering process that organizes the data into large groups, which contain smaller groups and so on.
- A hierarchical clustering may be drawn as a tree or dendrogram.
- The finest grouping is at the bottom of the dendrogram, each sample by itself forms a cluster.
- At the top of the dendrogram, where all samples are grouped into one cluster.

Hierarchical clustering

- Figure shown in figure illustrates hierarchical clustering.
- At the top level we have Animals... followed by sub groups...
- Do not have to assume any particular number of clusters.
- The representation is called dendrogram.
- Any desired number of clusters can be obtained by 'cutting' the **dendrogram** at the proper level.



Two types of Hierarchical Clustering

– Agglomerative:

- It is the most popular algorithm, It is popular than divisive algorithm.
- Start with the points as individual clusters
- It follows bottom up approach
- At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
- Ex: single-linkage, complete-linkage, Average linking algorithm etc.

– Divisive:

- Start with one, all-inclusive cluster
- At each step, split a cluster until each cluster contains a point (or there are k clusters)
- Traditional hierarchical algorithms use a similarity or distance matrix
 - Merge or split one cluster at a time

Agglomerative Clustering Algorithm

1. Compute the proximity matrix
2. Let each data point be a cluster
- 3. Repeat**
4. Merge the two closest clusters
5. Update the proximity matrix
- 6. Until** only a single cluster remains

Key operation is the computation of the proximity of two clusters

- Different approaches to defining the distance between clusters distinguish the different algorithms

Some commonly used criteria in Agglomerative clustering Algorithms

(The most popular distance measure used is Euclidean distance)

Single Linkage:

Distance between two clusters is the smallest pairwise distance between two observations/nodes, each belonging to different clusters.

Complete Linkage:

Distance between two clusters is the largest pairwise distance between two observations/nodes, each belonging to different clusters.

Mean or average linkage clustering:

Distance between two clusters is the average of all the pairwise distances, each node/observation belonging to different clusters.

Centroid linkage clustering:

Distance between two clusters is the distance between their centroids.

Single linkage algorithm

- Consider the following scatter plot points.
- In single link hierarchical clustering, we merge in each step the two clusters, whose two closest members have the smallest distance

	x	y
1	4	4
2	8	4
3	15	8
4	24	4
5	24	12

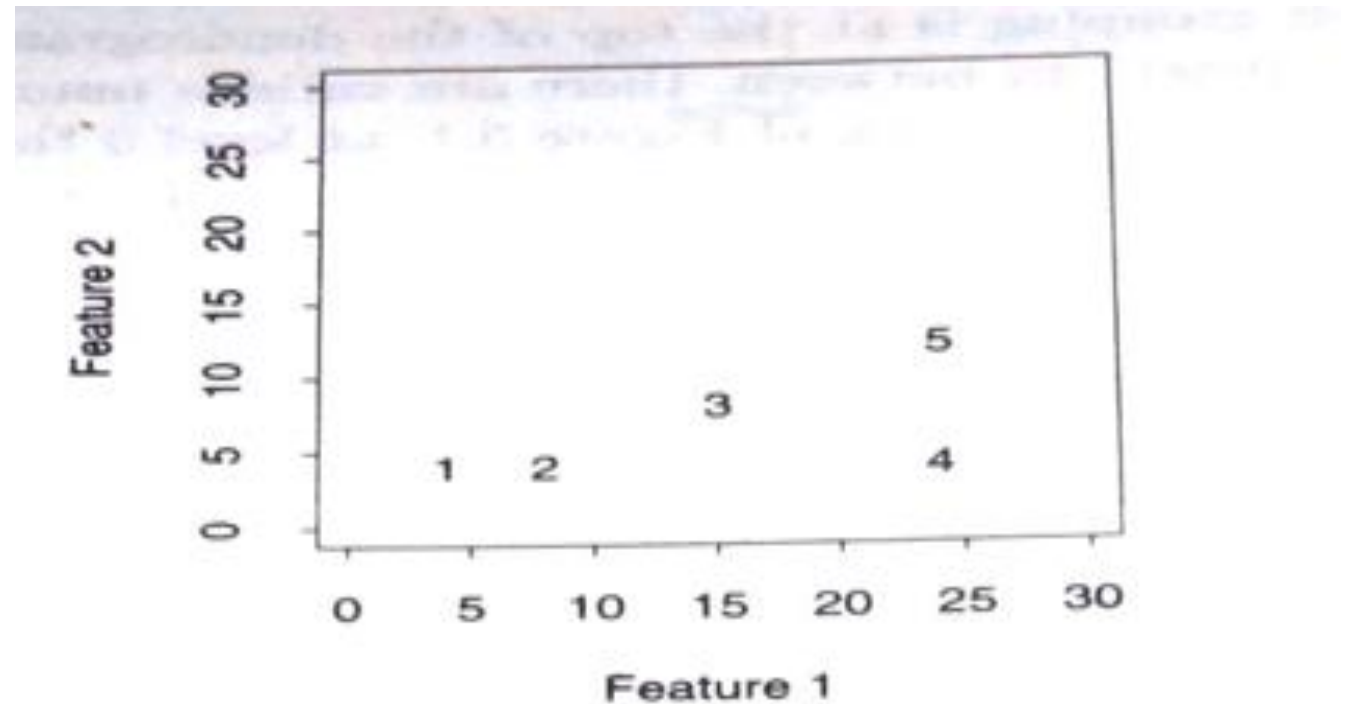


Figure 5.2: Samples for clustering.

Single linkage... Continued

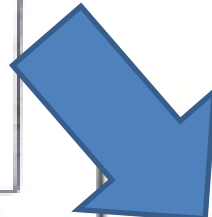
- The single linkage algorithm is also known as the minimum method and the nearest neighbor method.
- Consider C_i and C_j are two clusters.
- 'a' and 'b' are samples from cluster C_i and C_j respectively.

$$D_{SL}(C_i, C_j) = \min_{a \in C_i, b \in C_j} d(a, b),$$

- Where $d(a, b)$ represents the distance between 'a' and 'b'.

First level of distance computation D1 (Euclidean distance used)

	x	y
1	4	4
2	8	4
3	15	8
4	24	4
5	24	12

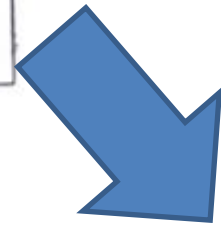


	1	2	3	4	5
1	—	4.0	11.7	20.0	21.5
2	4.0	—	8.1	16.0	17.9
3	11.7	8.1	—	9.8	9.8
4	20.0	16.0	9.8	—	8.0
5	21.5	17.9	9.8	8.0	—

- Use Euclidean distance for distance between samples.
- The table shown in the previous slide gives feature values for each sample and the distance d between each pair of samples.
- The algorithm begins with five clusters, each consisting of one sample.
- The two nearest clusters are then merged.
- The smallest number is 4 which is the distance between (1 and 2), so they are merged. Merged matrix is as shown in next slide.

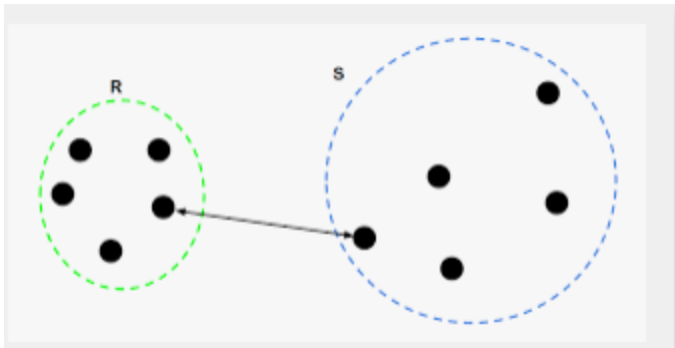
D2 matrix

	1	2	3	4	5
1	—	4.0	11.7	20.0	21.5
2	4.0	—	8.1	16.0	17.9
3	11.7	8.1	—	9.8	9.8
4	20.0	16.0	9.8	—	8.0
5	21.5	17.9	9.8	8.0	—



$\{1, 2\}, \{3\}, \{4\}, \{5\}.$

Next obtain the matrix that gives the distances between these clusters:

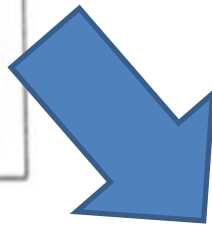


	$\{1,2\}$	3	4	5
$\{1,2\}$	—	8.1	16.0	17.9
3	8.1	—	9.8	9.8
4	16.0	9.8	—	8.0
5	17.9	9.8	8.0	—

- In the next level, the smallest number in the matrix is 8
- It is between 4 and 5.
- Now the cluster 4 and 5 are merged.
- With this we will have 3 clusters: $\{1,2\}$, $\{3\}$, $\{4,5\}$
- The matrix is as shown in the next slide.

D3 distance

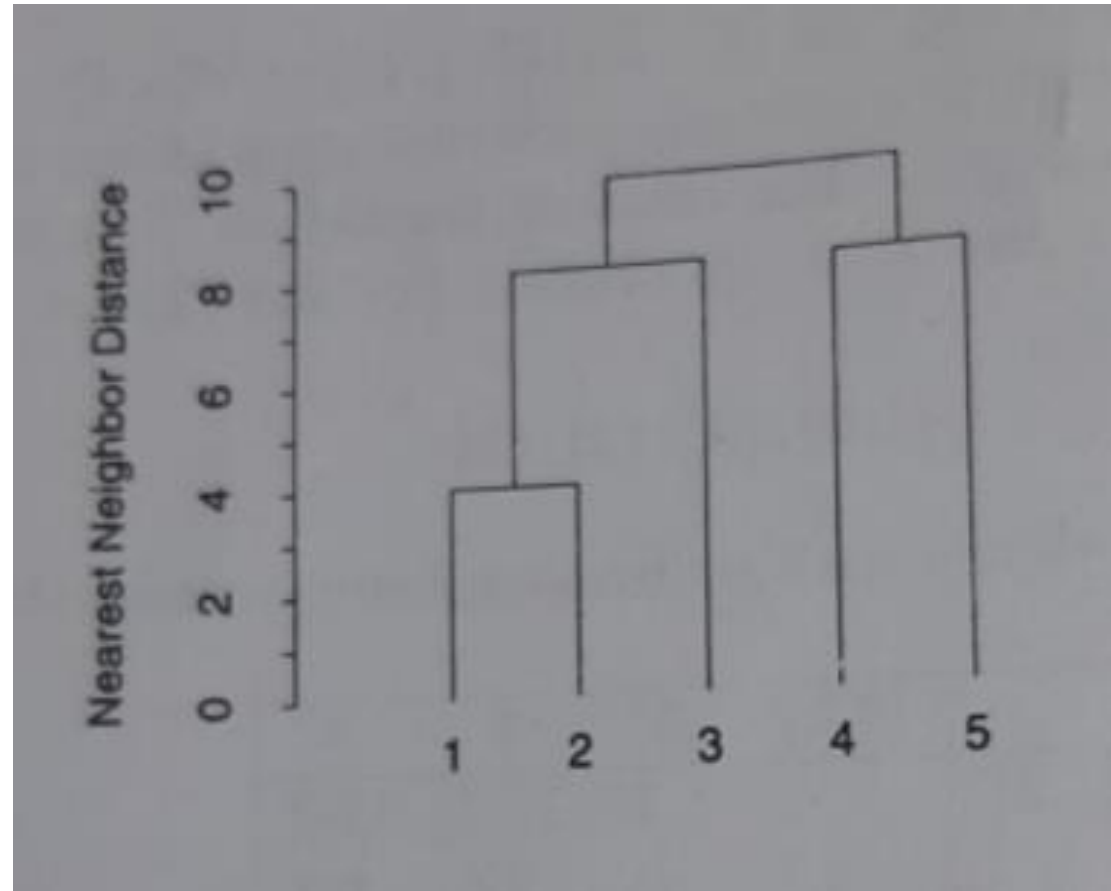
	{1,2}	3	4	5
{1,2}	—	8.1	16.0	17.9
3	8.1	—	9.8	9.8
4	16.0	9.8	—	8.0
5	17.9	9.8	8.0	—



	{1,2}	3	{4,5}
{1,2}	—	8.1	16.0
3	8.1	—	9.8
{4,5}	16.0	9.8	—

- In the next step $\{1,2\}$ will be merged with $\{3\}$.
- Now we will have two cluster $\{1,2,3\}$ and $\{4,5\}$
- In the next step.. these two are merged to have single cluster.
- Dendrogram is as shown here.
- Height of the dendrogram is decided based on the merger distance.

For example: 1 and 2 are merged at the least distance 4. hence the height is 4.



The complete linkage Algorithm

- It is also called the **maximum method or the farthest neighbor method**.
- It is obtained by defining the distance between two clusters to be largest distance between a sample in one cluster and a sample in the other cluster.
- If C_i and C_j are clusters, we define:

$$D_{CL}(C_i, C_j) = \max_{a \in C_i, b \in C_j} d(a, b).$$

Example : Complete linkage algorithm

- Consider the same samples used in single linkage:
- Apply Euclidean distance and compute the distance.
- Algorithm starts with 5 clusters.
- As earlier samples 1 and 2 are the closest, they are merged first.

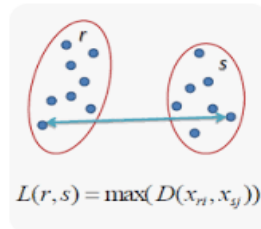
	x	y
1	4	4
2	8	4
3	15	8
4	24	4
5	24	12



	1	2	3	4	5
1	—	4.0	11.7	20.0	21.5
2	4.0	—	8.1	16.0	17.9
3	11.7	8.1	—	9.8	9.8
4	20.0	16.0	9.8	—	8.0
5	21.5	17.9	9.8	8.0	—

- While merging the maximum distance will be used to replace the distance/ cost value.
- For example, the distance between 1&3 = 11.7 and 2&3=8.1. This algorithm selects 11.7 as the distance.
- In complete linkage hierarchical clustering, the distance between two clusters is defined as **the longest distance between two points in each cluster.**

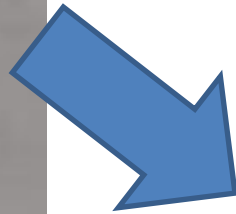
	1	2	3	4	5
1	—	4.0	11.7	20.0	21.5
2	4.0	—	8.1	16.0	17.9
3	11.7	8.1	—	9.8	9.8
4	20.0	16.0	9.8	—	8.0
5	21.5	17.9	9.8	8.0	—



	{1,2}	3	4	5
{1,2}	—	11.7	20.0	21.5
3	11.7	—	9.8	9.8
4	20.0	9.8	—	8.0
5	21.5	9.8	8.0	—

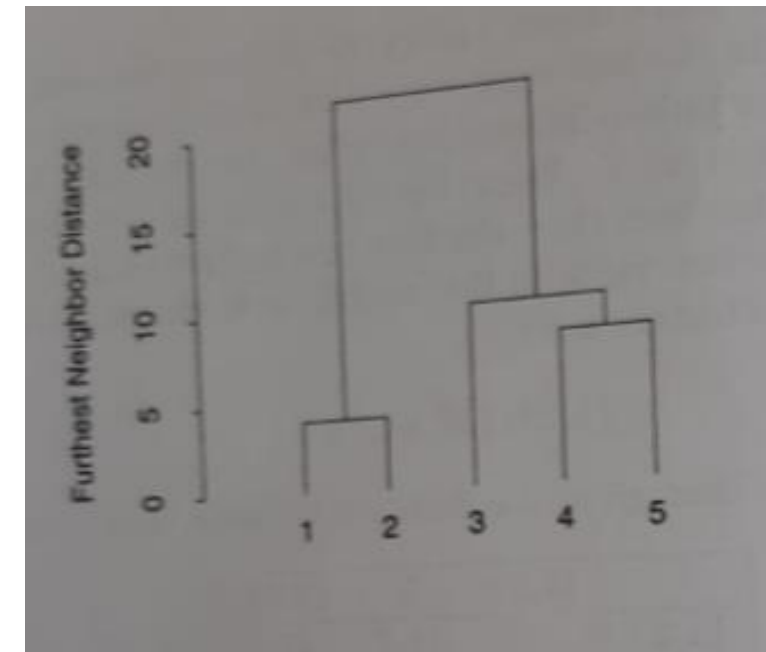
- In the next level, the smallest distance in the matrix is 8.0 between 4 and 5. Now merge 4 and 5.

	{1,2}	3	4	5
{1,2}	—	11.7	20.0	21.5
3	11.7	—	9.8	9.8
4	20.0	9.8	—	8.0
5	21.5	9.8	8.0	—



	{1,2}	3	{4,5}
{1,2}	—	11.7	21.5
3	11.7	—	9.8
{4,5}	21.5	9.8	—

- In the next step, the smallest distance is 9.8 between 3 and {4,5}, they are merged.
- At this stage we will have two clusters {1,2} and {3,4,5}.
- Notice that these clusters are different from those obtained from single linkage algorithm.
- At the next step, the two remaining clusters will be merged.
- The hierarchical clustering will be complete.
- The dendrogram is as shown in the figure.



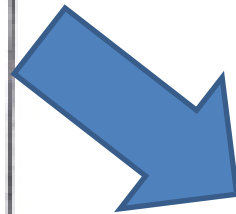
The Average Linkage Algorithm

- The average linkage algorithm, is an attempt to compromise between the extremes of the single and complete linkage algorithm.
- It is also known as the **unweighted pair group method using arithmetic averages**.

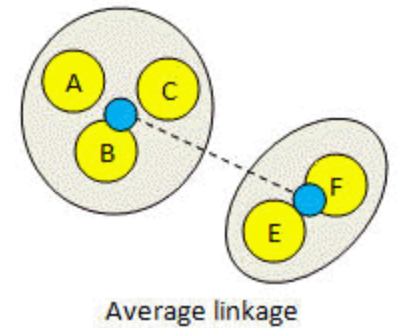
Example: Average linkage clustering algorithm

- Consider the same samples: compute the Euclidian distance between the samples

	x	y
1	4	4
2	8	4
3	15	8
4	24	4
5	24	12



	1	2	3	4	5
1	—	4.0	11.7	20.0	21.5
2	4.0	—	8.1	16.0	17.9
3	11.7	8.1	—	9.8	9.8
4	20.0	16.0	9.8	—	8.0
5	21.5	17.9	9.8	8.0	—



- In the next step, cluster 1 and 2 are merged, as the distance between them is the least.
- The distance values are computed based on the average values.
- For example distance between 1 & 3 =11.7 and 2&3=8.1 and the average is 9.9. This value is replaced in the matrix between {1,2} and 3.

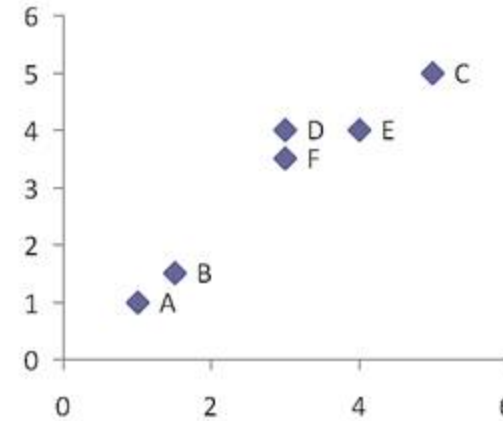
	{1,2}	3	4	5
{1,2}	—	9.9	18.0	19.7
3	9.9	—	9.8	9.8
4	18	9.8	—	8.0
5	19.7	9.8	8.0	—

- In the next stage 4 and 5 are merged:

	{1,2}	3	{4,5}
{1,2}	—	9.9	18.9
3	9.9	—	9.8
{4,5}	18.9	9.8	—

Example 2: Single Linkage

	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5



Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Min Distance (Single Linkage)

Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	?	4.24
B	0.71	0.00	4.95	?	3.54
C	5.66	4.95	0.00	?	1.41
D, F	?	?	?	0.00	?
E	4.24	3.54	1.41	?	0.00

Then, the updated distance matrix becomes

Min Distance (Single Linkage)

Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	3.20	4.24
B	0.71	0.00	4.95	2.50	3.54
C	5.66	4.95	0.00	2.24	1.41
D, F	3.20	2.50	2.24	0.00	1.00
E	4.24	3.54	1.41	1.00	0.00

Then the updated distance matrix is

Min Distance (Single Linkage)

Dist	A,B	C	(D, F)	E
A,B	0	4.95	2.50	3.54
C	4.95	0	2.24	1.41
(D, F)	2.50	2.24	0	1.00
E	3.54	1.41	1.00	0

Min Distance (Single Linkage)

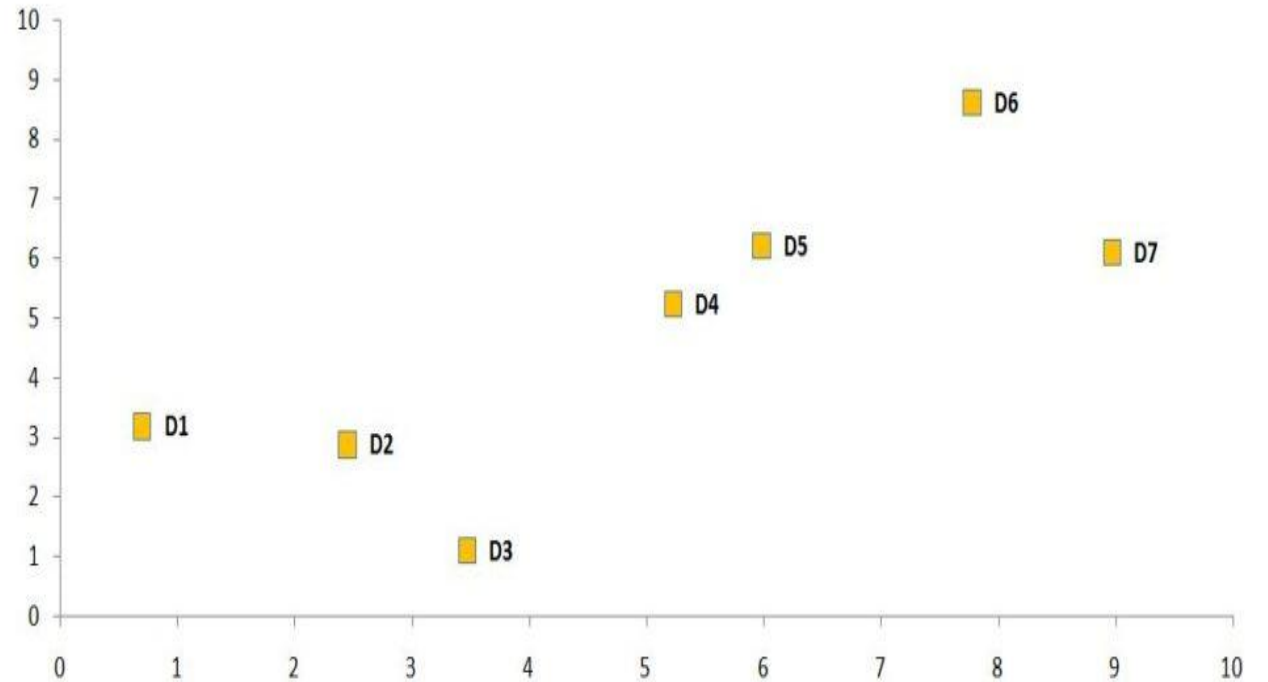
Dist	(A,B)	C	(D, F), E
(A,B)	0.00	4.95	2.50
C	4.95	0.00	1.41
(D, F), E	2.50	1.41	0.00

Min Distance (Single Linkage)

Dist	(A,B)	(D, F), E),C
(A,B)	0.00	2.50
((D, F), E),C	2.50	0.00

Example 3: Single linkage

Data Points	X	Y
D1	0.7	3.2
D2	2.45	2.89
D3	3.47	1.12
D4	5.23	5.24
D5	5.98	6.23
D6	7.778	8.63
D7	8.97	6.12

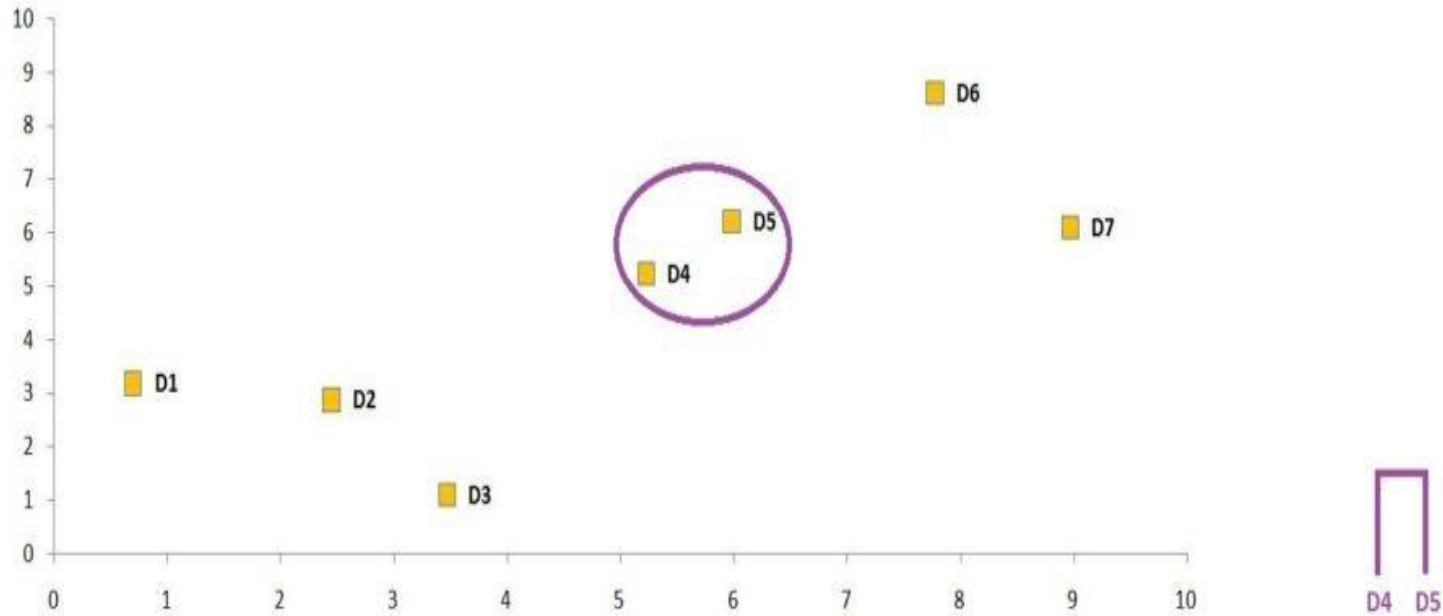


	D1	D2	D3	D4	D5	D6	D7
D1	0	1.78	3.46	4.97	6.09	8.92	8.77
D2	1.78	0	2.04	3.64	4.86	7.83	7.28
D3	3.46	2.04	0	4.48	5.69	8.66	7.43
D4	4.97	3.64	4.48	0	1.24	4.24	3.84
D5	6.09	4.86	5.69	1.24	0	3.00	2.99
D6	8.92	7.83	8.66	4.24	3.00	0	2.78
D7	8.77	7.28	7.43	3.84	2.99	2.78	0

	D1	D2	D3	D4	D5	D6	D7
D1	0	1.78	3.46	4.97	6.09	8.92	8.77
D2	1.78	0	2.04	3.64	4.86	7.83	7.28
D3	3.46	2.04	0	4.48	5.69	8.66	7.43
D4	4.97	3.64	4.48	0	1.24	4.24	3.84
D5	6.09	4.86	5.69	1.24	0	3.00	2.99
D6	8.92	7.83	8.66	4.24	3.00	0	2.78
D7	8.77	7.28	7.43	3.84	2.99	2.78	0

	D1	D2	D3	D4	D5	D6	D7
D1							
D2	1.78						
D3	3.46	2.04					
D4	4.97	3.64	4.48				
D5	6.09	4.86	5.69	1.24			
D6	8.92	7.83	8.66	4.24	3.00		
D7	8.77	7.28	7.43	3.84	2.99	2.78	

	D1	D2	D3	D4	D5	D6	D7
D1							
D2	1.78						
D3	3.46	2.04					
D4	4.97	3.64	4.48				
D5	6.09	4.86	5.69	1.24			
D6	8.92	7.83	8.66	4.24	3.00		
D7	8.77	7.28	7.43	3.84	2.99	2.78	

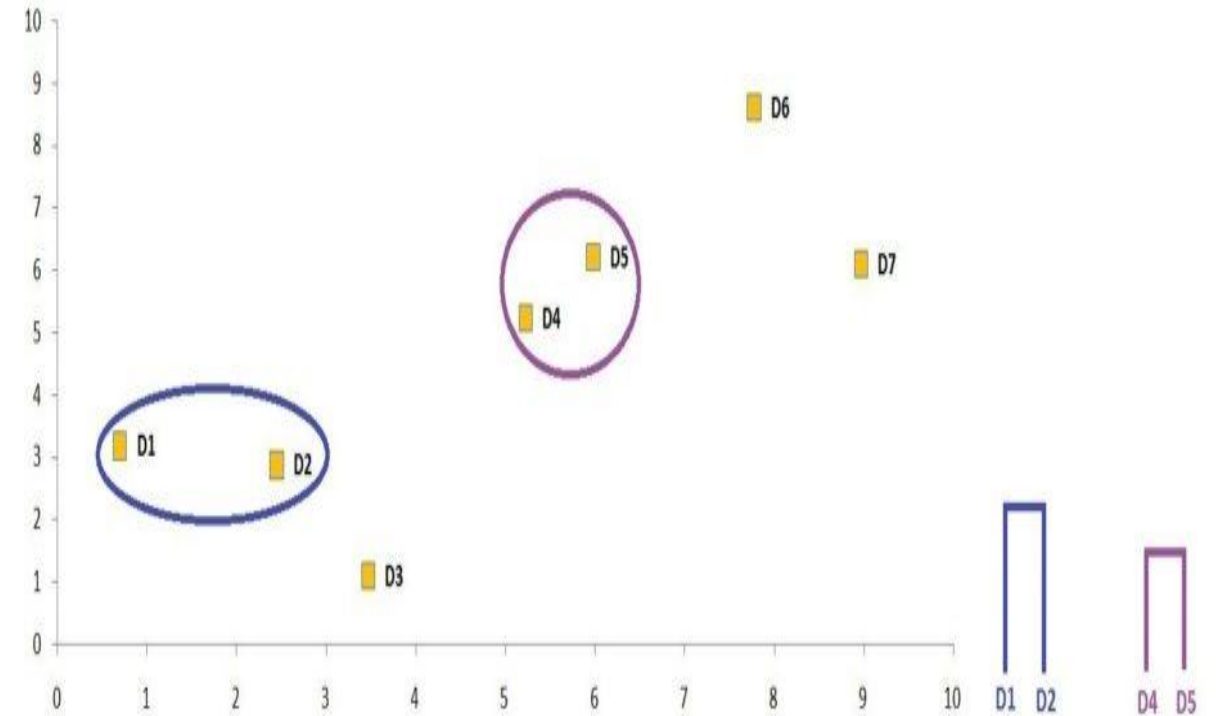


	D1	D2	D3	D4,D5	D6	D7
D1						
D2	1.78					
D3	3.46	2.04				
D4,D5	4.97	3.64	4.48			
D6	8.92	7.83	8.66	3.00		
D7	8.77	7.28	7.43	2.99	2.78	

As we are using single linkage, we choose the minimum distance, therefore, we choose 4.97 and consider it as the distance between the D1 and D4, D5. If we were using complete linkage then the maximum value would have been selected as the distance between D1 and D4, D5 which would have been 6.09. If we were to use Average Linkage then the average of these two distances would have been taken. Thus, here the distance between D1 and D4, D5 would have come out to be 5.53 $(4.97 + 6.09 / 2)$.

From now on we will simply repeat Step 2 and Step 3 until we are left with one cluster. We again look for the minimum value which comes out to be 1.78 indicating that the new cluster which can be formed is by merging the data points D1 and D2.

	D1	D2	D3	D4,D5	D6	D7
D1						
D2	1.78					
D3	3.46	2.04				
D4,D5	4.97	3.64	4.48			
D6	8.92	7.83	8.66	3.00		
D7	8.77	7.28	7.43	2.99	2.78	

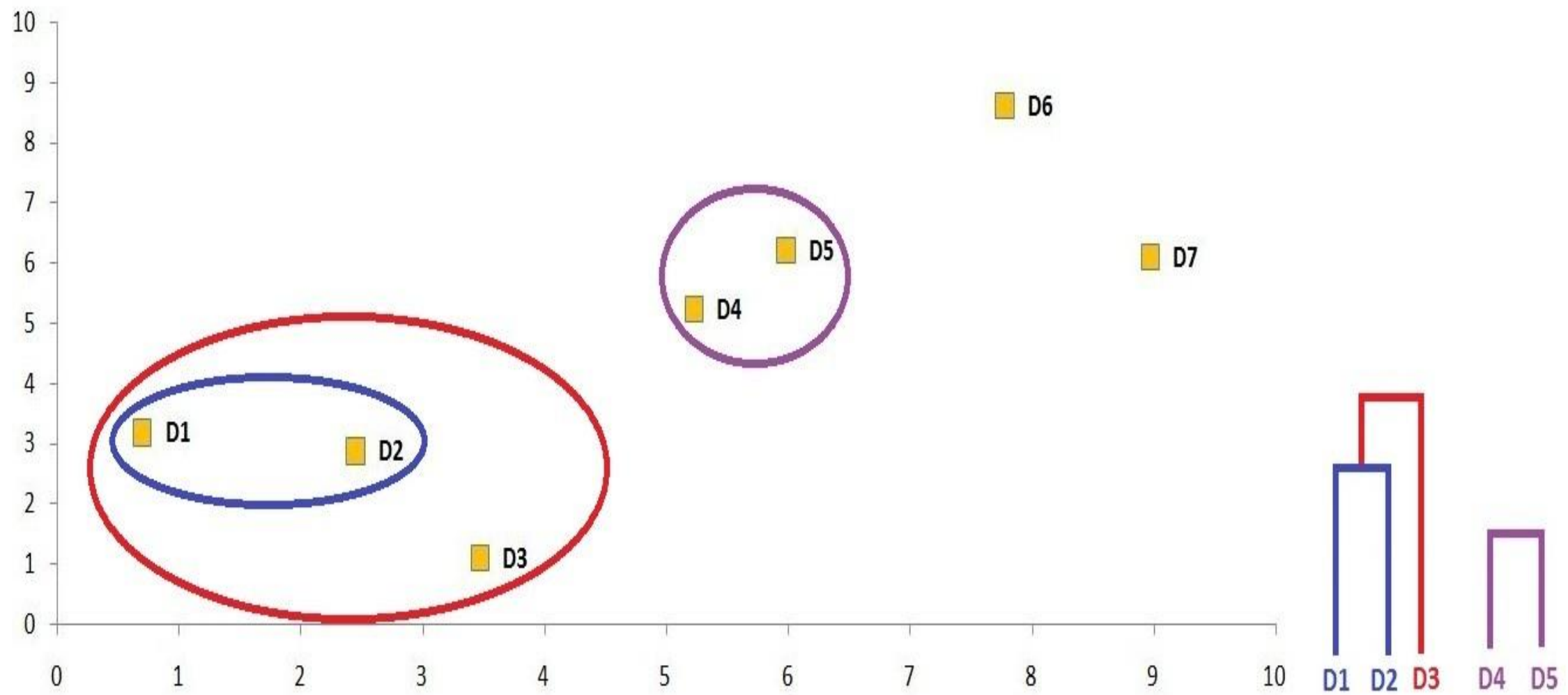


	D1,D2	D3	D4,D5	D6	D7
D1,D2					
D3	1.78				
D4,D5	3.64	4.48			
D6	7.83	8.66	3.00		
D7	7.28	7.43	2.99	2.78	

Similar to what we did in Step 3, we again recalculate the distance this time for cluster D1, D2 and come up with the following updated distance matrix.

	D1,D2	D3	D4,D5	D6	D7
D1,D2					
D3	1.78				
D4,D5	3.64	4.48			
D6	7.83	8.66	3.00		
D7	7.28	7.43	2.99	2.78	

We repeat what we did in step 2 and find the minimum value available in our distance matrix. The minimum value comes out to be 1.78 which indicates that we have to merge D3 to the cluster D1, D2.

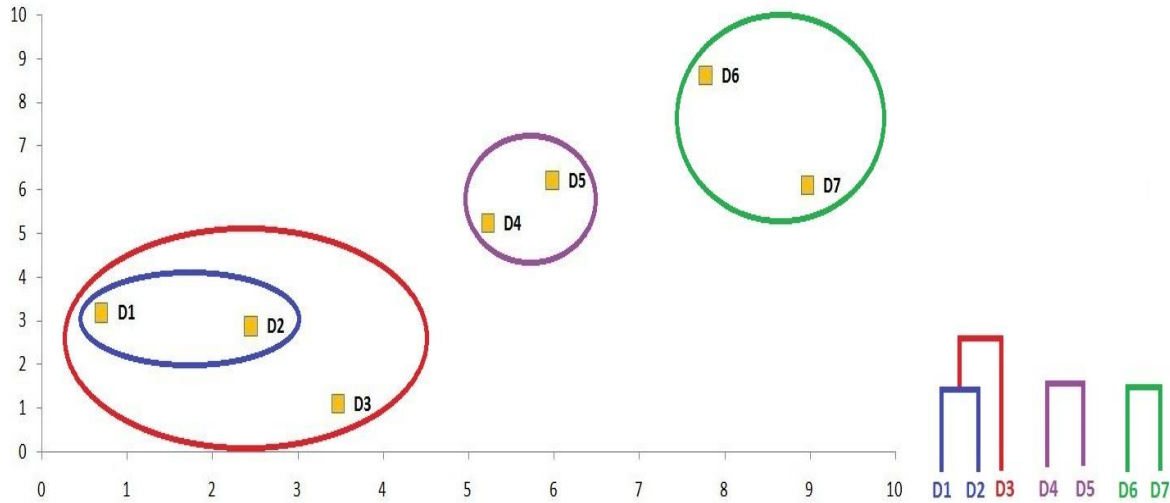


	D1,D2,D3	D4,D5	D6	D7
D1,D2,D3				
D4,D5	3.64			
D6	7.83	3.00		
D7	7.28	2.99	2.78	

Update the distance matrix using Single Link method.

	D1,D2,D3	D4,D5	D6	D7
D1,D2,D3				
D4,D5	3.64			
D6	7.83	3.00		
D7	7.28	2.99	2.78	

Find the minimum distance in the matrix.



Merge the data points accordingly and form another cluster.

	D1,D2,D3	D4,D5	D6,D7
D1,D2,D3			
D4,D5	3.64		
D6,D7	7.28	2.99	

	D1,D2,D3	D4,D5	D6,D7
D1,D2,D3			
D4,D5	3.64		
D6,D7	7.28	2.99	

Update the distance matrix using Single Link method.

Ward's Algorithm

This is also called minimum variance method. Begins with one cluster for each individual sample point.

- At each iteration, among all pairs of clusters, it merges pairs with least squared error
- The squared error for each cluster is defined as follows
- If a cluster contains m samples $x_1, x_2, x_3, \dots, x_m$ where x_i is the feature vector $(x_{i1}, x_{i2}, \dots, x_{id})$,
- the squared error for sample x_i , which is the squared Euclidean distance from the mean: $\sum_{j=1}^d (x_{ij} - \mu_j)^2$ (Variance)
- Where μ_j is the mean of the feature j for the values in the cluster

given by :

$$\mu_j = \frac{1}{m} \sum_{i=1}^m (x_{ij})$$

Ward's Algorithm... Continued

- The squared error E for the entire cluster is the sum of the squared errors for the samples
- $E = \sum_{i=1}^m \sum_{j=1}^d (x_{ij} - \mu_j)^2 = m \sigma^2$
- The vector composed of the means of each feature, $(\mu_1, \dots, \dots, \mu_d) = \mu$, is called the mean of the vector or centroid of the cluster
- The squared error is thus the total variance of the cluster σ^2 times the number of samples m .

Clusters	Squared Error, E
{1,2},{3},{4},{5}	8.0
{1,3},{2},{4},{5}	68.5
{1,4},{2},{3},{5}	200.0
{1,5},{2},{3},{4}	232.0
{2,3},{1},{4},{5}	32.5
{2,4},{1},{3},{5}	128.0
{2,5},{1},{3},{4}	160.0
{3,4},{1},{2},{5}	48.5
{3,5},{1},{2},{4}	48.5
{4,5},{1},{2},{3}	32.0

	x	y
1	4	4
2	8	4
3	15	8
4	24	4
5	24	12

5.5: Squared errors for each way of creating four clusters.

Clusters	Squared Error, E
{1,2,3},{4},{5}	72.7
{1,2,4},{3},{5}	224.0
{1,2,5},{3},{4}	266.7
{1,2},{3,4},{5}	56.5
{1,2},{3,5},{4}	56.5
{1,2},{4,5},{3}	40.0

Figure 5.6: Squared errors for three clusters.

In the Given example: for {1,2} the features are (4,4) and (8,4). Mean of these two: $\{(4+8)/2 = 6, (4+4)/2 = 4\} = \{6,4\}$
 Squared Error = between {1,2} is { square of (4-6) + square of (8-6) + square of (4-4) + square of (4-4)} = 8

Clusters	Squared Error, E
$\{1,2,3\}, \{4,5\}$	104.7
$\{1,2,4,5\}, \{3\}$	380.0
$\{1,2\}, \{3,4,5\}$	94.0

Figure 5.7: Squared errors for two clusters.

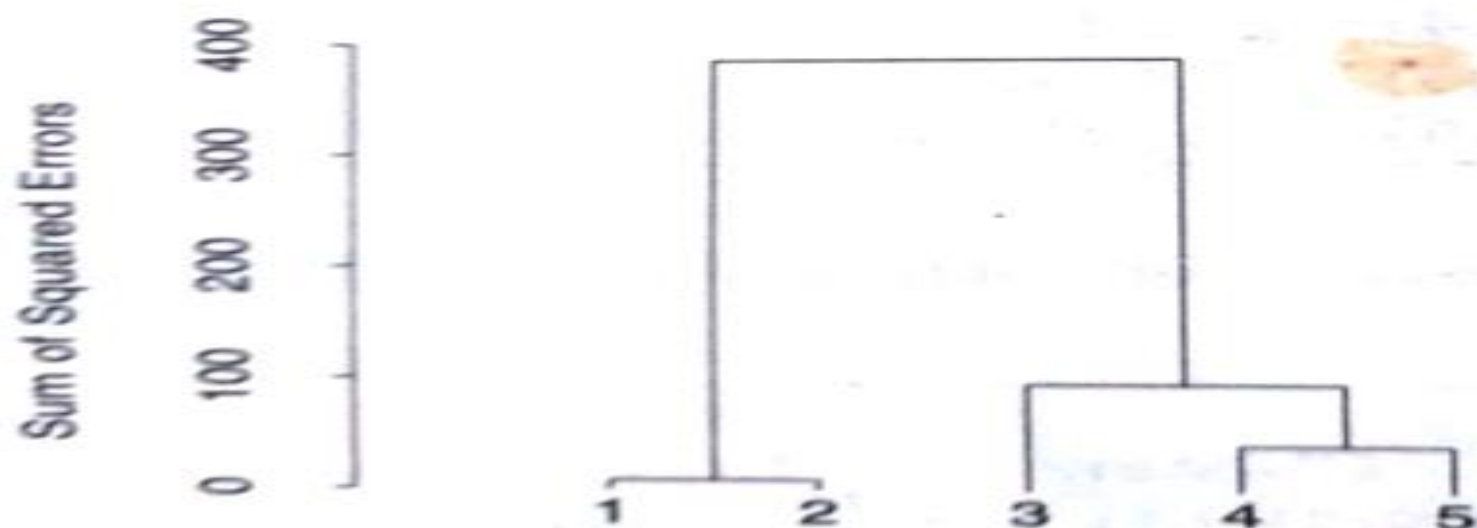


Figure 5.8: Dendrogram for Ward's method.

One Hot Encoding

- Popularly used in classification problem.
- One hot encoding creates new (binary) columns, indicating the presence of each possible value from the original data.
- It is good only when less number of classes
- It is illustrated through an example (next slide)

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29



One-Hot Encoded Data

Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29

Divisive approach

- At the top, there is a single cluster consisting of all the eight patterns.

$$\{A, B, C, D, E, F, G, H\}$$

- By considering all possible 2-partitions ($2^7 - 1 = 127$ partitions), the best 2-partition given by

$$\{A, B, C, H\}, \{D, E, F, G\}$$

- At the next level, the cluster $\{D, E, F, G\}$ is selected to split into two clusters.

$$\{D, E\}, \{F, G\}$$

- At the next level, we partition the cluster $\{A, B, C, H\}$ into two clusters.

$$\{A, B, C\}, \{H\}$$

- At the subsequent level, the cluster $\{A, B, C\}$ is split into two clusters.

$$\{A\}, \{B, C\}$$

- We have now 5 clusters given by

$$\{A\}, \{B, C\}, \{H\}, \{D, E\}, \{F, G\},$$

- Similarly, a dendrogram depicts partitions having 6, 7 and 8 clusters of the data at successive levels.
- Final level:** Each cluster has only one point (i.e., singleton cluster).

How to divide

- Fix some condition.
- Example: In this example, after computing the distance/ cost matrix, the least two will be put into one group (D,E), and others into another group.

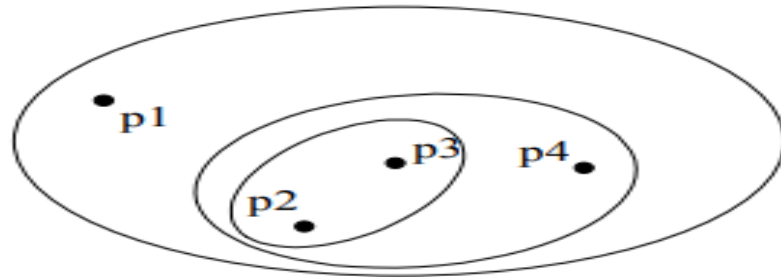
Example 4: Consider the 2-D data set of 8 patterns given by $A = (0.5, 0.5)$, $B = (2, 1.5)$, $C = (2, 0.5)$, $D = (5, 1)$, $E = (5.75, 1)$, $F = (5, 3)$, $G = (5.5, 3)$, $H = (2, 3)$.

	A	B	C	D	E	FG	H
A	0						
B	1.803	0					
C	1.5	1	0				
D	4.528	3.041	3.041	0			
E	5.274	3.783	3.783	0.75	0		
FG	$\max(5.148, 5.59)$ = 5.59	$\max(3.354, 3.808)$ = 3.808	$\max(3.905, 4.301)$ = 4.301	$\max(2, 2.062)$ = 2.062	$\max(2.136, 2.016)$ = 2.136	0	
H	2.915	1.5	2.5	3.606	4.25	$\max(3, 3.5)$ = 3.5	0

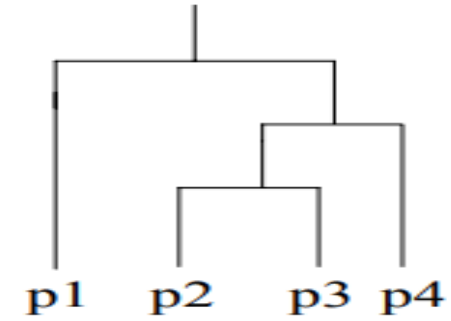
$\{D, E\}$

Hierarchical clustering: cluster is usually placed inside another cluster...follows tree structure

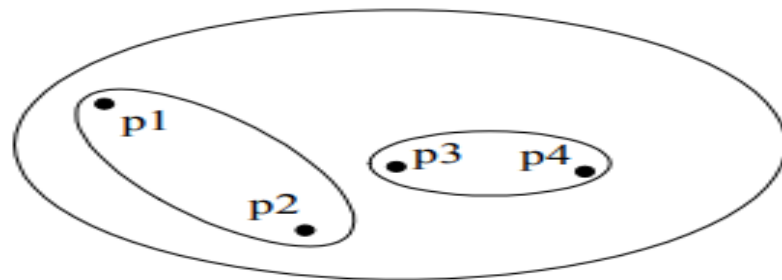
Hierarchical clustering



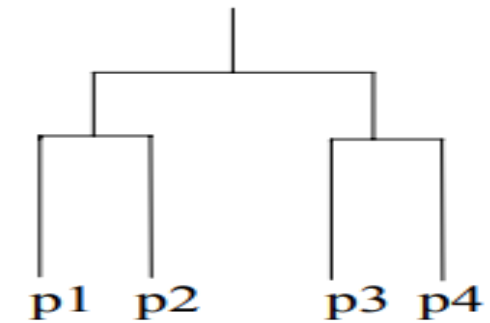
Hierarchical Clustering



Dendrogram



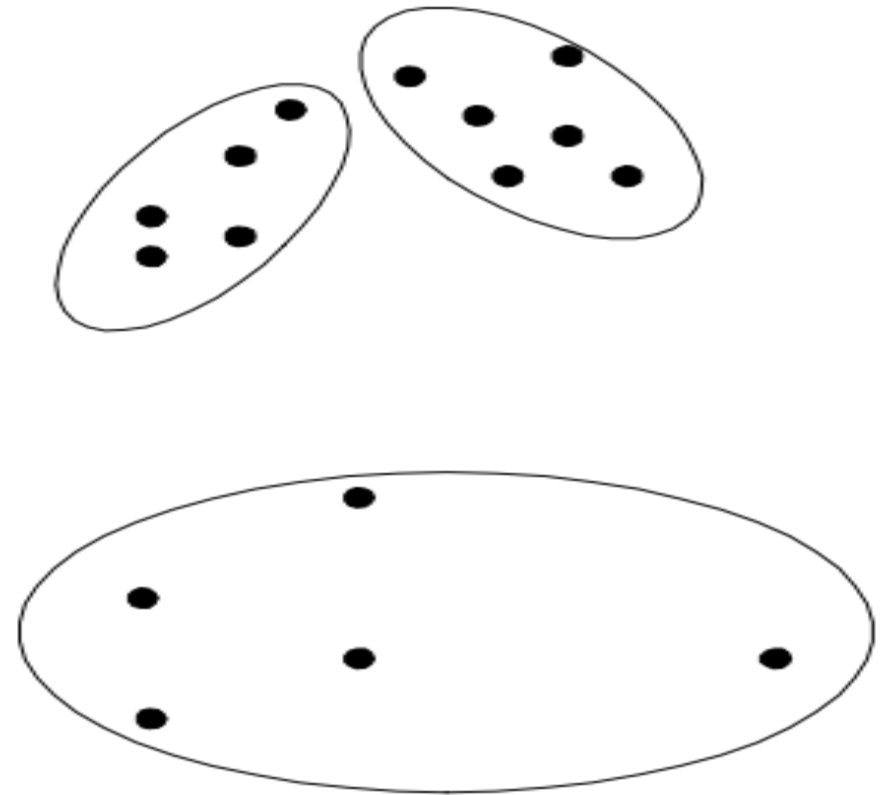
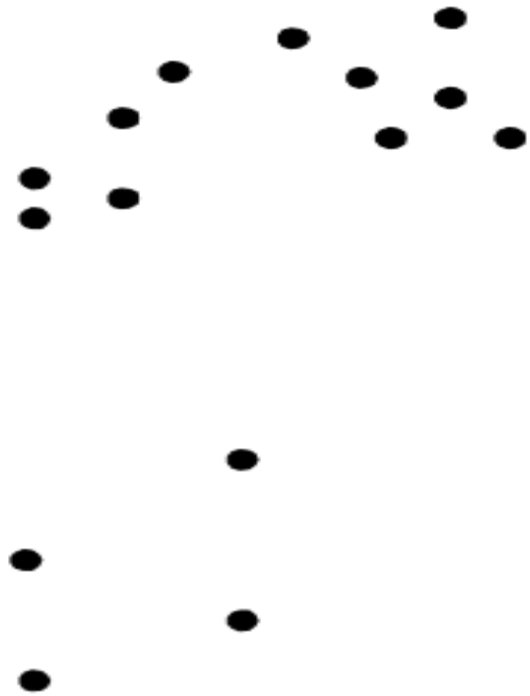
Hierarchical Clustering



Dendrogram

Partitional clustering: A sample belongs to exactly one cluster : No tree structure, no dendrogram representation

Partitional clustering



Partitional Clustering:

Agglomerative clustering creates a series of Nested clusters.

In partitional clustering the goal is to usually create one set of clusters that partitions the data into similar groups.

Samples close to one another are assumed to be in one cluster. This is the goal of partitional clustering.

Partitional clustering creates 'k' clusters for the given 'n' samples.

The number of clusters 'k' is also to be given in advance.

Forgy's Algorithm

One of the simplest partitioning algorithms is the Forgy's algorithm.

Apart from the data, the input to the algorithm is 'k', the number of clusters to be constructed

'k' samples are called seed points.

The seed points could be chosen randomly, or some knowledge of the desired could be used to guide their selection.

Forgy's Algorithm

1. Initialize the cluster centroid to the seed points.
2. For each sample, find the cluster centroid nearest to it. Put the sample in the nearest cluster identified with the cluster centroid.
3. If no samples changed the clusters in step 2
4. Compute the centroids of the resulting clusters and go to step 2.

Consider the Data points listed in the table and set $k = 2$ to produce two clusters
Use the first two samples (4,4) and (8,4) as the seed points.
Now applying the algorithm by computing the distance from each
cluster centroid and assigning them to the clusters:

Data Points	X	Y
1	4	4
2	8	4
3	15	8
4	24	4
5	24	12

Sample	Nearest Cluster Centroid
(4,4)	(4,4)
(8,4)	(8,4)
(15,8)	(8,4)
(24,4)	(8,4)
(24,12)	(8,4)

The clusters $\{(4,4)\}$ and $\{(8,4),(15,8),(24,4),(24,12)\}$ are formed.

Now re-compute the cluster centroids

New centroids are:

The first cluster $(4,4)$ and

The second cluster centroid is

$$x = (8+15+24+24)/4 = 17.75$$

$$y = (4+8+4+12)/4 = 7$$

Sample	Nearest Cluster Centroid
(4,4)	(4,4)
(8,4)	(4,4)
(15,8)	(17.75,7)
(24,4)	(17.75,7)
(24,12)	(17.75,7)

The clusters $\{(4,4),(8,4)\}$ and $\{(15,8),(24,4),(24,12)\}$ are formed.

Now re-compute the cluster centroids

The first cluster centroid

$$x = (4+8)/2 = 6 \text{ and } y = (4+4)/2 = 4$$

The second cluster centroid is

$$x = (15+24+24)/3 = 21$$

$$y = (8+4+12)/4 = 12$$

In the next step notice that the cluster centroid does not change

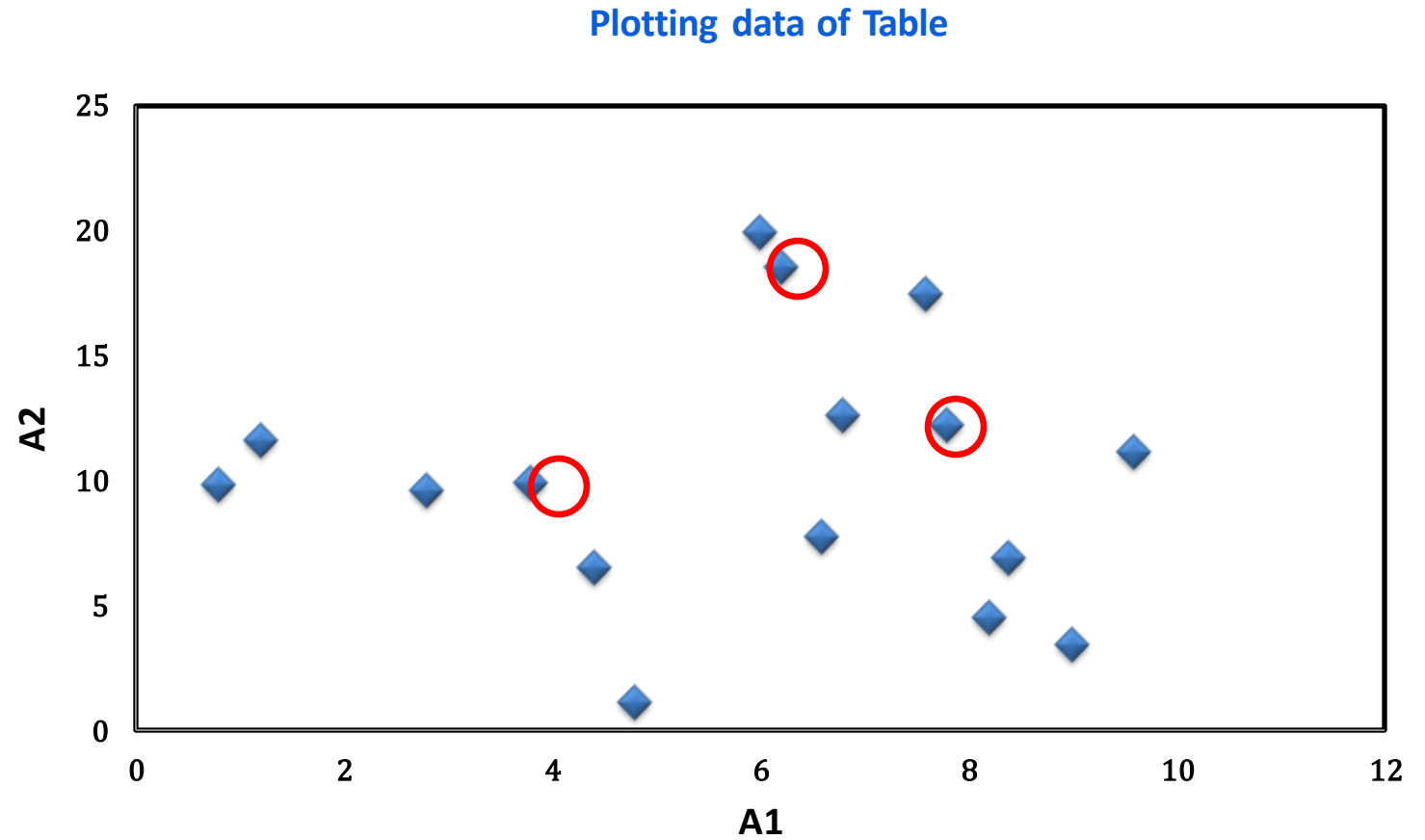
And samples also do not change the clusters.

Algorithm terminates.

Sample	Nearest Cluster Centroid
(4,4)	(6,4)
(8,4)	(6,4)
(15,8)	(21,12)
(24,4)	(21,12)
(24,12)	(21,12)

Example-2 Illustration Forgy's clustering algorithms

A_1	A_2
6.8	12.6
0.8	9.8
1.2	11.6
2.8	9.6
3.8	9.9
4.4	6.5
4.8	1.1
6.0	19.9
6.2	18.5
7.6	17.4
7.8	12.2
6.6	7.7
8.2	4.5
8.4	6.9
9.0	3.4
9.6	11.1



Example 2: Forgy's clustering algorithms

- Suppose, $k=3$. Three objects are chosen at random shown as circled. These three centroids are shown below.

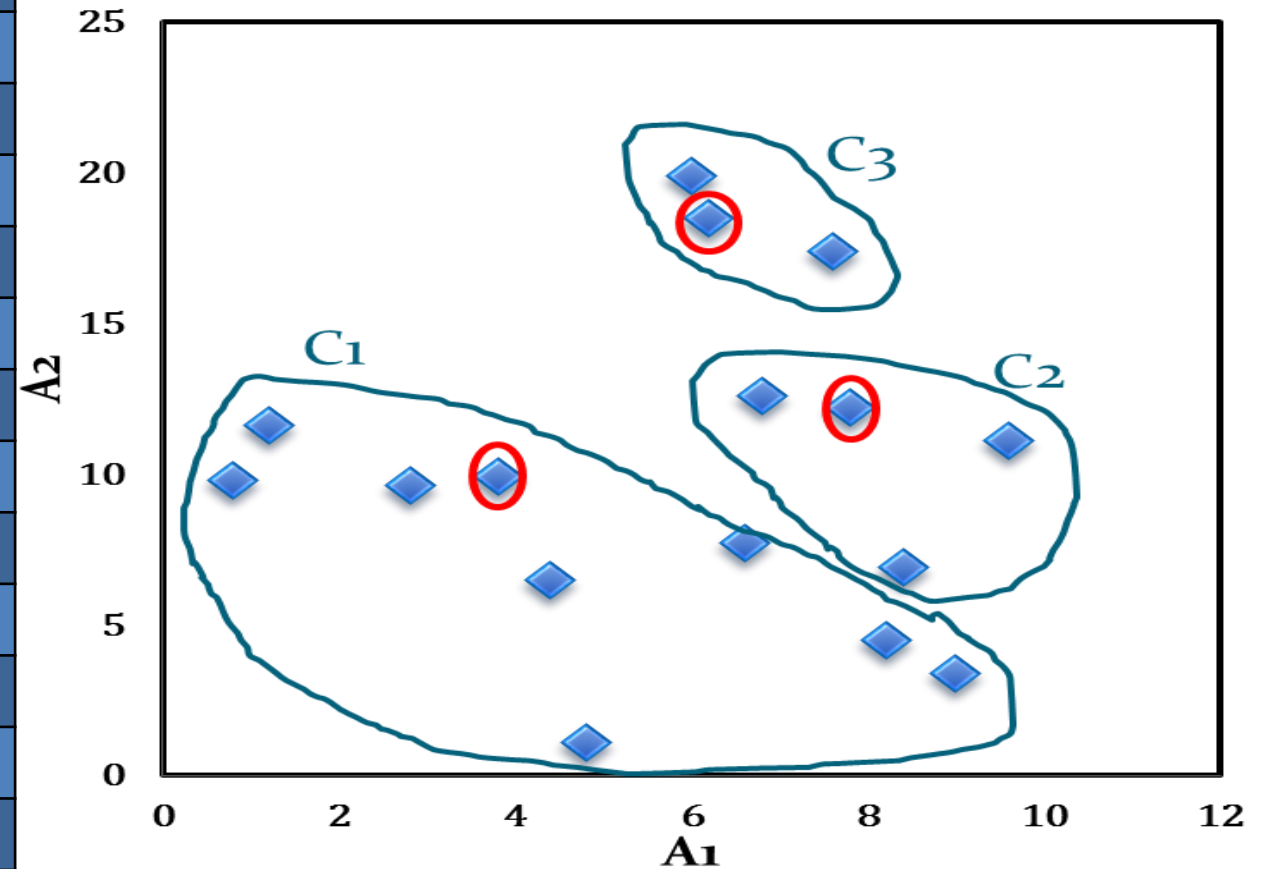
Initial Centroids chosen randomly

Centroid	Objects	
	A1	A2
c_1	3.8	9.9
c_2	7.8	12.2
c_3	6.2	18.5

- Let us consider the Euclidean distance measure (L2 Norm) as the distance measurement in our illustration.
- Let d_1 , d_2 and d_3 denote the distance from an object to c_1 , c_2 and c_3 respectively. The distance calculations are shown in Table
- Assignment of each object to the respective centroid is shown in the right-most column and the clustering so obtained is shown in Figure.

Example 2: Forgy's clustering algorithms

A_1	A_2	d_1	d_2	d_3	cluster
6.8	12.6	4.0	1.1	5.9	2
0.8	9.8	3.0	7.4	10.2	1
1.2	11.6	3.1	6.6	8.5	1
2.8	9.6	1.0	5.6	9.5	1
3.8	9.9	0.0	4.6	8.9	1
4.4	6.5	3.5	6.6	12.1	1
4.8	1.1	8.9	11.5	17.5	1
6.0	19.9	10.2	7.9	1.4	3
6.2	18.5	8.9	6.5	0.0	3
7.6	17.4	8.4	5.2	1.8	3
7.8	12.2	4.6	0.0	6.5	2
6.6	7.7	3.6	4.7	10.8	1
8.2	4.5	7.0	7.7	14.1	1
8.4	6.9	5.5	5.3	11.8	2
9.0	3.4	8.3	8.9	15.4	1
9.6	11.1	5.9	2.1	8.1	2

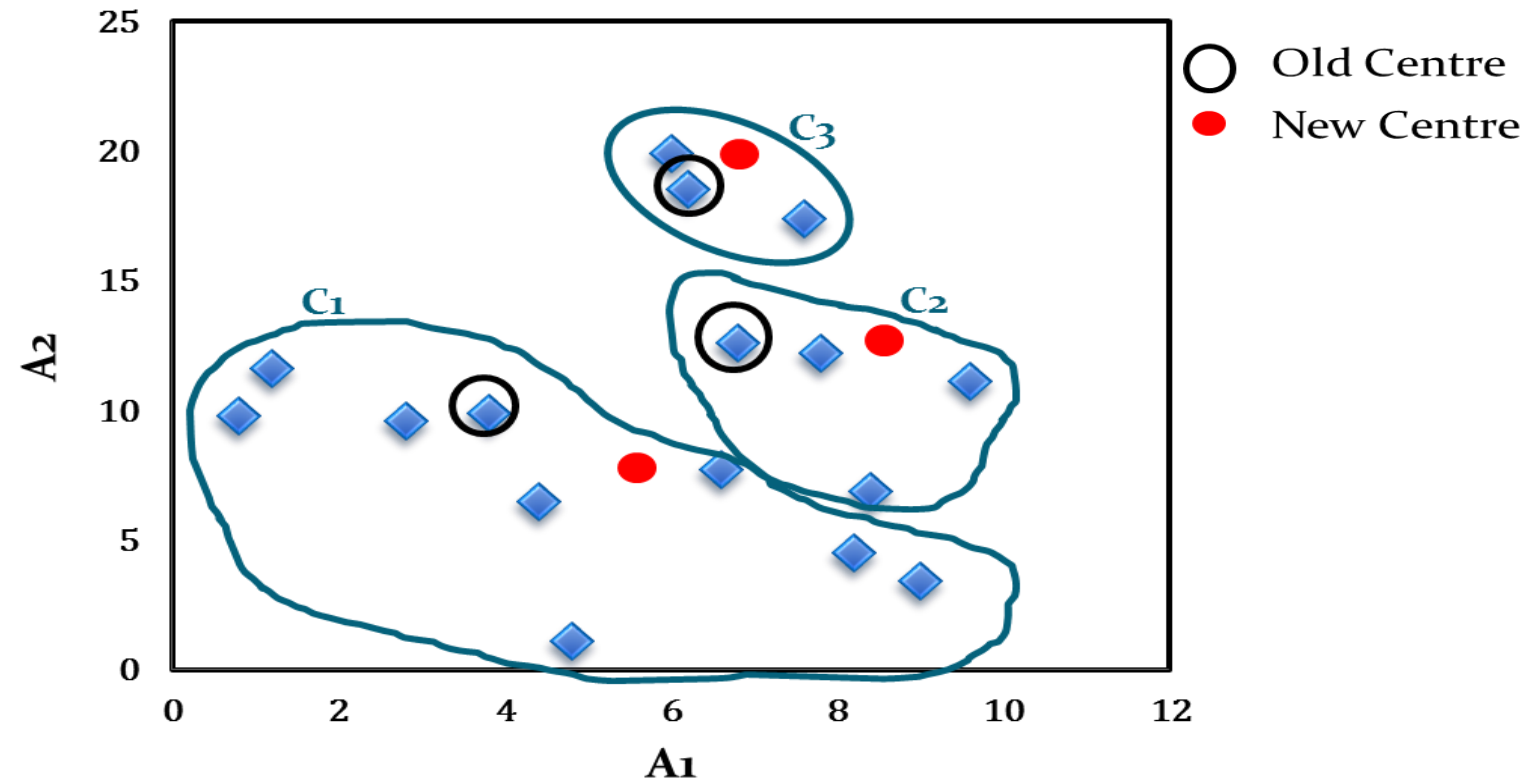


Example 2: Forgy's clustering algorithms

The calculation new centroids of the three cluster using the mean of attribute values of A1 and A2 is shown in the Table below. The cluster with new centroids are shown in Figure.

Calculation of new centroids

New Centroid	Objects	A1	A2
c_1		4.6	7.1
c_2		8.2	10.7
c_3		6.6	18.6

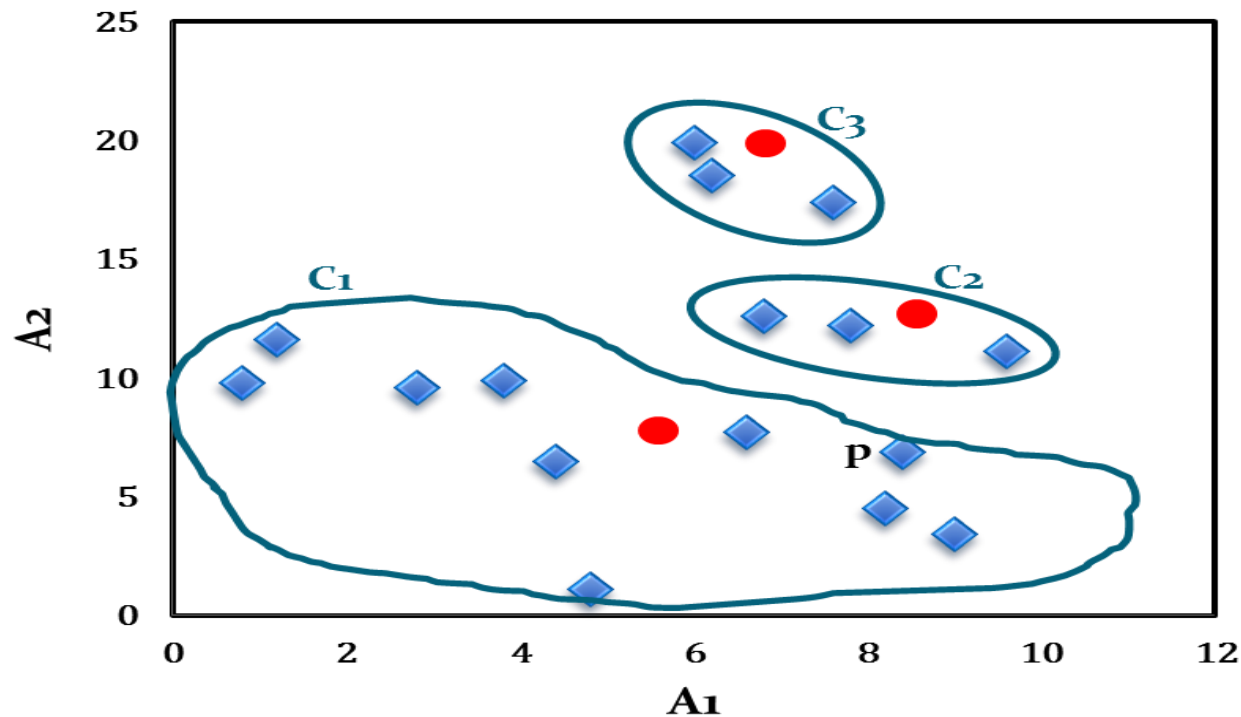


Next cluster with new centroids

Example 2: of Forgy's clustering algorithms

We next reassign the 16 objects to three clusters by determining which centroid is closest to each one. This gives the revised set of clusters shown in.

Note that point p moves from cluster C2 to cluster C1.

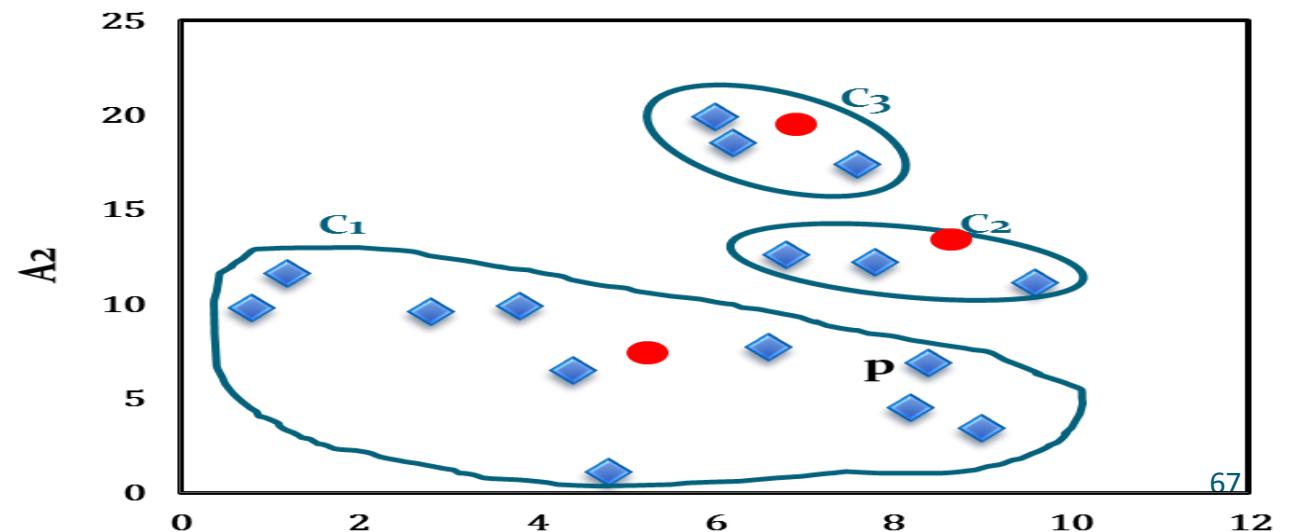


Cluster after first iteration

Example 2: of Forgy's clustering algorithms

- The newly obtained centroids after second iteration are given in the table below. Note that the centroid c_3 remains unchanged, where c_2 and c_1 changed a little.
- With respect to newly obtained cluster centres, 16 points are reassigned again. These are the same clusters as before. Hence, their centroids also remain unchanged.
- Considering this as the termination criteria, the algorithm stops here.

Centro id	Revised Centroids	
	A1	A2
c_1	5.0	7.1
c_2	8.1	12.0
c_3	6.6	18.6



Apply Forgy's algorithm for the following dataset with $K = 2$

Sample	X	Y
1	0.0	0.5
2	0.5	0.0
3	1.0	0.5
4	2.0	2.0
5	3.5	8.0
6	5.0	3.0
7	7.0	3.0

Pros

Simple, fast to compute

Converges to local minimum of within-cluster squared error

Cons

Setting k

Sensitive to initial centres

Sensitive to outliers

Detects spherical clusters

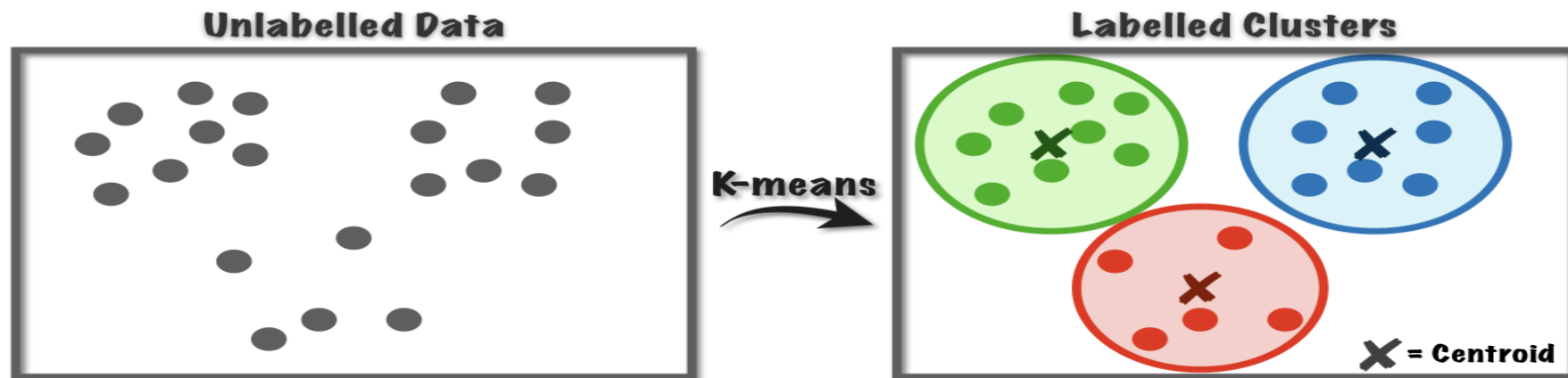
Assuming means can be computed

K-Means Algorithm

It is similar to Forgy's algorithm.

The k-means algorithm differs from Forgy's algorithm in that the centroids of the clusters are recomputed as soon as sample joins a cluster.

Also unlike Forgy's algorithm which is iterative in nature, the k-means **only two passes through the data set.**



The K-Means Algorithm

1. Input for this algorithm is K (the number of clusters) and 'n' samples, x_1, x_2, \dots, x_n .
1. Identify the centroids c_1 to c_k from the random locations. That is randomly select 'k' samples as centroids. (note: n should be greater than k)
2. For each remaining (n-k) samples, find the centroid nearest it. Put the sample in the cluster identified with this nearest centroid. **After each sample is assigned, re-compute the centroid of the altered cluster.**
3. Go through the data a second time. For each sample, find the centroid nearest it. Put the sample in the cluster identified with the nearest cluster. **(During this step do not recompute the centroid)**

Apply k-means Algorithm on the following sample points

Data Points	X	Y
1	4	4
2	8	4
3	15	8
4	24	4
5	24	12

Begin with two clusters $\{(8,4)\}$ and $\{(24,4)\}$ with the centroids $(8,4)$ and $(24,4)$

For each remaining samples, find the nearest centroid and put it in that cluster.

Then re-compute the centroid of the cluster.

The next sample $(15,8)$ is closer to $(8,4)$ so it joins the cluster $\{(8,4)\}$.

The centroid of the first cluster is updated to $(11.5,6)$.

$$(8+15)/2 = 11.5 \text{ and } (4+8)/2 = 6.$$

The next sample is $(4,4)$ is nearest to the centroid $(11.5,6)$ so it joins the cluster $\{(8,4),(15,8),(4,4)\}$.

Now the new centroid of the cluster is $(9,5.3)$

The next sample $(24,12)$ is closer to centroid $(24,4)$ and joins the cluster $\{(24,4),(24,12)\}$.

Now the new centroid of the second cluster is updated to $(24,8)$.

At this point step1 is completed.

For step2 examine the samples one by one and put each sample in the identified with the nearest cluster centroid.

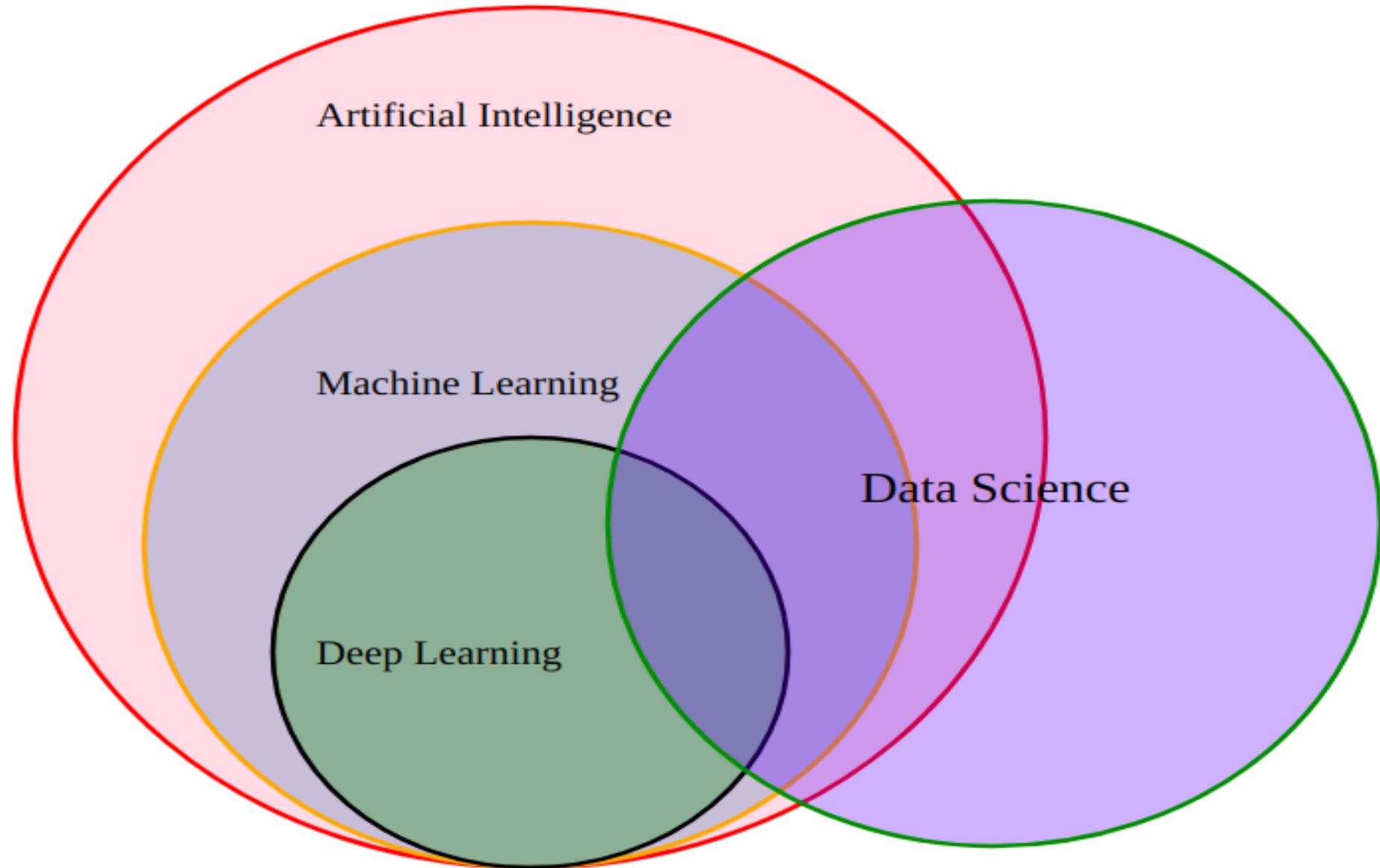
Sample	Distance to centroid (9,5.3)	Distance to centroid (24,8)
(8,4)	1.6	16.5
(24,4)	15.1	4.0
(15,8)	6.6	9.0
(4,4)	6.6	40.0
(24,12)	16.4	4.0

Example: $\text{Sqrt}(\text{square of } (9-8) + \text{square of } (4-5.3)) = 1.6$

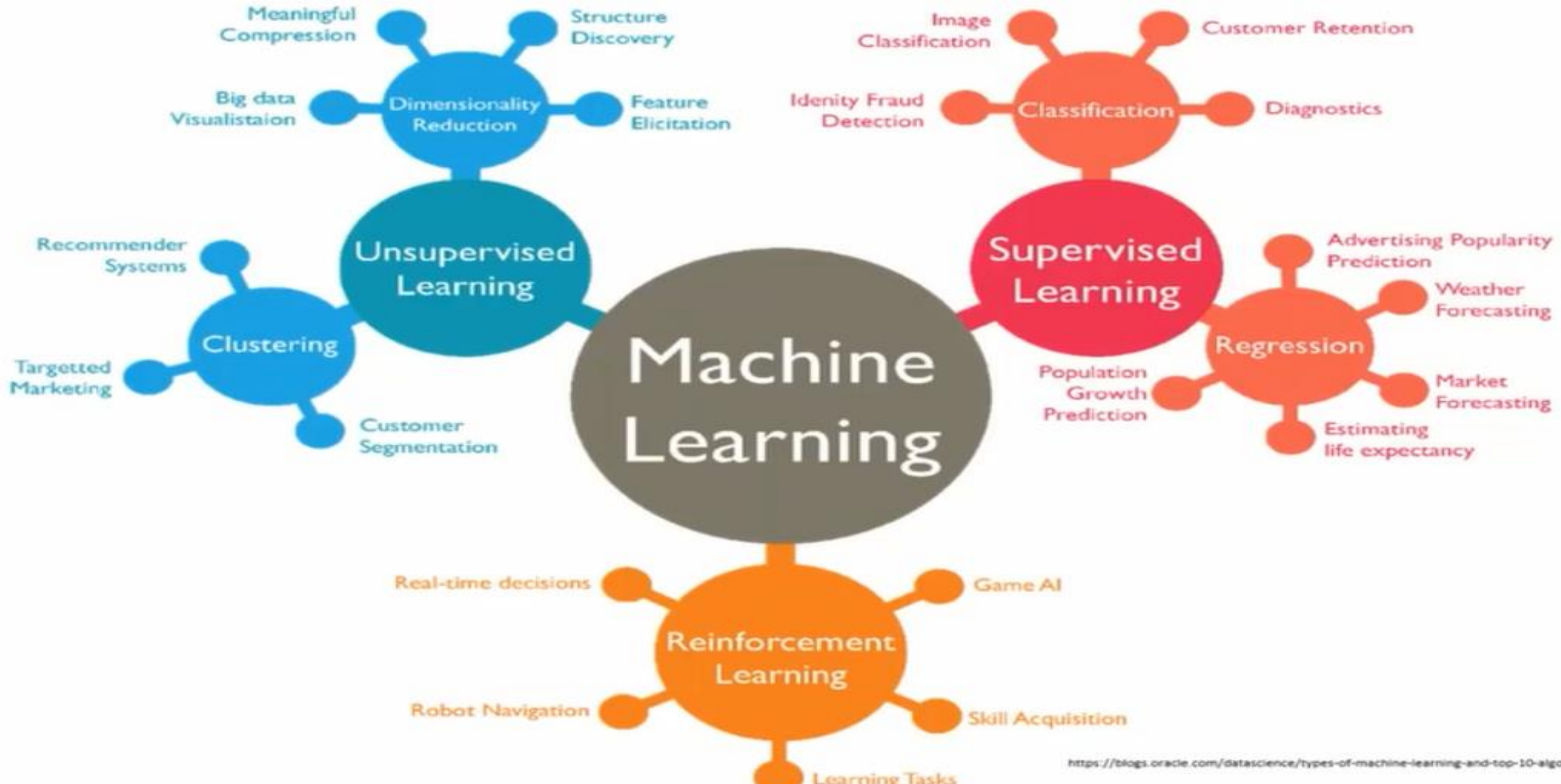
Final clusters of K-Means algorithms

- So the new clusters are :
- $C1 = \{ (8,4), (15,8), (4,4) \}$
- $C2 = \{ (24,4), (24,12) \}$
- K-Means algorithms ends.

AI-ML-DL and Data Science



Summary of Machine Learning Algorithms learnt so far...



End of Unit 4

Dimensionality Reduction

Unit-5

Dr. Srinath.S

Syllabus

- **Dimensionality Reduction:**
 - Singular Value Decomposition
 - Principal Component Analysis
 - Linear Discriminated Analysis
 - Independent Component Analysis.

What is Dimensionality Reduction?

- The number of input features, variables, or columns present in a given dataset is known as dimensionality, and the process to reduce these features is called dimensionality reduction.
- A dataset contains a huge number of input features in various cases, which makes the predictive modeling task more complicated, for such cases, dimensionality reduction techniques are required to use.

Dimensionality Reduction...?

- Dimensionality reduction technique can be defined as, *"It is a way of converting the higher dimensions dataset into lesser dimensions dataset ensuring that it provides similar information."*
- These techniques are widely used in Machine Learning for obtaining a better fit predictive model while solving the classification and regression problems.
- Handling the high-dimensional data is very difficult in practice, commonly *known as the curse of dimensionality.*

Benefits of Dimensionality Reduction..

- By reducing the dimensions of the features, the space required to store the dataset also gets reduced.
- Less Computation training time is required for reduced dimensions of features.
- Reduced dimensions of features of the dataset help in visualizing the data quickly.
- It removes the redundant features (if present).

Two ways of Dimensionality Reduction

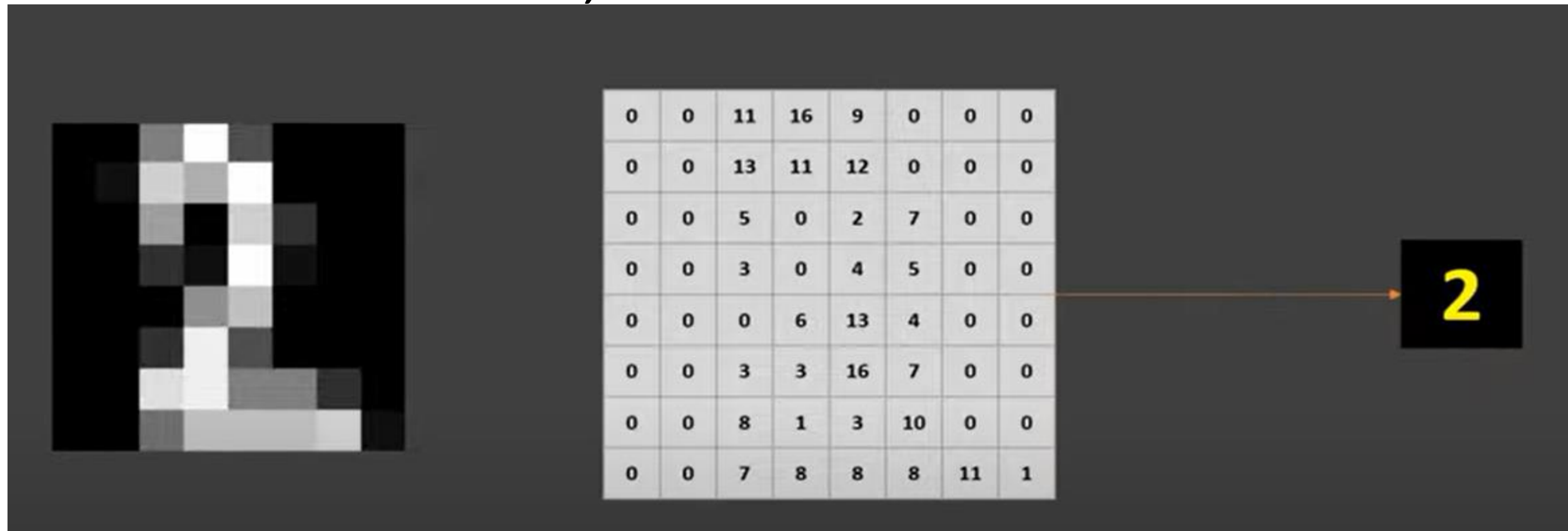
- 1. Feature Selection
- 2. Feature Extraction

Feature Selection

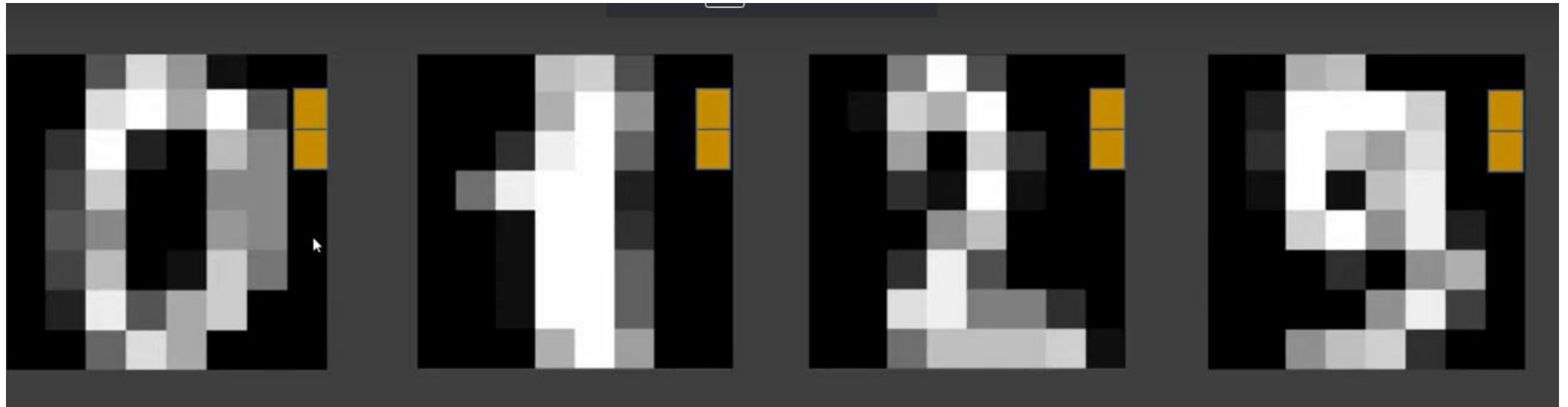
- Feature selection is the process of selecting the subset of the relevant features and leaving out the irrelevant features present in a dataset to build a model of high accuracy. In other words, it is a way of selecting the optimal features from the input dataset.

General – features reduction technique

- In this example number 2 has 64 features... but many of them are of no importance to decide the characteristics of 2, are removed first.



Remove features which are of no importance



Feature Selection – 3 Methods

- 1. Filter Method
 - Correlation
 - Chi-Square Test
 - ANOVA
 - Information Gain, etc.
- 2. Wrapper Method
 - Forward Selection
 - Backward Selection
 - Bi-directional Elimination
- 3. Embedded Method
 - LASSO
 - Elastic Net
 - Ridge Regression, etc.

Feature Extraction

- Feature extraction is the **process of transforming the space containing many dimensions into space with fewer dimensions.**
- This approach is useful when we want to keep the whole information but use fewer resources while processing the information.

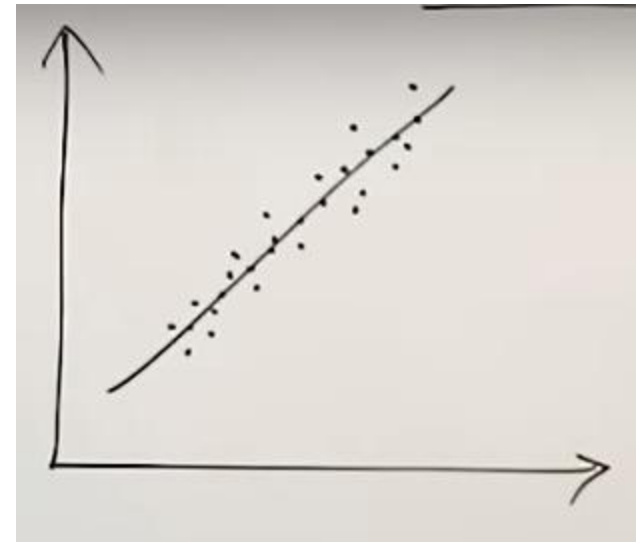
Some common feature extraction techniques are:

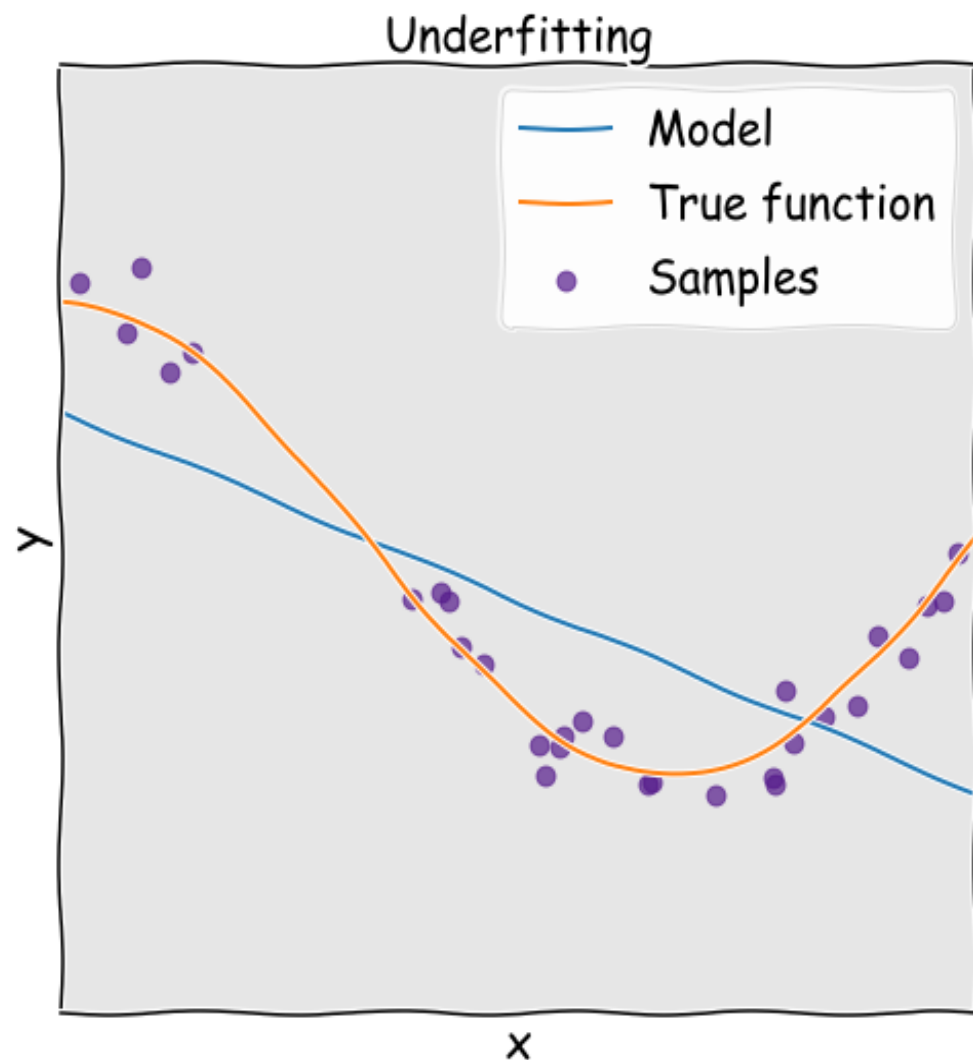
1. Principal Component Analysis (PCA)
2. Linear Discriminant Analysis (LDA)
3. Kernel PCA
4. Quadratic Discriminant Analysis (QDA)etc.

ML Model design

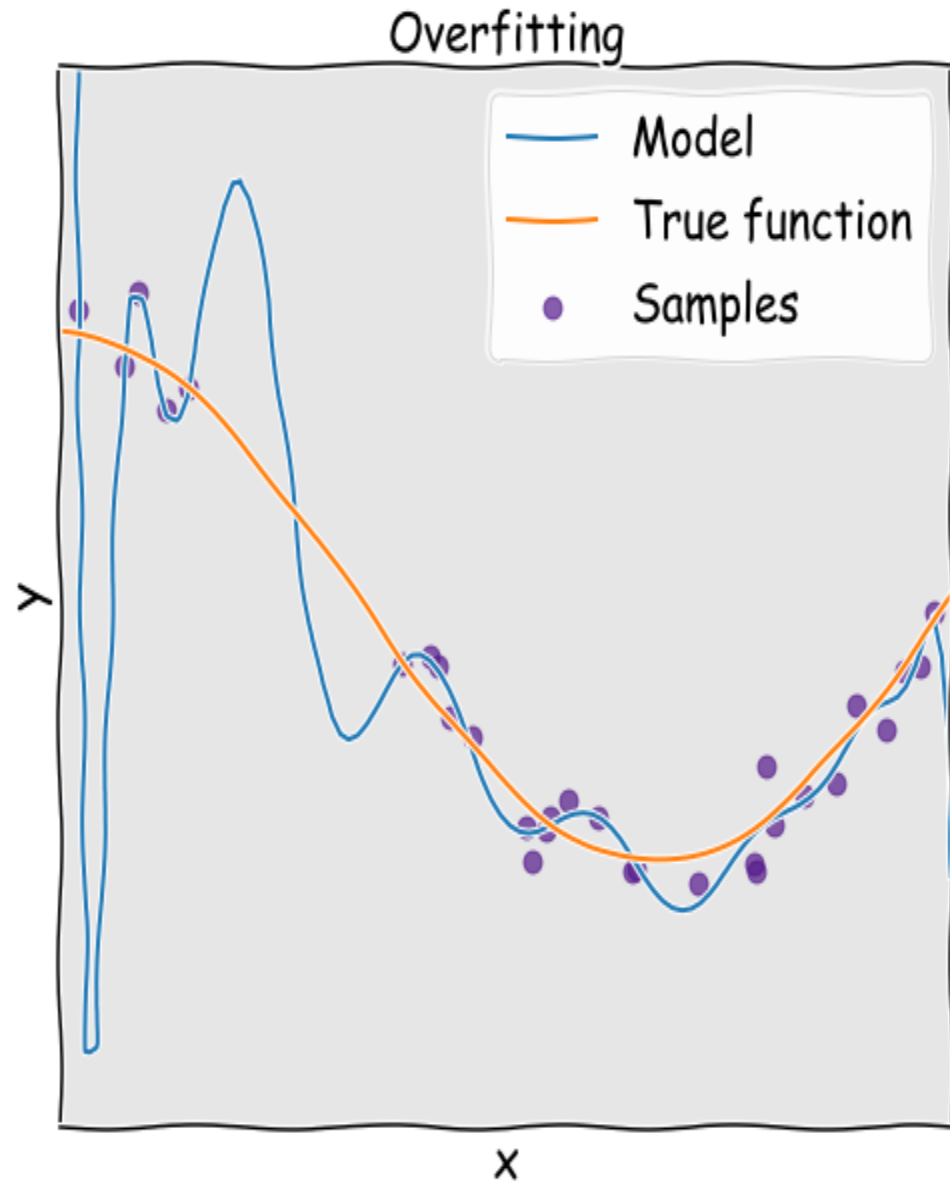
- Consider the line passing through the samples in the diagram.
- It (line) is the model/function/hypothesis generated after the training phase.

The line is trying to reach all the samples as close as possible.

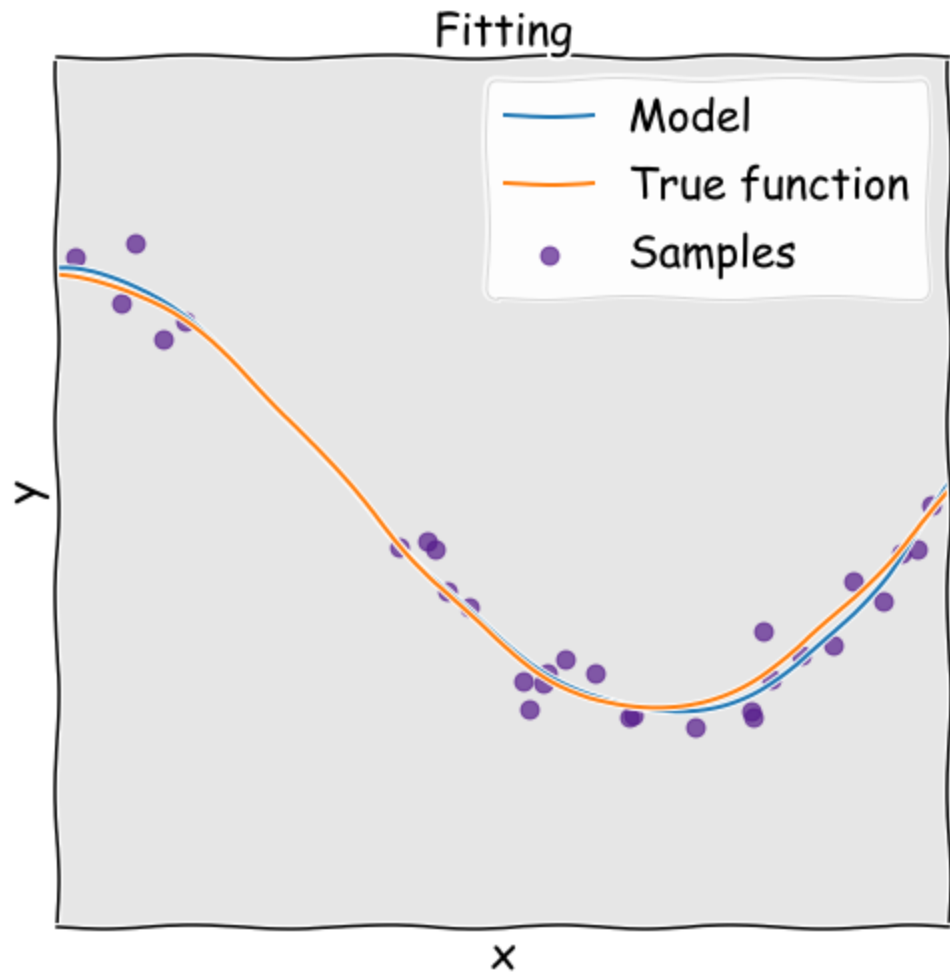




- If we have an underfitted model, this means that we do not have enough parameters to capture the trends in the underlying system.
- In general, in underfitting, model fails during testing as well as training.



- In this a complex model is built using too many features.
- During training phase, model works well. But it fails during testing.



- Under/Overfitting can be solved in different ways.
- One of the solutions for overfitting is dimensionality reduction.
- Diagram shows that model neither suffers from under or overfitting.

Example to show requirement of Dimensionality reduction

- In this example important features to decide the price are town, area and plot size. Features like number of bathroom and trees nearby may not be significant, hence can be dropped.

town	area	bathroom	plot	trees nearby	price
monroe	2600	2	8500	2	550000
monroe	3000	3	9200	2	565000
monroe	3200	3	8750	2	610000
monroe	3600	4	10200	2	680000
monroe	4000	4	15000	2	725000
west windsor	2600	2	7000	2	585000
west windsor	2800	3	9000	2	615000
west windsor	3300	4	10000	1	650000
west windsor	3600	4	10500	1	710000

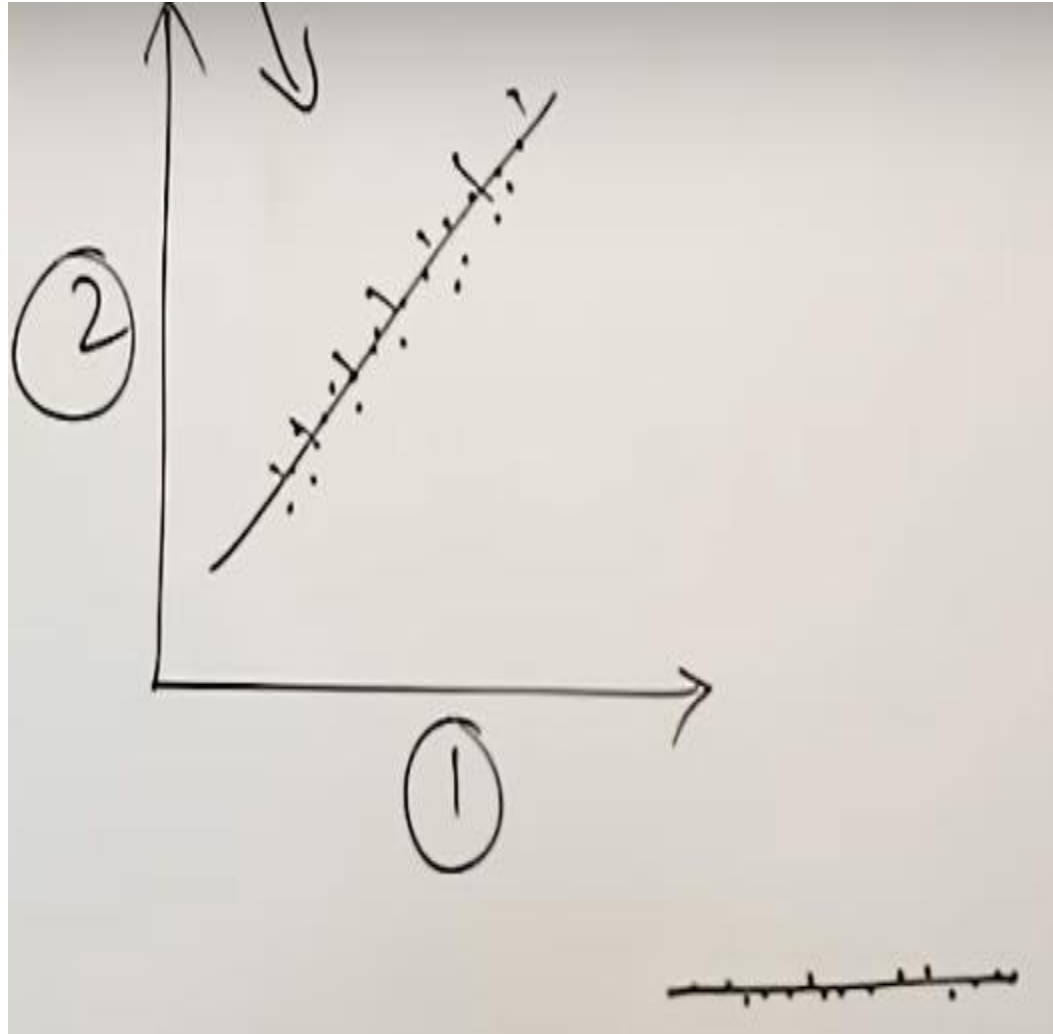
PCA

- PCA is a method of Dimensionality Reduction.
- PCA is a process of identifying Principal Components of the samples.
- It tries to address the problem of overfitting.

Example for PCA (from SK learn (SciKit Learn) library)

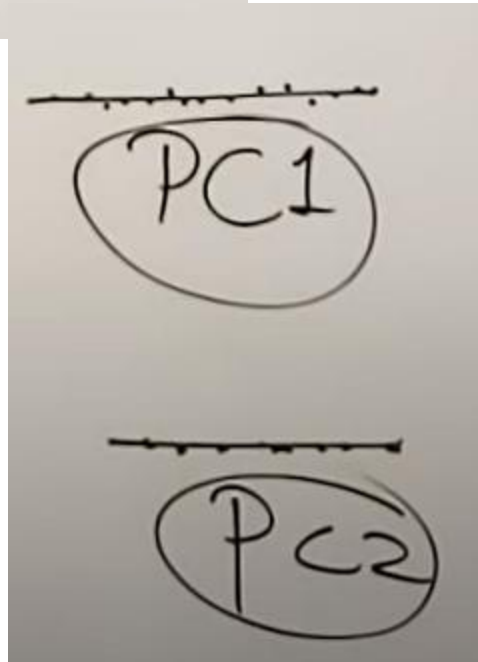
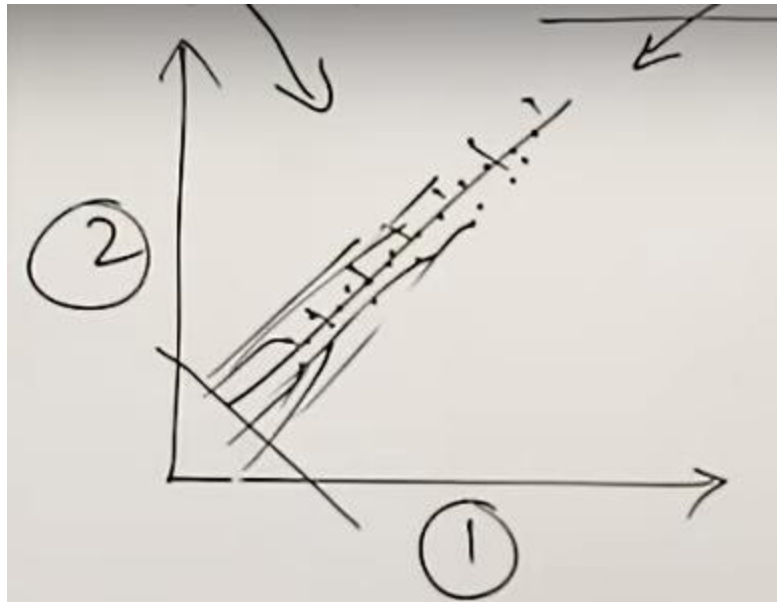


What does PCA do?



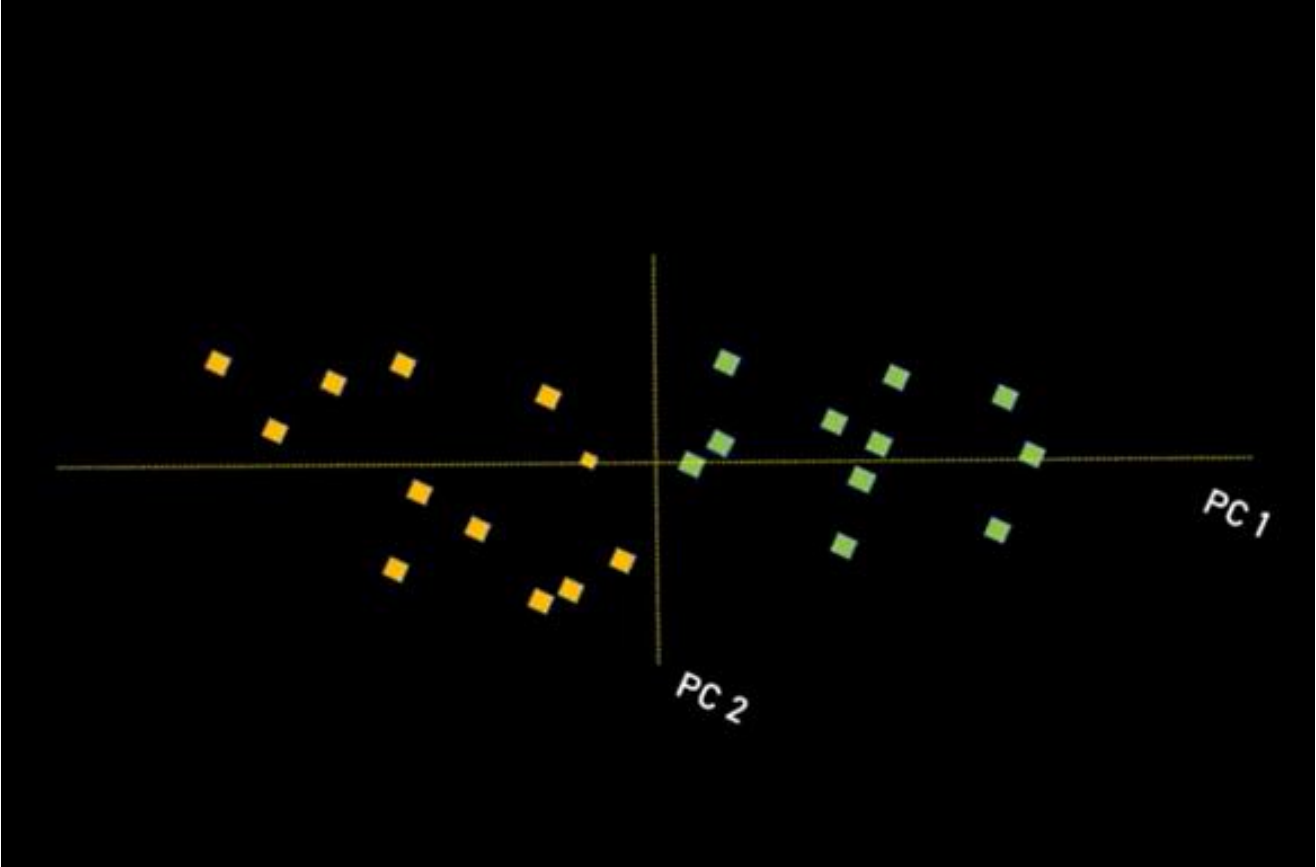
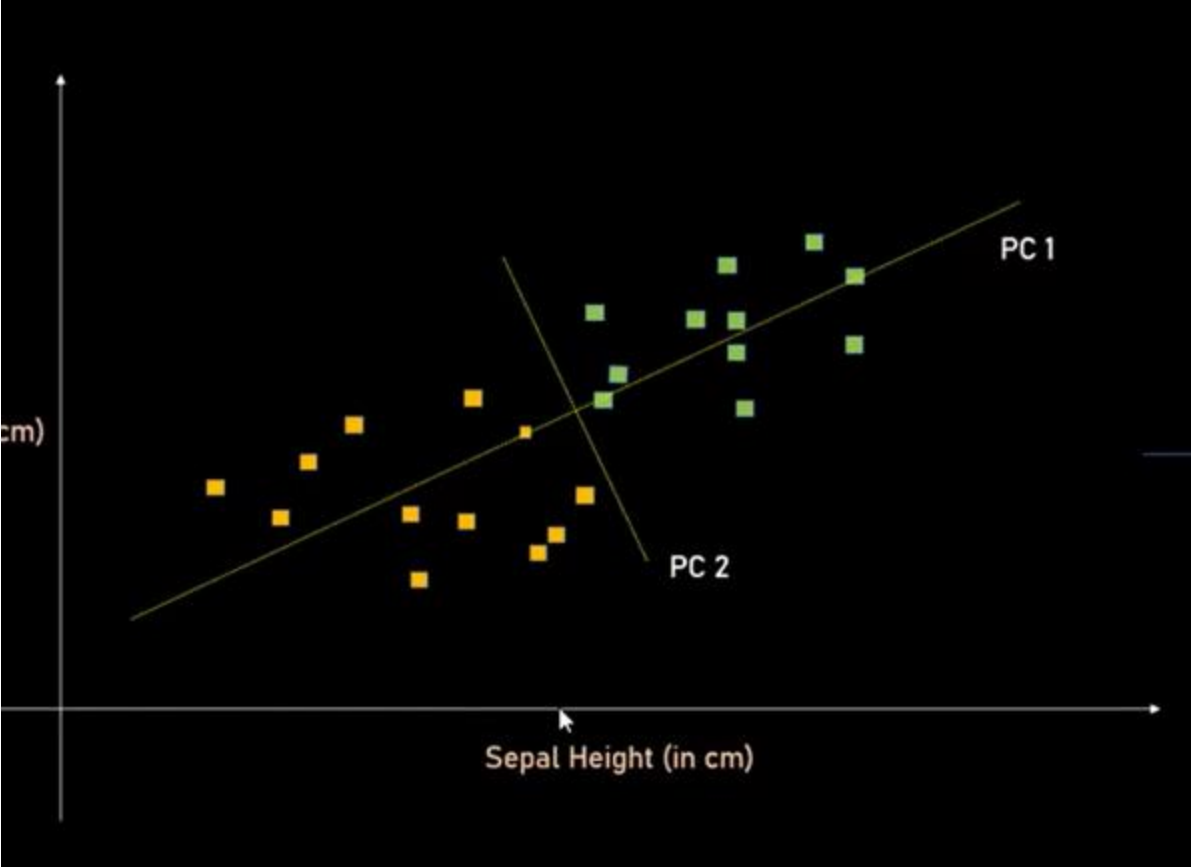
- To address overfitting, reduce the dimension, without losing the information.
- In this example two dimension is reduced to single dimension.
- But in general, their can be multiple dimensions... and will be reduced.
- When the data is viewed from one angle, it will be reduced to single dimension and the same is shown at the bottom right corner, and this will be Principal Component 1.

Similarly compute PC2



- Figure shows the representation of PC1 and PC2.
- Like this we have several principal components...
- Say PC1, PC2, PC3... and so on..
- In that PC1 will be of top priority.
- Each Principal Components are independent and are orthogonal. It means one PC does not depends on another...all of them are independent.

Another Example

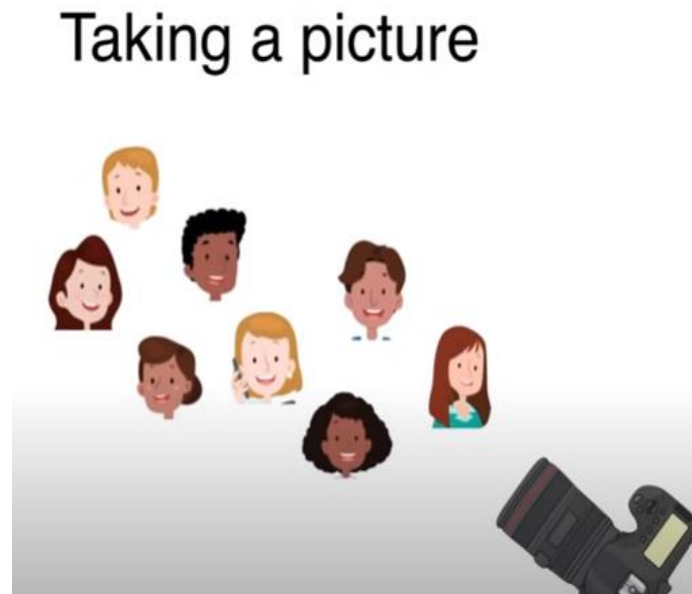


Example to illustrate the PC

Taking a picture



Multiple angles in which picture can be captured



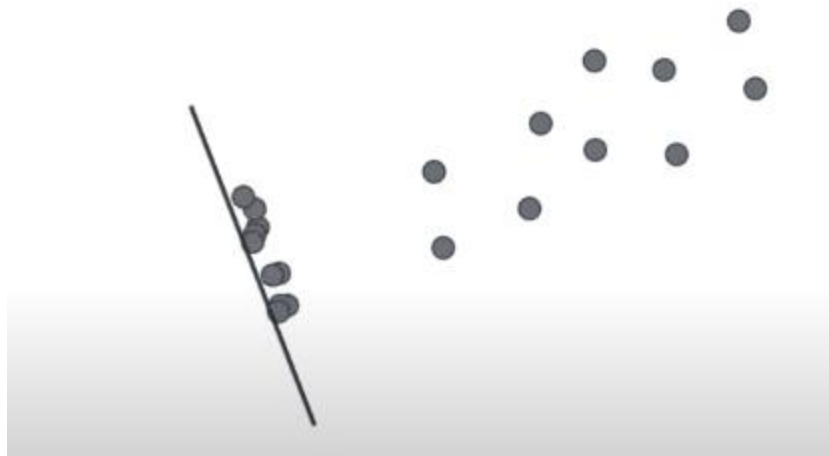
- In previous slide, the last picture gives the right angle to take the picture.
- It means, you have to identify a better angle to collect the data without losing much information.
- The angle shown in the last picture will capture all the faces, without much overlapping and without losing information.

In this example the second one is the best angle to project :

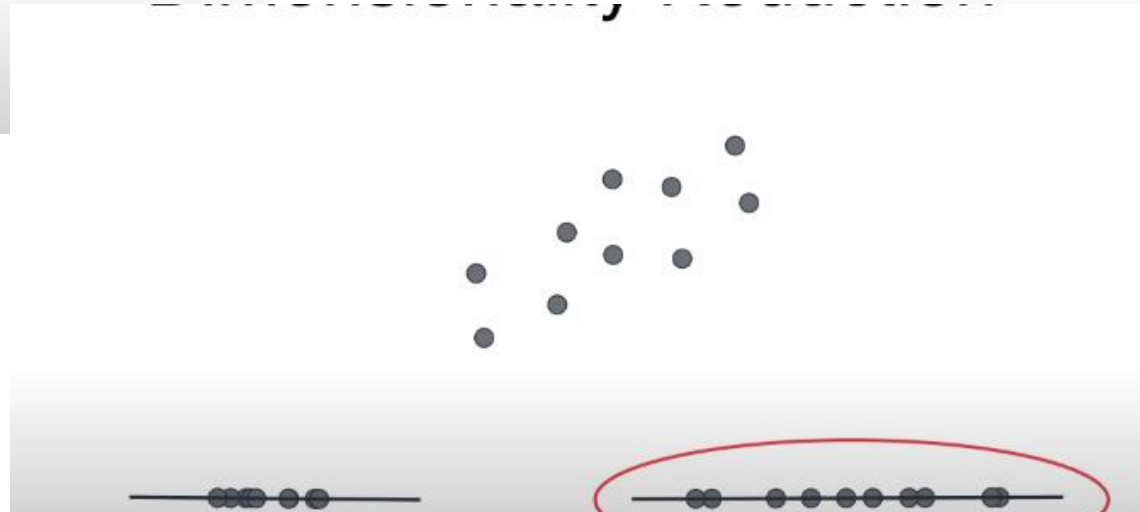
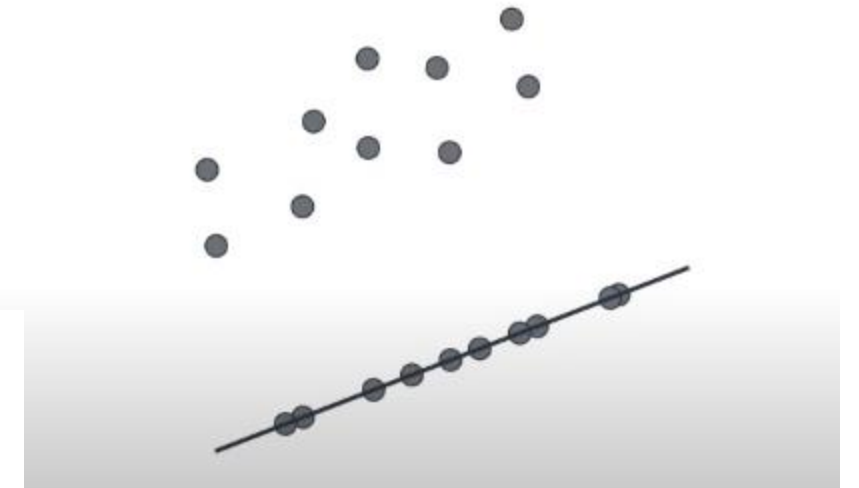
<https://www.youtube.com/watch?v=g-Hb26agBFg> (reference video)

<https://www.youtube.com/watch?v=MLaJbA82nzk>

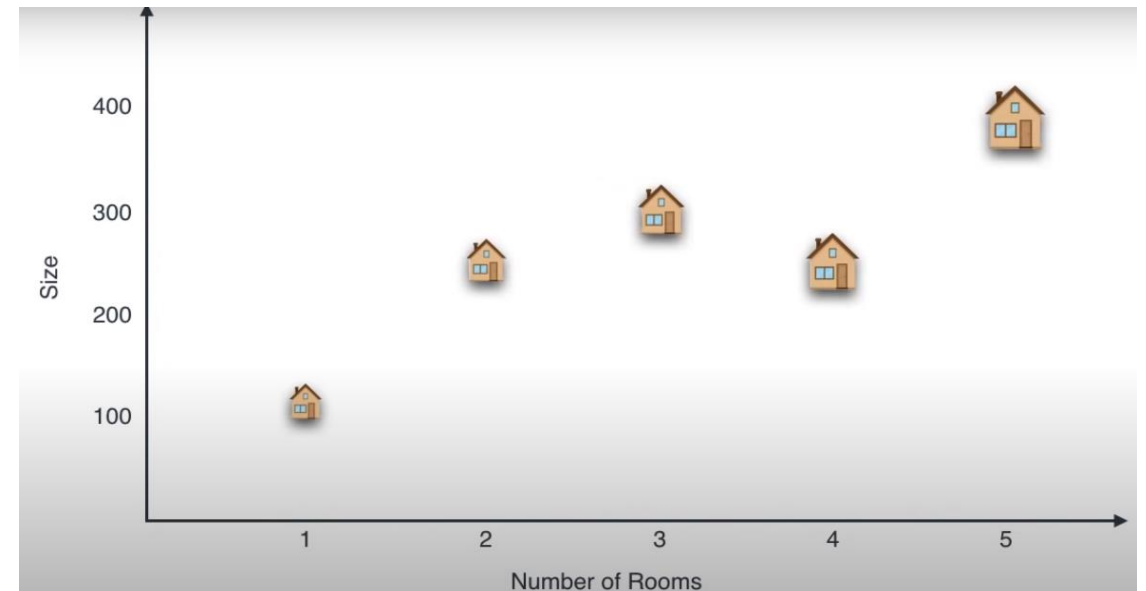
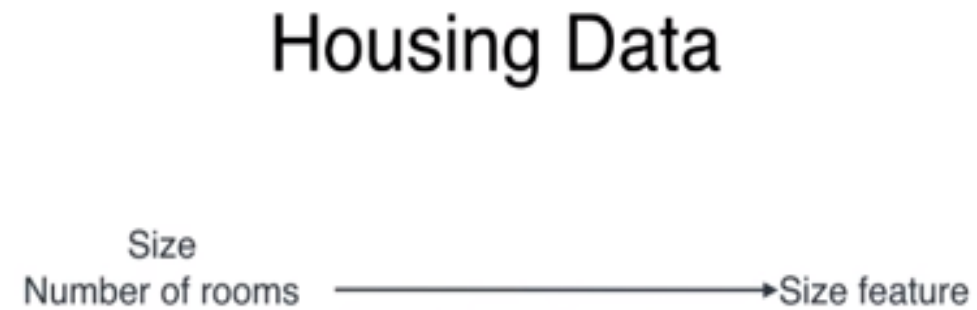
Dimensionality Reduction



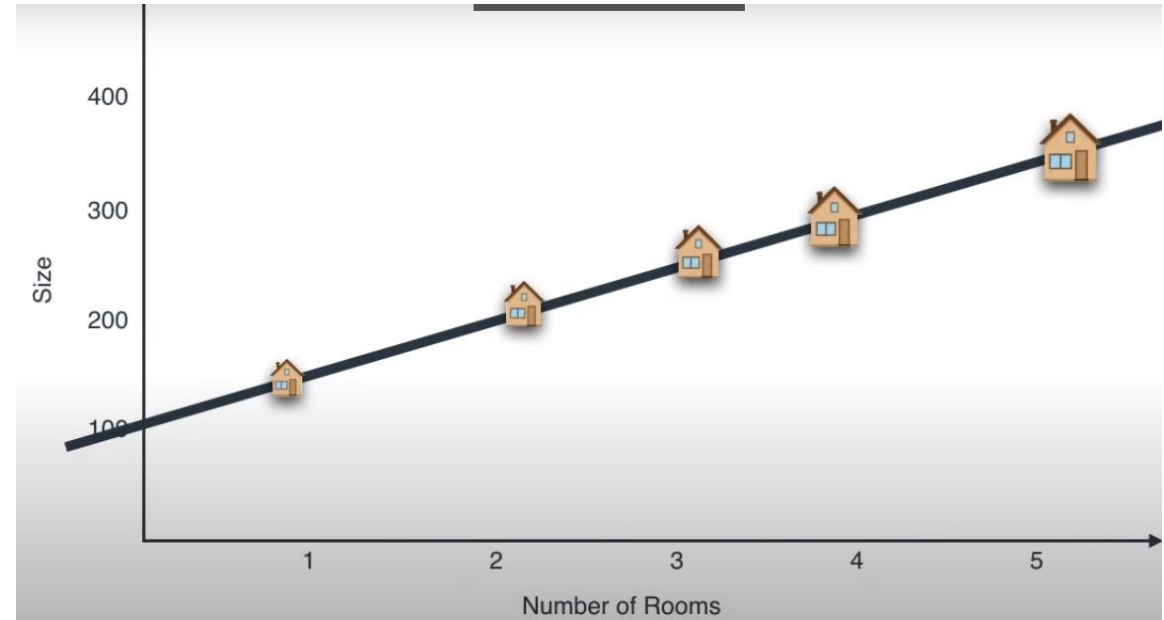
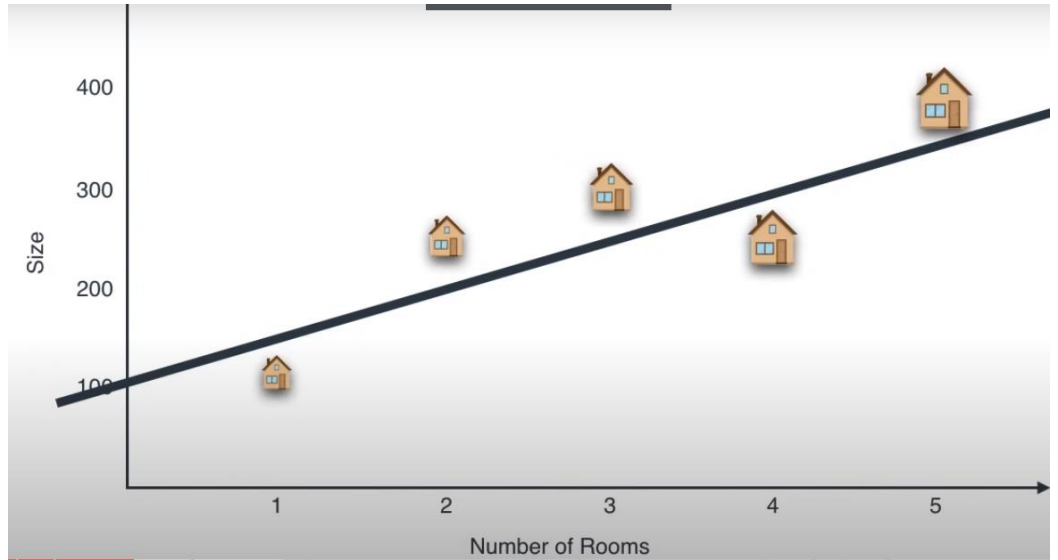
Dimensionality Reduction



Housing Example: More rooms..more the size



Two dimension is reduced to single dimension



- PCA is a method of dimensionality reduction.
- Example shows how to convert a two dimension to one dimension.

How to compute PCA?

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

- Consider the Samples given in the table (10 Samples).
- Compute the mean of X and mean of Y independently. Similar computation has to be done for each features. (In this example only two features).
- **Mean of X = 1.81 and Mean of Y = 1.91**

Next Step is to compute Co-Variance Matrix.

- Covariance between (x, y) is computed as given below:

$$\text{Cov}(x, y) = \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- The following covariance Matrix to be computed is:

$$C = \begin{bmatrix} \text{Cov}(x, x) & \text{Cov}(x, y) \\ \text{Cov}(y, x) & \text{Cov}(y, y) \end{bmatrix}$$

Covariance between (x and x)

X	Y	(X- Mean(X))	(x-mean(x) * (x-Mean(x))
2.5	2.4	0.69	0.476
0.5	0.7	-1.31	1.716
2.2	2.9	0.39	0.152
1.9	2.2	0.09	0.008
3.1	3	1.29	1.664
2.3	2.7	0.49	0.24
2	1.6	0.19	0.036
1	1.1	-0.81	0.656
1.5	1.6	-0.31	0.096
1.1	0.9	-0.71	0.504
			Total= 5.549
			Total/9 0.617

- Similarly compute co variance between (x,y) , (y,x) and (y,y) .
- Computed Co-Variance matrix is given in next slide

Final co-variance matrix

$$C = \begin{bmatrix} \text{Cov}(x, x) & \text{Cov}(x, y) \\ \text{Cov}(y, x) & \text{Cov}(y, y) \end{bmatrix}$$

$$= \begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix}$$

Alternate Method to compute Co-variance matrix

1	Original Data			Mean Centered Data	
2	2.5	2.4		0.69	0.49
3	0.5	0.7		-1.31	-1.21
4	2.2	2.9		0.39	0.99
5	1.9	2.2		0.09	0.29
6	3.1	3		1.29	1.09
7	2.3	2.7		0.49	0.79
8	2	1.6		0.19	-0.31
9	1	1.1		-0.81	-0.81
10	1.5	1.6		-0.31	-0.31
11	1.1	0.9		-0.71	-1.01
12					
13	1.81	1.91			
14	(mean of X)	Mean of Y			

Consider Mean centered Matrix as A and now compute Transpose of A * A to get the Covariance matrix: Divide the resultant matrix by (n-1)

Matrix A input

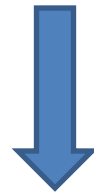
	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇	A ₈	A ₉	A ₁₀
1	0.69	-1.31	0.39	0.09	1.29	0.49	0.19	-0.81	-0.31	-0.71
2	0.49	-1.21	0.99	0.29	1.09	0.79	-0.31	-0.81	-0.31	-1.01

Matrix B dimension: 10 X 2

Restore matrix

	B ₁	B ₂
1	0.69	0.49
2	-1.31	-1.21
3	0.39	0.99
4	0.09	0.29
5	1.29	1.09
6	0.49	0.79
7	0.19	-0.31
8	-0.81	-0.81
9	-0.31	-0.31
10	-0.71	-1.01

	C ₁	C ₂
1	5.549	5.539
2	5.539	6.449



0.616556	0.615444
0.615444	0.716556

Next Step is to Compute Eigen Values using
the Co-variance matrix

If A is the given matrix (in this case co-variance matrix)

We can calculate eigenvalues from the following equation:

$$|A - \lambda I| = 0$$

Where A is the given matrix

λ is the eigen value

I is the identity Matrix

$$|A - \lambda I| = 0$$

$$\Rightarrow \begin{vmatrix} 0.6165 - \lambda & 0.6154 \\ 0.6154 & 0.7165 - \lambda \end{vmatrix} = 0 \rightarrow \begin{vmatrix} - & 0 \\ 0 & - \end{vmatrix} = 0$$

$$\Rightarrow \left| \begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = 0$$

$$\Rightarrow \left| \begin{bmatrix} 0.6165 - \lambda & 0.6154 \\ 0.6154 & 0.7165 - \lambda \end{bmatrix} \right| = 0$$

Determinant computation and finally Eigen values

$$\begin{aligned} & [(0.6165 - \lambda)(0.7165 - \lambda) - (0.6154)(0.6154)] = 0 \\ \Rightarrow & (0.6165 \times 0.7165) - (0.6165\lambda) - (0.7165\lambda) + \lambda^2 \\ & \quad - (0.6154) \times (0.6154) = 0 \\ \Rightarrow & \boxed{\lambda^2 - 1.333\lambda + 0.0630 = 0} \\ & \boxed{\begin{array}{l} a = 1 \\ b = -1.33 \\ c = 0.0630 \end{array}} \end{aligned}$$

Quadratic Formula Calculator

$$ax^2 + bx + c = 0$$

a =

b =

c =

Answer:

$$\begin{array}{l} x = 1.28081 \\ x = 0.0491875 \end{array}$$

$$\begin{array}{l} \lambda_1 = 1.2840 \\ \lambda_2 = 0.490 \end{array}$$

- Compute Eigen vector for each of the eigen value.

$$C V = \lambda V$$

- Consider the first eigen value $\lambda_1 = 1.284$
- C is the covariance matrix
- V is the eigen vector to be computed.

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 1.2840 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 1.2840 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

$$0.6165x_1 + 0.6154y_1 = 1.2840x_1$$

$$0.6154x_1 + 0.7165y_1 = 1.2840y_1$$

$$0.6154y_1 = 1.2840x_1 - 0.6165x_1$$

$$0.6154y_1 = 0.6675x_1$$

$$0.6675x_1 = 0.6154y_1$$

$$x_1 = \frac{0.6154}{0.6675} y_1$$

$$x_1 = 0.9219 y_1$$

$$y_1 = 1$$

$$x_1 = 0.9219$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 0.9219 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.9219/p \\ 1/p \end{bmatrix}$$

$$p = \sqrt{(0.9219)^2 + 1^2}$$

$$p = 1.360$$

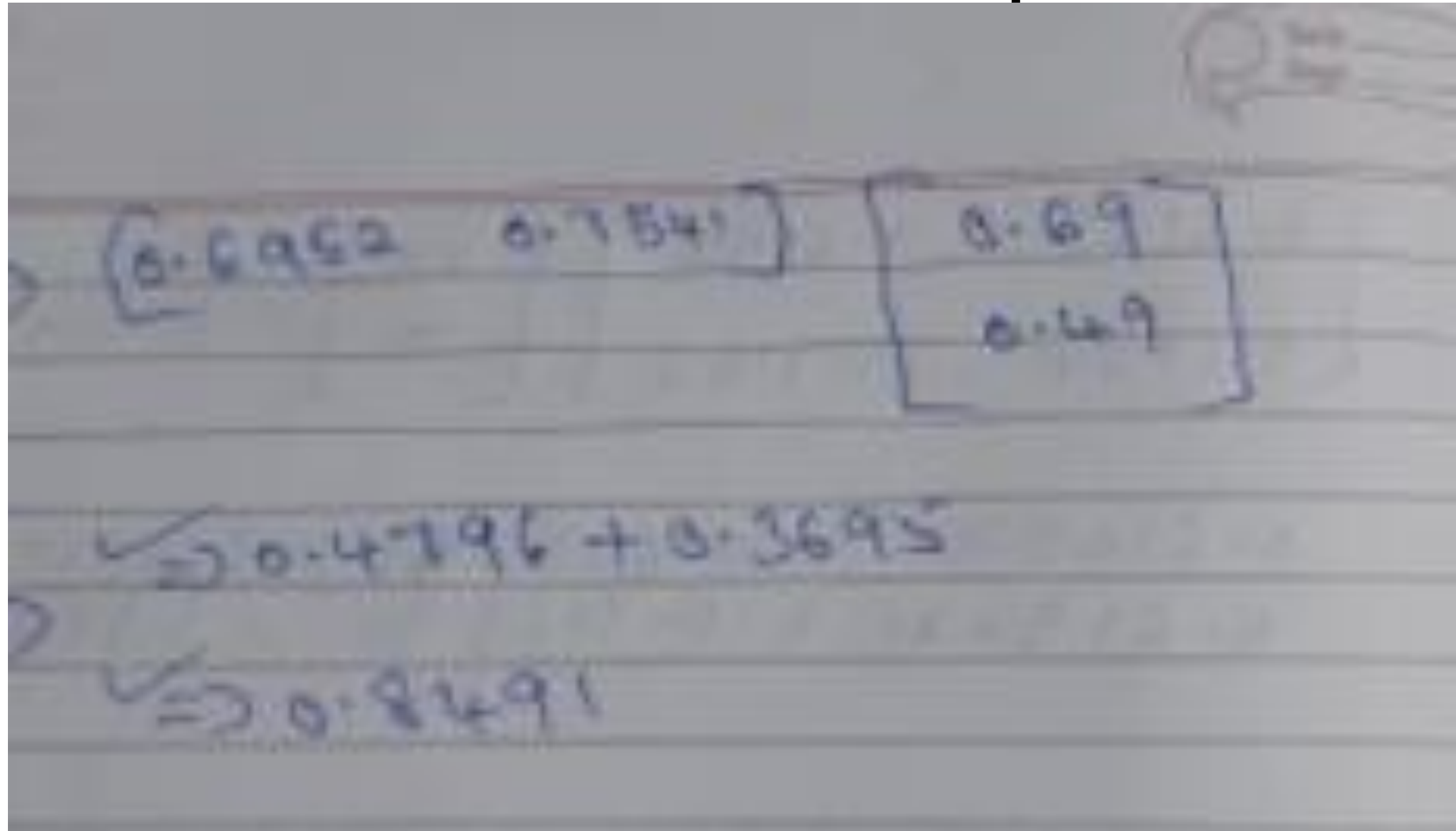
$$\begin{bmatrix} 0.6952 \\ 0.7541 \end{bmatrix}$$

Now convert the two dimension data to single dimension

Sample 1:

$$\begin{array}{l} x = 19.5 \\ y = 9.4 \end{array} \quad \Bigg| \quad \begin{array}{l} \bar{x} = 1.81 \\ \bar{y} = 1.91 \end{array}$$
$$\Rightarrow \begin{bmatrix} x - \text{mean}(x) \\ y - \text{mean}(y) \end{bmatrix} \Rightarrow \text{new + transformed data}$$

Final step



- Compute Eigen vector for the second eigen value.

$$CV = \lambda V$$

- Consider the first eigen value $\lambda_2 = 0.0490$
- C is the covariance matrix
- V is t|

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} \begin{bmatrix} X_1 \\ Y_1 \end{bmatrix} = 0.0490 \begin{bmatrix} X_1 \\ Y_1 \end{bmatrix}$$

- Using this we can have two linear equations:

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0.0490 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

$$\begin{aligned} 0.6165 x_1 + 0.6154 y_1 &= 0.0490 x_1 \\ 0.6154 x_1 + 0.7165 y_1 &= 0.0490 y_1 \end{aligned}$$

$$0.5674 x_1 = -0.6154 y_1$$

$$0.6154 x_1 = -0.6674 y_1$$

- Use any one of the following equation... final result remains same.

$$0.5674 x_1 = -0.6154 y_1$$
$$0.6154 x_1 = -0.6674 y_1$$

- $0.5674 x_1 = -0.6154 y_1$
- Divide both side by 0.5674.
- You will get : $x_1 = -1.0845 y_1$

- **$x_1 = -1.0845 y_1$**
- **If $y_1=1$, then x_1 will be -1.0845**
- So in that case (x_1, y_1) will be $(-1.0845, 1)$. This will be the initial eigen vector. Needs normalization to get the final value.
- To normalize, take square-root of sum of square of each eigen vector values, and consider this as 'x'
- Finally divide each eigen vector values by 'x' to get the final eigen vector.

eigen vectors are generated for the eigen value : 0.490

$$X_1 = -1.0845 Y_1$$
$$\begin{bmatrix} -1.0845 \\ 1 \end{bmatrix} = \frac{1^2 + 7614}{\sqrt{2 \cdot 17614}} + 1$$
$$= \frac{1.47517}{1}$$
$$\Rightarrow \begin{bmatrix} -0.7351 \\ 0.6778 \end{bmatrix}$$

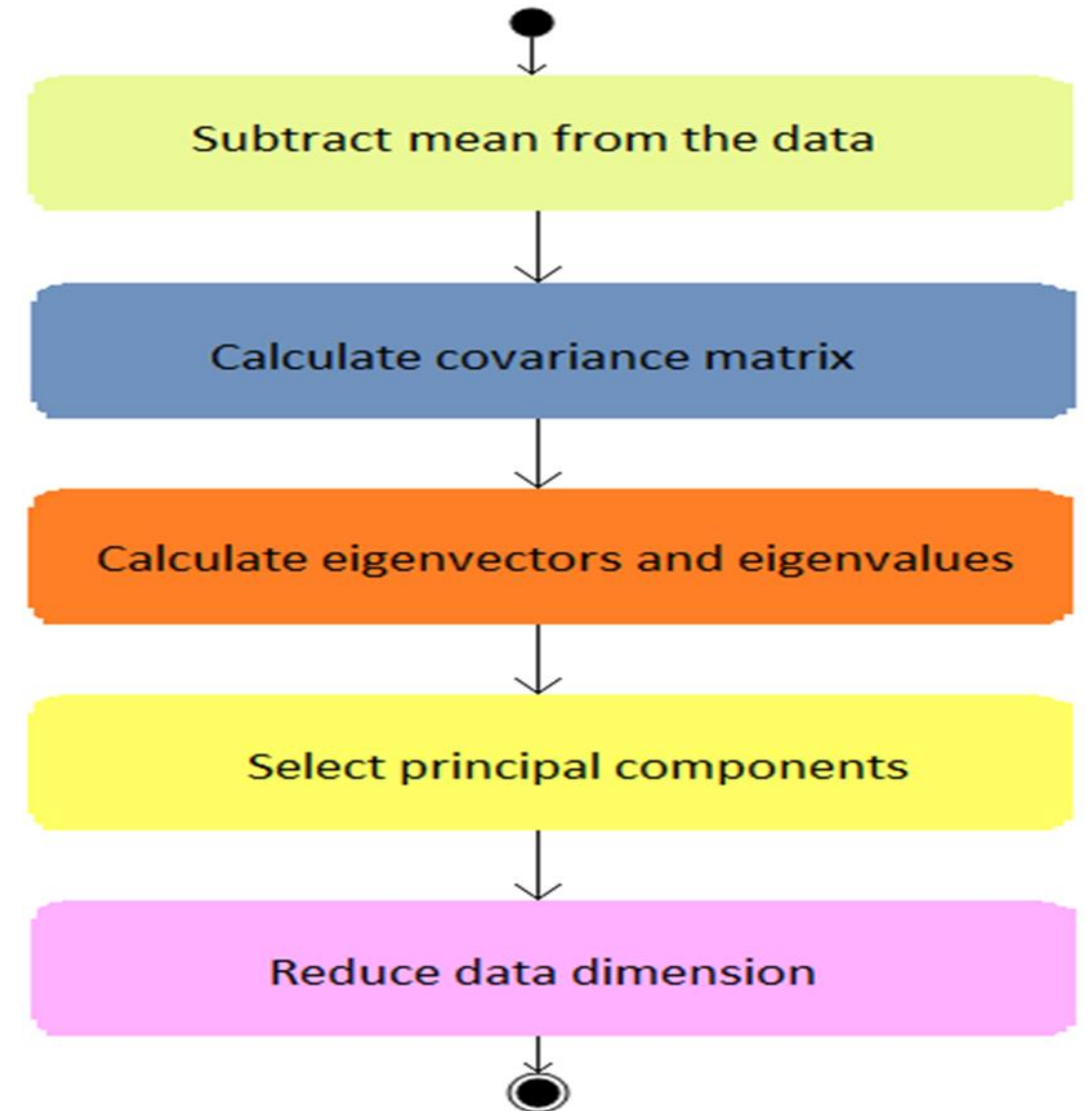
$$X_2 = 0.92194 Y_2$$
$$\begin{bmatrix} 0.92194 \\ 1 \end{bmatrix} = \frac{0.8499 + 1}{\sqrt{1.8499}}$$
$$= 1.3601$$
$$\Rightarrow \begin{bmatrix} 0.6778 \\ 0.7351 \end{bmatrix}$$

PCA

Theory – Algorithms – steps explained

Steps/ Functions to perform PCA

- Subtract mean.
- Calculate the covariance matrix.
- Calculate eigenvectors and eigenvalues.
- Select principal components.
- Reduce the data dimension.



- Principal components is a form of multivariate statistical analysis and is one method of studying the correlation or covariance structure in a set of measurements on m variables for n observations.
- Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.
- Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.
- So to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

- What do the covariances that we have as entries of the matrix tell us about the correlations between the variables?
- It's actually the sign of the covariance that matters
- if positive then : the two variables increase or decrease together (correlated)
- if negative then : One increases when the other decreases (Inversely correlated)
- Now, that we know that the covariance matrix is not more than a table that summaries the correlations between all the possible pairs of variables, let's move to the next step.

Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the principal components of the data.

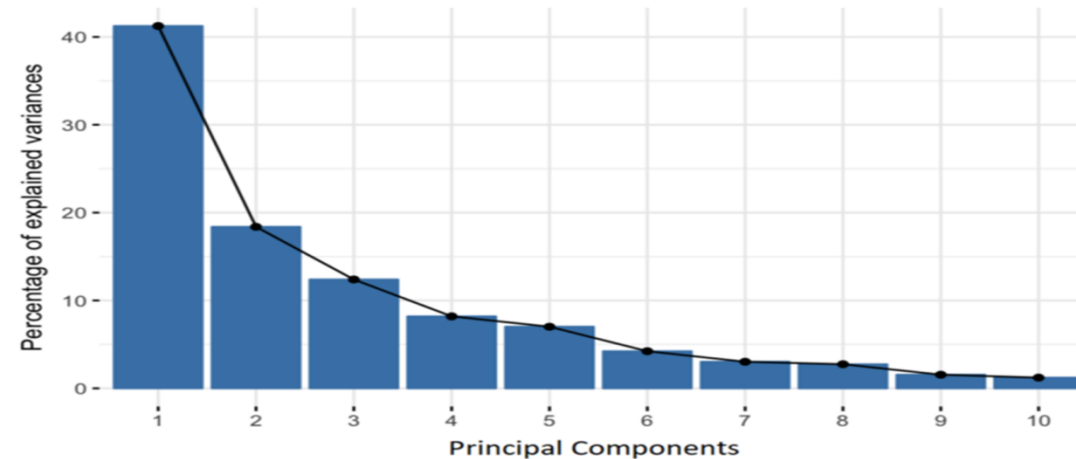
Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables.

These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components.

So, the idea is 10-dimensional data gives you 10 principal components, but PCA tries to put maximum possible information in the first component.

Then maximum remaining information in the second and so on, until having something like shown in the scree plot below.

- As there are as many principal components as there are variables in the data, principal components are constructed in such a manner that the first principal component accounts for the largest possible variance in the data set.



- Organizing information in principal components this way, will allow you to reduce dimensionality without losing much information, and this by discarding the components with low information and considering the remaining components as your new variables.
- An important thing to realize here is that, the principal components are less interpretable and don't have any real meaning since they are constructed as linear combinations of the initial variables.

**Characteristic Polynomial and characteristic equation
and**

Eigen Values and Eigen Vectors

Computation for 2×2 and 3×3 Square Matrix

Eigen Values and Eigen Vectors

Definition

Let A be an $n \times n$ matrix. A scalar λ is called an **eigenvalue** of A if there exists a nonzero vector \mathbf{x} in \mathbf{R}^n such that

$$A\mathbf{x} = \lambda\mathbf{x}.$$

The vector \mathbf{x} is called an **eigenvector** corresponding to λ .

The eigenvectors \mathbf{x} and eigenvalues λ of a matrix A satisfy

$$A\mathbf{x} = \lambda\mathbf{x}$$

If A is an $n \times n$ matrix, then \mathbf{x} is an $n \times 1$ vector, and λ is a constant.

The equation can be rewritten as $(A - \lambda I)\mathbf{x} = 0$, where I is the $n \times n$ identity matrix.

Solving the equation $|A - \lambda I_n| = 0$ for λ leads to all the eigenvalues of A .

On expanding the determinant $|A - \lambda I_n|$, we get a polynomial in λ .

This polynomial is called the **characteristic polynomial** of A .

The equation $|A - \lambda I_n| = 0$ is called the **characteristic equation** of A .

2 X 2 Example : Compute Eigen Values

$$A = \begin{bmatrix} 1 & -2 \\ 3 & -4 \end{bmatrix} \quad \text{so } A - \lambda I = \begin{bmatrix} 1 - \lambda & -2 \\ 3 & -4 - \lambda \end{bmatrix}$$

$$\begin{aligned} \det(A - \lambda I) &= (1 - \lambda)(-4 - \lambda) - (3)(-2) \\ &= \lambda^2 + 3\lambda + 2 \end{aligned}$$

Set $\lambda^2 + 3\lambda + 2$ to 0

$$\text{Then } \lambda = (-3 \pm \sqrt{9-8})/2$$

So the two values of λ are -1 and -2.

Example 1: Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} -4 & -6 \\ 3 & 5 \end{bmatrix}$$

Solution

Let us first derive the characteristic polynomial of A .

We get

$$A - \lambda I_2 = \begin{bmatrix} -4 & -6 \\ 3 & 5 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -4 - \lambda & -6 \\ 3 & 5 - \lambda \end{bmatrix}$$

$$|A - \lambda I_2| = (-4 - \lambda)(5 - \lambda) + 18 = \lambda^2 - \lambda - 2$$

We now solve the characteristic equation of A .

$$\lambda^2 - \lambda - 2 = 0 \implies (\lambda - 2)(\lambda + 1) = 0 \implies \lambda = 2 \text{ or } -1$$

The eigenvalues of A are 2 and -1 .

The corresponding eigenvectors are found by using these values of λ in the equation $(A - \lambda I_2)\mathbf{x} = \mathbf{0}$.

There are many eigenvectors corresponding to each eigenvalue.

For $\lambda = 2$

We solve the equation $(A - 2I_2)\mathbf{x} = \mathbf{0}$ for \mathbf{x} .

The matrix $(A - 2I_2)$ is obtained by subtracting 2 from the diagonal elements of A .

We get

$$\begin{bmatrix} -6 & -6 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{0}$$

This leads to the system of equations

$$\begin{aligned} -6x_1 - 6x_2 &= 0 \\ 3x_1 + 3x_2 &= 0 \end{aligned}$$

giving $x_1 = -x_2$. The solutions to this system of equations are $x_1 = -r$, $x_2 = r$, where r is a scalar.

Thus the eigenvectors of A corresponding to $\lambda = 2$ are nonzero vectors of the form

$$\mathbf{v}_1 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_2 \begin{bmatrix} -1 \\ 1 \end{bmatrix} = r \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

For $\lambda = -1$

We solve the equation $(A + 1I_2)x = 0$ for x .

The matrix $(A + 1I_2)$ is obtained by adding 1 to the diagonal elements of A . We get

$$\begin{bmatrix} -3 & -6 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{0}$$

This leads to the system of equations

$$-3x_1 - 6x_2 = 0$$

$$3x_1 + 6x_2 = 0$$

Thus $x_1 = -2x_2$. The solutions to this system of equations are $x_1 = -2s$ and $x_2 = s$, where s is a scalar. Thus the **eigenvectors** of A corresponding to $\lambda = -1$ are nonzero vectors of the form

$$\mathbf{v}_2 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_2 \begin{bmatrix} -2 \\ 1 \end{bmatrix} = s \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

- **Example 2** Calculate the eigenvalue equation and eigenvalues for the following matrix –

$$\begin{matrix} 1 & 0 & 0 \\ 0 & -1 & 2 \\ 2 & 0 & 0 \end{matrix}$$

Solution : Let $A = \begin{matrix} 1 & 0 & 0 \\ 0 & -1 & 2 \\ 2 & 0 & 0 \end{matrix}$ and $A - \lambda I = \begin{matrix} 1 - \lambda & 0 & 0 \\ 0 & -1 - \lambda & 2 \\ 2 & 0 & 0 - \lambda \end{matrix}$

We can calculate eigenvalues from the following equation:

$$\begin{aligned} |A - \lambda I| &= 0 & (1 - \lambda) [(-1 - \lambda)(-\lambda) - 0] - 0 + 0 &= 0 \\ & & \lambda (1 - \lambda) (1 + \lambda) &= 0 \end{aligned}$$

From this equation, we are able to estimate eigenvalues which are –
 $\lambda = 0, 1, -1$.

Example2 : Eigenvalues 3x3 Matrix

Find the eigenvalues of

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & -4 & 2 \\ 0 & 0 & 7 \end{bmatrix}$$

Solution:

$$A - \lambda I_n = \begin{bmatrix} 1 & 2 & 3 \\ 0 & -4 & 2 \\ 0 & 0 & 7 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 - \lambda & 2 & 3 \\ 0 & -4 - \lambda & 2 \\ 0 & 0 & 7 - \lambda \end{bmatrix}$$

$$\det(A - \lambda I_n) = 0 \rightarrow \det \begin{bmatrix} 1 - \lambda & 2 & 3 \\ 0 & -4 - \lambda & 2 \\ 0 & 0 & 7 - \lambda \end{bmatrix} = 0$$

$$(1 - \lambda)(-4 - \lambda)(7 - \lambda) = 0$$

$$\lambda = \{1, -4, 7\}$$

Example 3: Eigenvalues and Eigenvectors

Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 5 & 4 & 2 \\ 4 & 5 & 2 \\ 2 & 2 & 2 \end{bmatrix}$$

Solution

The matrix $A - \lambda I_3$ is obtained by subtracting λ from the diagonal elements of A . Thus

$$A - \lambda I_3 = \begin{bmatrix} 5 - \lambda & 4 & 2 \\ 4 & 5 - \lambda & 2 \\ 2 & 2 & 2 - \lambda \end{bmatrix}$$

The characteristic polynomial of A is $|A - \lambda I_3|$. Using row and column operations to simplify determinants, we get

Alternate Solution

$$|A - \lambda I_3| = 0$$

$$A = \begin{bmatrix} 5 & 4 & 1 \\ 2 & 4 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

λ = eigen values

For 3×3 matrix eigen values can be computed using the following equations

$$\lambda^3 - S_1\lambda^2 + S_2\lambda - S_3 = 0$$

S_1 = Sum of principal diagonal elements

S_2 = Sum of minors of principal diagonals

S_3 = Determinant of A

$$S_1 = \begin{bmatrix} 5 & & & \\ 2 & 5 & & \\ & 2 & 5 & \\ & & 2 & 5 \end{bmatrix}$$

$$S_1 = 5 + 5 + 2$$

$$S_2 = \begin{vmatrix} 5 & 2 \\ 2 & 2 \end{vmatrix} + \begin{vmatrix} 5 & 2 \\ 2 & 2 \end{vmatrix} + 2 \begin{vmatrix} 5 & 4 \\ 4 & 5 \end{vmatrix}$$
$$= 6 + 6$$

$$S_2 = \begin{vmatrix} 5 & 2 \\ 2 & 2 \end{vmatrix} + \begin{vmatrix} 5 & 2 \\ 2 & 2 \end{vmatrix} + \begin{vmatrix} 5 & 4 \\ 4 & 5 \end{vmatrix}$$
$$= (10 - 4) + (10 - 4) + (25 - 16)$$
$$= 6 + 6 + 9$$
$$= 21$$

$$S_2 = 21$$

S_3 = Determinant of A

$$S_3 = |A| = \begin{vmatrix} 5 & 4 & 2 \\ 4 & 5 & 2 \\ 2 & 2 & 2 \end{vmatrix}$$

$$= 5 \times \begin{vmatrix} 5 & 2 \\ 2 & 2 \end{vmatrix} - 4 \times \begin{vmatrix} 4 & 2 \\ 2 & 2 \end{vmatrix} + 2 \times \begin{vmatrix} 4 & 5 \\ 2 & 2 \end{vmatrix}$$

$$= 5 \times (10 - 4) - 4 \times (8 - 4) + 2 \times (8 - 10)$$

$$= 5 \times 6 - 4 \times 4 + 2 \times (-2)$$

$$= 30 - 16 - 4$$

$$\boxed{S_3 = 10}$$

$$\lambda^3 - 21\lambda^2 + 22\lambda - 22 = 0$$

$$\lambda^3 - 12\lambda^2 + 21\lambda - 10 = 0$$

Eigen values will be factors of C₂

Factors of 10 = 1, 2, 5, 10

check for sum +ve as well as -ve

In the above case we have

$$\lambda = 10 \& 1$$

$$\lambda_1 = 10$$

$$\lambda_2 = 1$$

Eigen vector:

$$(A - \lambda I_3) x = 0$$

$$A = \begin{bmatrix} 5 & 4 & 2 \\ 4 & 5 & 2 \\ 2 & 2 & 8 \end{bmatrix} \quad \lambda = \{10, 1\} \quad I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$\lambda = 0$

$$\begin{bmatrix} 5 & 4 & 2 \\ 4 & 5 & 2 \\ 2 & 2 & 8 \end{bmatrix} - 10 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

$$\Rightarrow \begin{bmatrix} -5 & 4 & 2 \\ 4 & -5 & 2 \\ 2 & 2 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

$$\Rightarrow \begin{bmatrix} -5 & 4 & 2 \\ 4 & -5 & 2 \\ 2 & 2 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

$$-5x_1 + 4x_2 + 2x_3 = 0$$

$$4x_1 - 5x_2 + 2x_3 = 0$$

$$2x_1 + 2x_2 + 8x_3 = 0$$

Consider any two equations

$$-5x_1 + 4x_2 + 3x_3 = 0$$

$$6x_1 - 5x_2 + 3x_3 = 0$$

$$\frac{x_1}{(3+5)-(1)} = \frac{-x_2}{-(15-8)} = \frac{x_3}{28-16}$$

$$\frac{x_1}{8+10} = \frac{x_2}{18} = \frac{x_3}{12}$$

$$\frac{x_1}{18} = \frac{x_2}{18} = \frac{x_3}{12} = \lambda$$

$$\Rightarrow \frac{x_1}{9} = \frac{x_2}{9} = \frac{x_3}{-1} = 2\lambda = \mu$$

$$\frac{x_1}{9} = \frac{x_2}{9} = \frac{x_3}{-1} \quad ; \quad \begin{matrix} x_1 = 9\mu \\ x_2 = 9\mu \\ x_3 = -\mu \end{matrix}$$

$$\text{Hence } X_1 = \begin{bmatrix} 9\mu \\ 9\mu \\ -\mu \end{bmatrix} \\ = \mu \begin{bmatrix} 9 \\ 9 \\ -1 \end{bmatrix}$$

- $\lambda_2 = 1$

Let $\lambda = 1$ in $(A - \lambda I_3)\mathbf{x} = \mathbf{0}$. We get

$$(A - 1I_3)\mathbf{x} = \mathbf{0}$$
$$\begin{bmatrix} 4 & 4 & 2 \\ 4 & 4 & 2 \\ 2 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{0}$$

The solution to this system of equations can be shown to be $x_1 = -s - t$, $x_2 = s$, and $x_3 = 2t$, where s and t are scalars. Thus the eigenspace of $\lambda_2 = 1$ is the space of vectors of the form.

$$\begin{bmatrix} -s - t \\ s \\ 2t \end{bmatrix}$$

Separating the parameters s and t , we can write

$$\begin{bmatrix} -s - t \\ s \\ 2t \end{bmatrix} = s \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -1 \\ 0 \\ 2 \end{bmatrix}$$

Thus the eigenspace of $\lambda = 1$ is a two-dimensional subspace of \mathbf{R}^3 with basis

$$\left\{ \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 2 \end{bmatrix} \right\}$$

If an eigenvalue occurs as a k times repeated root of the characteristic equation, we say that it is of **multiplicity** k . Thus $\lambda=10$ has multiplicity 1, while $\lambda=1$ has multiplicity 2 in this example.

Linear Discriminant Analysis (LDA)

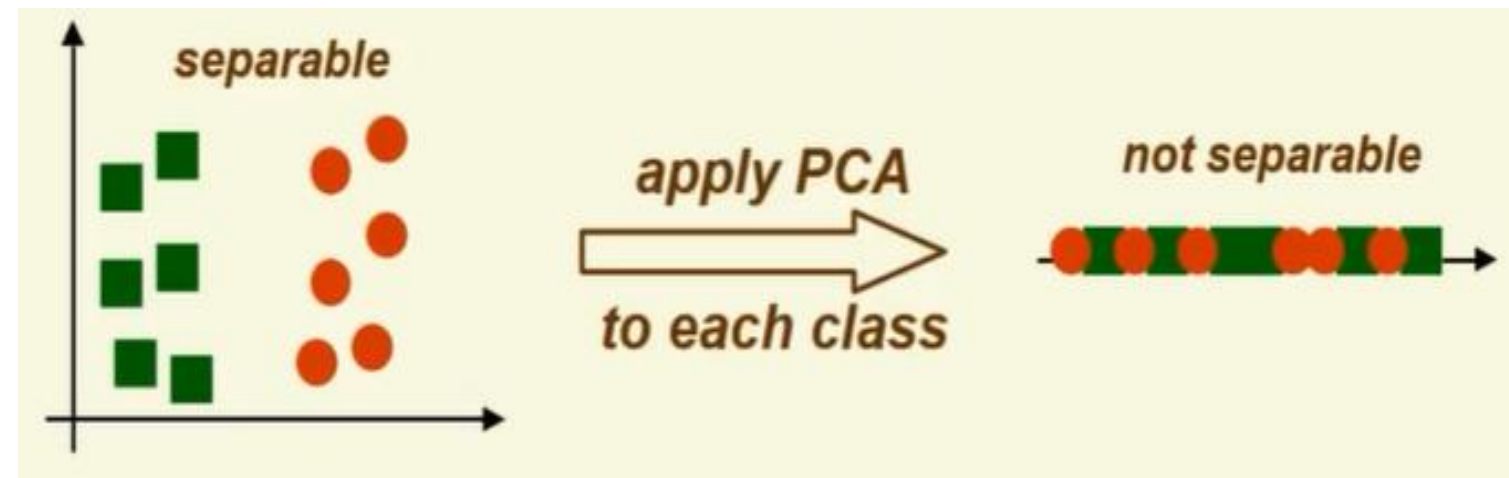
Data representation vs. Data Classification

Difference between PCA vs. LDA

- PCA finds the most accurate data representation in a lower dimensional space.
- Projects the data in the directions of maximum variance.
- However the directions of maximum **variance may be useless for classification**
- In such condition LDA which is also called as Fisher LDA works well.
- LDA is similar to PCA but LDA in addition finds the axis that maximizes the separation between multiple classes.

LDA Algorithm

- PCA is good for dimensionality reduction.
- However Figure shows how PCA fails to classify. (because it will try to project this points which maximizes variance and minimizes the error)



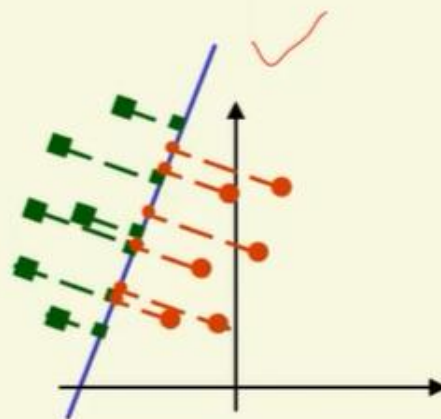
- Fisher Linear Discriminant Project to a line which reduces the dimension and also maintains the class discriminating information.

Projection of the samples in the second picture is the best:

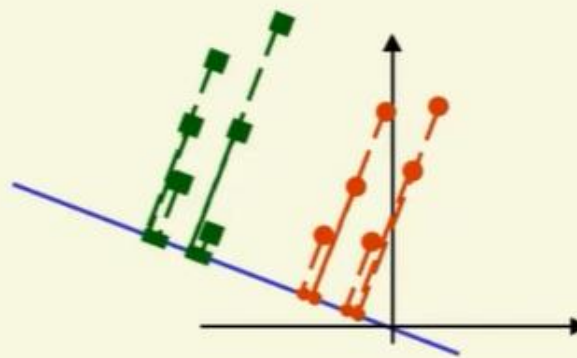
Fisher Linear Discriminant

- **Main idea:** find projection to a line s.t. samples from different classes are well separated

Example in 2D



*bad line to project to,
classes are mixed up*



*good line to project to,
classes are well separated*

Describe the algorithm with an example:

- Consider a 2-D dataset
- $C1 = X1 = (x1, x2) = \{(4,1), (2,4), (2,3), (3,6), (4,4)\}$
- $C2 = X2 = (x1, x2) = \{(9,10), (6,8), (9,5), (8,7), (10,8)\}$

Step 1: Compute within class scatter matrix(S_w)

- $S_w = s_1 + s_2$
- s_1 is the covariance matrix for class 1 and
- s_2 is the covariance matrix for s_2 .

- Note : Covariance matrix is to be computed on the Mean Centered data
- For the given example: mean of $C_1 = (3, 3.6)$ and
- mean of $C_2 = (8, 4, 7.6)$
- $S_1 = \text{Transpose of mean centred data} * \text{Mean centred data}$

$$X = \text{Transpose of } A * A ; X / (n-1)$$

Matrix A dimension: 2 X 5

Matrix B dimension: 5 X 2

Matrix A input

Insert matrix Restore matrix

	A ₁	A ₂	A ₃	A ₄	A ₅
1	1	-1	-1	0	1
2	-2.6	0.4	-0.6	2.4	0.4

Clear Fill empty cells with zero

Matrix B input

Insert matrix Restore matrix

Complex numbers (more)

Decimal *i*

	B ₁	B ₂
1	1	-2.6
2	-1	0.4
3	-1	-0.6
4	0	2.4
5	1	0.4

Clear Fill empty cells with zero

Calculate

1. The main condition of matrix multiplication is that the number of rows of the 2nd one.

2. As a result of multiplication you will get a new matrix has and the same quantity of columns as the 2nd one.

3. For example if you multiply a matrix of 'n' x 'k' by 'k' x 'm' you will get a matrix of 'n' x 'm'.

To understand matrix multiplication better input any example and examine the solution.

	C ₁	C ₂
1	4	-2
2	-2	13.2

	1	-0.5
	-0.5	3.3

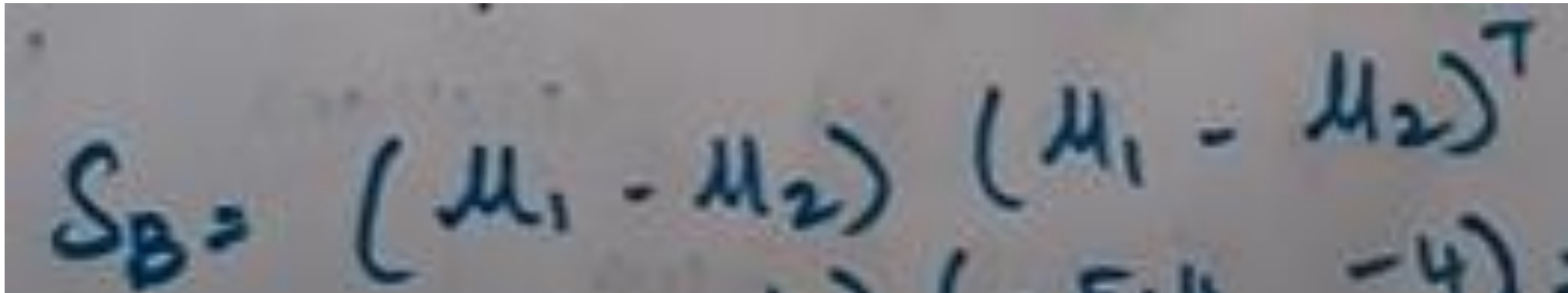
Computed values s_1, s_2 and S_w

$$S_1 = \begin{bmatrix} 0.8 & -0.4 \\ -0.4 & 2.6 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 1.84 & -0.04 \\ -0.04 & 2.64 \end{bmatrix}$$

$$S_w = S_1 + S_2$$
$$S_w = \begin{bmatrix} 2.64 & -0.44 \\ -0.44 & 5.28 \end{bmatrix}$$

Step 2: Compute between class scatter
Matrix(S_b)



A photograph of a handwritten mathematical formula on a piece of paper. The formula is $S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$. The handwriting is in blue ink on a light-colored background. The text is slightly blurred and tilted.

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

- Mean 1 (M1) = (3, 3.6)
- Mean 2 (M2) = (8, 4, 7.6)
- $(M1 - M2) = (3 - 8.4, 3.6 - 7.6) = (-5.4, 4.0)$

$$S_B = (\mu_1 - \mu_2) (\mu_1 - \mu_2)^T \\ = \begin{pmatrix} -5.4 \\ -4 \end{pmatrix} \begin{pmatrix} -5.4 & -4 \end{pmatrix} = \begin{pmatrix} 29.16 & 21.6 \\ 21.6 & 16.00 \end{pmatrix}$$

Step 3: Find the best LDA projection vector

- To do this ..compute the Eigen values and eigen vector for the largest eigen value, on the matrix which is the product of :

$$S_W^{-1} S_B = \begin{bmatrix} 11.89 & 8.81 \\ 5.08 & 3.76 \end{bmatrix}$$

- In this example, highest eigen value is : 15.65 ($\lambda = 15.65$)

Compute inverse of S_w^{-1}

• =

$$S_w^{-1} = \begin{bmatrix} 2.64 & -0.44 \\ -0.44 & 5.28 \end{bmatrix}$$

S_w^{-1} is found by using the formula

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$S_w^{-1}$$

$$\text{So, } S_w = \begin{bmatrix} 2.64 & -0.44 \\ -0.44 & 5.28 \end{bmatrix}$$

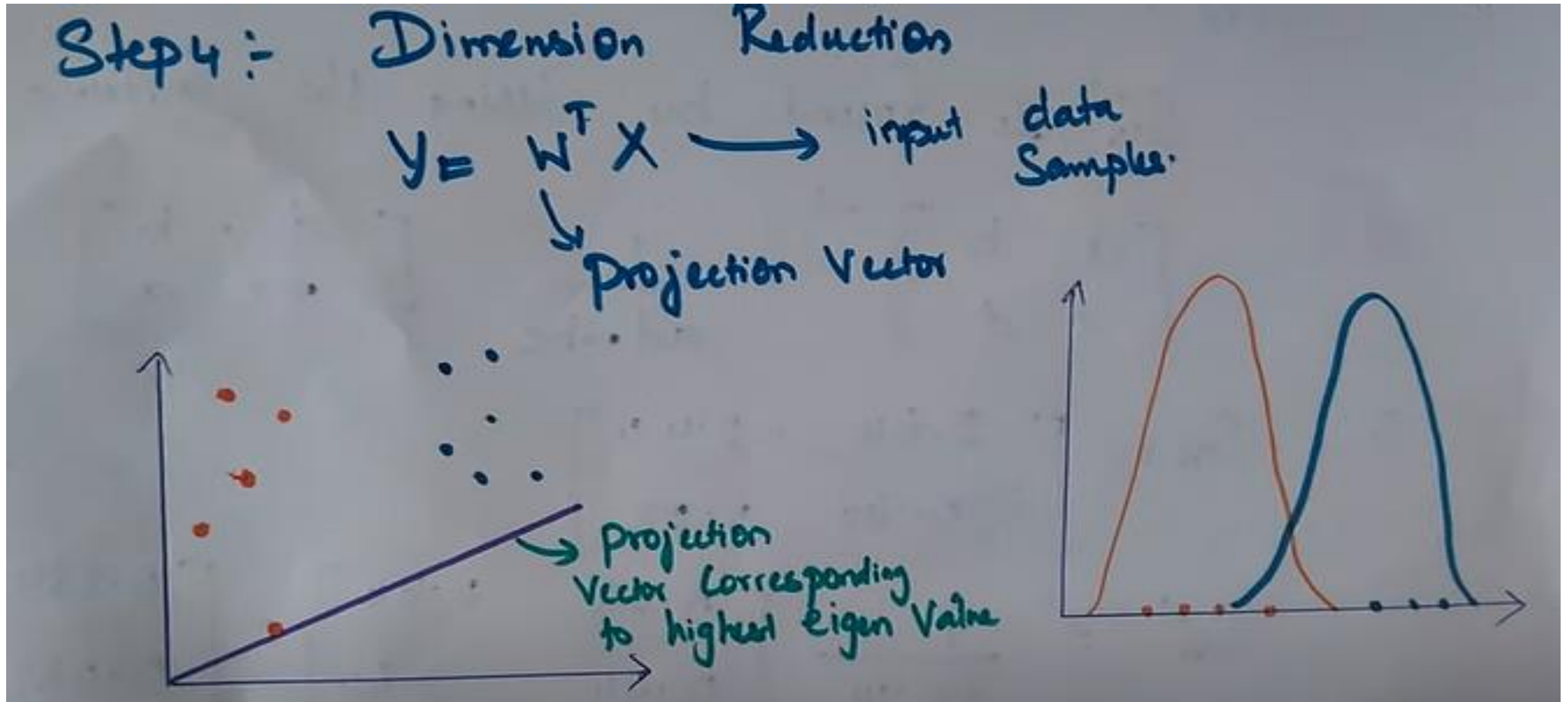
$$S_w^{-1} = \frac{1}{13.74} \begin{bmatrix} 5.28 & 0.44 \\ 0.44 & 2.64 \end{bmatrix} = \begin{bmatrix} 0.384 & 0.032 \\ 0.032 & 0.192 \end{bmatrix}$$

Eigen vector computed for Eigen value: 15.65

$$\begin{bmatrix} 11.89 & 8.81 \\ 5.08 & 3.76 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 15.65 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

we get $\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0.91 \\ 0.39 \end{bmatrix}$

Step 4: Dimension Reduction



Summary of the Steps

- Step 1 - Computing the within-class and between-class scatter matrices.
- Step 2 - Computing the eigenvectors and their corresponding eigenvalues for the scatter matrices.
- Step 3 - Sorting the eigenvalues and selecting the top k .
- Step 4 - Creating a new matrix that will contain the eigenvectors mapped to the k eigenvalues.
- Step 5 - Obtaining new features by taking the dot product of the data and the matrix from Step 4.

Singular Value Decomposition (SVD)

What is singular value decomposition

explain with example?

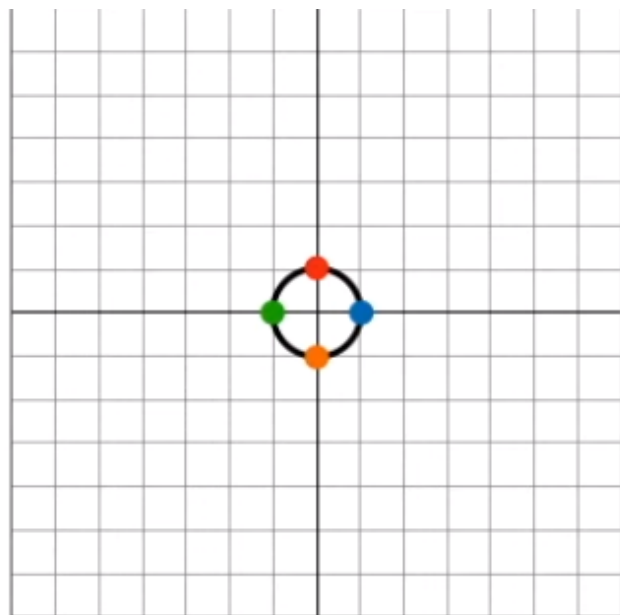
- The singular value decomposition of a matrix A is **the factorization of A into the product of three matrices $A = UDV^T$ where the columns of U and V^T are orthonormal and the matrix D is diagonal with positive real entries.** The SVD is useful in many tasks.
- Calculating the SVD consists of finding the eigenvalues and eigenvectors of AA^T and $A^T A$.
- The eigenvectors of $A^T A$ make up the columns of V , the eigenvectors of AA^T make up the columns of U .
- Also, the singular values in S are square roots of eigenvalues from AA^T or $A^T A$.
- The singular values are the diagonal entries of the S matrix and are arranged in descending order. The singular values are always real numbers.
- If the matrix A is a real matrix, then U and V are also real.

where:

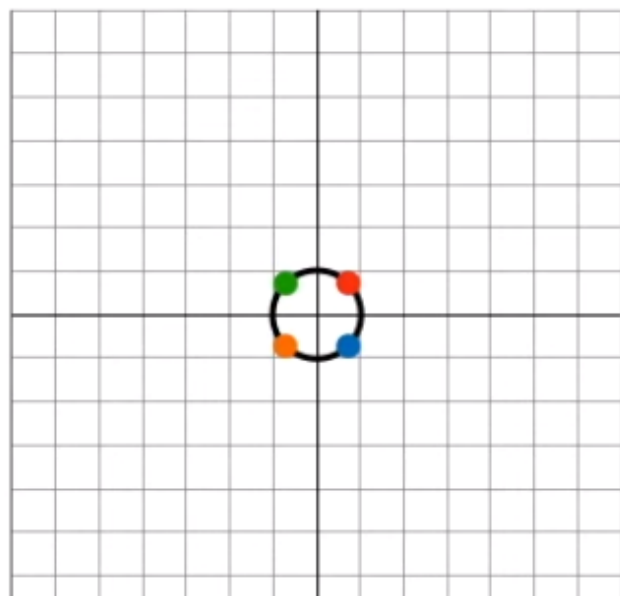
- U : $m \times r$ matrix of the orthonormal eigenvectors of AA^T .
- V^T : transpose of a $r \times n$ matrix containing the orthonormal eigenvectors of $A^T A$.
- W : a $r \times r$ diagonal matrix of the singular values which are the square roots of the eigenvalues of AA^T and $A^T A$.

Singular decomposition analysis(SVD)

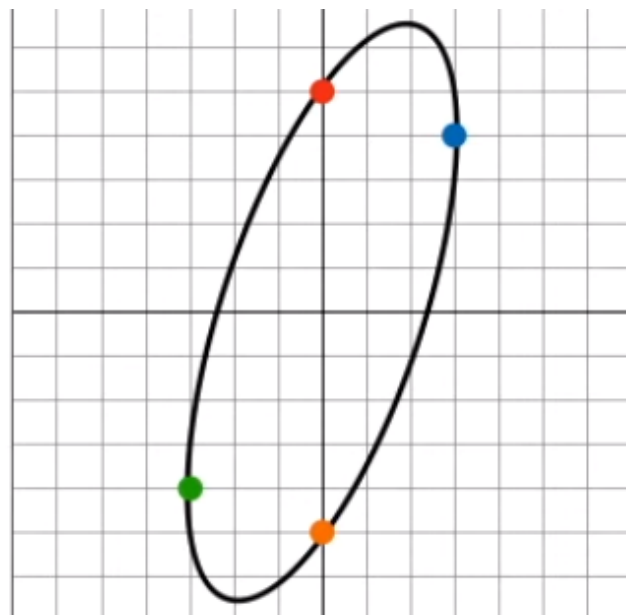
$$C_{m \times n} = U_{m \times r} \times \Sigma_{r \times r} \times V_{r \times n}^T$$



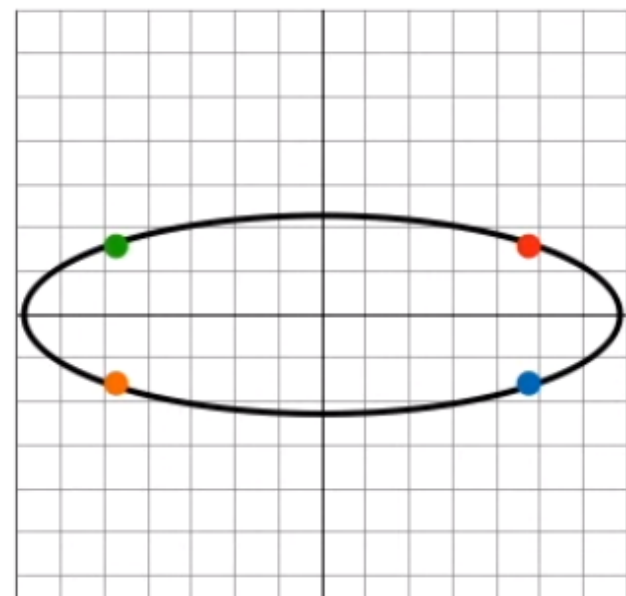
V^\dagger ↓ ↻



A →

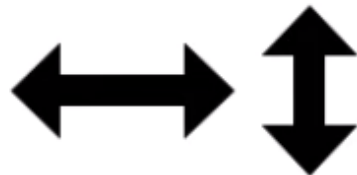


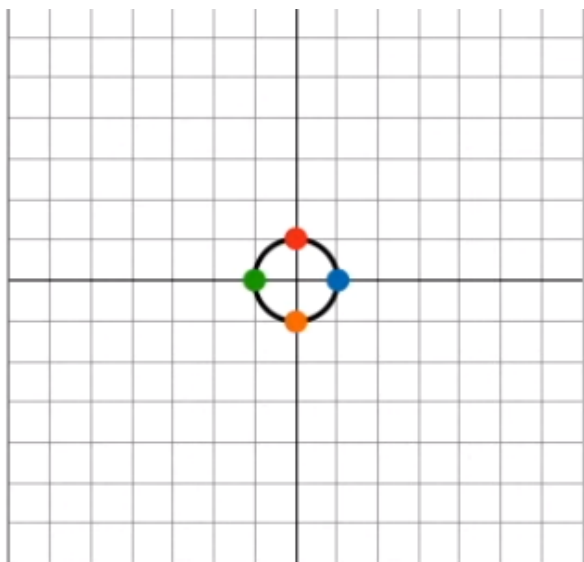
U ↑ ↻



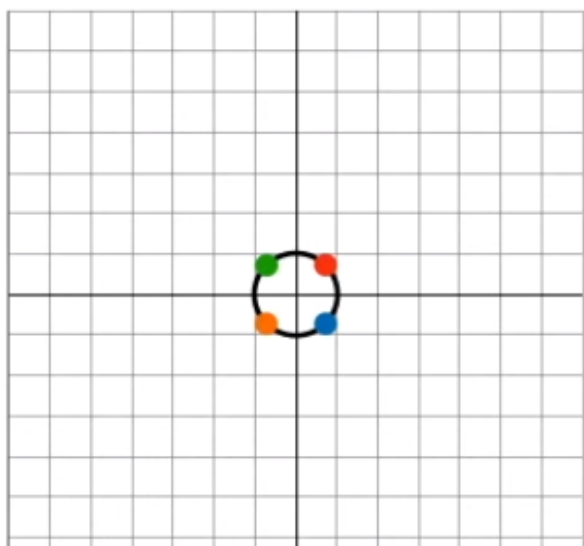
$$A = U \Sigma V^\dagger$$

Σ →



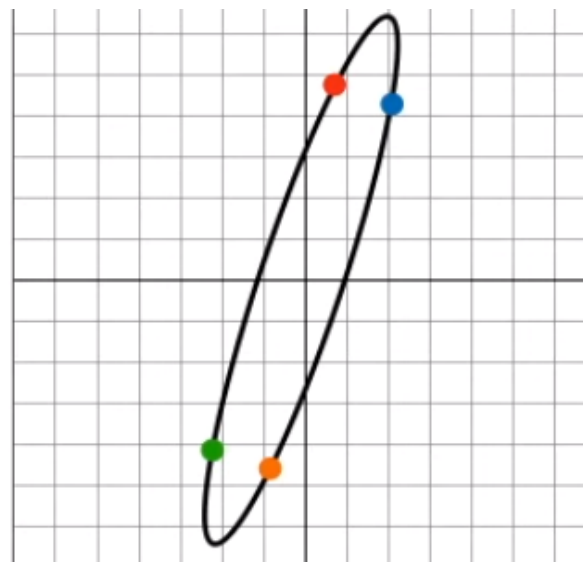


$$\begin{bmatrix} 0.7071 & 0.7071 \\ -0.7071 & 0.7071 \end{bmatrix} \downarrow V^\dagger$$



$$\begin{bmatrix} 1.8 & 1.2 \\ 4.4 & 4.6 \end{bmatrix} \xrightarrow{A}$$

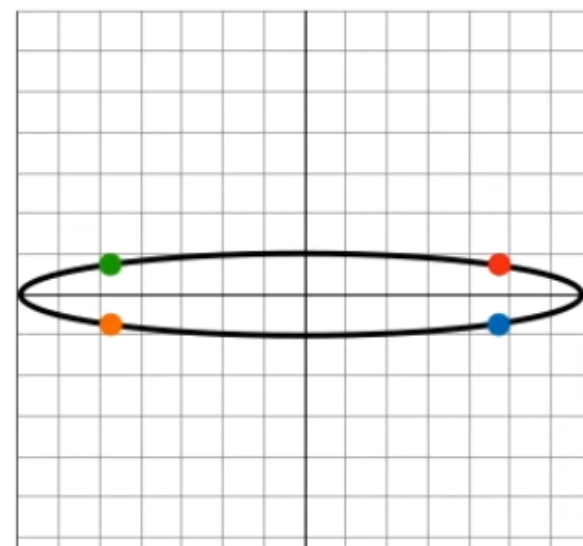
$$A = U \Sigma V^\dagger$$



$$U \uparrow \begin{bmatrix} 0.316 & -0.949 \\ 0.949 & 0.316 \end{bmatrix}$$

$$\Sigma \begin{bmatrix} 6.71 & 0 \\ 0 & 0.44 \end{bmatrix}$$

\longleftrightarrow \updownarrow
 6.71 0.44



$$C = \begin{pmatrix} 5 & 5 \\ -1 & 7 \end{pmatrix}$$

$$\text{SVD of } C = U \Sigma V^T$$

$$C^T C = \begin{pmatrix} 5 & -1 \\ 5 & 7 \end{pmatrix} \begin{pmatrix} 5 & 5 \\ -1 & 7 \end{pmatrix} = \begin{pmatrix} 26 & 18 \\ 18 & 74 \end{pmatrix}$$

COMPUTE EIGEN VALUES

~~QED~~

$$|C^T C - \lambda I| = \begin{vmatrix} 26-\lambda & 18 \\ 18 & 74-\lambda \end{vmatrix}$$

$$= \lambda^2 - 100\lambda + 1600$$

$$\begin{aligned} a &= 1 \\ b &= -100 \\ c &= 1600 \end{aligned}$$

$$\boxed{\begin{aligned} \lambda_1 &= 20 \\ \lambda_2 &= 80 \end{aligned}}$$

EIGEN VECTORS

$$\lambda_1 = 20$$

$$(C^T C - 20I) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 6 & 18 \\ 18 & 54 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$v_1 = \begin{pmatrix} -3/\sqrt{10} \\ 1/\sqrt{10} \end{pmatrix}$$

$$v_2 = \begin{pmatrix} 1/\sqrt{10} \\ 3/\sqrt{10} \end{pmatrix}$$

$$\text{So } Z = \begin{pmatrix} -\frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \\ \frac{1}{\sqrt{10}} & \frac{3}{\sqrt{10}} \end{pmatrix}$$

λ_1 = Square roots of eigen values
of $Z^T Z$ in the diagonal position

$$= \begin{pmatrix} \sqrt{20} & 0 \\ 0 & \sqrt{80} \end{pmatrix} = \begin{pmatrix} 2\sqrt{5} & 0 \\ 0 & 4\sqrt{5} \end{pmatrix}$$

$$C = U \Sigma V^T$$

$$CV = U \Sigma V^T V$$

$$\boxed{CV = U \Sigma}$$

$$\begin{pmatrix} 5 & 5 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} -\frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \\ \frac{1}{\sqrt{10}} & \frac{3}{\sqrt{10}} \end{pmatrix} = C \Sigma$$

$$\Rightarrow \begin{pmatrix} -\sqrt{10} & 2\sqrt{10} \\ \sqrt{10} & 2\sqrt{10} \end{pmatrix} = C \begin{pmatrix} 2\sqrt{5} & 0 \\ 0 & 4\sqrt{5} \end{pmatrix}$$

End of Unit 5

End of the Syllabus : Pattern Recognition

CS745

Thank you and all the best