Machine Learning (19CSE305)

Syllabus

Unit 1

Foundations of supervised learning - Decision trees and inductive bias, Regression Vs Classification, Supervised: Linear Regression, Logistic Regression, Generalisation, Training, Validation and Testing, Problem of Overfitting, Bias vs Variance, Performance metrics, Decision Tree, Random Forest, Perceptron, Beyond binary classification

Unit 2

Advanced supervised learning - Naive Bayes, Bayesian Belief Network, K-Nearest Neighbour, Support vector machines, Markov model, Hidden Markov Model, Parameter Estimation : MLE and Bayesian Estimate, Expectation Maximisation, Neural Networks

Unit 3

Unsupervised Learning : Curse of Dimensionality, Dimensionality Reduction Techniques, Principal component analysis, Linear Discriminant Analysis Clustering: K-means, Hierarchical, Spectral, subspace clustering, association rule mining. Case Study: Recommendation systems

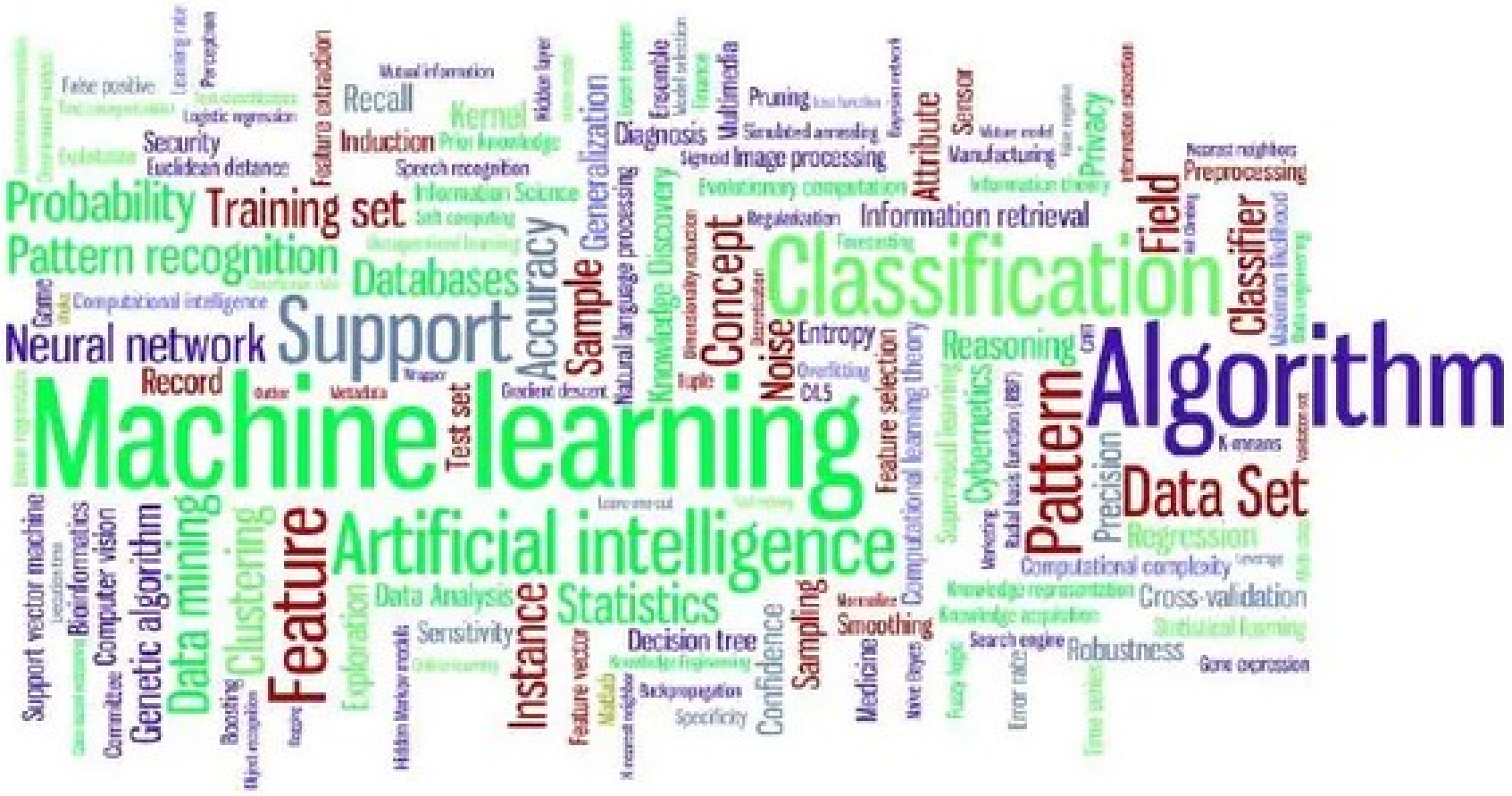
Text Books

Tom Mitchell. Machine Learning. McGraw Hill; 2017

Reference(s)

Christopher M Bishop. Pattern Recognition and Machine Learning. Springer 2010
Richard O. Duda, Peter E. Hart, David G. Stork. Pattern Classification. Wiley, Second Edition; 2007
Kevin P. Murphey. Machine Learning, a probabilistic perspective. The MIT Press Cambridge, Massachusetts, 2012.

Introduction to Machine Learning



Machine Learning

- **Herbert Alexander Simon**: "Learning is any process by which a system improves performance from experience."
- "Machine Learning is concerned with computer programs that automatically improve their performance through experience. "

Herbert Simon
Turing Award 1975
Nobel Prize in Economics 1978

Why now?

- Flood of available data (especially with the advent of the Internet)
- Increasing computational power
- Growing progress in available algorithms and theory developed by researchers
- Increasing support from industries

ML Applications

The concept of learning in a ML system

- Learning = Improving with experience at some task
  - Improve over task *T*,
  - With respect to performance measure, *P*
  - Based on experience, *E*.

Motivating Example
Learning to Filter Spam

Example: Spam Filtering

Spam - is all email the user does not want to receive and has not asked to receive

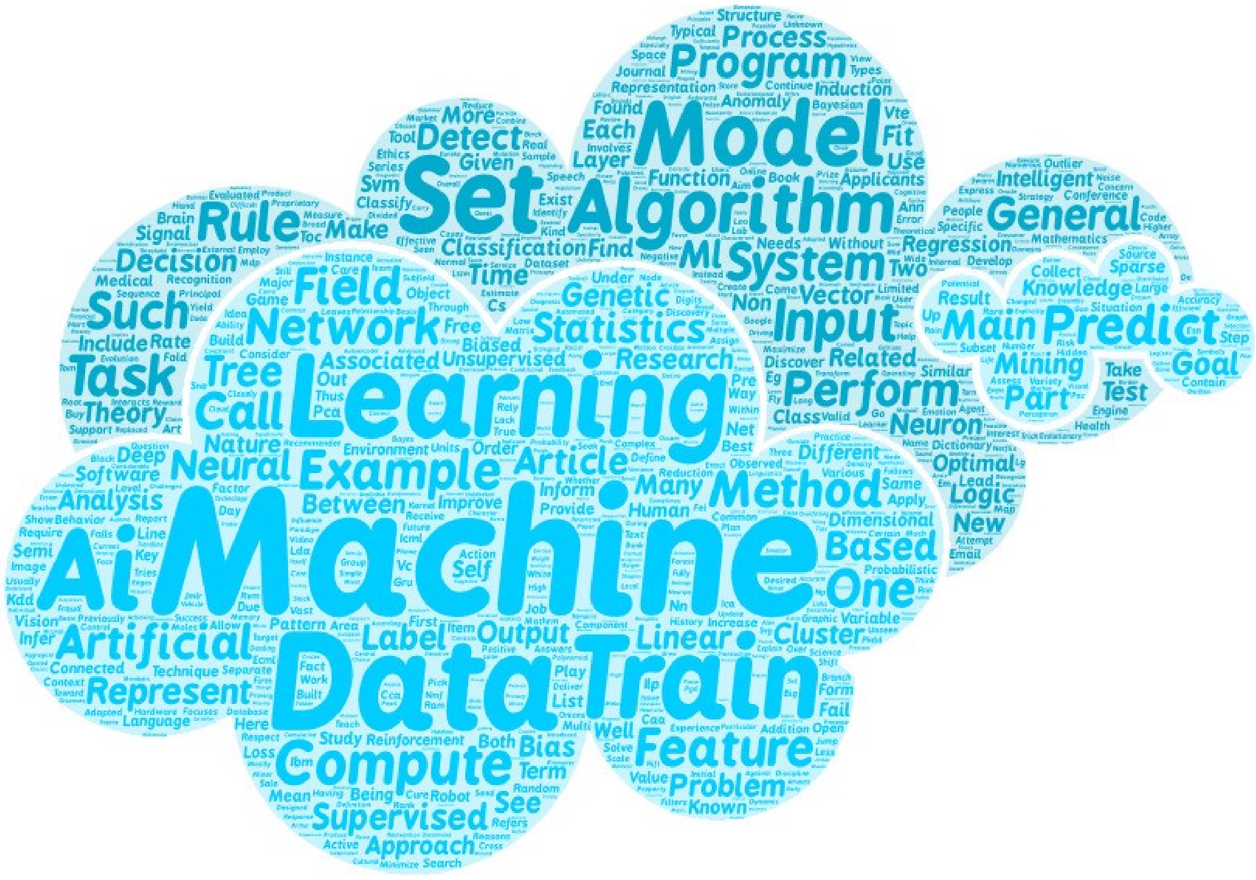
T: Identify Spam Emails

P:

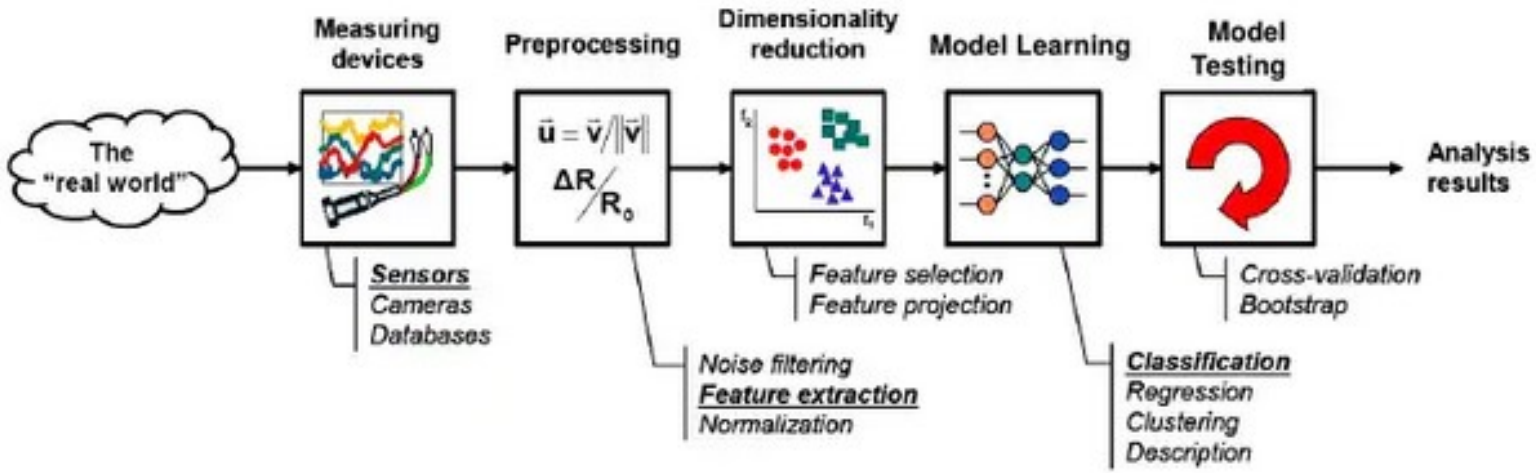
% of spam emails that were filtered

% of ham/ (non-spam) emails that were incorrectly filtered-out

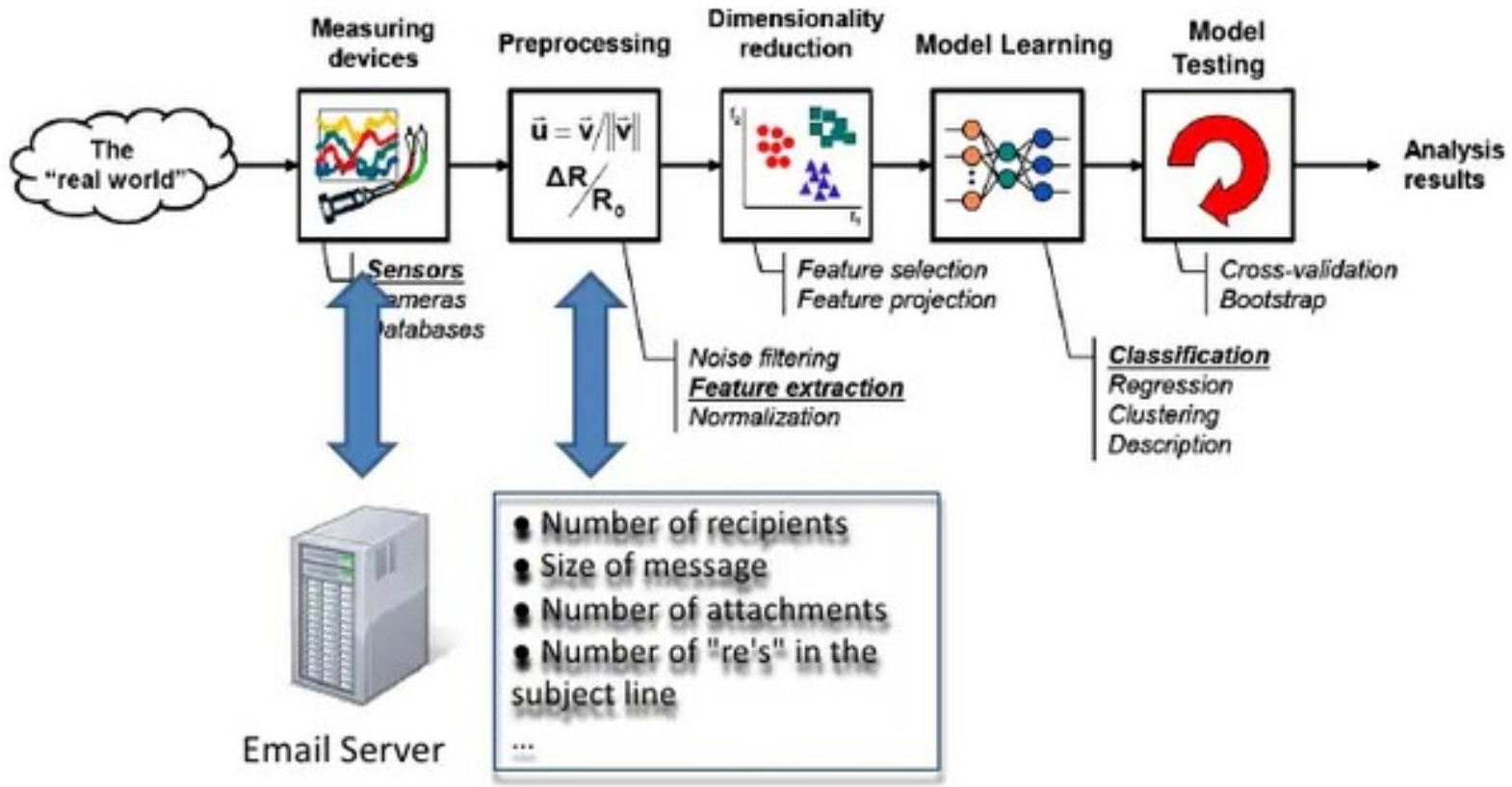
E: a database of emails that were labelled by users


Machine
Data
Train
Learning
Model
Set
Algorithm
Ai
Artificial
Compute
Feature
Method
Input
Predict
System
Program
Example
Network
Statistics
Field
Task
Such
Rule
Decision
Neural
Supervised
Problem
Perform
Main
General
Label
Output
Linear
Cluster
Represent
Approach
Process
Detect
Genetic
Research
Associated
Unsupervised
Classification
Regression
Knowledge
Mining
Goal
Test
Part
Based
One
Logic
New
Neuron
Optimal
Article
Tree
Call
Theory
Analysis
Software
Deep
Nature
Between
Reinforcement
Bias
Term
See

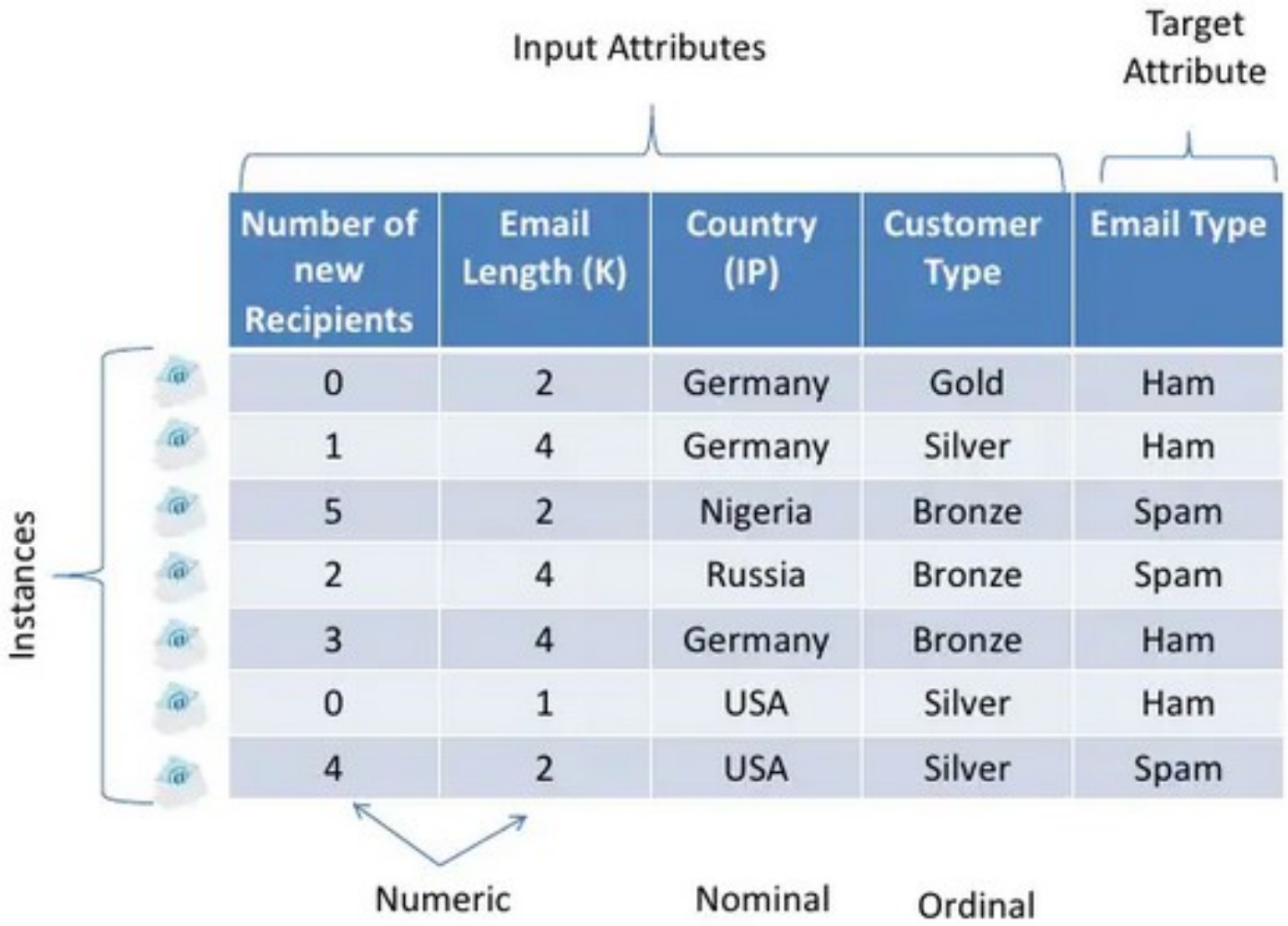
The Learning Process



The Learning Process in our Example



Data Set



Number of new Recipients	Email Length (K)	Country (IP)	Customer Type	Email Type
0	2	Germany	Gold	Ham
1	4	Germany	Silver	Ham
5	2	Nigeria	Bronze	Spam
2	4	Russia	Bronze	Spam
3	4	Germany	Bronze	Ham
0	1	USA	Silver	Ham
4	2	USA	Silver	Spam

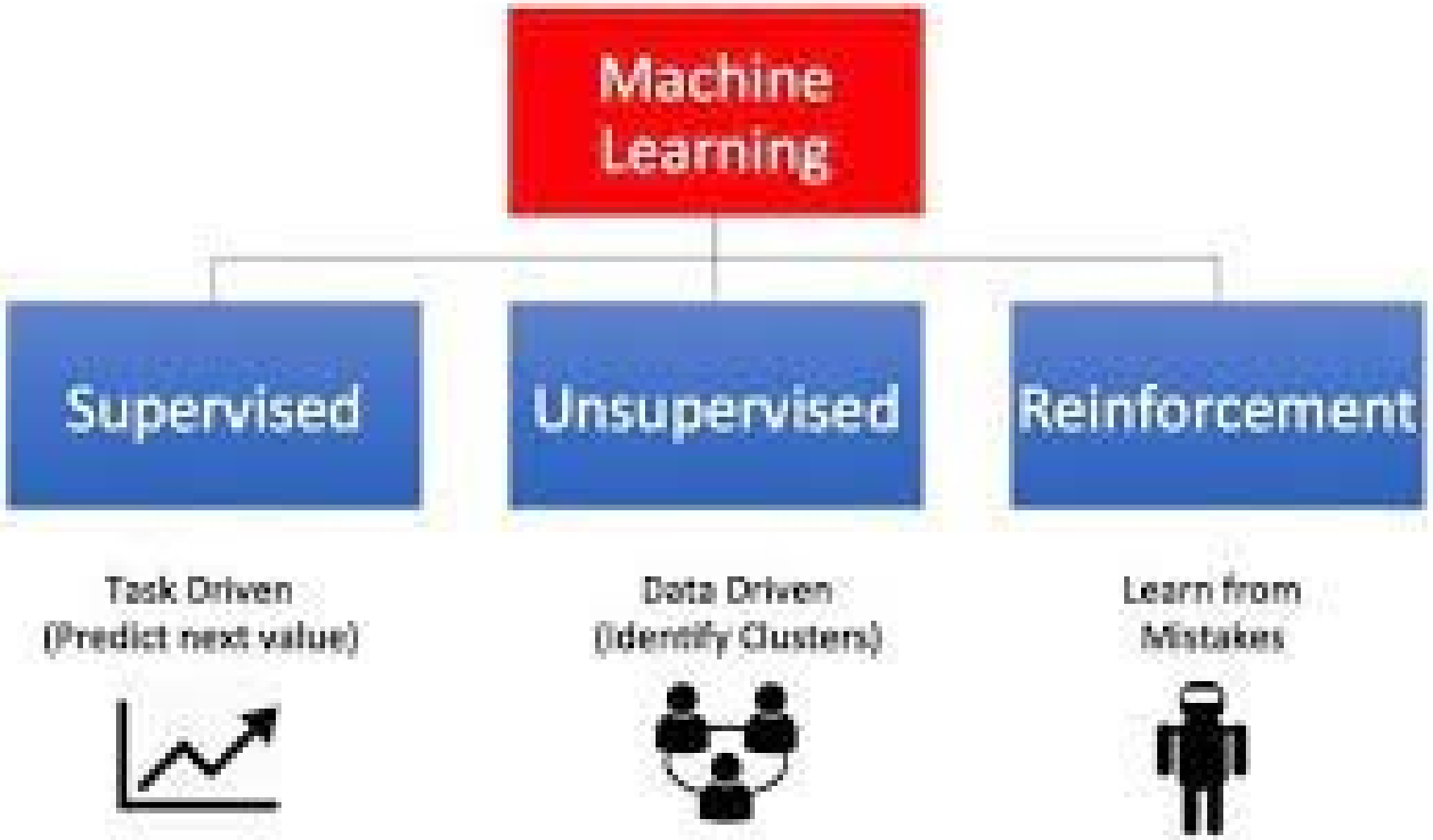
Face Recognition

Training examples of a person

Test images

AT&T Laboratories, Cambridge UK
http://www.uk.research.att.com/facedatabase.html

Types of Machine Learning

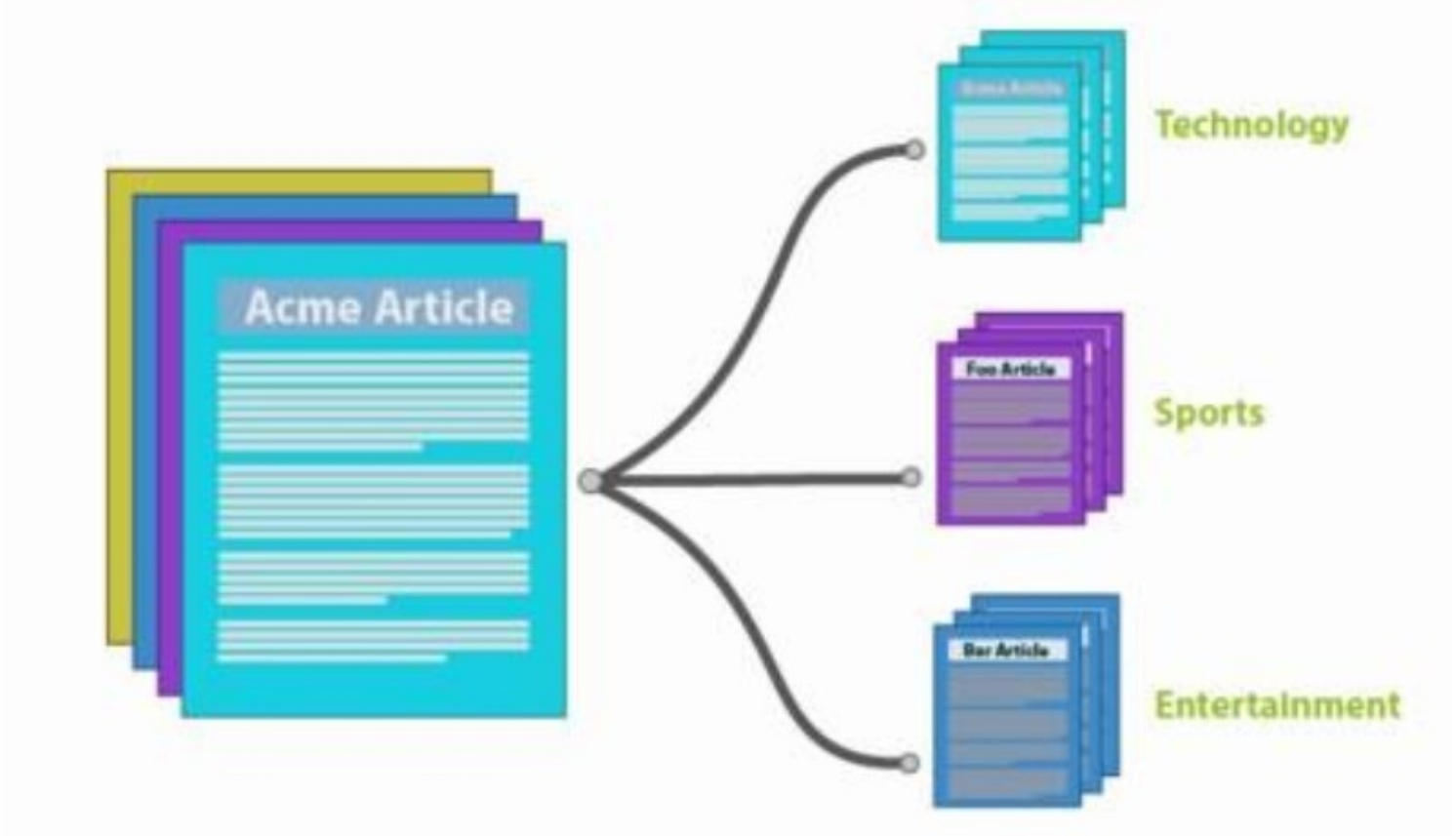


Predictive/supervised learning

- Goal is to learn a mapping from inputs x to outputs y, given a labeled set of input-output pairs
  - D is the training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$
  - N is the number of training exan[illegible]
  - x_i is the ith input vector of M dimensions in the training set
    - Each dimension represents a feature (or attribute).
- Each yi is a single value
- Classification
  - yi is categorical or nominal variable from some finite set
    - {male, female}
    - {spam, not-spam}
    - {cloudy, sunny, rainy}
    - {rose,jasmine,lotus,hibiscus}
- Regression
  - yi is numerical
    - Temperature
    - share price

Some applications of supervised learning

- Document classification



Some applications of supervised learning

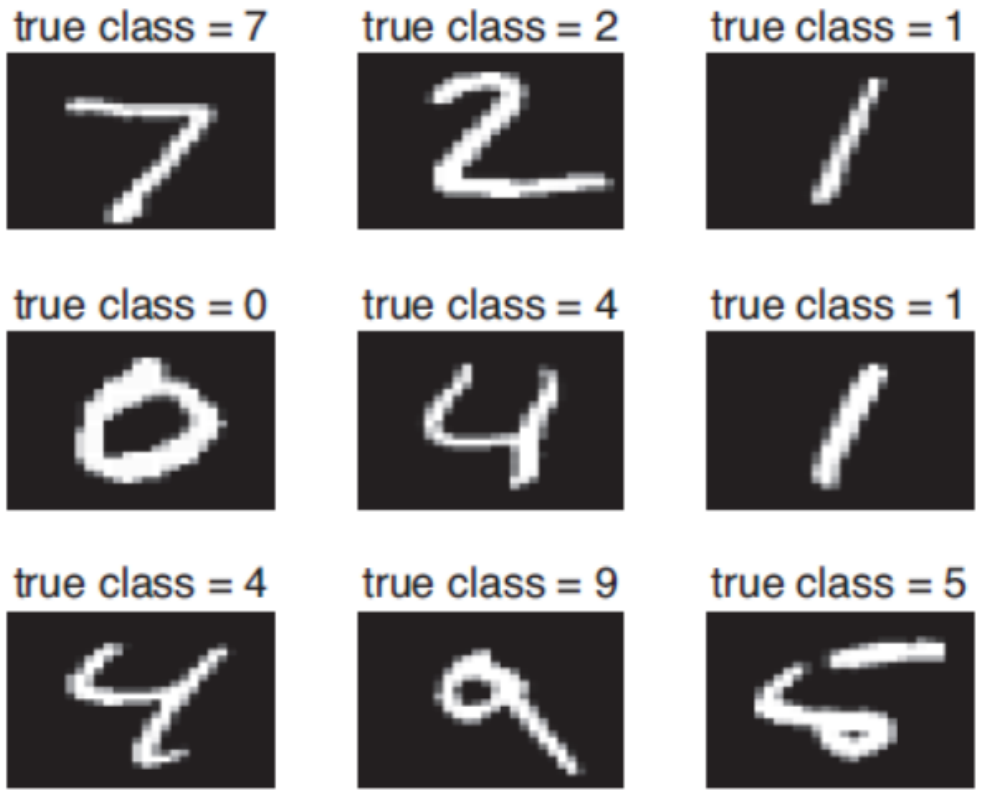
- Email spam filtering

Good email

spam

Some applications of supervised learning

- Handwriting Recognition



Some applications of supervised learning

- Face detection

(a) (b)

Figure 1.6 Example of face detection. (a) Input image (Murphy family, photo taken 5 August 2010). Used with kind permission of Bernard Diedrich of Sherwood Studios. (b) Output of classifier, which detected 5 faces at different poses. This was produced using the online demo at `http://demo.pittpatt.com/`. The

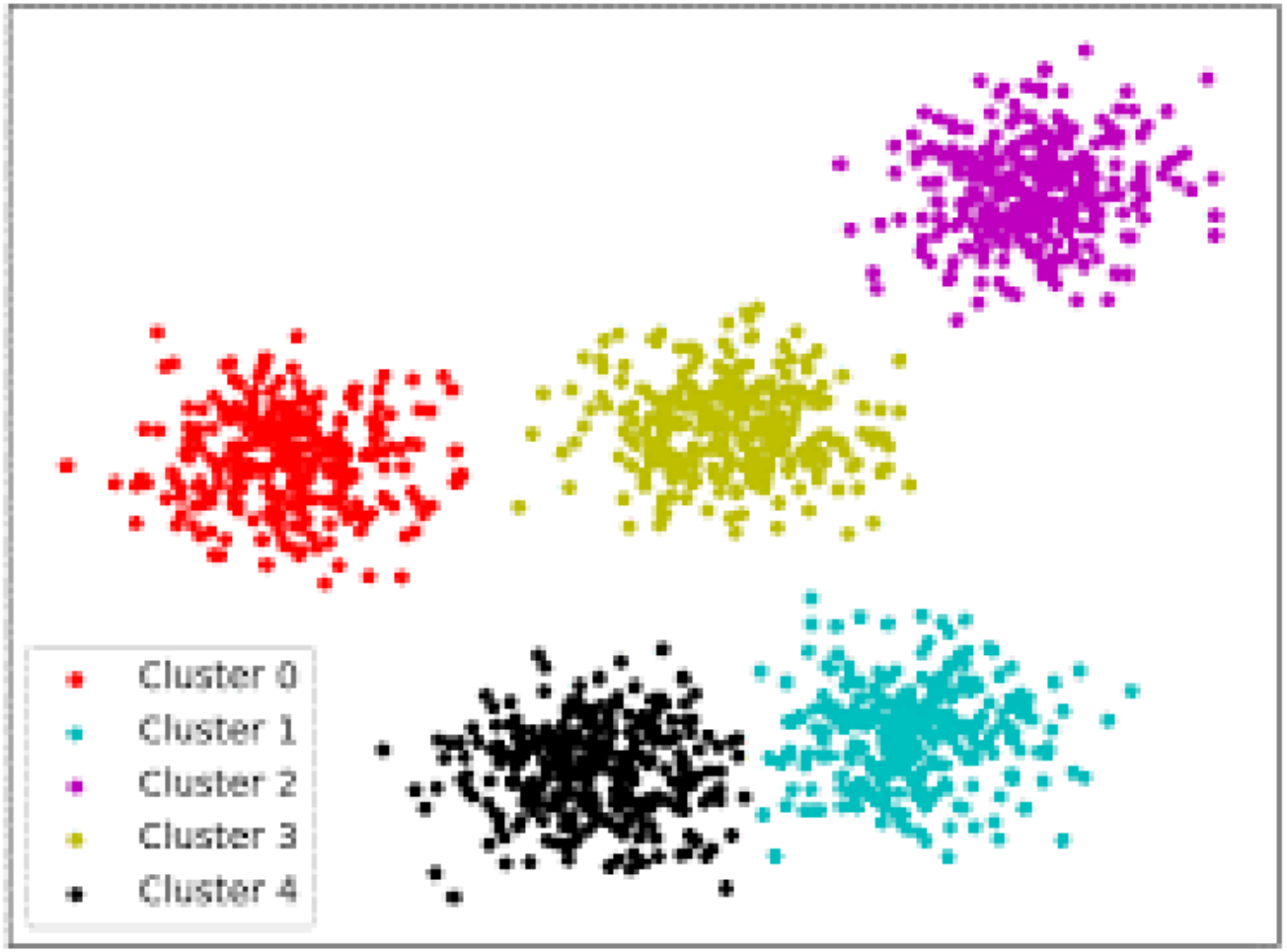
Regression - Examples

- Predict tomorrow's stock market price given current market conditions and other possible side information.
- Predict the age of a viewer watching a given video on YouTube.
- Predict the location in 3d space of a robot arm end effector, given control signals (torques) sent to its various motors.
- Predict the amount of prostate specific antigen (PSA) in the body as a function of a number of different clinical measurements.
- Predict the temperature at any location inside a building using weather data, time, door sensors, etc.

- Training set is not available
- Knowledge discovery
- Find “interesting patterns” in the data

- Clustering
- Recommender Systems
- Market basket analysis

Clustering



Find “interesting patterns”

- Recommender Systems
  - Collaborative filtering
- Market basket analysis

Reinforcement learning

- Somewhere between supervised and unsupervised
- Uses occasional reward or punishment signals
- Examples
  - Playing chess
  - Robot navigation

Machine Learning (19CSE305)

Linear Algebra

Dr. Peeta Basa Pati
Ms. Priyanka V
Department of Computer Science & Engineering,
Amrita School of Engineering, Bengaluru

Topics

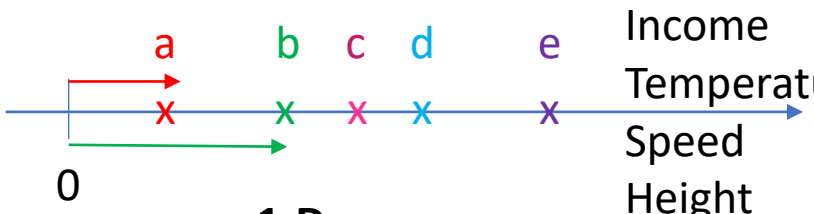
- Matrices
- Vectors
- Rank
- Invertibility

Question

A seller has 3 products to choose from and buy. We are provided with the purchase details (count of items selected and payment made) of 10 customers.

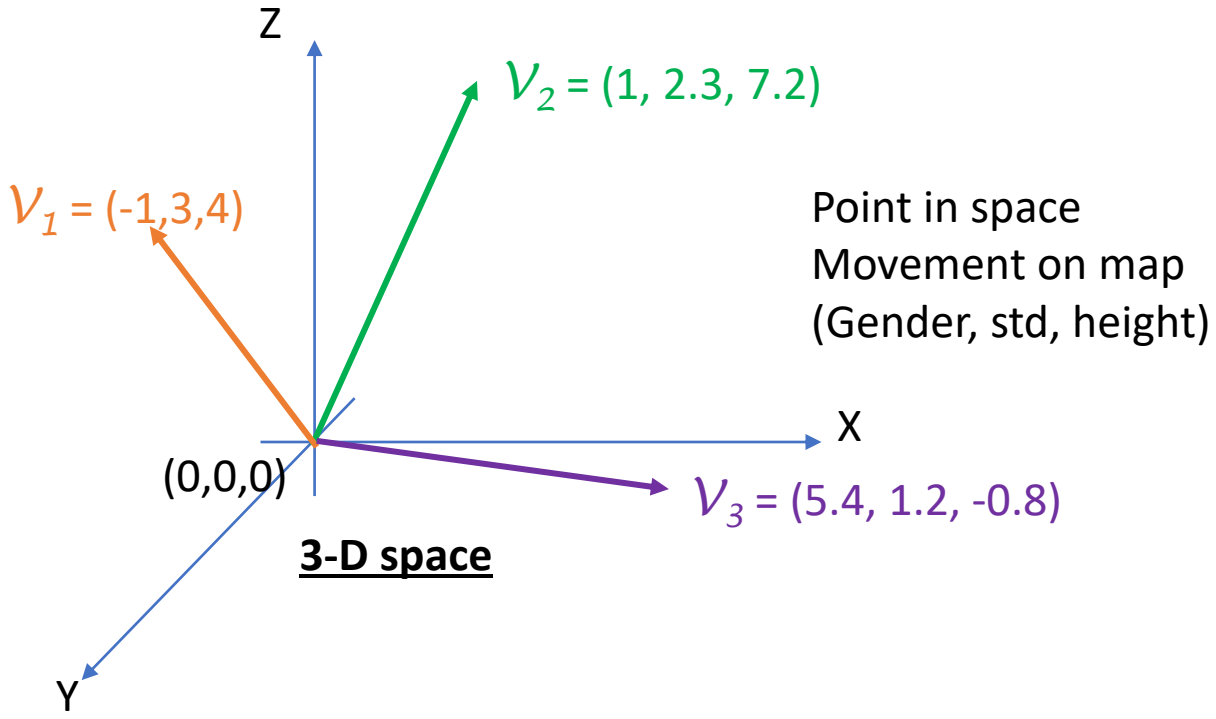
Find the cost of each product.

Vectors



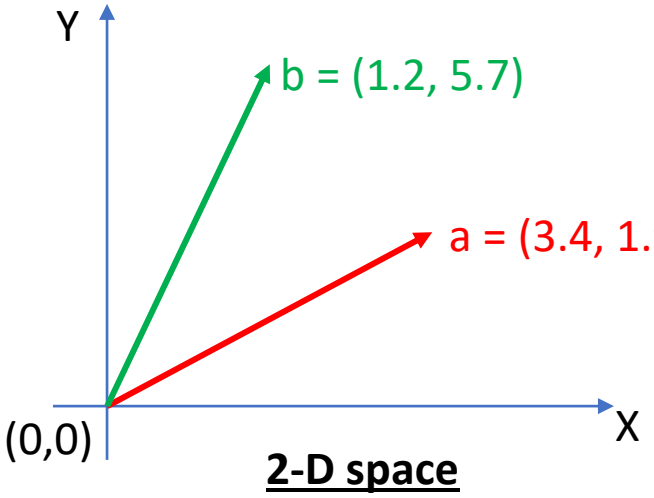
Income
Temperature
Speed
Height

1-D space



Point in space
Movement on map
(Gender, std, height)

3-D space



Map location
Velocity
Date & Time
Height & Weight

2-D space

n-D space

Mathematical representation ➔

$V = (x_1, x_2, \ldots, x_n)$

Vectors

{2}, {13.7}, {x}… {z} ← $\mathfrak{R}^1$

{3,1.5}, {12, 5.6}, … {x,y} ← $\mathfrak{R}^2$

{1,2,3}, {56, 32, 111}, … {x,y,z} ← $\mathfrak{R}^3$

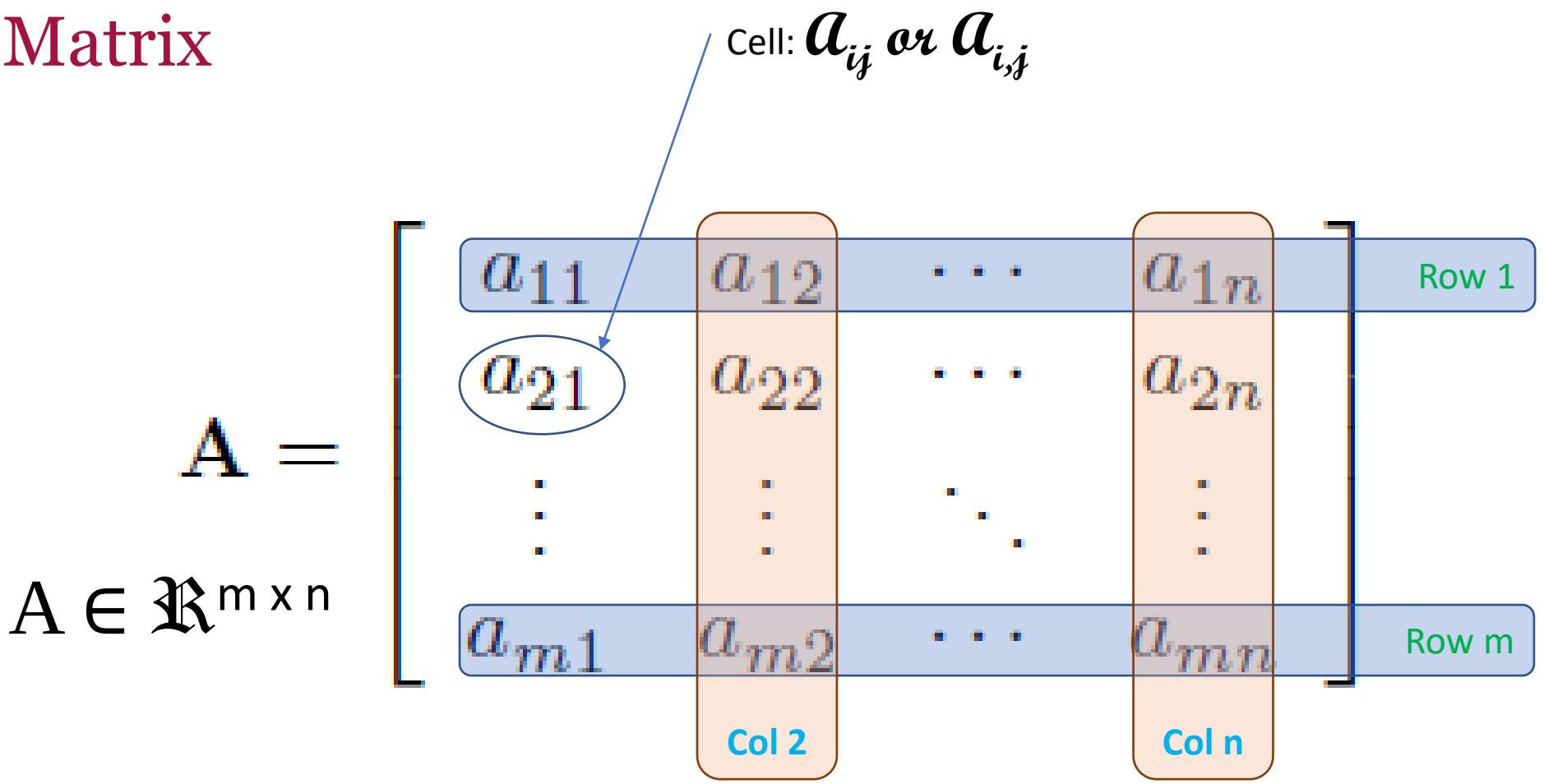
$\{p_1, p_2, p_3, \ldots, p_n\}, \{q_1, q_2, q_3, \ldots, q_n\}$ ← $\mathfrak{R}^n$

{1,2,3}, {56, 32, 3}, … {x,y,3} ← $\mathfrak{R}^?$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

A vector is usually written as a column.

Matrix



When m = n, the matrix is called a Square matrix.

Matrix Addition

2 matrices A & B can be added only when:

$$A, B \in \mathbb{R}^{m \times n}$$

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 7 & 9 \end{bmatrix}$$

$$B = \begin{bmatrix} 5 & 9 & 2 \\ 4 & 3 & 6 \end{bmatrix}$$

$$A + B = D \in \mathbb{R}^{m \times n}$$

$$A + B = D = \begin{bmatrix} 1+5 & 2+9 & 4+2 \\ 3+4 & 7+3 & 9+6 \end{bmatrix}$$

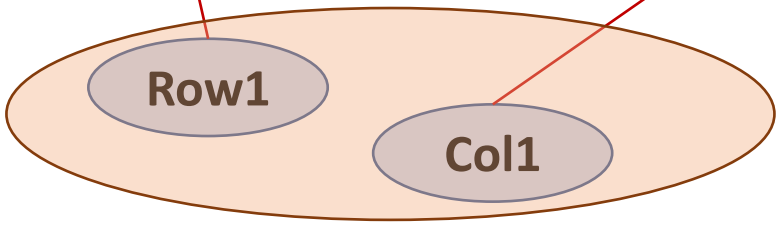
Matrix Multiplication

$$A \in \mathbb{R}^{m \times n}$$

$$C \in \mathbb{R}^{p \times q}$$

- A & C can be multiplied only when n = p
- The sequence of multiplication matters

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 7 & 9 \end{bmatrix} \quad C = \begin{bmatrix} 5 & 2 \\ 3 & 7 \\ 1 & 9 \end{bmatrix}$$



$$AC = D \in \mathbb{R}^{m \times q}$$

- Pre-multiplication / post-multiplication
- In this example, CA is invalid
- Only when A & C are square matrices of same size (m x m), both pre and post-multiplication possible

$$D_{1,1} = A_{1,1} \times C_{1,1} + A_{1,2} \times C_{2,1} + A_{1,3} \times C_{3,1}$$

$$AC = D = \begin{bmatrix} 15 & 52 \\ 3 & 7 \end{bmatrix}$$

Matrix Transpose

1. $A^T \in \mathbb{R}^{n \times m}$

2. $(\mathbf{A}^\mathsf{T})_{ij} = A_{ji}$

3. $A = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 7 & 9 \end{bmatrix}$ $A^T = \begin{bmatrix} 1 & 3 \\ 2 & 7 \\ 4 & 9 \end{bmatrix}$

4. $(\mathbf{A}^\mathsf{T})^\mathsf{T} = \mathbf{A}$

   $(\mathbf{AB})^\mathsf{T} = \mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T}$

   $(\mathbf{A} + \mathbf{B})^\mathsf{T} = \mathbf{A}^\mathsf{T} + \mathbf{B}^\mathsf{T}$

5. $A = A^T \Rightarrow A$ is symmetric

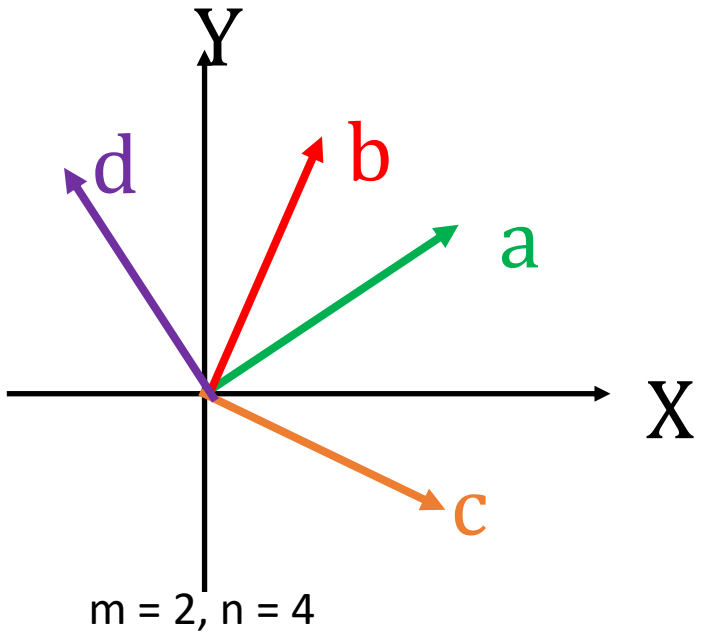
Rank of a Matrix

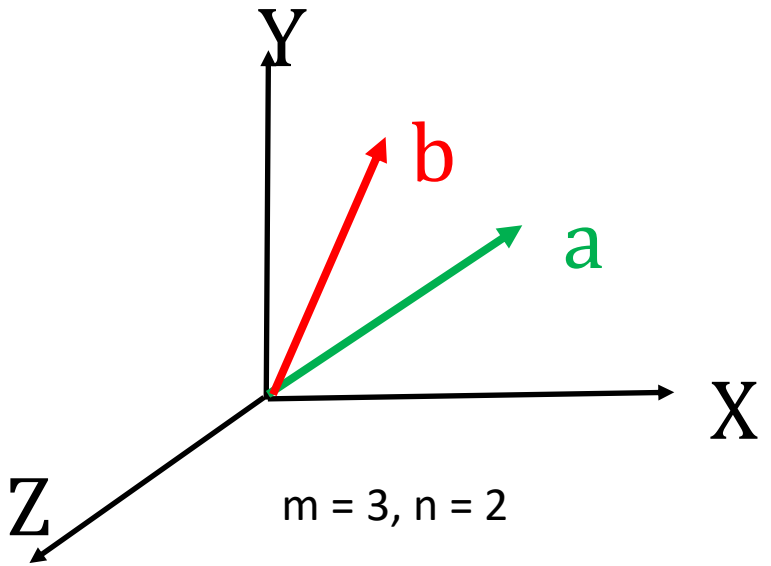
- Rank of a matrix is the number of independent vectors present in the matrix
- Its also the number of non-zero rows present in a row-echelon matrix

- It indicates the dimensionality of the space spanned by the vectors in the matrix
- When the rank(A) < m; the vectors present don't span the $\mathbb{R}^m$. This indicates the vectors cover a subspace.
- rank(A) < m, the data set is a good candidate for dimensionality reduction (to be covered later).

$$A \in \mathbb{R}^{m \times n}$$

$$\text{rank}(A) \leq \min(m,n)$$





Matrix Determinant

Determinants

$$\det A = |A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

(* $\mathbb{R}^{2\times 2}$ case)

$$\det(A) = |A| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

What happens to determinant when rank(A) < m?
What is a singular matrix?

Matrix Inversion

Inverse of a number is one when multiplied to the number, the product becomes 1. It's also called reciprocal.
n x (1/n) = 1

Similarly, for a square matrix A:

$$A^{-1}A = A\,A^{-1} = I$$

$$I_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$A^{-1} = \frac{1}{det\,(A)}\,adj\,(A)$$

Linear Equations & Matrices

$$x + 2y + 4z = 33$$
$$4x + 3y - z = 29$$
$$-x - y + 2z = -2$$

$$\begin{bmatrix} 1 & 2 & 4 \\ 4 & 3 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ -1 \end{bmatrix}$$

$$A \quad X \quad C$$

$$AX = C$$

$$\left[\begin{array}{ccc|c} 1 & 2 & 4 & 33 \\ 4 & 3 & -1 & 29 \\ -1 & -1 & 2 & -2 \end{array}\right]$$ Row 1, Row 2, Row 3

- Row2 – 4 x Row1
- Row3 + Row1

$$\left[\begin{array}{ccc|c} 1 & 2 & 4 & 33 \\ 0 & -5 & -17 & -103 \\ 0 & 1 & 6 & 31 \end{array}\right]$$

- Row2 & Row3 swap
- Row3 + 5 x Row2

$$\left[\begin{array}{ccc|c} 1 & 2 & 4 & 33 \\ 0 & 1 & 6 & 31 \\ 0 & 0 & 13 & 52 \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} 1 & 2 & 4 & 33 \\ 0 & 1 & 6 & 31 \\ 0 & 0 & 1 & 4 \end{array}\right]$$

- Row3 / 13

Solution with Matrix Inversion

$$A^{-1}A = A\,A^{-1} = I$$

$$A^{-1} = \frac{1}{det\,(A)}\,adj\,(A)$$

$$x + 2y + 4z = 33$$
$$4x + 3y - z = 29$$
$$-x - y + 2z = -2$$

$$\begin{bmatrix} 1 & 2 & 4 \\ 4 & 3 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ -1 \end{bmatrix}$$

$$AX = C$$
$$(A^{-1}A)X = A^{-1}C$$
$$IX = X = A^{-1}C = \begin{bmatrix} 3 \\ 7 \\ 4 \end{bmatrix}$$

What happens to inverse when rank(A) < m?
What happens to inverse of a rectangular matrix?

Rectangular Matrix Inversion

$$\begin{pmatrix} 2 & 5 & 7 & 8 \\ 1 & 2 & 3 & 1 \\ 4 & 5 & 0 & 1 \end{pmatrix}$$

No inverse exists mathematically.
Engineers are smart... pseudo-inverse helps

Principal Component Analysis (PCA) is the approach.
This works on the principle of minimum error.

Singular Value Decomposition (SVD) is the algorithm.

Real life example:
Customers of a store can choose to purchase from 3 available products. We have the item count & payment data for 10 purchases made. Find the cost of each product.

Questions

- What happens when a matrix is transposed? Think from space point of view.
- What is the relationship between column and row ranks?
- What information does the rank convey about the matrix? Think from space span point of view.
- When the rank of a matrix is much lower than *min(m,n)*, how does it impact the design of our ML systems?

Thank you !!!!!

Machine Learning (19CSE305)

Statistics & Probability

Dr. Peeta Basa Pati
Ms. Priyanka V
Department of Computer Science & Engineering,
Amrita School of Engineering, Bengaluru

Agenda

- Recap of last session
- Variables
- Mean, median & mode
- Standard Deviation and variance
- Probability & MLE
- Conditional Probability
- Joint Probability

Linear Equations & Matrices

$$A^{-1}A = A\,A^{-1} = I$$

$$A^{-1} = \frac{1}{det\,(A)}\,adj\,(A)$$

$$\begin{aligned} x + 2y + 4z &= 33 \\ 4x + 3y - z &= 29 \\ -x - y + 2z &= -2 \end{aligned}$$

$$\begin{bmatrix} 1 & 2 & 4 \\ 4 & 3 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ -1 \end{bmatrix}$$

$$\left[\begin{array}{ccc|c} 1 & 2 & 4 & 33 \\ 4 & 3 & -1 & 29 \\ -1 & -1 & 2 & -2 \end{array}\right]$$

Row 1

Row 2

Row 3

$$AX = C$$

$$(A^{-1}A)X = A^{-1}C$$

$$IX = X = A^{-1}C = \begin{bmatrix} 3 \\ 7 \\ 4 \end{bmatrix}$$

Rectangular Matrix Inversion

$$\begin{pmatrix} 2 & 5 & 7 & 8 \\ 1 & 2 & 3 & 1 \\ 4 & 5 & 0 & 1 \end{pmatrix}$$

No inverse exists mathematically.
Engineers are smart... pseudo-inverse helps

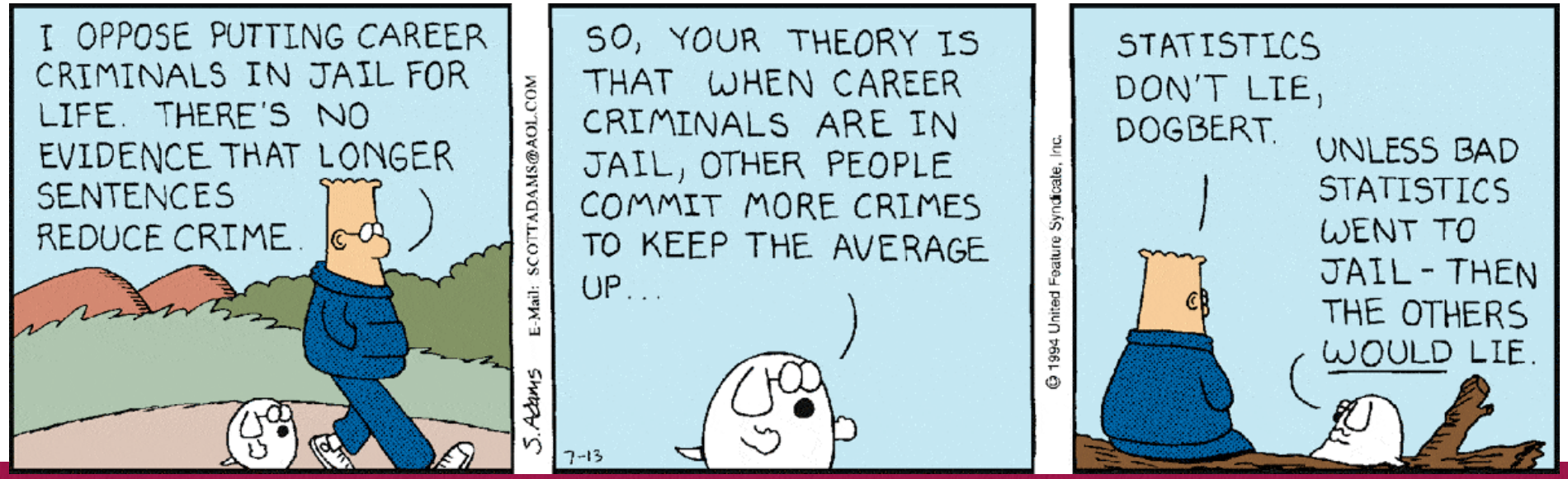
Principal Component Analysis (PCA) is the approach.
This works on the principle of minimum error.

Singular Value Decomposition (SVD) is the algorithm.

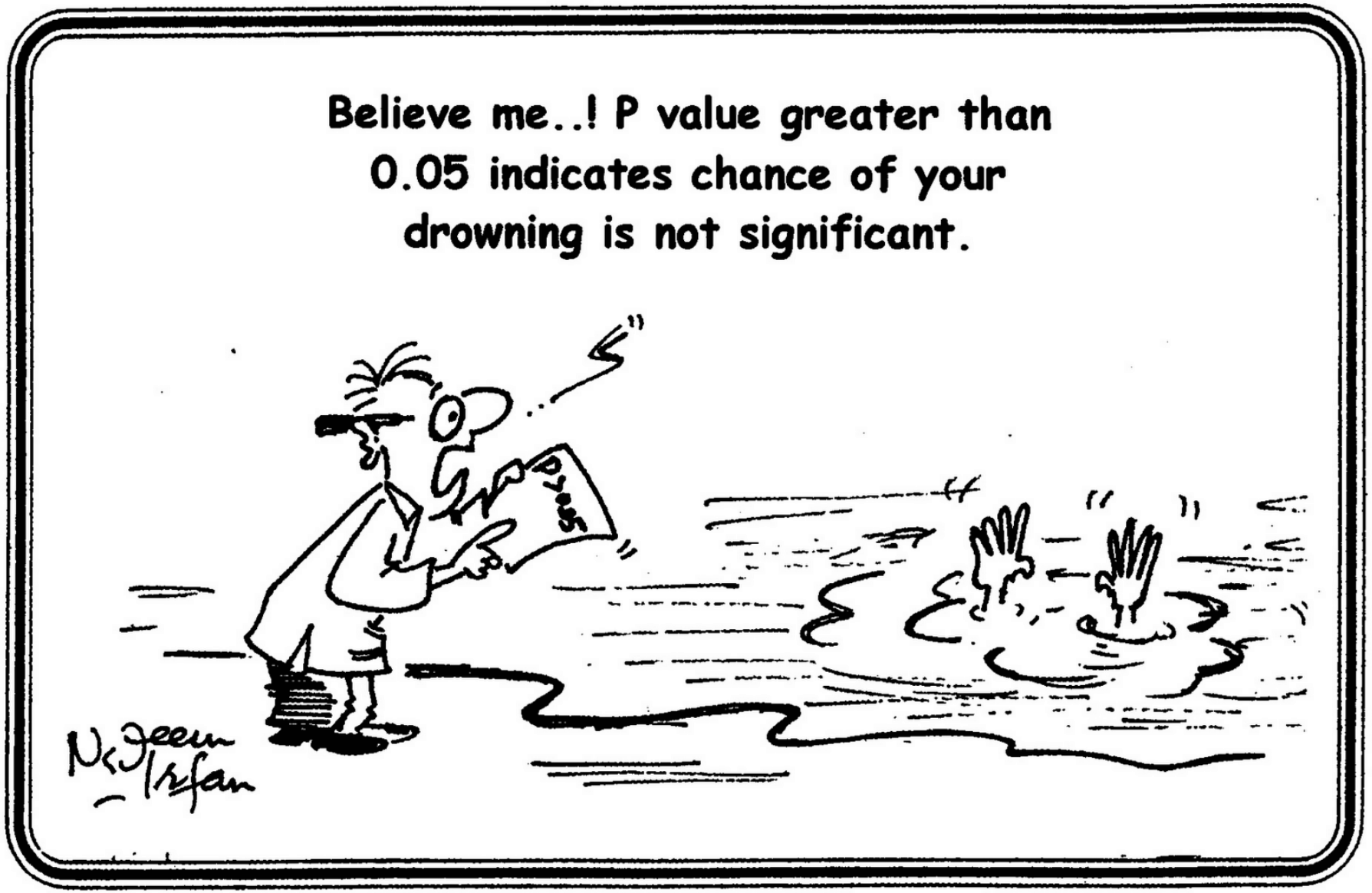
Real life example:
Customers of a store can choose to purchase from 3 available products. We have the item count & payment data for 10 purchases made. Find the cost of each product.

And today we deal with lots of data…

Fun with Data



Fun with Data



Probability vs Statistics

Statistics

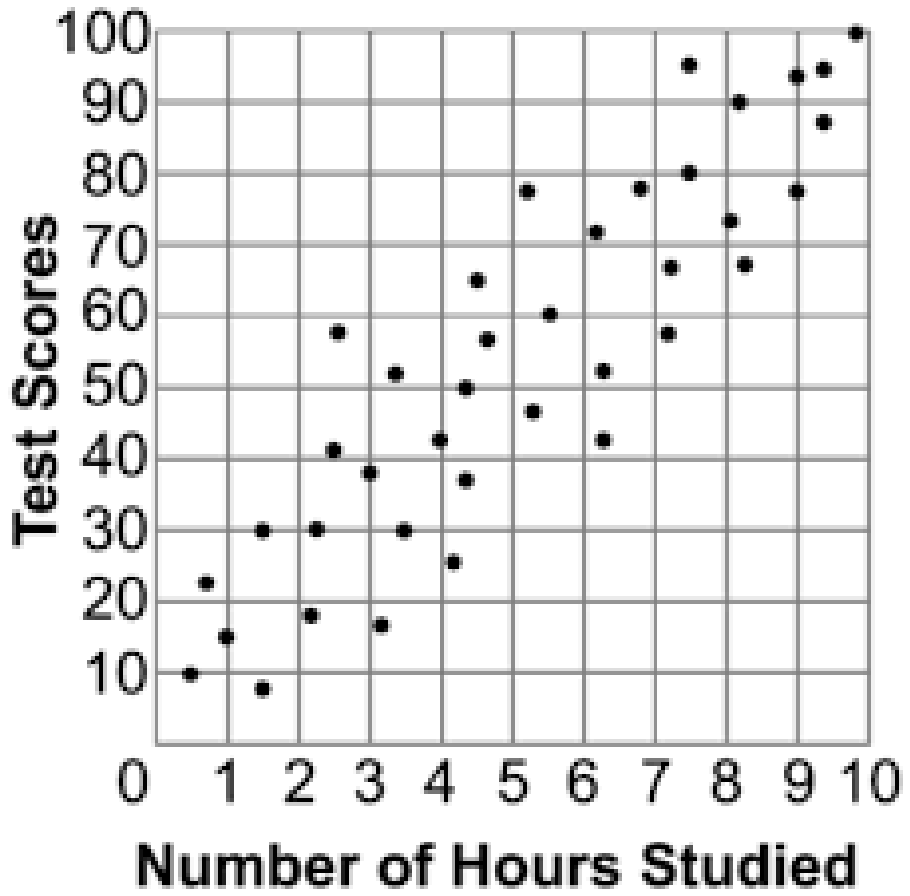
- Analyze and interpret available data
- Deals with events that has happened (past)
- Science of Data

Probability

- Science of prediction and chances (future)
- Distributions and implications
- Deals with mathematical formulas

Collection of Data

Student	Height (inches)
S_1	35
S_2	36
S_3	30
S_4	31
S_5	36
S_6	35
S_7	32
S_8	35
S_9	32
S_10	34
S_11	34
S_12	34
S_13	34
S_14	32
S_15	32
S_16	34
S_17	35
S_18	34
S_19	30
S_20	36
S_21	34
S_22	31



Mean / Average & Variance

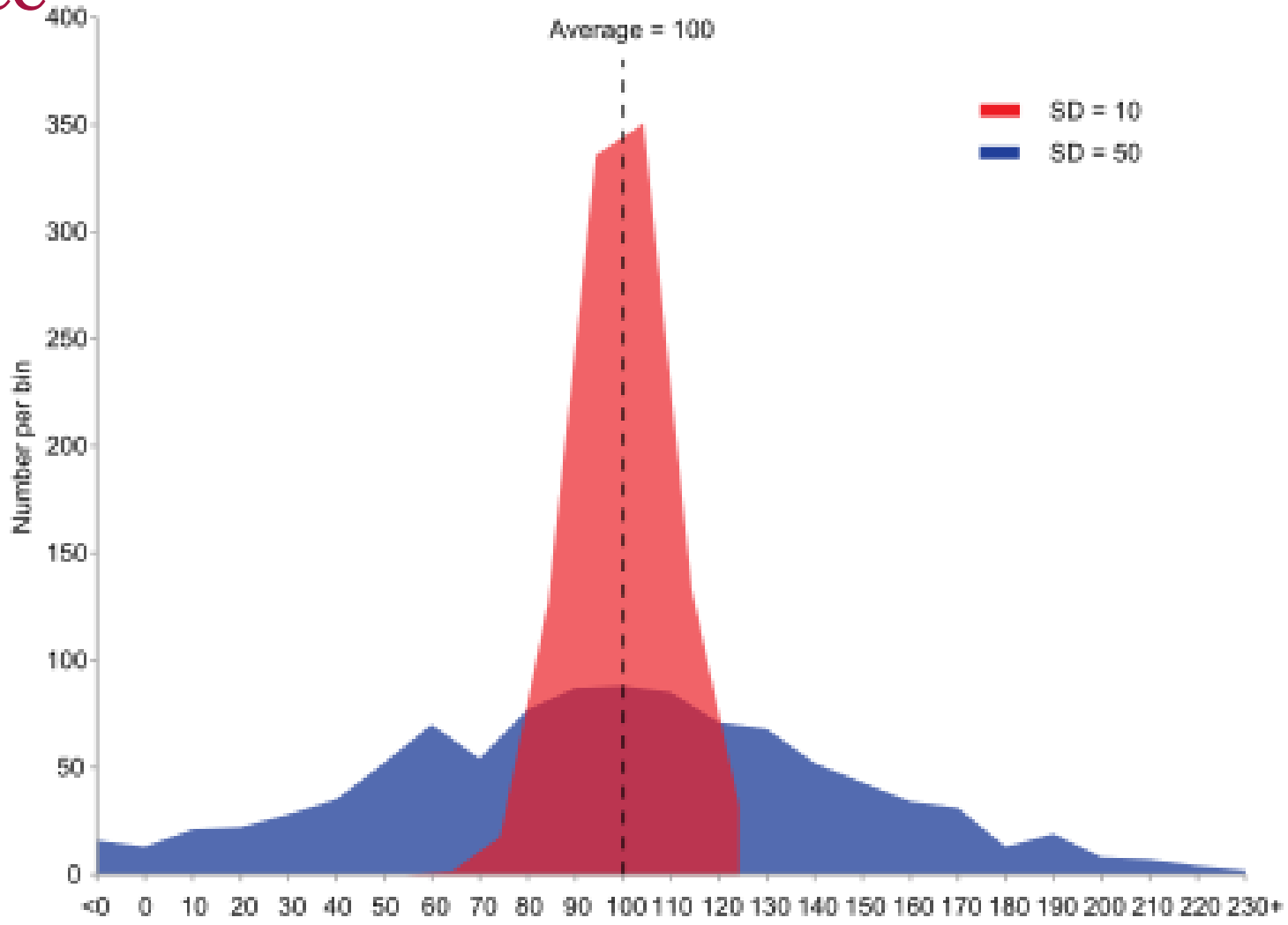
Mean or Average

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{\sum X}{N}$$

Variance

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Standard deviation (σ) is square root of variance.



Mean, Median, Mode

Student	Height (inches)
S_3	30
S_19	30
S_4	31
S_22	31
S_7	32
S_9	32
S_14	32
S_15	32
S_10	34
S_11	34
S_12	34
S_13	34
S_16	34
S_18	34
S_21	34
S_1	35
S_6	35
S_8	35
S_17	35
S_2	36
S_5	36
S_20	36

Mean → average of a data set

$$\text{Mean } (\mu) = \frac{sum\ of\ all\ terms}{count\ of\ terms}$$

μ = 33.45455

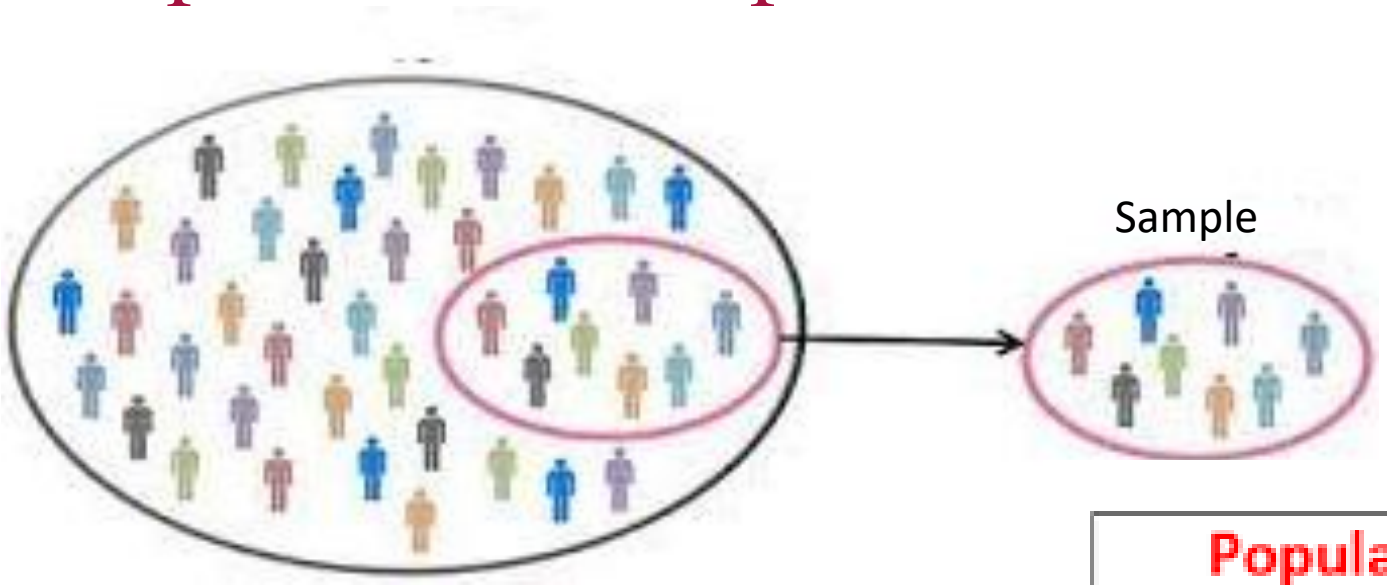
Median → middle value of the set of numbers.

Median = 34

Mode → most common number in a data set.

Mode = 34

Population & Sample



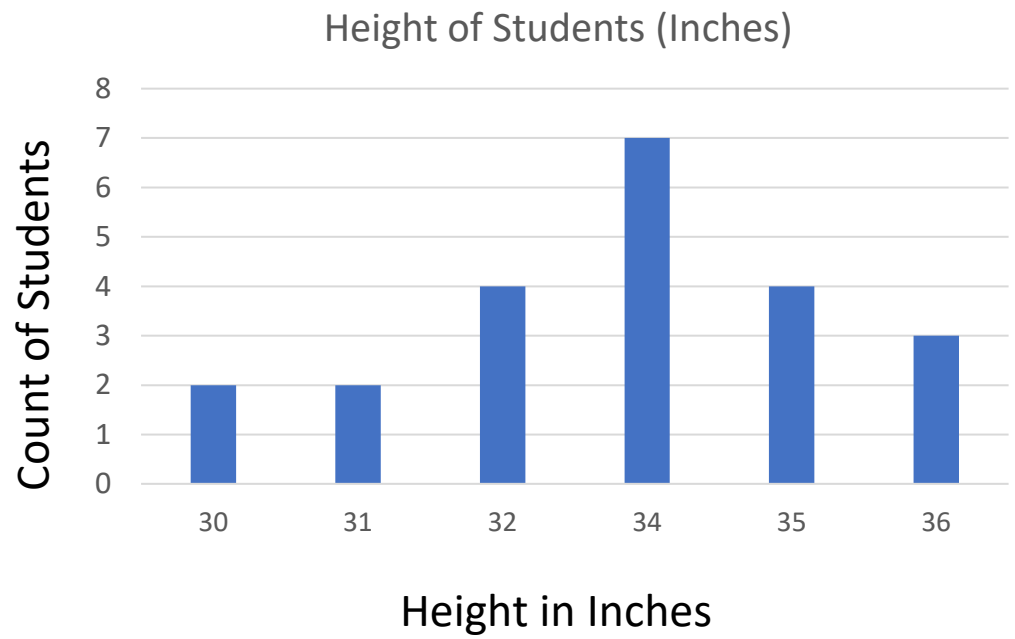
Population

- For blood CBC, would you draw few drops or drain the blood completely from a person?
- If you want to measure the average length of your hair, would you take a few hair samples or shave your head off?
- Average weekly hours spent in watching TV by Indians.

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

Histogram

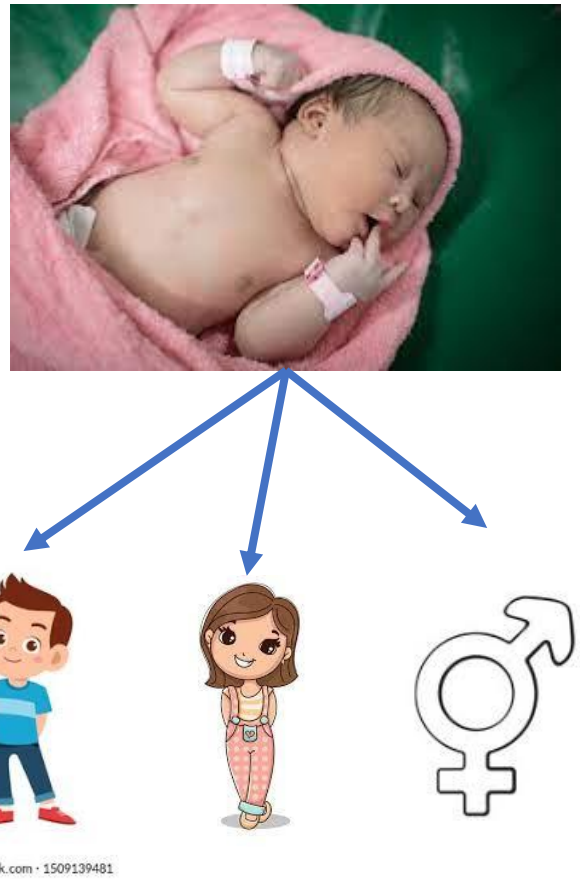
- Frequency distribution of data
- If continuous data, create buckets and find the membership in each bucket

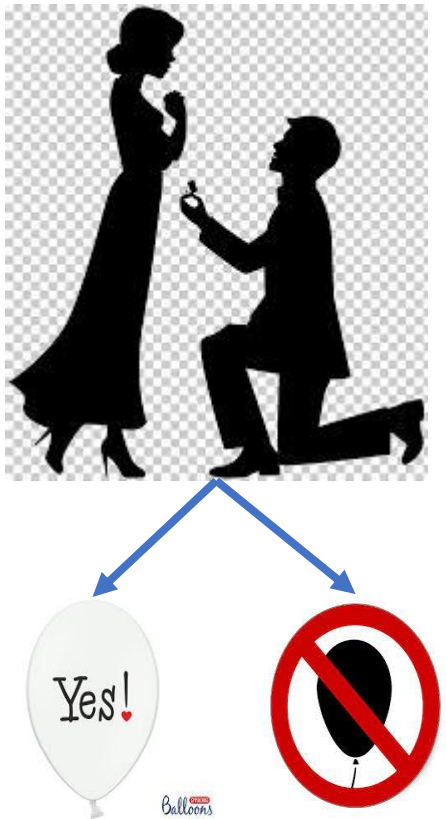


How shall we deal with continuous data?

Student	Height (inches)
S_3	30
S_19	30
S_4	31
S_22	31
S_7	32
S_9	32
S_14	32
S_15	32
S_10	34
S_11	34
S_12	34
S_13	34
S_16	34
S_18	34
S_21	34
S_1	35
S_6	35
S_8	35
S_17	35
S_2	36
S_5	36
S_20	36

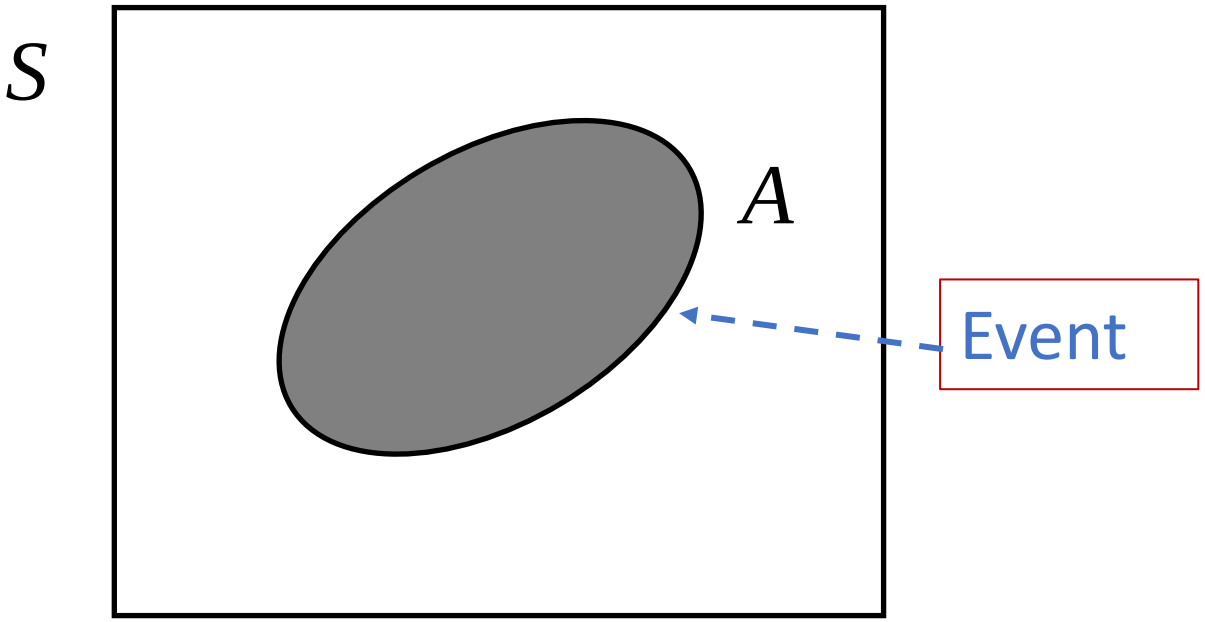
Observations and Events

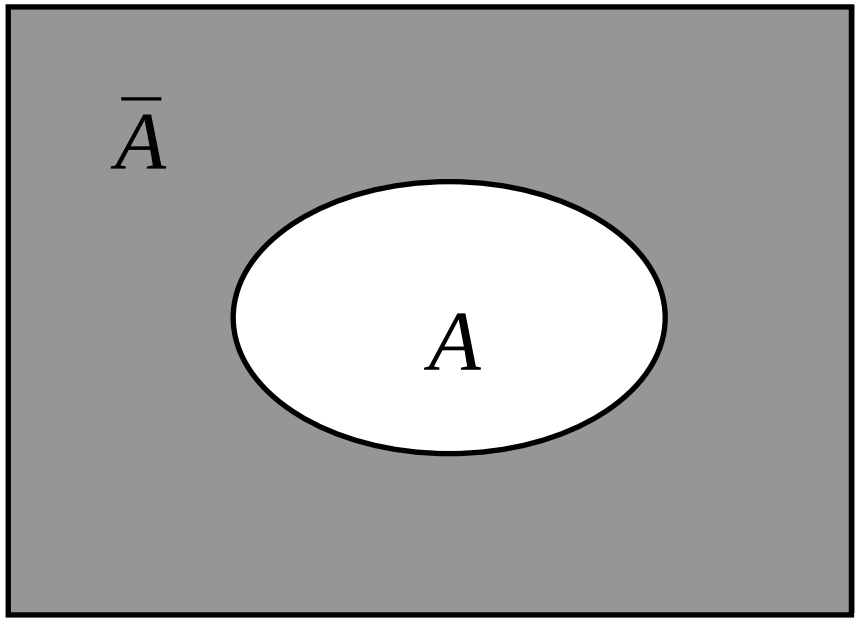




Venn Diagram (Sample Space & Event)

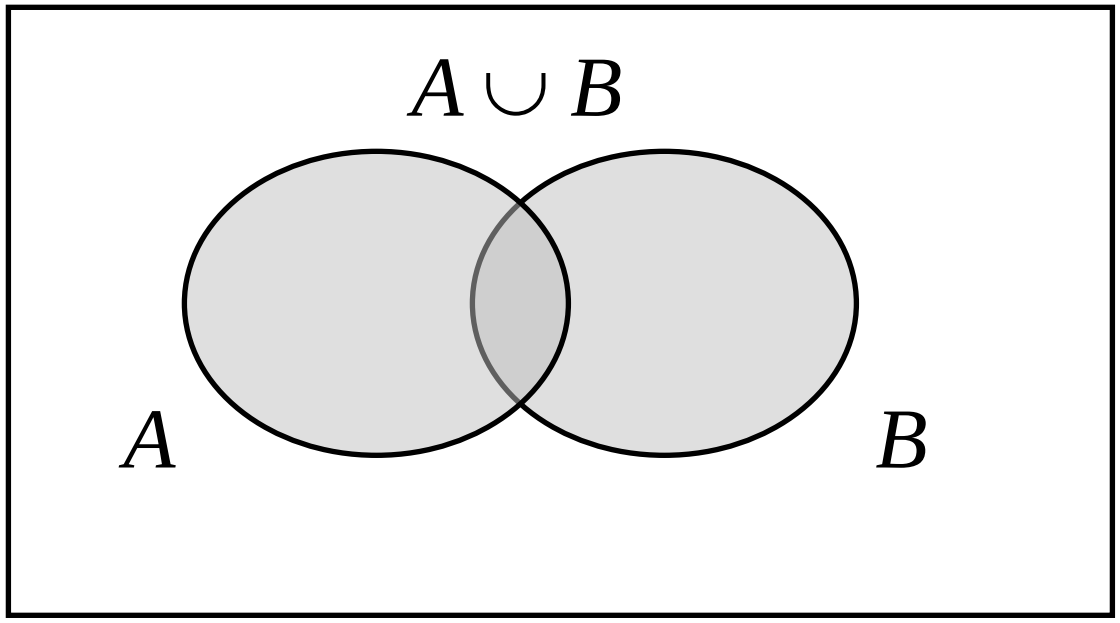
- All observations made form the sample space
- Event is the observation category
  - Baby is girl
  - Price of a stock increased
  - “YES” to marriage proposal

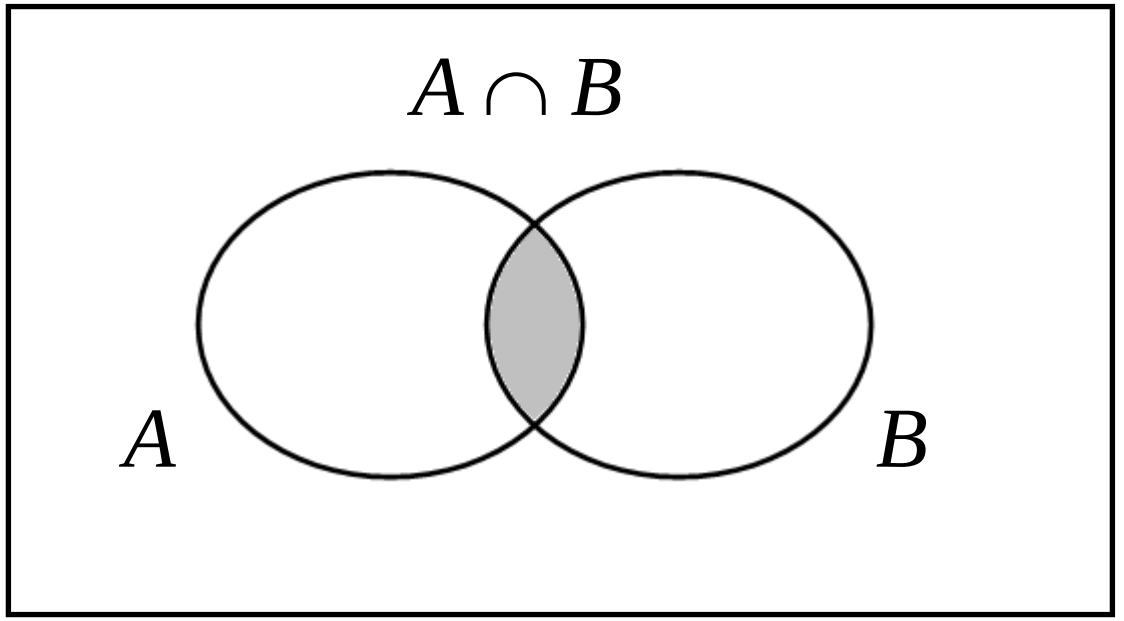




Venn diagram

Operations on Sets

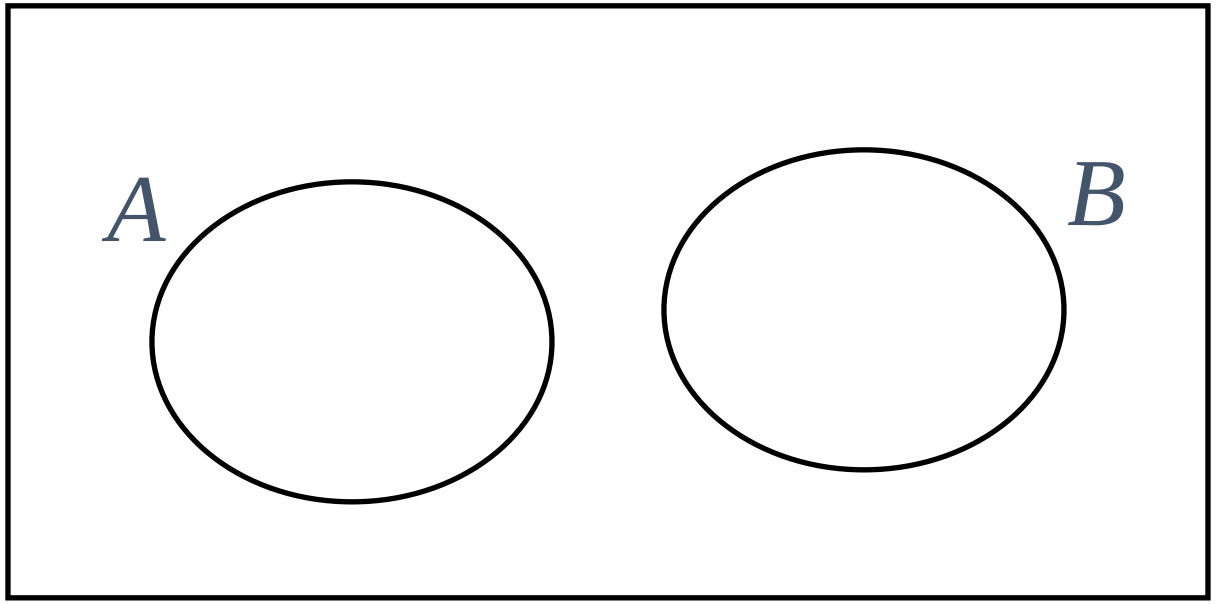




$A \cup B$ = {event belongs either to A **or** B}

$A \cap B$ = {event belongs both A **and** B}

Mutually Exclusive Sets



$$A \cap B = \phi$$

Null Set

Probability

500 newborn inspected

272

225

3

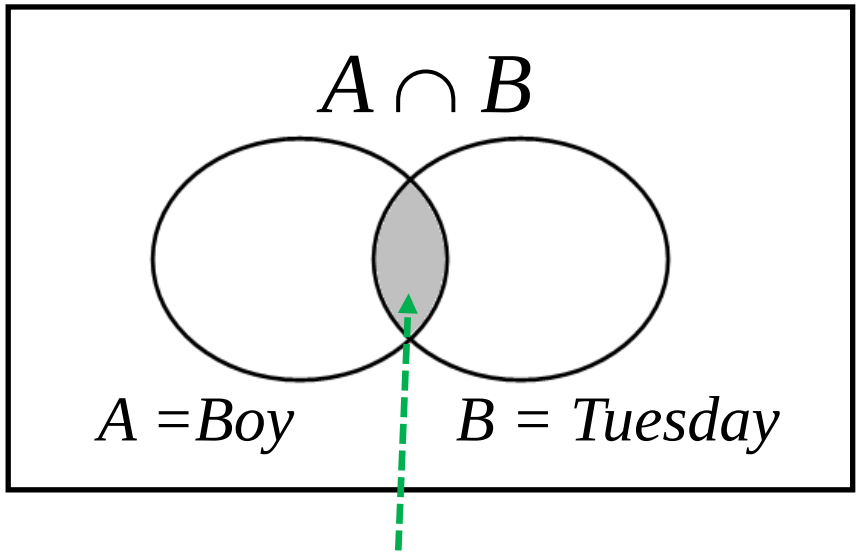
$$P[E] = \frac{n(E)}{n(S)} = \frac{n(E)}{N} = \frac{\text{no. of outcomes in } E}{\text{total no. of outcomes}}$$

P (Boy) = $^{272}/_{500}$ = 0.54

P (Girl) = $^{225}/_{500}$ = 0.45

P(Trans) = = $^{3}/_{500}$ = 0.01

Conditional Probability



What is the probability of a new born being a "***Boy***" given that today is "***Tuesday***"?

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

Conditional Probability

	Monday	Tuesday	Friday	Total	Prob
Boys	272	232	176	680	0.53
Girls	225	248	137	610	0.47
Trans	3	0	0	3	0.00
Total	500	480	313	1293	
Prob	0.39	0.37	0.24		

Total babies on all 3 days = 1293

Boys born on Tuesday = 232

$P(A \cap B) = {}^{232}/_{1293} = 0.18$

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

Prob of a new born being boy given that today is Tuesday

→ P(Boy | Tuesday)

$= {}^{0.18}/_{0.37} = \mathbf{0.49}$

Joint Probability

Joint probability of 2 events is said to be the probability of both events happening at the same time.

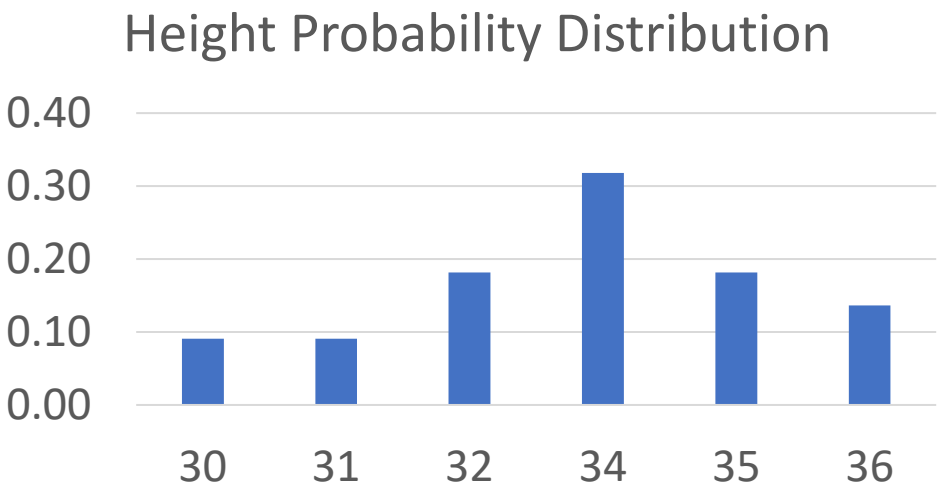
Ex: You randomly visit the hospital to inspect. The probability of the new born is boy and today is Tuesday is referred as joint probability.

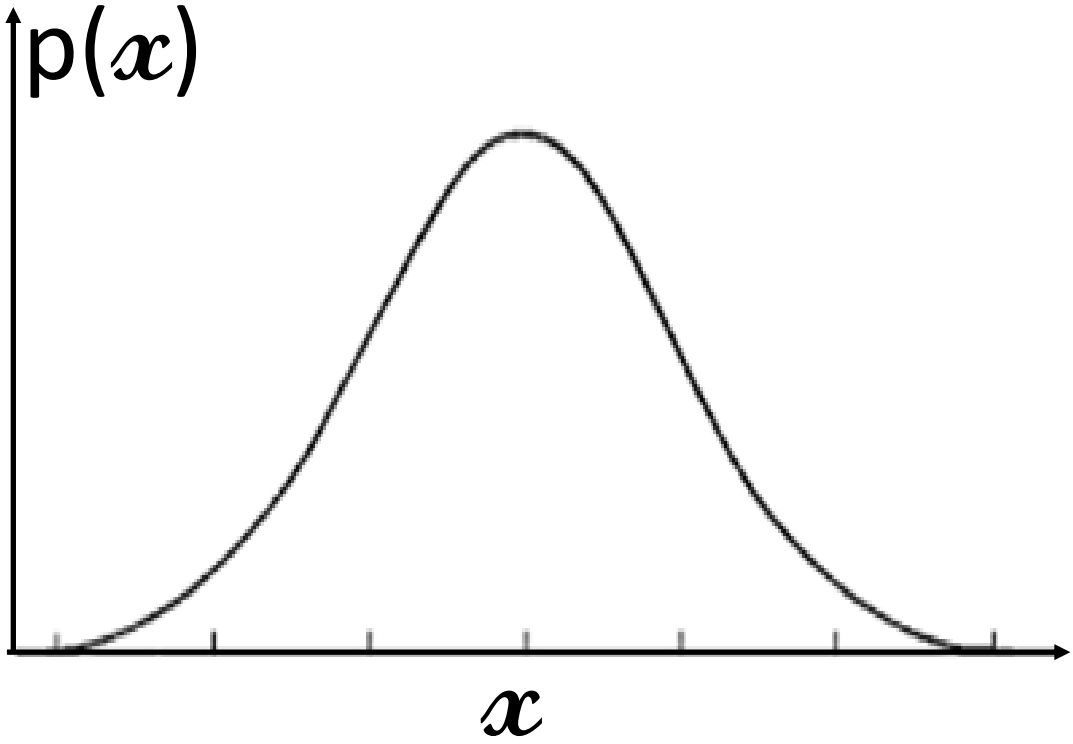
$$P(A \text{ and } B) = P(A|B)\,P(B) = P(A \cap B)$$

P(A and B) is also denoted as P(A,B)

Probabilities for discrete and continuous variables

Student	Height (inches)
S_3	30
S_19	30
S_4	31
S_22	31
S_7	32
S_9	32
S_14	32
S_15	32
S_10	34
S_11	34
S_12	34
S_13	34
S_16	34
S_18	34
S_21	34
S_1	35
S_6	35
S_8	35
S_17	35
S_2	36
S_5	36
S_20	36





Thank you !!!!!

Machine Learning (19CSE305)

Features, Distance & Similarity

AMRITA VISHWA VIDYAPEETHAM

DEEMED TO BE UNIVERSITY

Dr. Peeta Basa Pati
Ms. Priyanka V
Department of Computer Science & Engineering,
Amrita School of Engineering, Bengaluru

Topics

- Recap of covered sessions
- Features / attributes
- Types of Data & Characteristics
- Distances & Similarities

Linear Equations & Matrices

$$x + 2y + 4z = 33$$
$$4x + 3y - z = 29$$
$$-x - y + 2z = -2$$

$$A \quad X \quad = \quad C$$

$$\begin{bmatrix} 1 & 2 & 4 \\ 4 & 3 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ -1 \end{bmatrix}$$

$$A^{-1}A = A\,A^{-1} = I$$

$$A^{-1} = \frac{1}{det\,(A)}\,adj\,(A)$$

$$\begin{pmatrix} 2 & 5 & 7 & 8 \\ 1 & 2 & 3 & 1 \\ 4 & 5 & 0 & 1 \end{pmatrix}$$

No inverse exists mathematically for rectangular matrices. However, Singular Value Decomposition (SVD) algorithm (and such others) work on principle of error minimization and help us arrive at a pseudo-inverse. This pseudo-inverse meets most of day-to-day needs.

Statistics & Probability

Mean → average of a data set
Median → middle value of the set of numbers.
Mode → most common number in a data set.

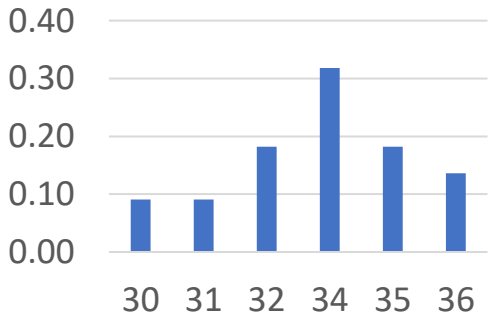
Variance & **Standard Deviation** & their relation

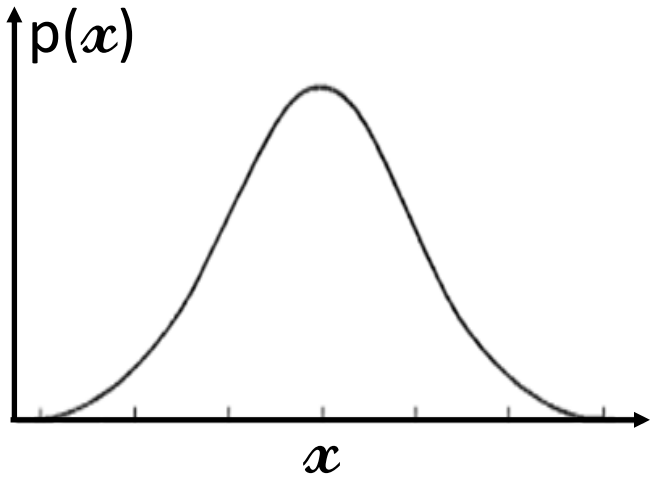
Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

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

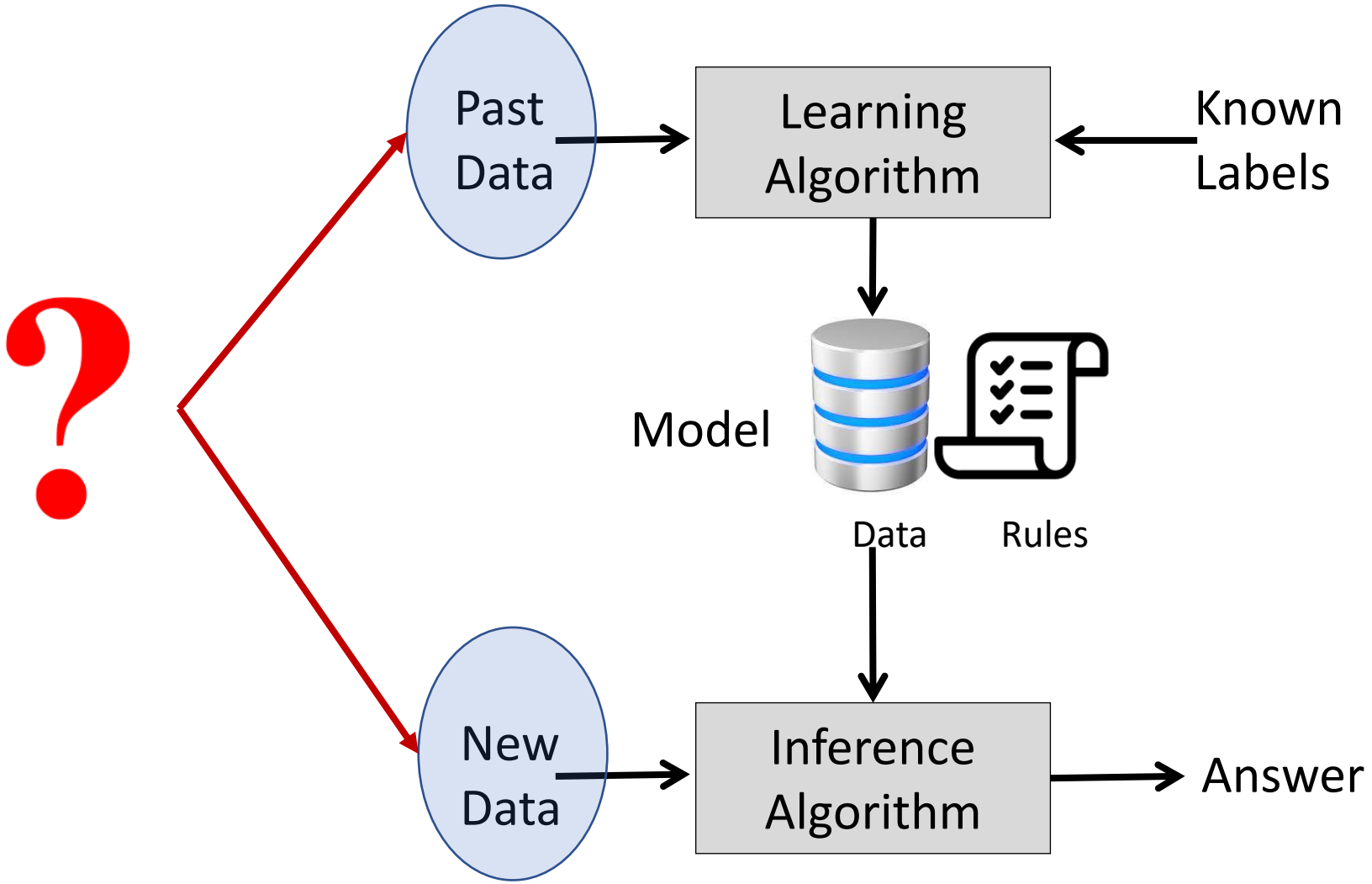
$$P(A|B) = P(A \cap B)/P(B)$$

$$P(A \text{ and } B) = P(A|B)\,P(B) = P(A \cap B)$$





Machine Learning System



Features

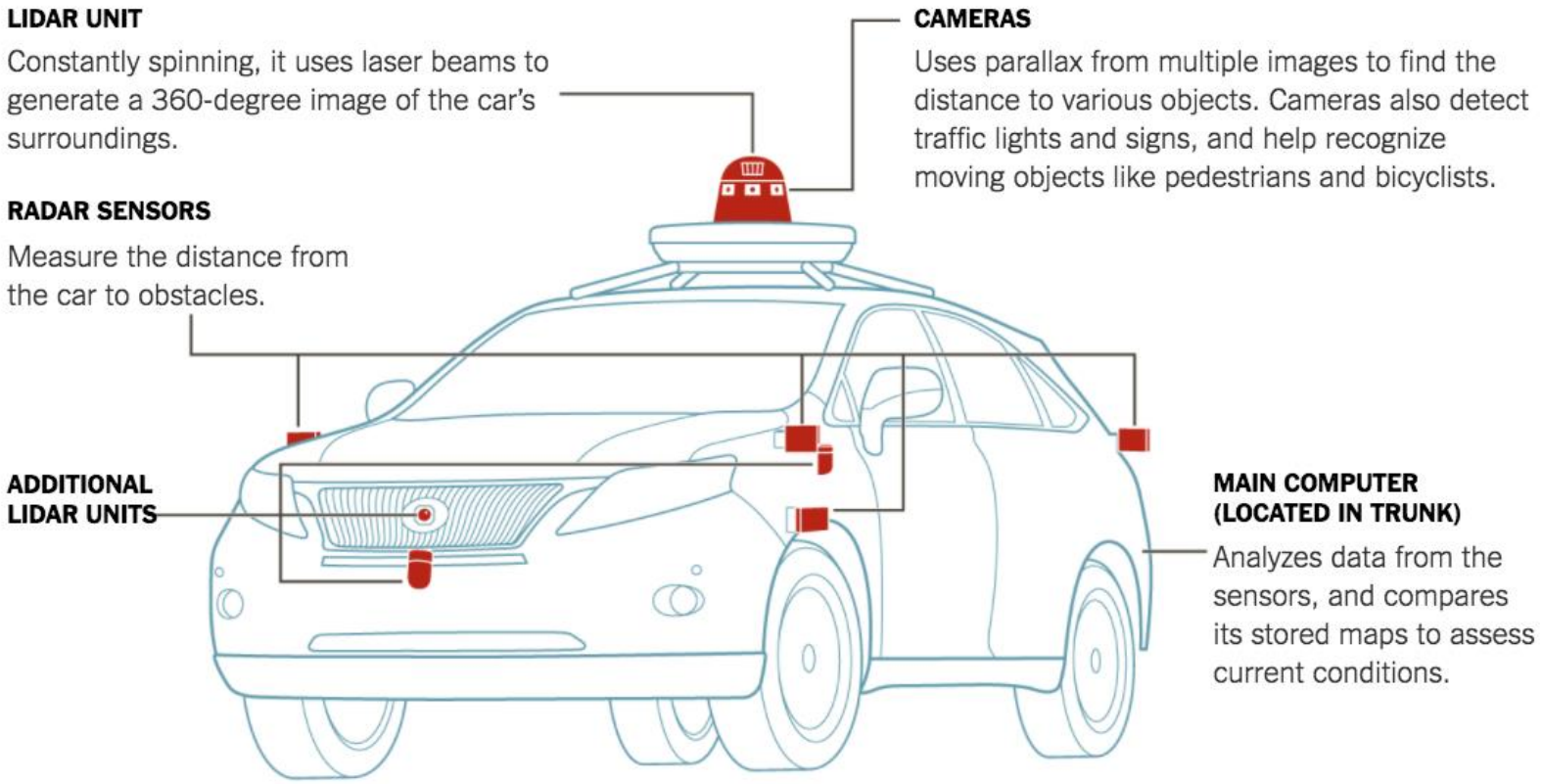
Do I have fever?

Will it rain today?

How much will my money grow?

Will I be able to reach college for my job interview?
How long would it take for me to pay my loan?
Will this medicine create adverse reactions?
How can I reduce my monthly expenses?
Will the monsoon be good this year?
Will I be happy with this person?
Will this borrower default?
Shall I get a heart stroke?

How does an autonomous vehicle move?



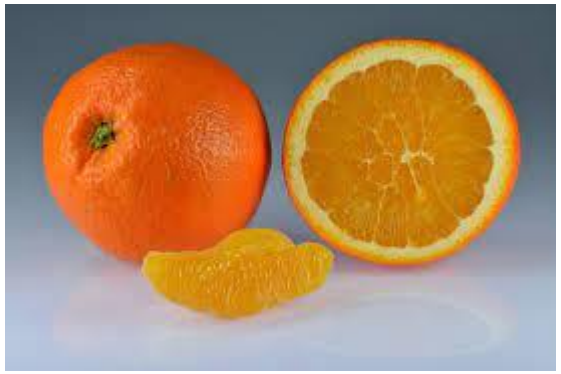
What is important?

Sensors or decision makers?

Feature

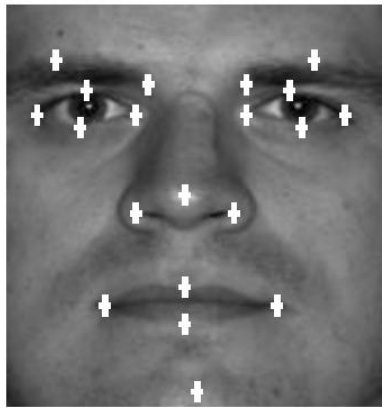
Feature is an attribute (Qualitative or Quantitative) that represents an object.

- Attribute is a property or characteristic of the object; varies from object to another
- Qualitative features: Color, Shape etc.
- Quantitative features: measured values such as height, weight, frequency etc.



Color: Orange
Shape: spherical
Texture: dimpled
Hardness: Soft & moist

Color, pattern, shape,
Rings, spots, stripes



Contours
Distances
Lengths & widths
Skin color etc

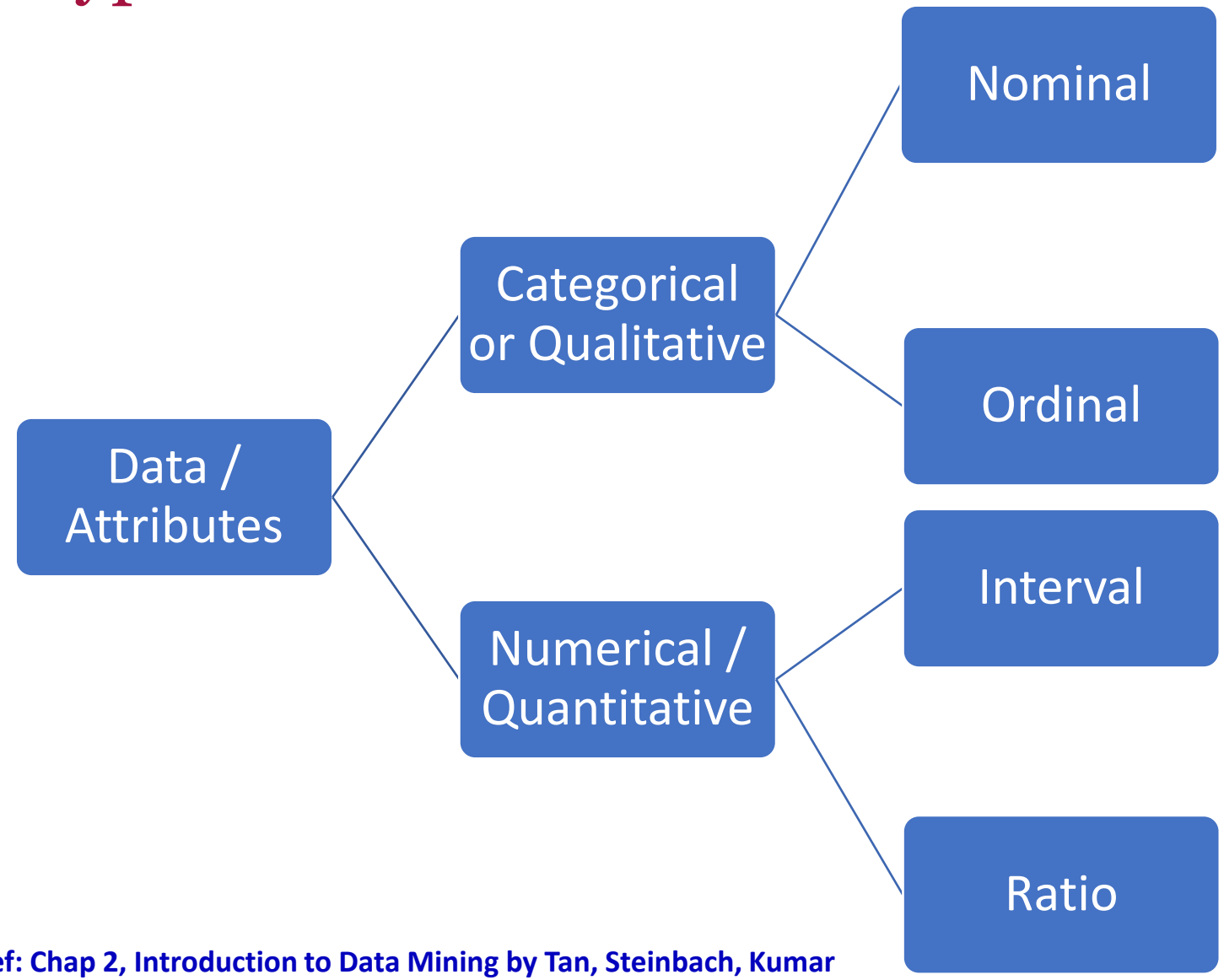
Data

- Referred to as vectors, points, instances, objects, entities, patterns, events, cases etc.
- Real life data is far from being perfect
  - Issues with capture
  - Improper management
  - Noise presence during capture
- Various pre-processing steps employed

W's of Dataset

- Who (or what) does the observation represent? The most important question to be addressed.
- What are the variables?
- Why was the data collected?
- How was the data collected?
- When/Where was the data collected?

Types of Data



- Just different names or values
- No order between two values
- Operations allowed: = or ≠
- Examples
  - people, places, gender
  - PIN / ZIP codes, Phone, Aadhaar numbers

- Sequencing / ordering possible
- Operations allowed: =, ≠, <, >
- Examples
  - Good, better, best or
  - Satisfaction rankings
  - Designations or levels
  - Street numbers

- Measured differences are uniform units
- Operations allowed: =, ≠, <, >, +, -
- Examples
  - Calendar days
  - Age in years
  - Distance in light years

- Measures in continuous scale
- Operations allowed: =, ≠, <, >, +, -, *, /
- Examples
  - Height / weight, BMI
  - Current / Voltage

Ref: Chap 2, Introduction to Data Mining by Tan, Steinbach, Kumar

Data Characteristics

- Dimensionality
  - # of attributes the object possesses
  - Count of parameters in the feature vector
  - Dimensionality of the vector space
- Sparsity
  - Amount of zero values present
  - Sparse data set allows for computational efficiency
  - May help in dimensionality reduction
- Resolution
  - Closeness of observation in time or space
  - Temperature measured every second or hour
  - Low resolution may lead to loss of information
    - *Battle tank movements may not be observed with 10 meter resolution camera on a satellite*
  - High resolution may add noise or lead to loss of data
    - *Stock price change every minute may not help in deciding on investment*

79	17	44	87	6	15	84
97	64	16	55	58	86	93
6	37	75	53	81	50	24
84	15	31	59	30	35	77
91	89	2	78	3	45	80
3	17	11	61	82	92	51
95	22	67	29	27	68	35
92	64	13	59	72	18	16
73	24	76	39	9	15	5
78	46	14	75	43	73	97
59	6	21	13	54	74	86

Dense Data Matrix

2	1	4	0	0	1	0
4	0	2	0	0	3	1
0	0	2	0	0	0	0
0	0	3	5	3	0	0
5	2	1	0	0	0	0
1	5	1	5	2	0	0
0	0	1	0	0	0	0
0	0	5	0	0	5	5
0	0	1	0	0	0	5
1	1	0	0	3	0	1
2	1	3	3	0	3	1

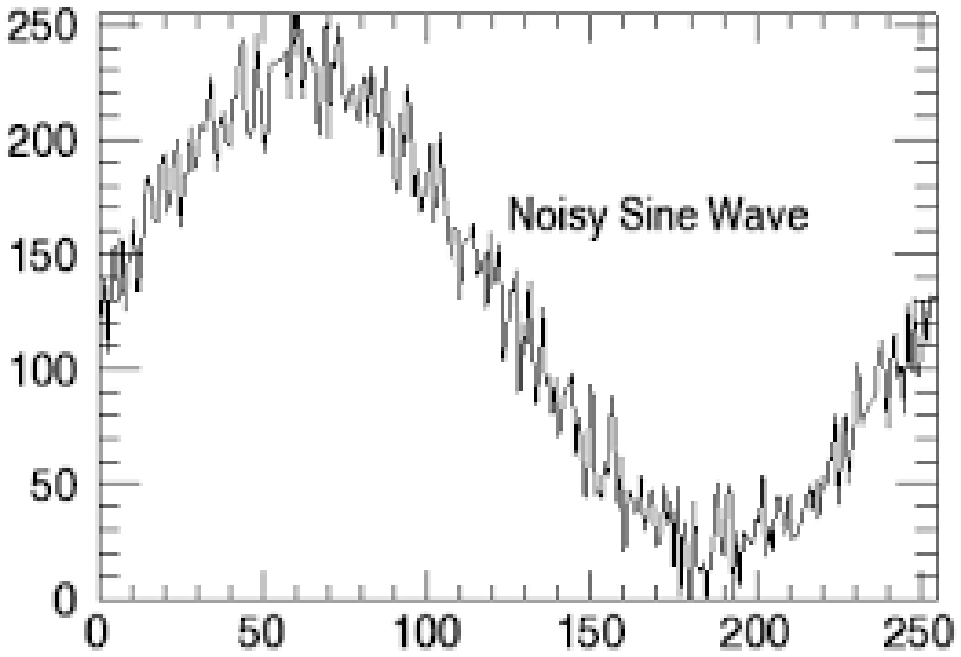
Sparse Data Matrix

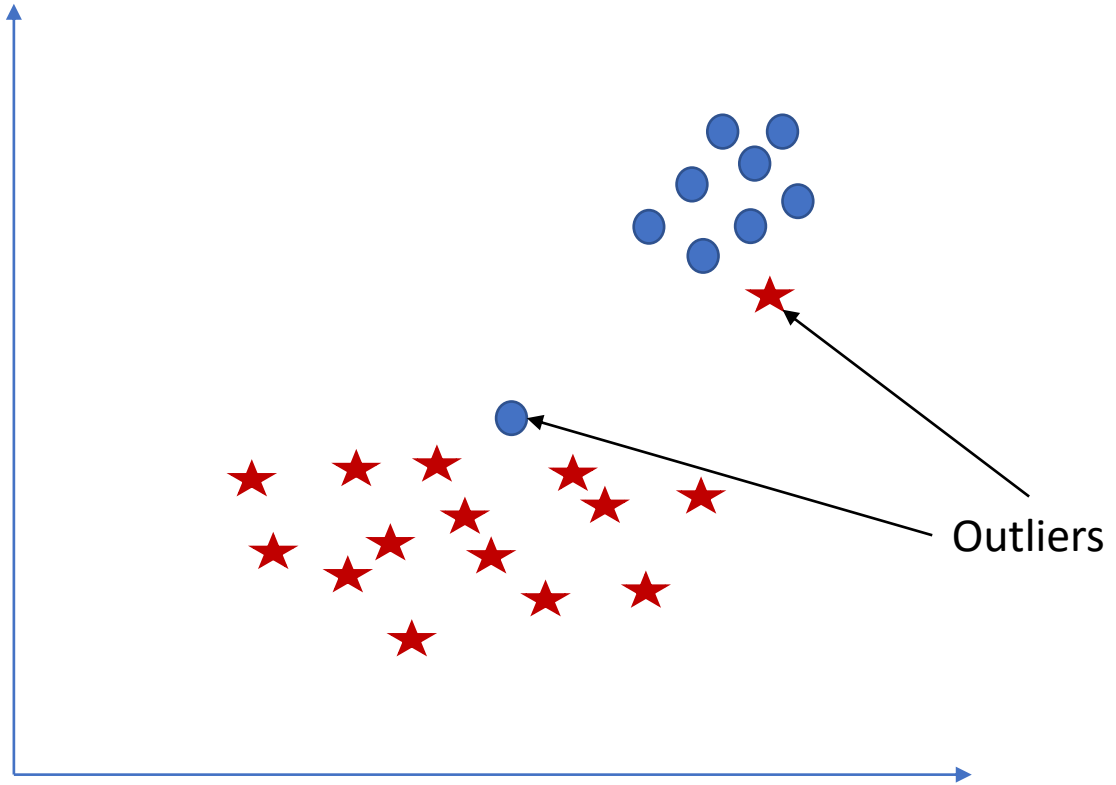
Feature Vector

Card Type	Transaction Amount	Place of residence	Place of Transaction	POS / On-line	Avg Tx Amount
Silver	350	Bangalore	Gurgaon	POS	235
Gold	12,499	Agra	Delhi	On-line	2546
Gold	450	Hyd	Hyd	On-line	1123
Silver	35,000	Chennai	Bhopal	POS	279

	prompt	happy	Satisfy	eager	quickly	kind	attentive	delay	rude	wait	sad	unhappy	bad	
Document 1	1	2						1						I am very happy with the prompt customer support I have received from you. My queries were resolved without much delay and I am so happy for that.
Document 2									1	1		1		You guys are pathetic. You made me wait for hours without any help and the agent was rude. I am so unhappy.
Document 3	1	1				1	1							Very happy with your prompt service. The executive was kind and attentive to my needs.

Noise and Outliers





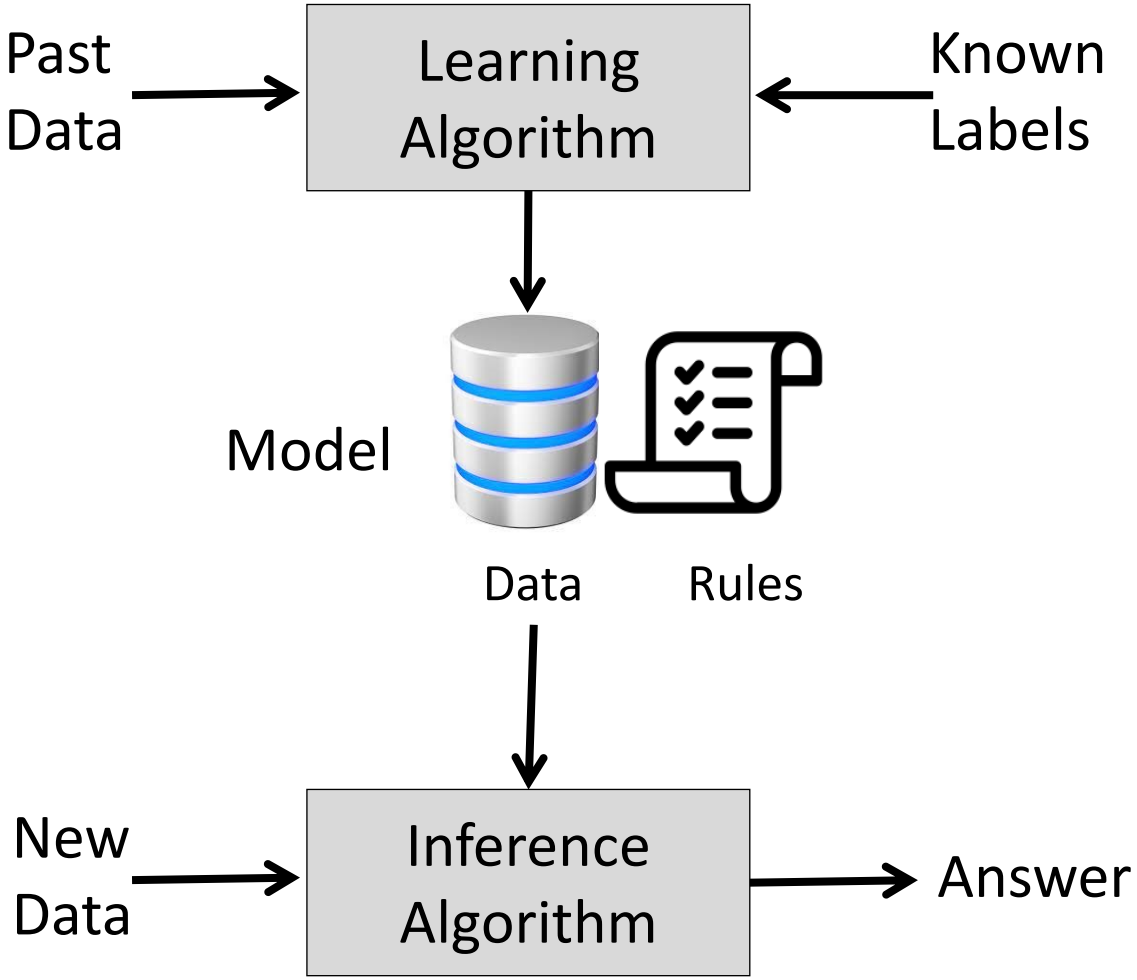
Thank you !!!!!

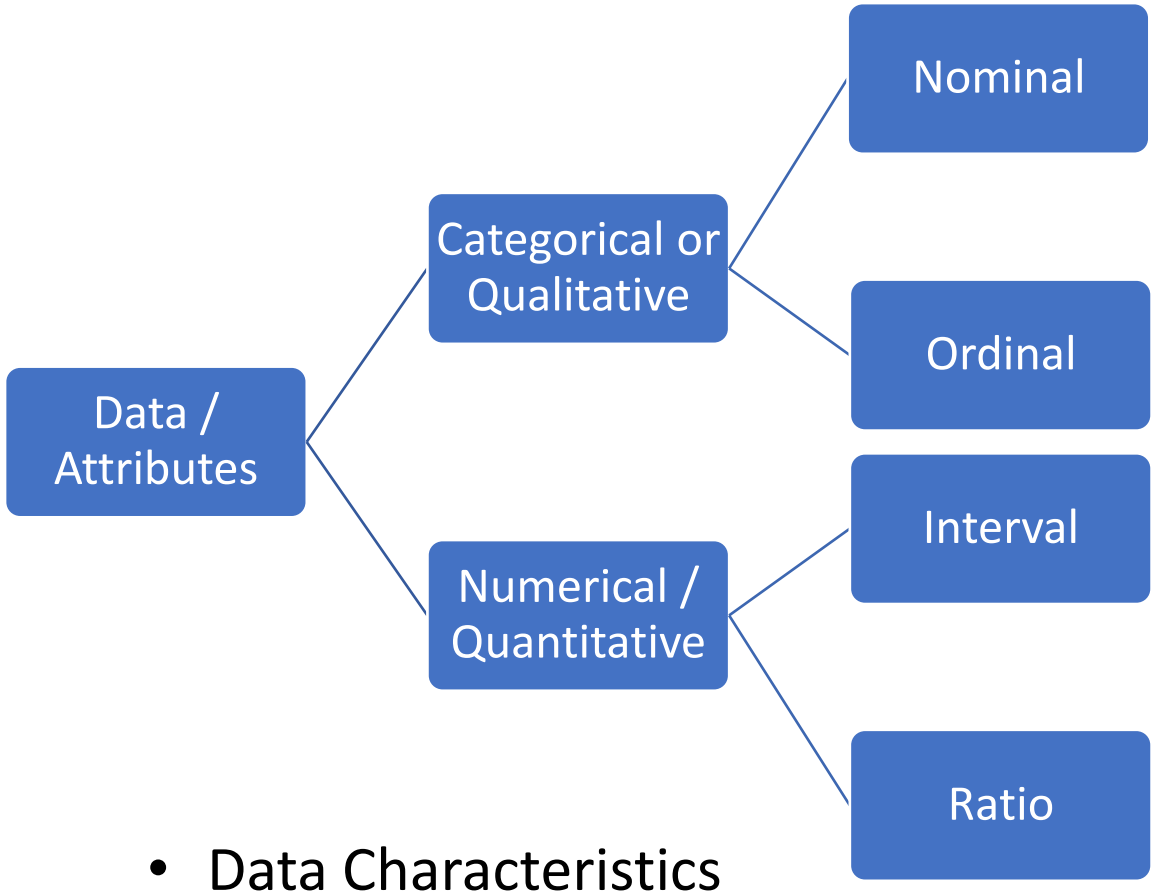
Machine Learning (19CSE305)

Distance, Similarity & Independence

Dr. Peeta Basa Pati
Ms. Priyanka V
Ms.Jyotsna
Department of Computer Science & Engineering,
Amrita School of Engineering, Bengaluru

Data, Types & Characteristics

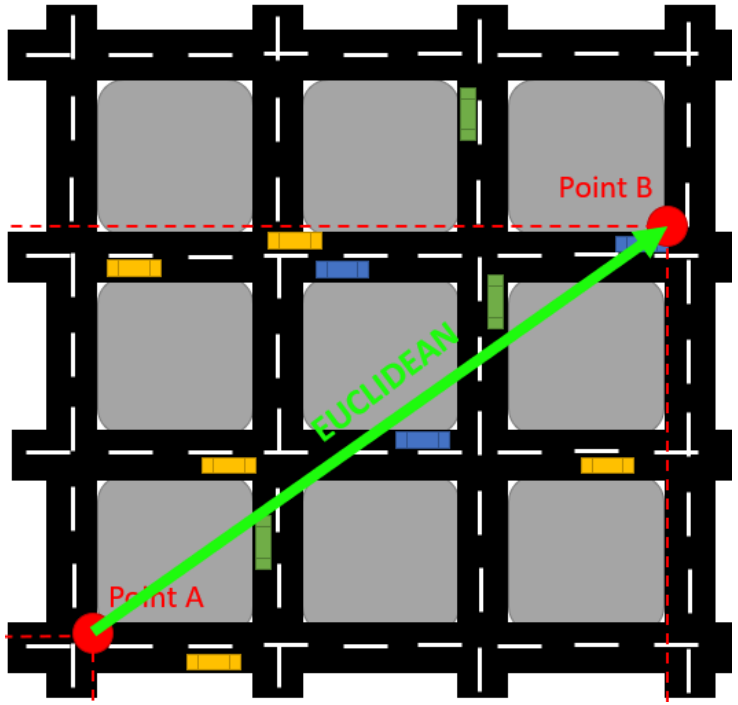




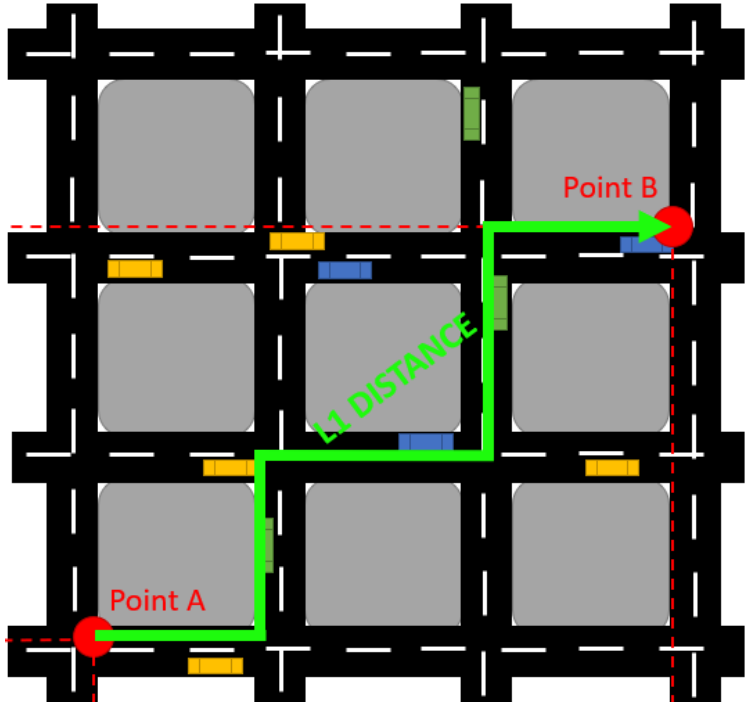
- Data Characteristics
  - Dimensionality
  - Sparsity
  - Resolution
- Noise & Outliers
- W's of the dataset

Data is as important as the algorithm that uses the data.

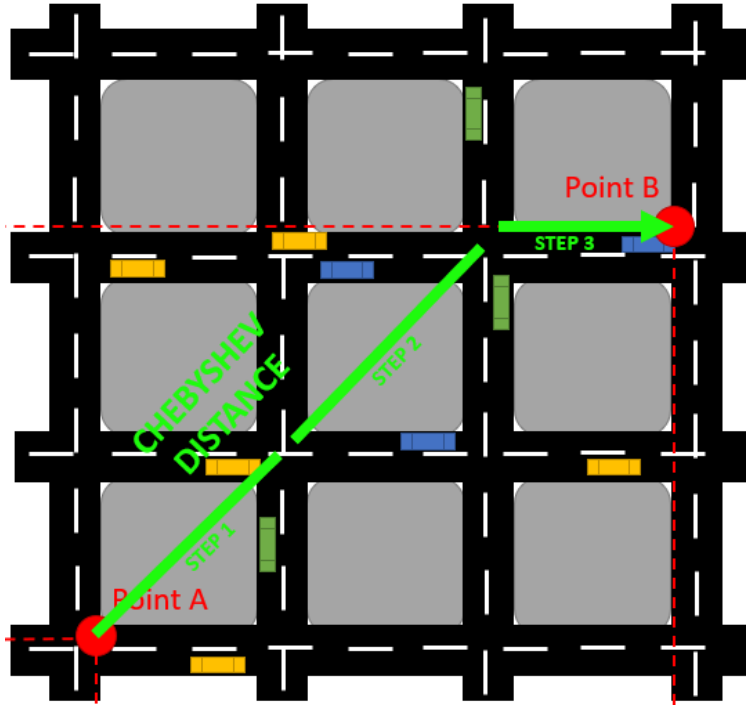
Distance Measures



Euclidean



City-block or Manhattan

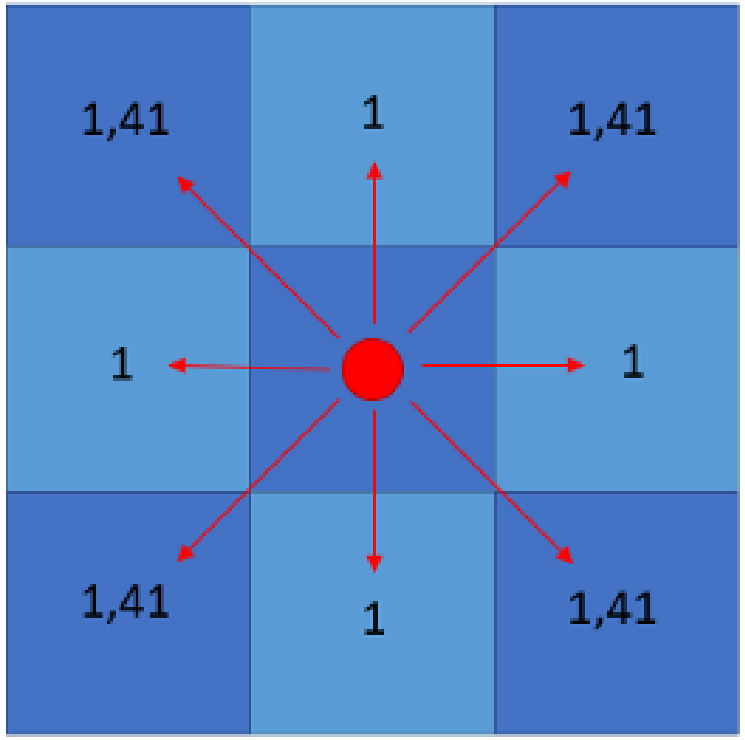


Chebyshev

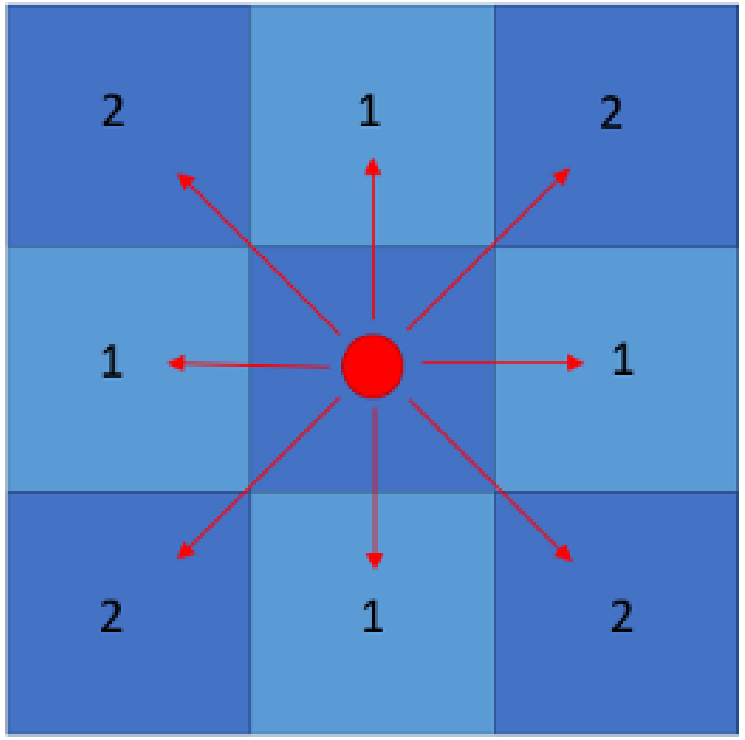
Source: https://towardsdatascience.com/

Distance Measures

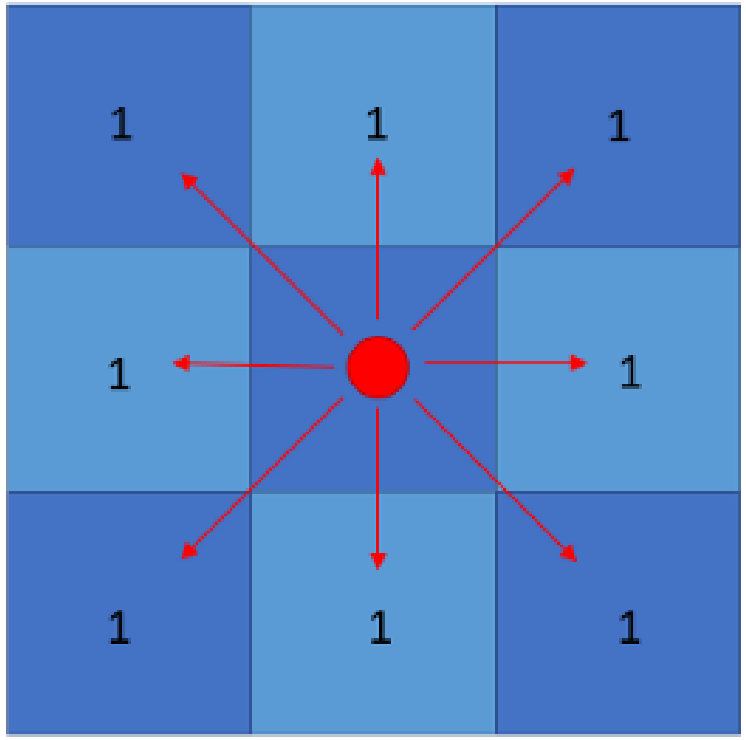
EUCLIDEAN (STRAIGHT LINE)



L1 (CITY BLOCK)

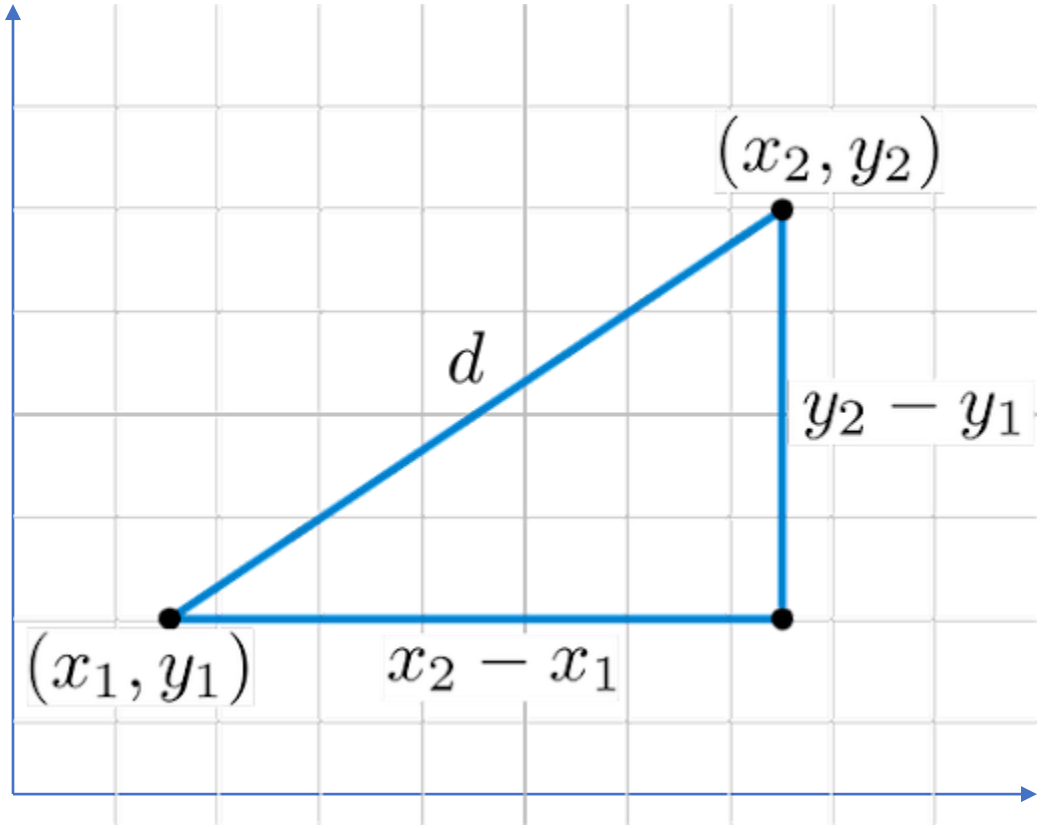


CHEBYSHEV (CHESSBOARD)



Source: https://towardsdatascience.com/

Distance Measures – Euclidean & Minkowski



$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

Euclidean Distance

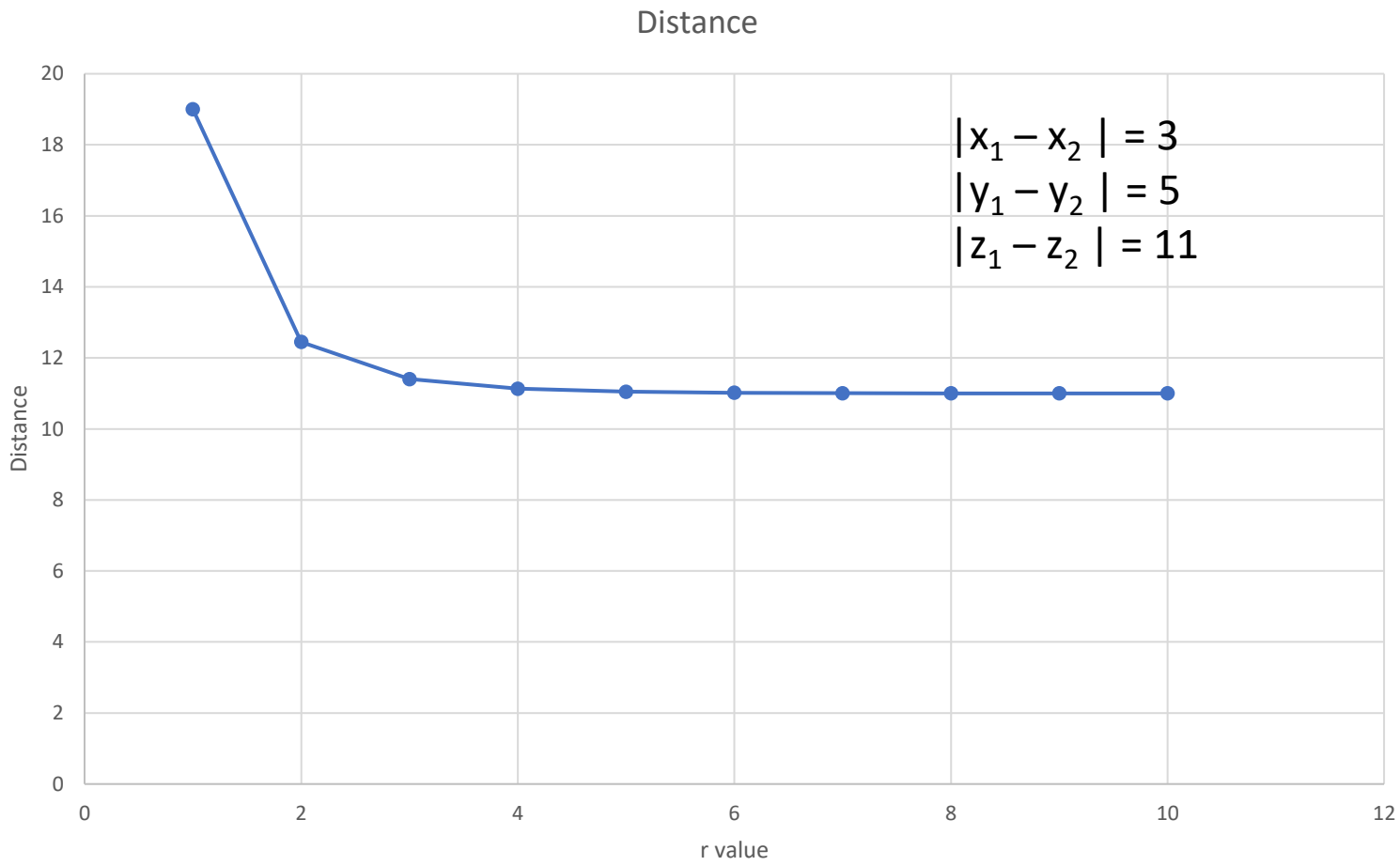
n = number of dimensions (attributes)
p and *q are data points in Euclidean space*

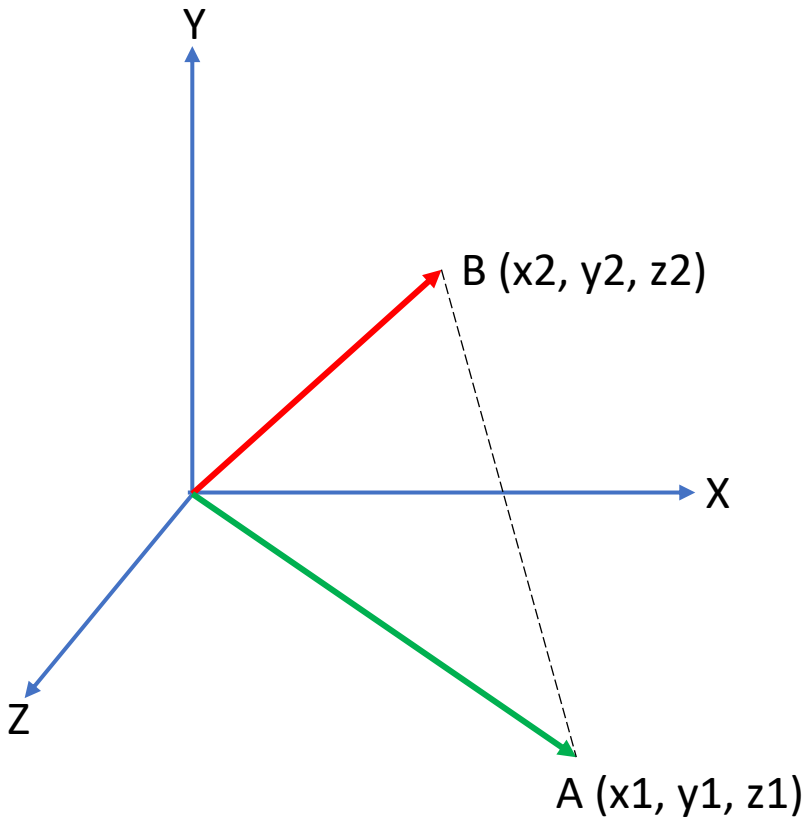
Minkowski Distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Minkowski distance is a generalization of Manhattan & Euclidean distances.
Minkowski with r=1 ➔ city-block / Manhattan
Minkowski with r=2 ➔ Euclidean

Minkwoski Distance with ‘r’ variation





Properties of Distances

- Distance properties

  - $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all x and y and $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.
  - $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)
  - $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all points $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. (Triangle Inequality)

- A distance that satisfies these properties is a metric

Similarity & Properties

- Similarity
  - measure of how alike two objects are
  - higher value when two objects are similar

- Similarity properties
  - $0 \leq s(\mathbf{x}, \mathbf{y}) \leq 1$ for all x and y
  - $s(\mathbf{x}, \mathbf{y}) = 1$ when $\mathbf{x} = \mathbf{y}$ (mostly*)
  - $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$ for all **x** and **y**. (Symmetry)

Binary Vectors: Simple Matching

XNOR

INPUT		OUTPUT
A	B	
0	0	1
1	0	0
0	1	0
1	1	1

f_{01} = number of attributes where **x** was 0 and **y** was 1
f_{10} = number of attributes where **x** was 1 and **y** was 0
f_{00}= number of attributes where **x** was 0 and **y** was 0
f_{11} = number of attributes where **x** was 1 and **y** was 1

Similarity Matching Coefficient (SMC)
= number of matches / number of attributes
= $(f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$

	b7	b6	b5	b4	b3	b2	b1	b0
X	1	0	0	0	1	0	0	0
Y	1	1	0	1	1	1	1	0
XNOR	1	0	1	0	1	0	0	1

$f_{01} = 4$
$f_{10} = 0$
$f_{00} = 2$
$f_{11} = 2$

SMC $= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$
$= (2 + 2) / 8 = 0.5$

Binary Vectors: Jaccard Similarity

2 - input AND gate

A	B	Output
0	0	0
0	1	0
1	0	0
1	1	1

Jaccard Coefficient (JC)

= number of 11 matches / # attributes with any value 1

= $f_{11} / (f_{01} + f_{10} + f_{11})$

	b7	b6	b5	b4	b3	b2	b1	b0
X	1	0	0	0	1	0	0	0
Y	1	1	0	1	1	1	1	0
XNOR	1	0	1	0	1	0	0	1

$f_{01} = 4$
$f_{10} = 0$
$f_{00} = 2$
$f_{11} = 2$

JC $= (f_{11}) / (f_{01} + f_{10} + f_{11})$

$= (2) / 6 = 0.33$

Cosine Similarity

$A = \{a_1, a_2, \ldots, a_n\}$ $B = \{b_1, b_2, \ldots, b_n\}$

If **A** and **B** are two document vectors, then

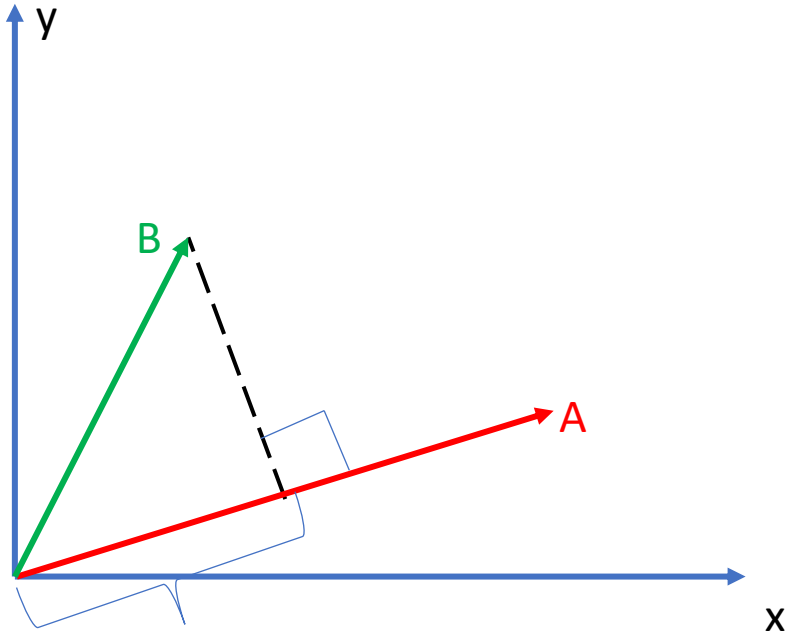
$$\cos(\mathbf{A}, \mathbf{B}) = \langle \mathbf{A}, \mathbf{B} \rangle / \|\mathbf{A}\| \|\mathbf{B}\|$$

$$< A, B >= \sum_{k=1}^{n} a_k * bk$$

$\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are lengths of vectors A & B

A = 3 2 0 5 0 0 0 2 0 0

B = 1 0 0 0 0 0 0 1 0 2



Cos(A,B)

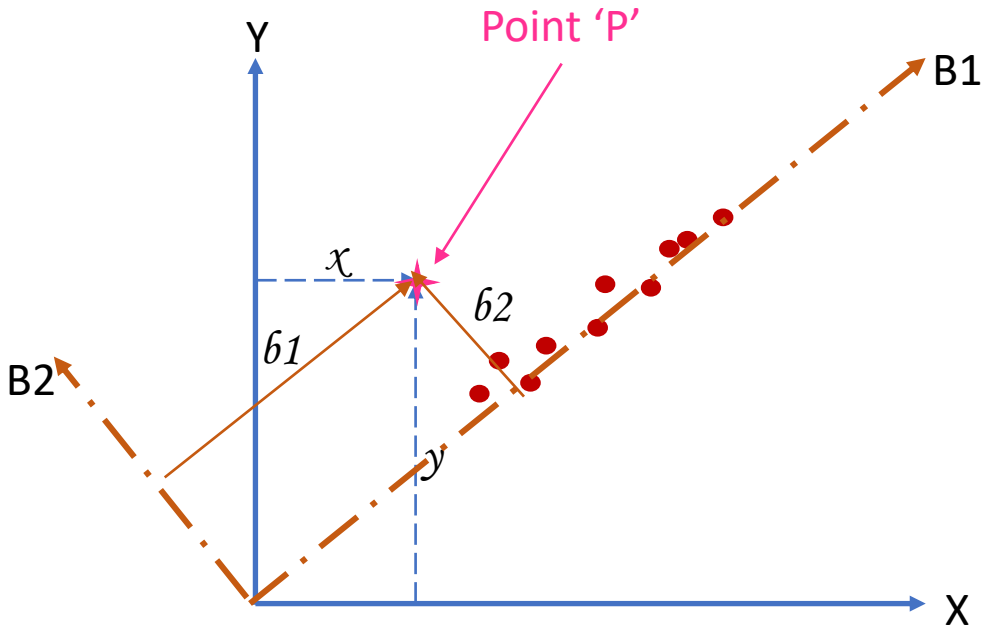
$\langle \mathbf{A}, \mathbf{B} \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$ ⟸ Also known as dot product

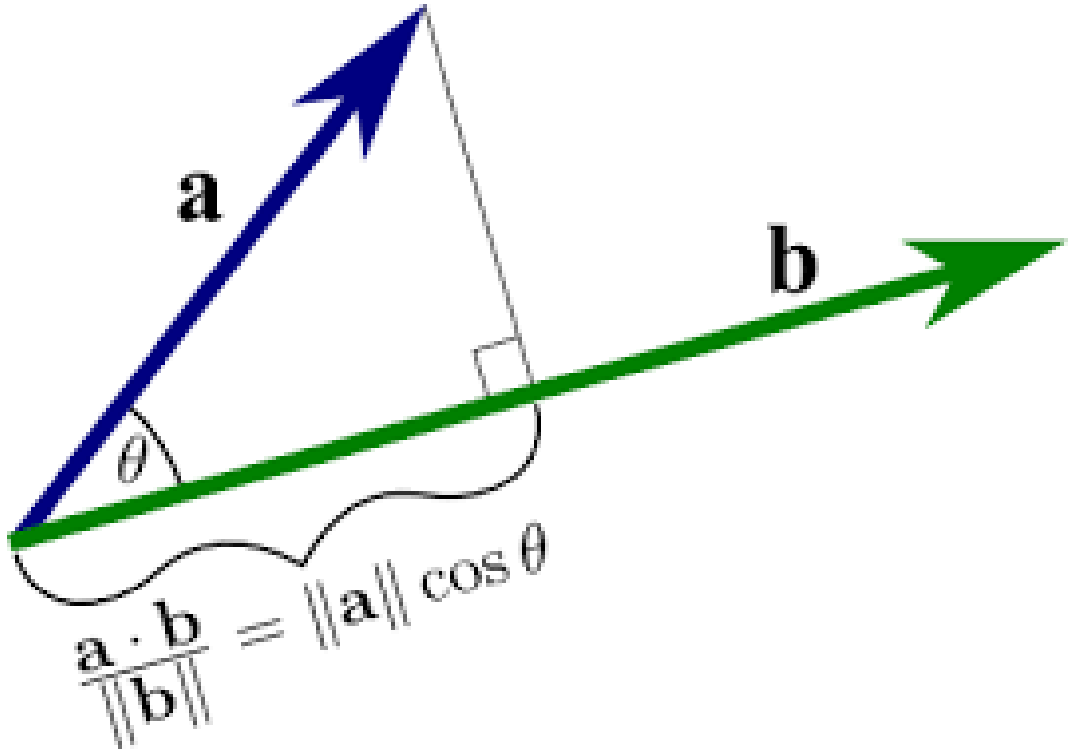
$\| \mathbf{A} \| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$\| \mathbf{B} \| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.449$

$\cos(\mathbf{A}, \mathbf{B}) = 0.3150$

Projections





Point 'P' is represented as (x,y) in X-Y coordinate system

The same point 'P' is presented as $(b1,b2)$ in B1-B2 coordinate system.

Thank you !!!!!

Machine Learning (19CSE305)

Seperability, Systems,Data Preprocessing

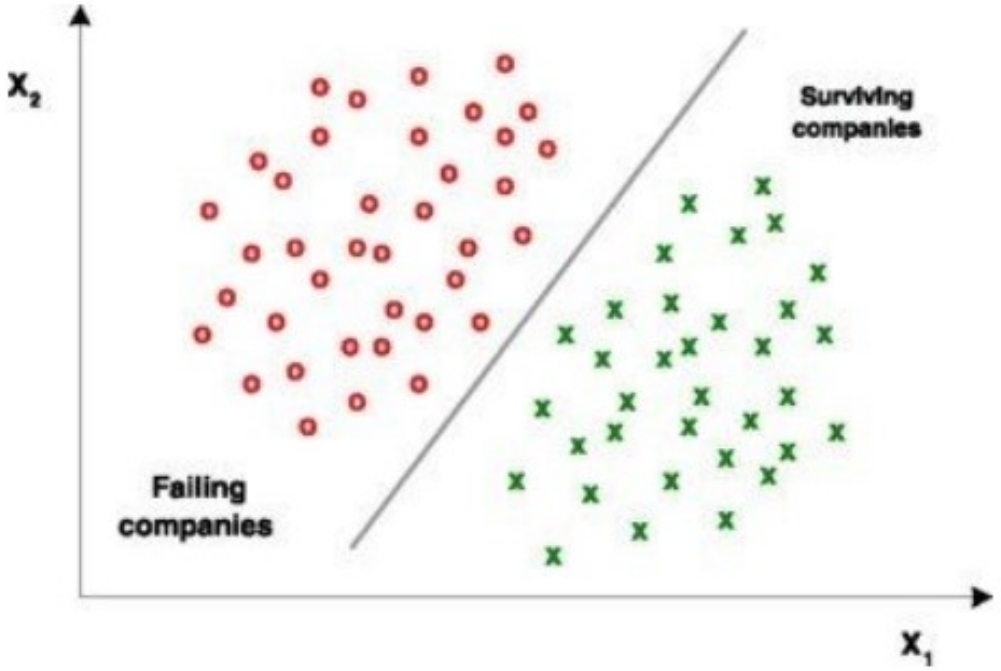
AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

Dr. Peeta Basa Pati
Ms. Priyanka V
Department of Computer Science & Engineering,
Amrita School of Engineering, Bengaluru

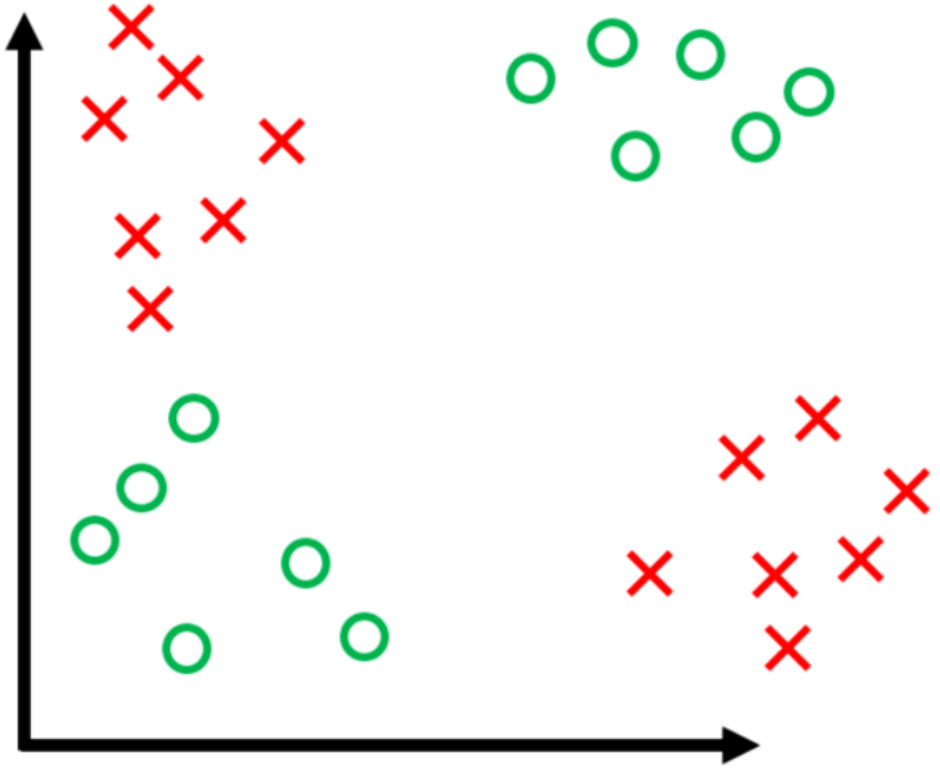
Topics

- Linearly seperable &non-seperable data
- Observable and Controllable
- Time Invariant Systems
- Data Preprocessing

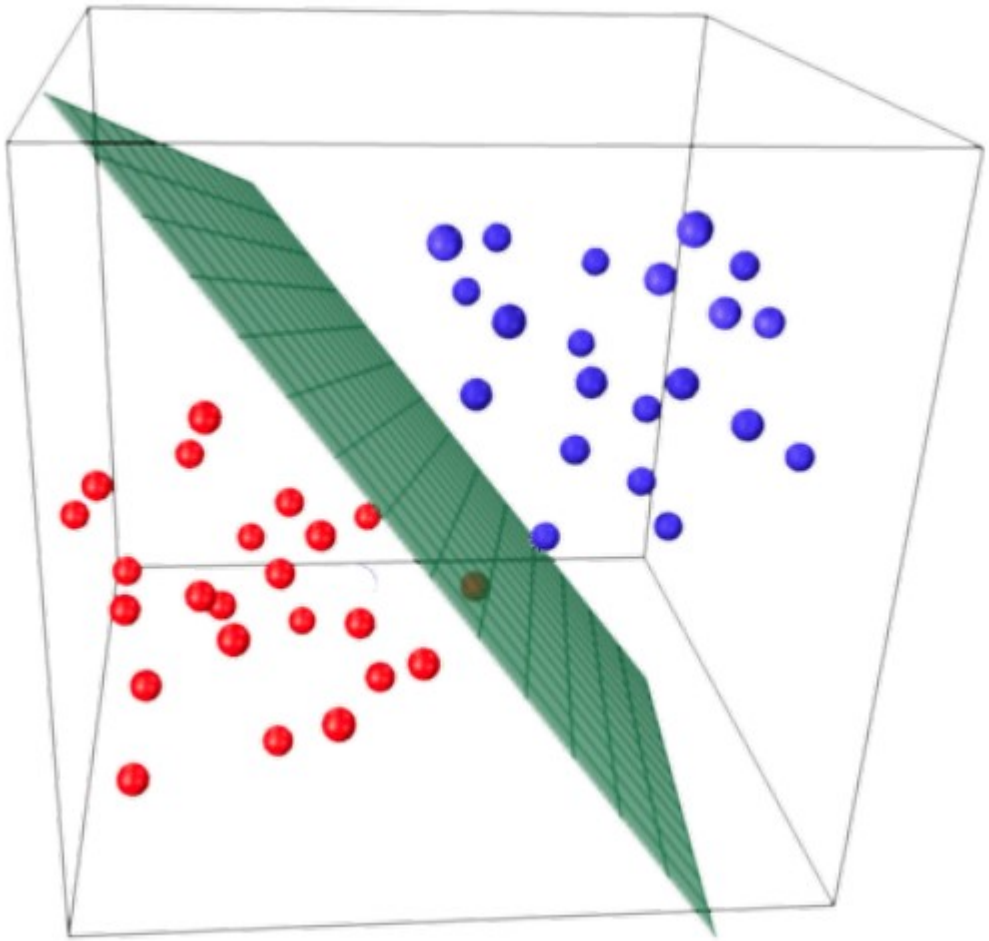
Linearly separable data



Linearly non separable data



Linearly seperable



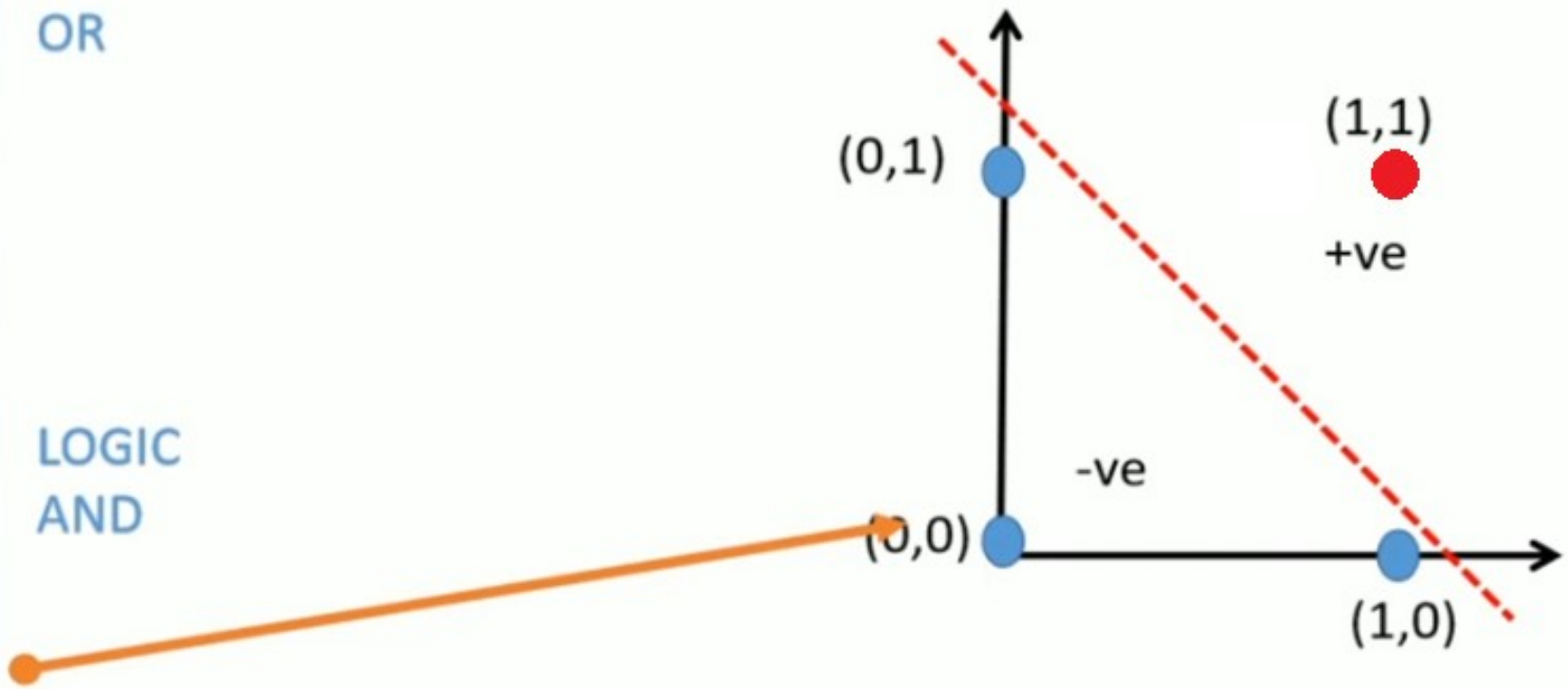
Linearly Separable

x1	x2	y
1	1	1
1	0	1
0	1	1
0	0	0

LOGIC OR

x1	x2	y
1	1	1
1	0	0
0	1	0
0	0	0

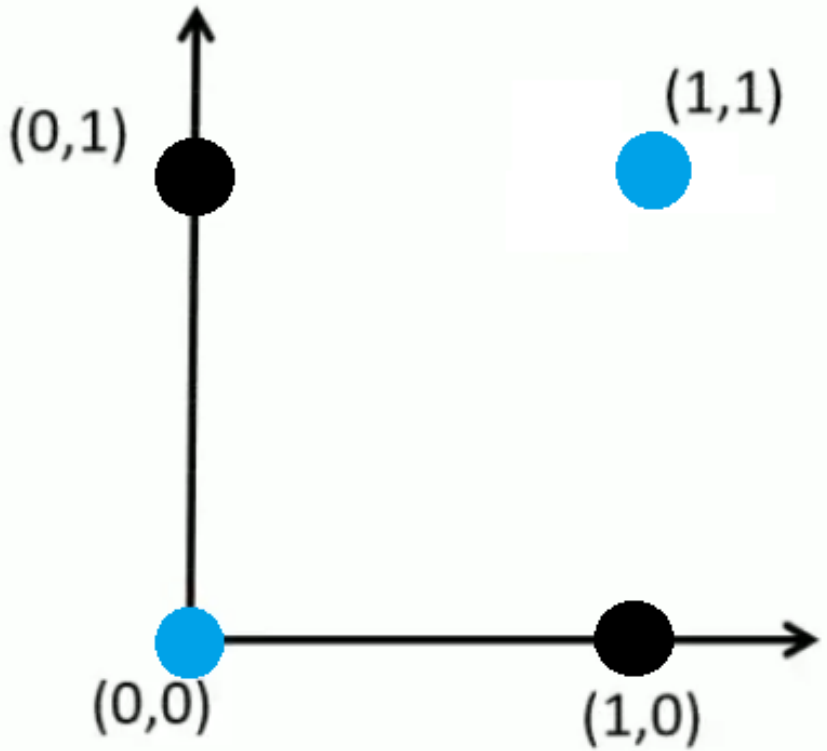
LOGIC AND



Linearly non-separable

x1	x2	y
1	1	0
1	0	1
0	1	1
0	0	0

LOGIC
XOR



can't separate using single line

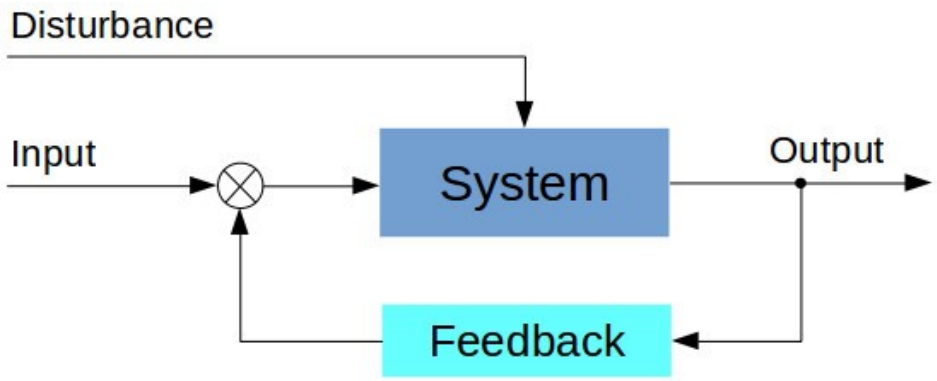
Observability & Controllability

Observable Systems

- System is said to be observable if the behavior of the system can be observed with input-output relationships
- Its also the ability to infer the internal states based on observed outputs

Controllable Systems

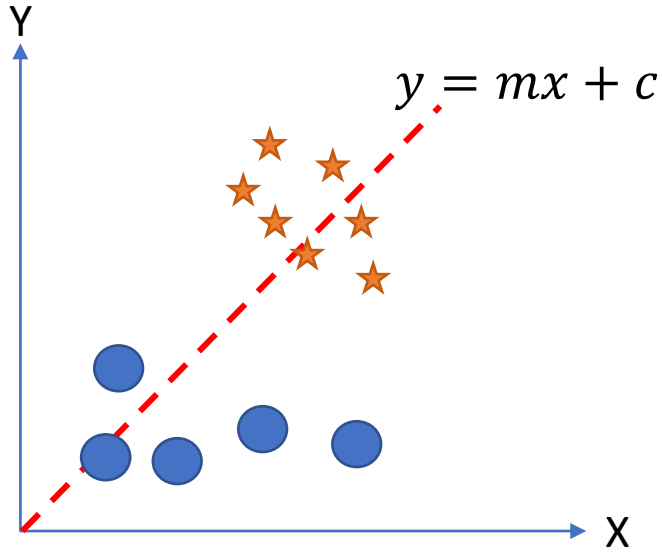
- Controllable system can move from an initial state to a desired end state in finite time

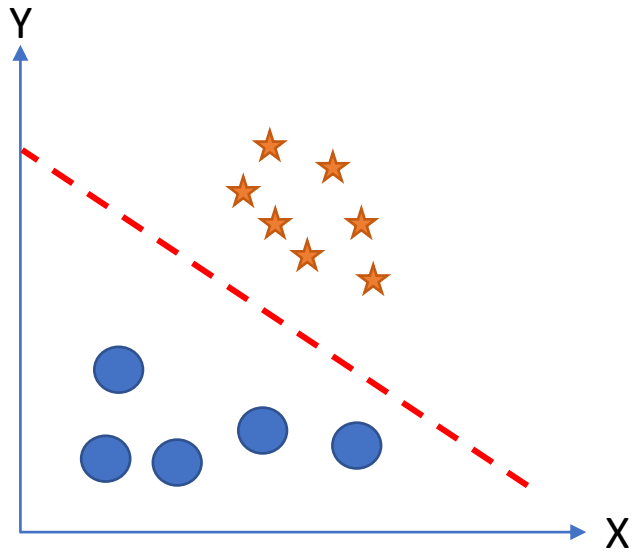


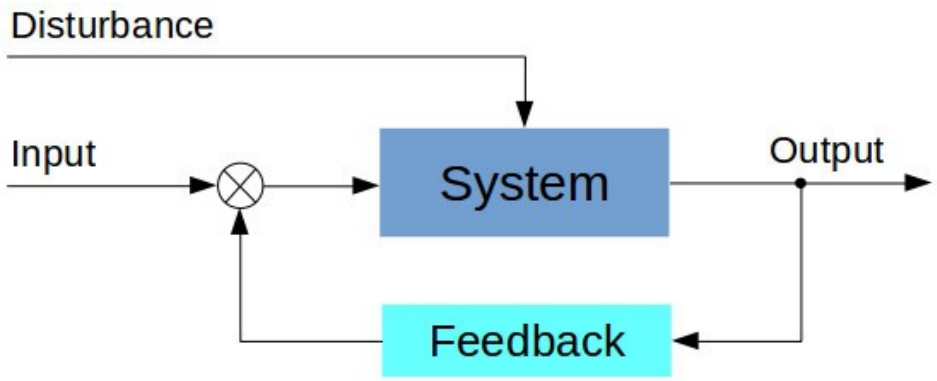
$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + B\mathbf{u}(t)$$
$$\mathbf{y}(t) = C\mathbf{x}(t) + D\mathbf{u}(t)$$

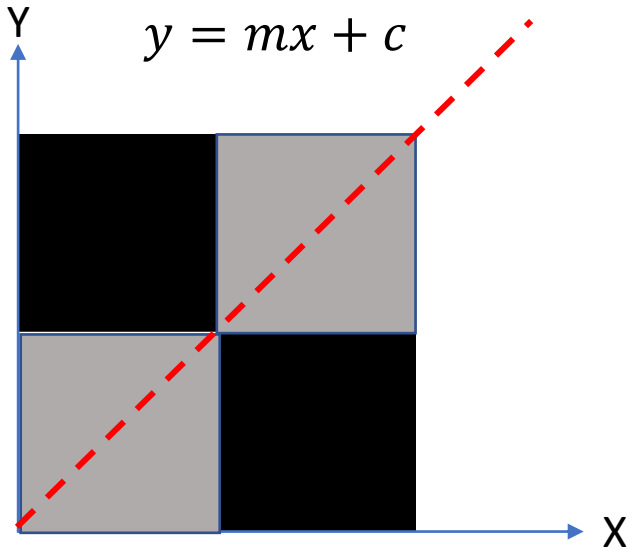
$$\mathbf{x}(k+1) = A\mathbf{x}(k) + B\mathbf{u}(k)$$

Observability & Controllability



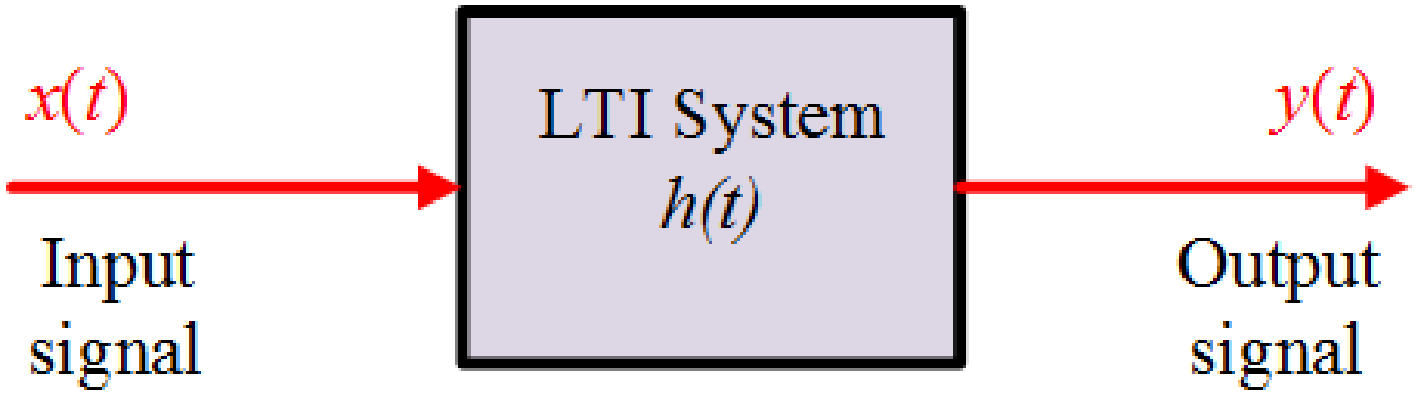






Linear Time-Invariant System

- The output of the system depends on the input and impulse response
- Causal system – current output is from current and past inputs
- Non-causal system – current output from future input



Data Preprocessing

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation*=“ ” (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., *Salary*=“−10” (an error)
  - inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age*=“42”, *Birthday*=“03/07/2010”
    - Was rating “1, 2, 3”, now rating “A, B, C”
    - discrepancy between duplicate records
  - Intentional (e.g., *disguised missing* data)
    - Jan. 1 as everyone’s birthday?

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - mode

Thank you !!!!!

Machine Learning (19CSE305)

K-Nearest neighbor

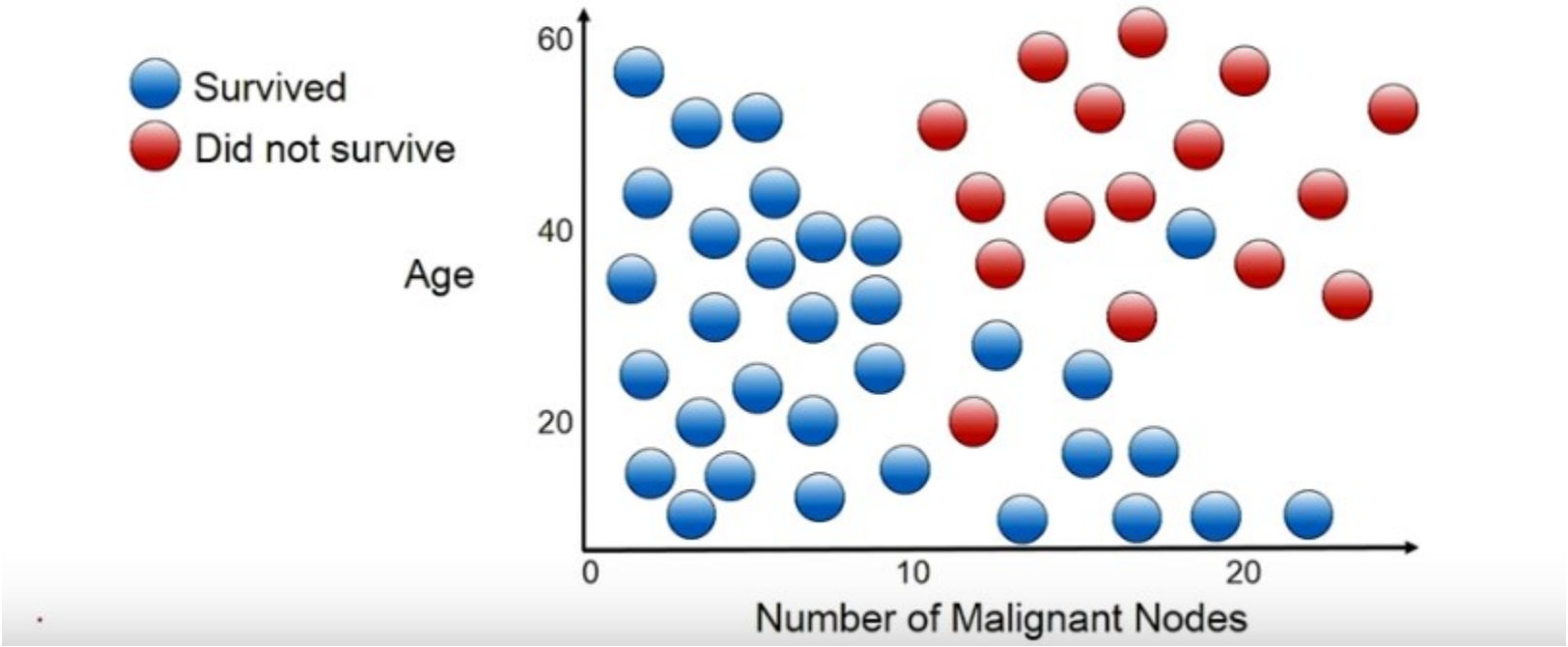
Dr. Peeta Basa Pati
Ms. Priyanka V
Department of Computer Science & Engineering,
Amrita School of Engineering, Bengaluru

Topics

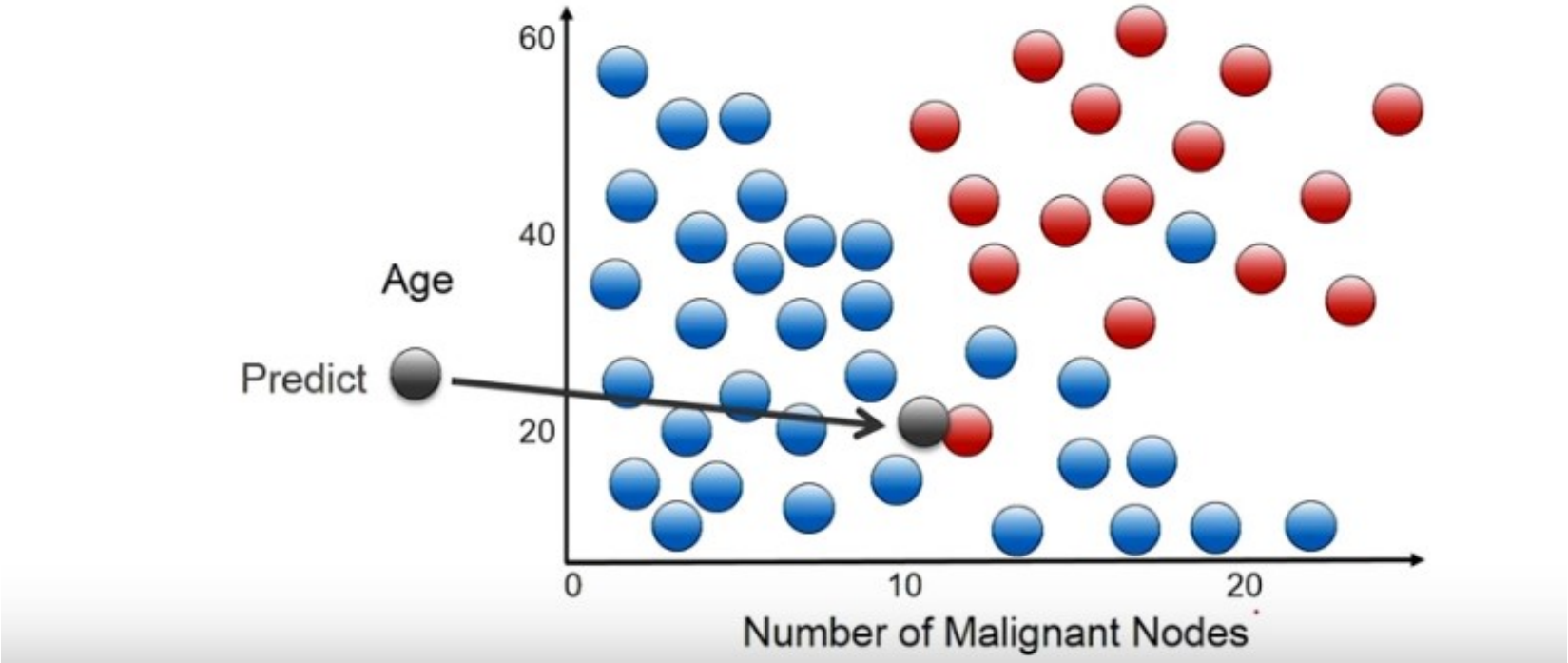
- K-nearest neighbour classifier

- Distance measure
- Voronoi diagram

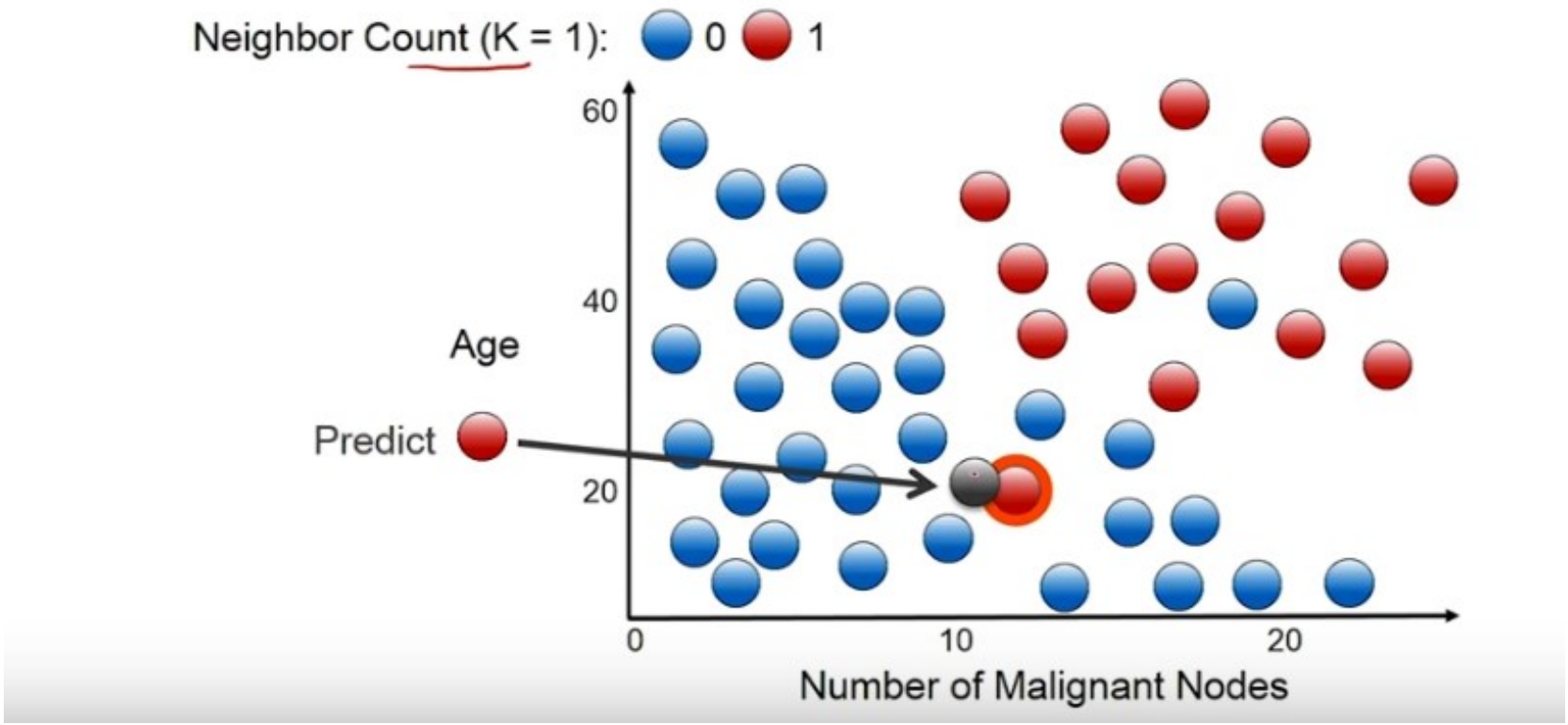
K-Nearest Neighbour- Classification



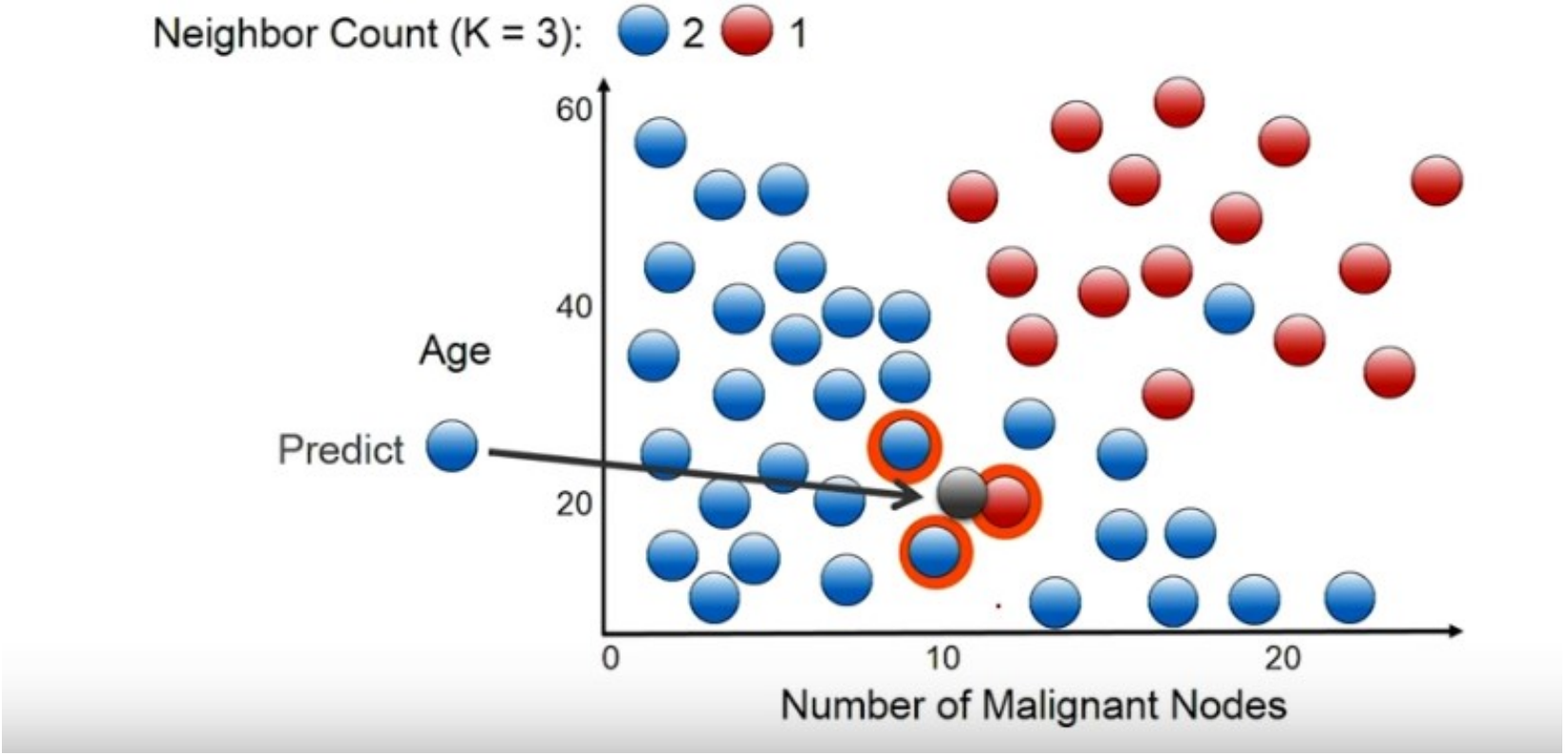
K-Nearest Neighbour- Classification



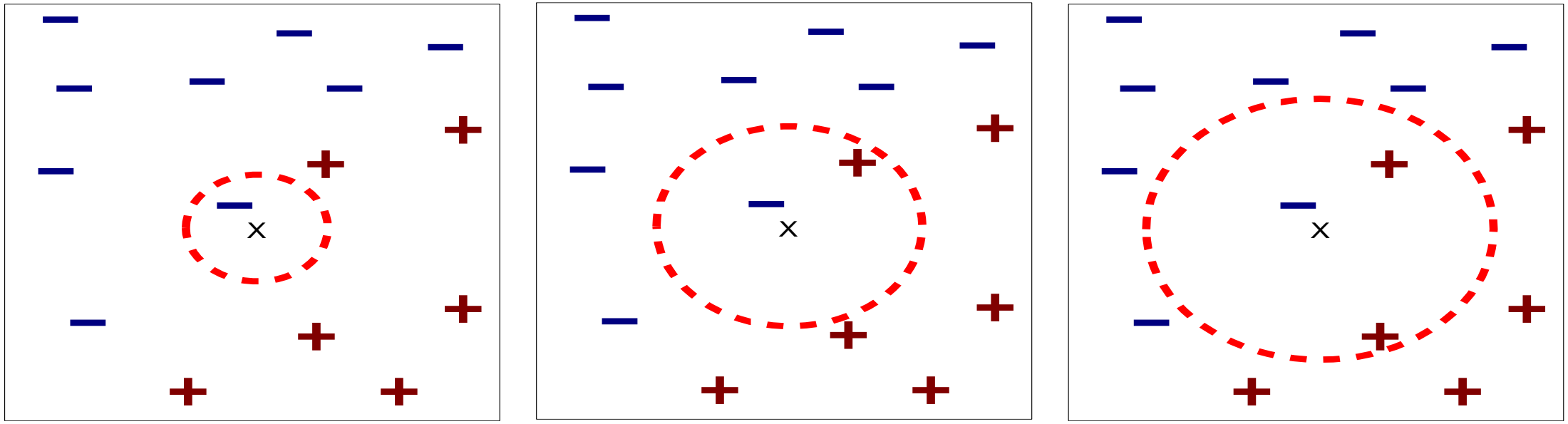
.. Classification



...Classification



Definition of Nearest Neighbor



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x are data points that have the k smallest distance to x

K-nearest neighbour- algorithm

- Training phase : Save the examples
- Prediction phase: Get the test instance

Find the k- training examples

$\{(x_1,y_1), (x_2,y_2), (x_3,y_3)..... (x_k,y_k)\}$ that is closest to x_t

Classification : predict the majority class from $\{y_1, y_2, y_3.... y_k\}$

Regression : predict the average of $\{y_1, y_2, y_3.... y_k\}$

Calculating Distance

- Euclidean
- Let A and B are represented by feature vectors $A = (x_1, x_2, ..., x_n)$ and $B = (y_1, y_2, ..., y_n)$, where n is the dimensionality of the feature space.

$$dist\,(A, B) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Example

We have data from the questionnaires survey (to ask people opinion) and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not. Here is four training samples

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Y = Classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

Now the factory produces a new paper tissue that pass laboratory test with X1 = 3 and X2 = 7. Without another expensive survey, can we guess what the classification of this new tissue is?

Assume k = 3

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	**Square Distance to query instance (3, 7)**
7	7	$(7-3)^2+(7-7)^2=16$
7	4	$(7-3)^2+(4-7)^2=25$
3	4	$(3-3)^2+(4-7)^2=9$
1	4	$(1-3)^2+(4-7)^2=13$

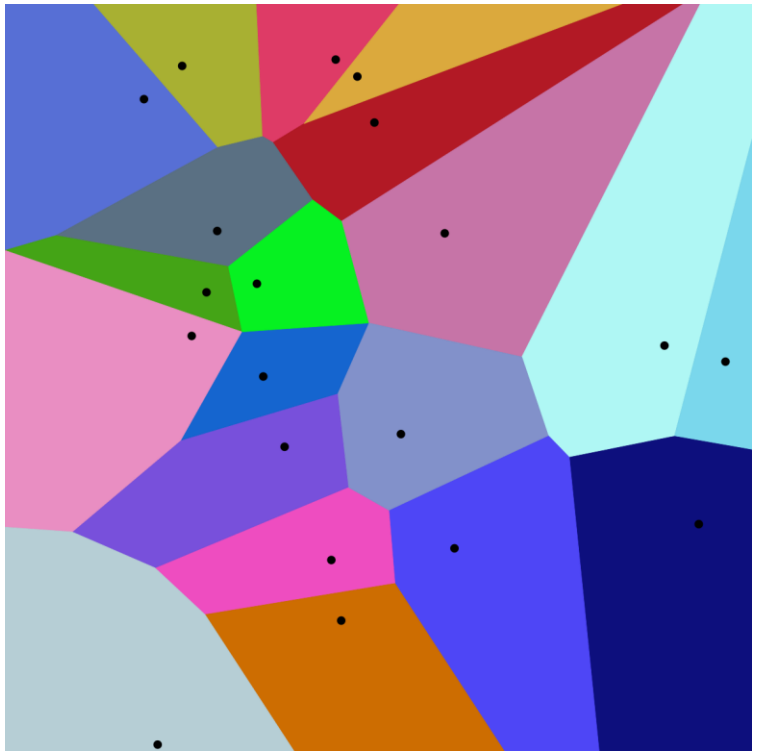
Assume k = 3

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Square Distance to query instance (3, 7)	**Rank minimum distance**	**Is it included in 3-Nearest neighbors?**
7	7	$(7-3)^2+(7-7)^2=16$	**3**	**Yes**
7	4	$(7-3)^2+(4-7)^2=25$	**4**	**No**
3	4	$(3-3)^2+(4-7)^2=9$	**1**	**Yes**
1	4	$(1-3)^2+(4-7)^2=13$	**2**	**Yes**

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Square Distance to query instance (3, 7)	Rank minimum distance	Is it included in 3-Nearest neighbors?	**Y = Category of nearest Neighbor**
7	7	$(7-3)^2+(7-7)^2=16$	3	Yes	**Bad**
7	4	$(7-3)^2+(4-7)^2=25$	4	No	-
3	4	$(3-3)^2+(4-7)^2=9$	1	Yes	**Good**
1	4	$(1-3)^2+(4-7)^2=13$	2	Yes	**Good**

We have 2 good and 1 bad, since 2>1 then we conclude that a new paper tissue that pass laboratory test with X1 = 3 and X2 = 7 is included in **Good** category.

Voronoi diagram k=1



Lazy vs Eager Learner

- Eager learners
  - when given a set of training tuples, will construct a generalization (i.e., classification) model before receiving new (e.g., test) tuples to classify

- Lazy learners-
  - waits until the last minute before doing any model construction to classify a given test tuple
  - simply stores it (or does only a little minor processing) and waits until it is given a test tuple.
  - Also referred as instance based learner

kNN - advantage

it is relatively straightforward to update the model when new labeled instances become available—we simply add them to the training dataset.

Nearest Neighbor Classification…

- Scaling issues
  - Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes
  - Example:
    - height of a person may vary from 1.5m to 1.8m
    - weight of a person may vary from 50kg to 110kg
    - income of a person may vary from ₹10K to ₹1crore
  - normalisation

If attributes are non numeric?

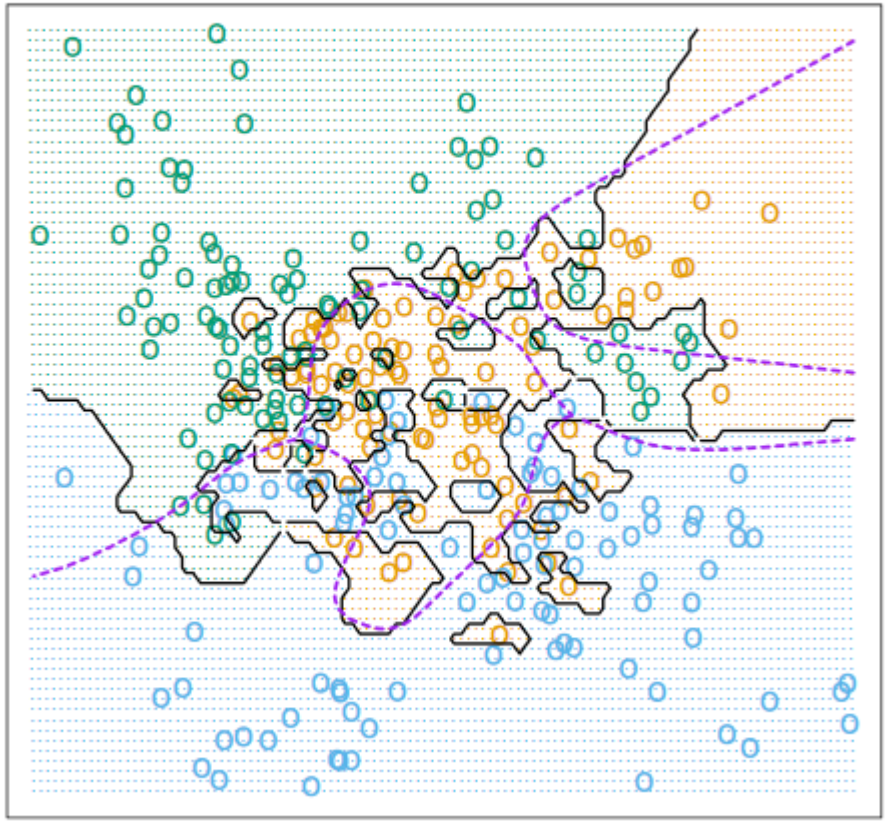
" color?"

- compare the corresponding value of the attribute in tuple X1 with that in tuple X2 .
- If the two are identical (e.g., tuples X1 and X2 both have the color blue), then the difference between the two is taken as 0.
- If the two are different (e.g., tuple X 1 is blue but tuple X 2
- is red), then the difference is considered to be 1.
- Other methods may incorporate more sophisticated schemes for differential grading (e.g., where a larger difference score is
- assigned, say, for blue and white than for blue and black).

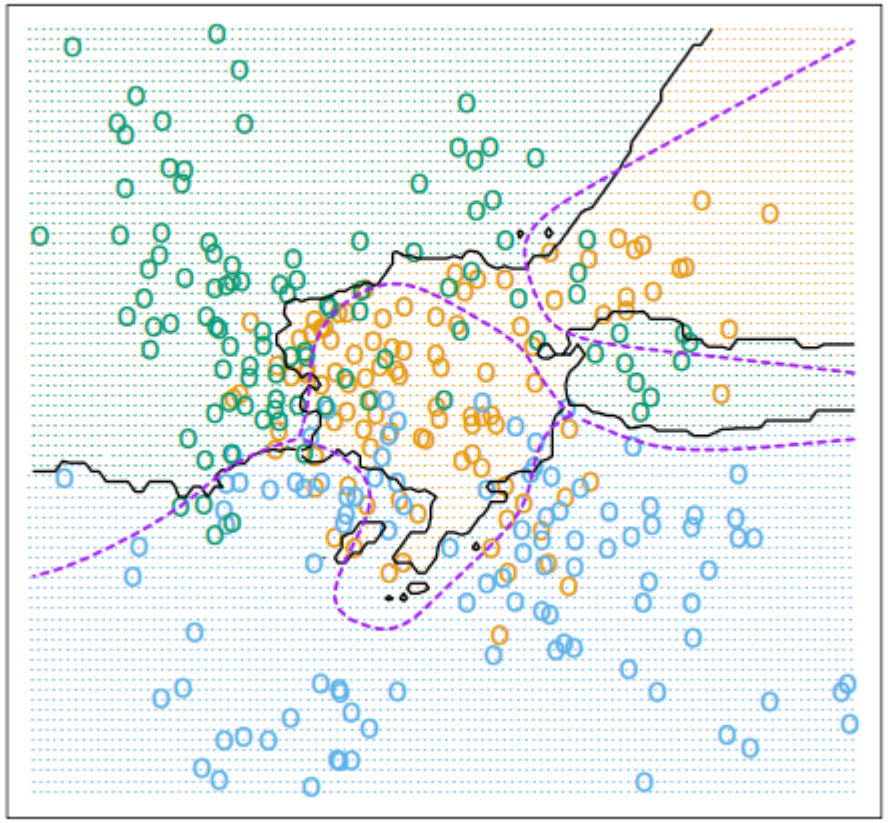
How to determine a good value for k?

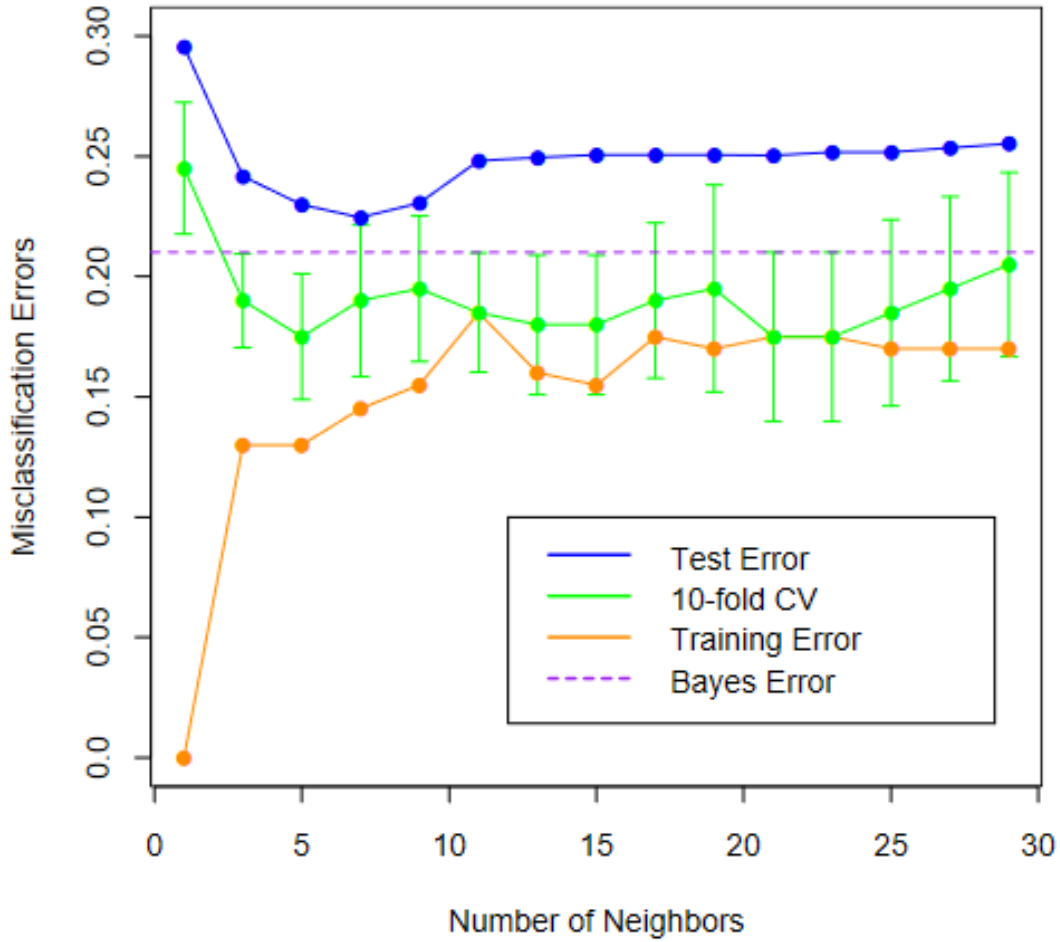
- Starting with k = 1, we use a test set to estimate the error rate of the classifier.
- This process can be repeated each time by incrementing k to allow for one more neighbor.
- The k value that gives the minimum error rate may be selected. the larger the number of training tuples, the larger the value of k will be (so that classification and numeric prediction decisions can be based on a larger portion of the stored tuples).

1-Nearest Neighbor

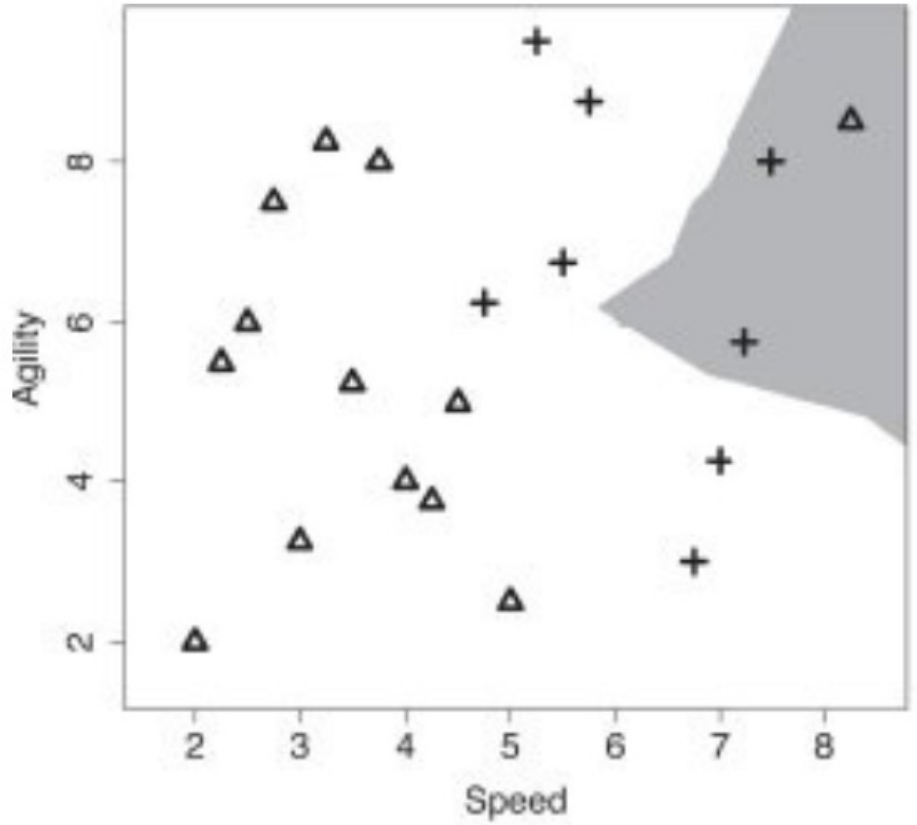


15-Nearest Neighbors





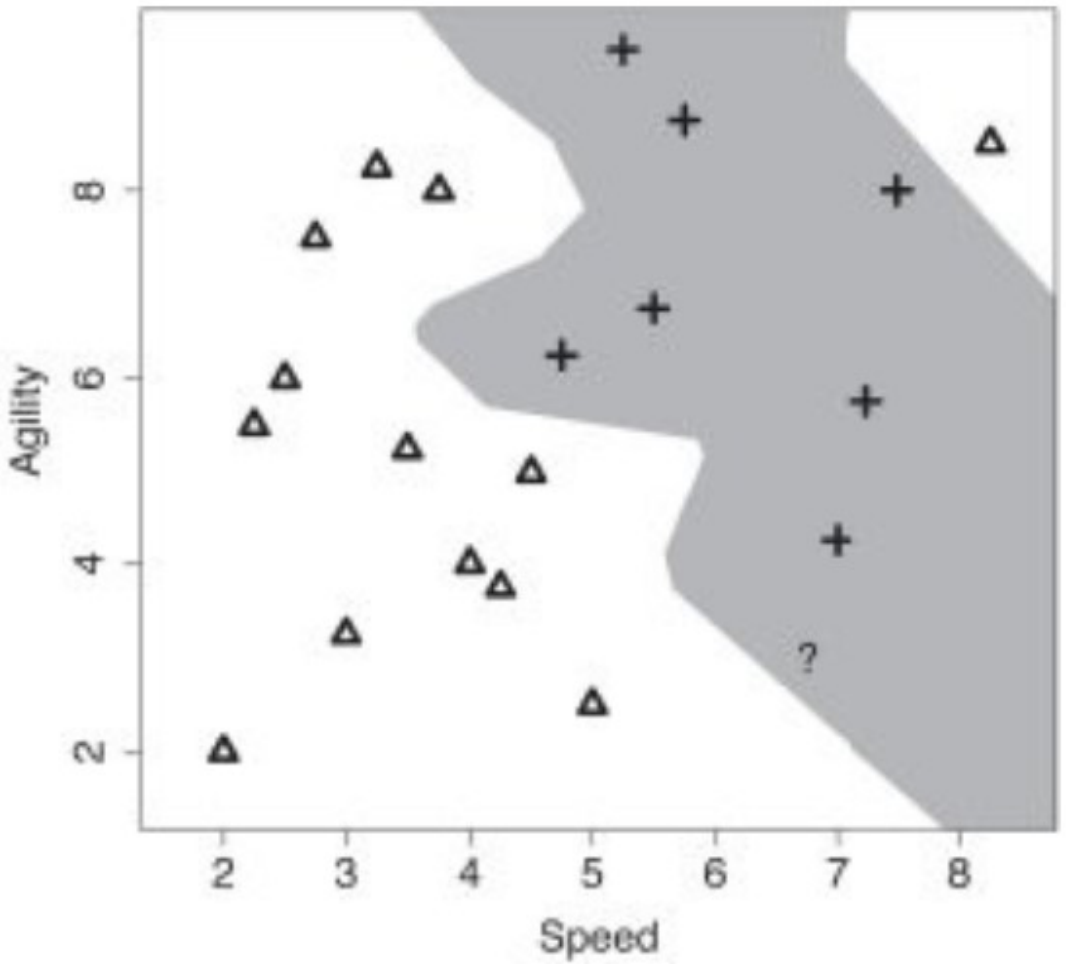
Imbalanced data

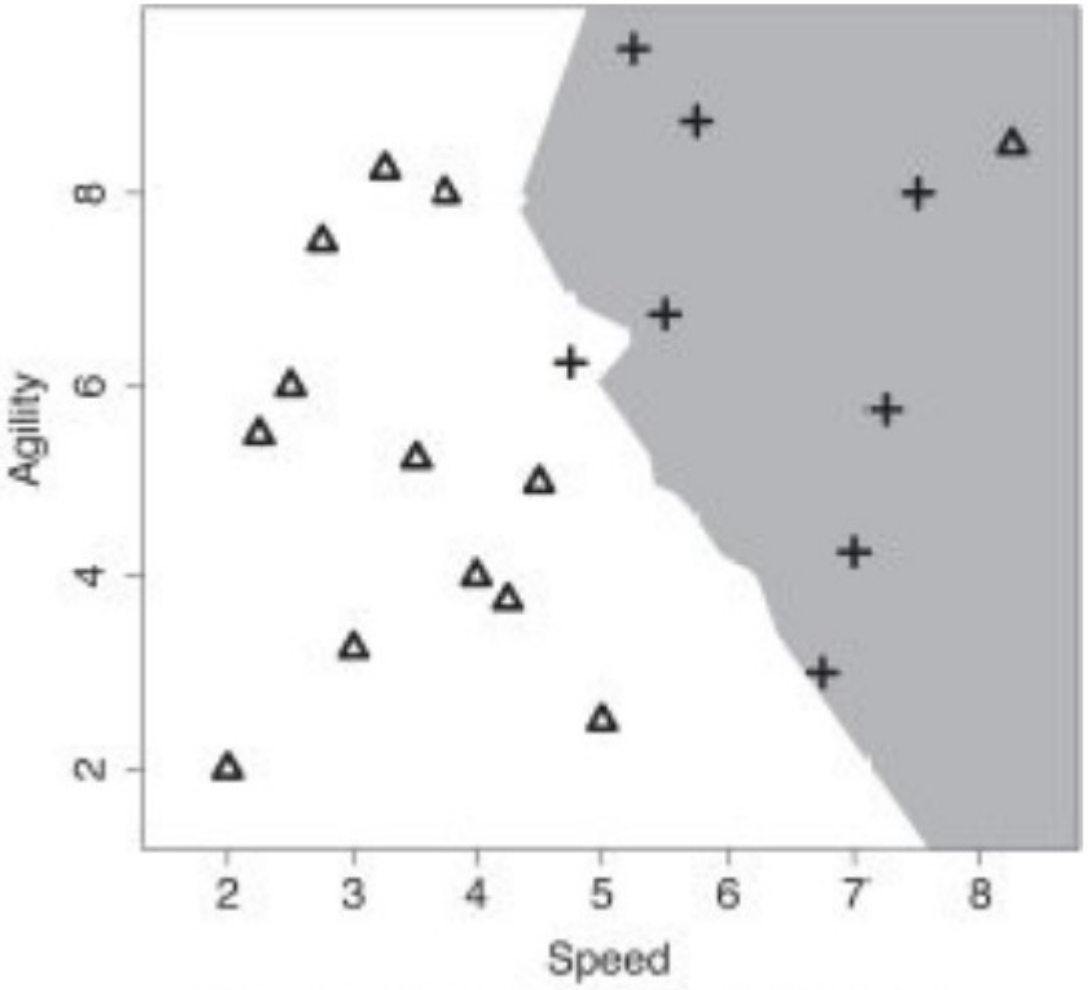


(a) Decision boundary ($k = 15$)

- The risks associated with setting k to a high value are particularly acute when we are dealing with an imbalanced dataset.
- as k increases, the majority target level begins to dominate the feature space

Noise in data





(b) Decision boundary ($k = 5$)

Distance-weighted nearest neighbor algorithm

- When a distance weighted k nearest neighbor approach is used, the contribution of each neighbor to the prediction is a function of the inverse distance between the neighbor and the query x_q
  - Give greater weight to closer neighbor

$$w \equiv \frac{1}{d(x_q, x_i)^2}$$

.

Issues with Distance-weighted knn

- if the dataset is very imbalanced, then even with a weighting applied to the contribution of the training instances, the majority target level may dominate.
- when the dataset is very large, which means that computing the reciprocal of squared distance between the query and all the training instances can become too computationally expensive to be feasible

Improvements

- Weighted Euclidean distance

$$D(c1,c2) = \sqrt{\sum_{i=1}^{N} w_i \cdot \left(attr_i(c1) - attr_i(c2)\right)^2}$$

- large weights => attribute is more important
- small weights => attribute is less important
- zero weights => attribute doesn't matter

Efficient Memory Search

- if we are working with a large dataset, the time cost in computing the distances between a query and all the training instances and retrieving the k nearest neighbors may be prohibitive.
- use k-d tree,(k-dimensional tree),.
- A k-d tree is a balanced binary tree in which each of the nodes in the tree (both interior and leaf nodes) index one of the instances in a training dataset.
- The tree is constructed so that nodes that are nearby in the tree index training instances that are nearby in the feature space

Refer for more details

- **FUNDAMENTALS OF MACHINE LEARNING FOR PREDICTIVE DATA ANALYTICS-** Algorithms, Worked Examples, and Case Studies ,John D. Kelleher, Brian Mac Namee, Aoife D'Arcy

Another approach

- speed up classification time include the use of partial distance calculations and editing the stored tuples.
- Compute the distance based on a subset of the n attributes.
- If this distance exceeds a threshold, then further computation for the given stored tuple is halted, and the process moves on to the next stored tuple.
- The editing method removes training tuples that prove useless.
- This method is also referred to as pruning or condensing because it reduces the total number of tuples stored.

Thank you !!!!!