Subject- Pattern Recognition & Machine Learning
Branch: B Tech Honour's (Data Science)

Department of CSE, UTD CSVTU, Bhilai

Introduction Pattern Recognition & Machine Learning

Shesh Narayan Sahu
Asst. Prof. CSE, UTD CSVTU

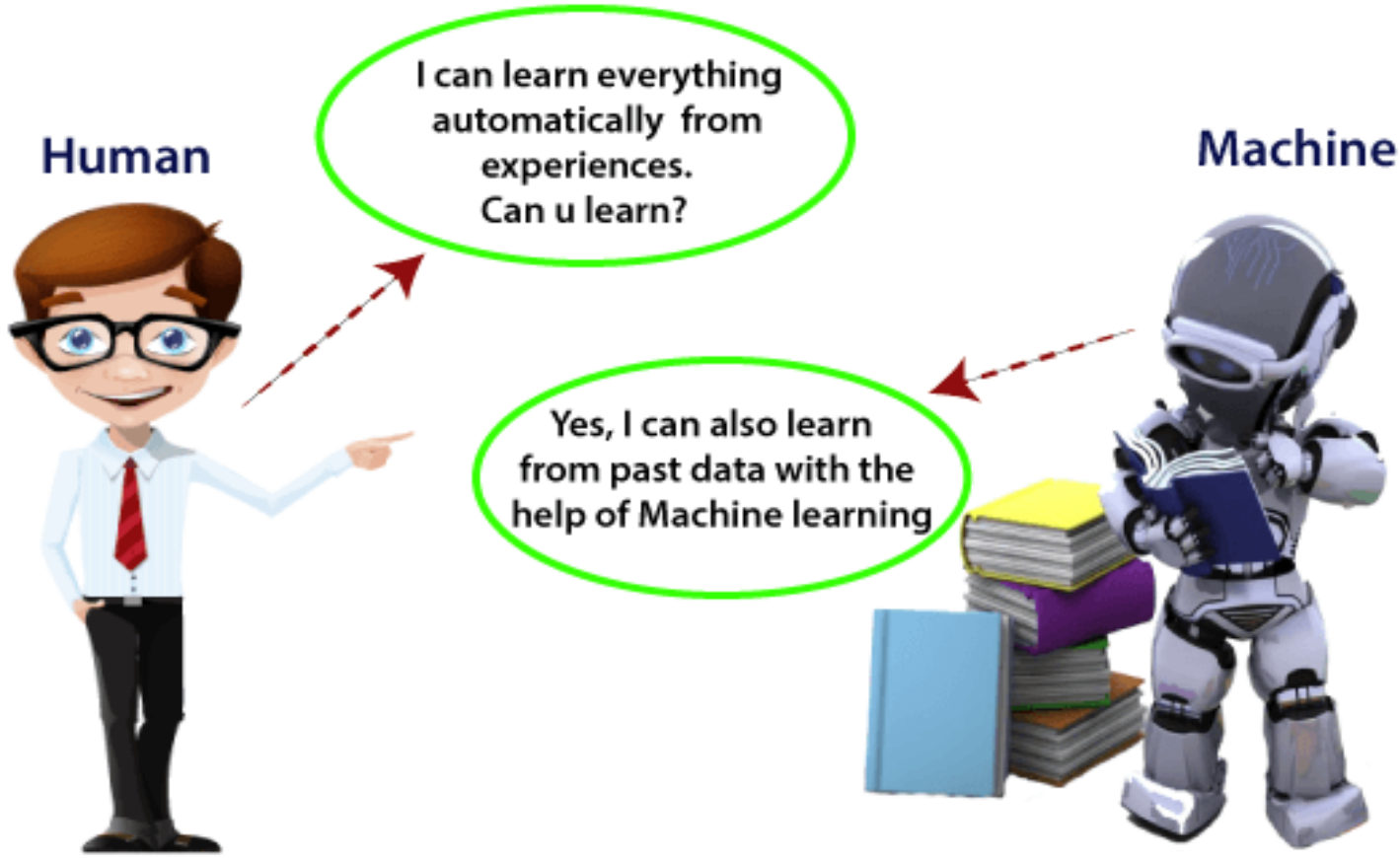
Basics of Machine Learning

- Introduction
- Concepts
- Evolution
- Definition by Tom Mitchel
- Traditional Vs ML system
- Use and classical example
- Concluding remarks

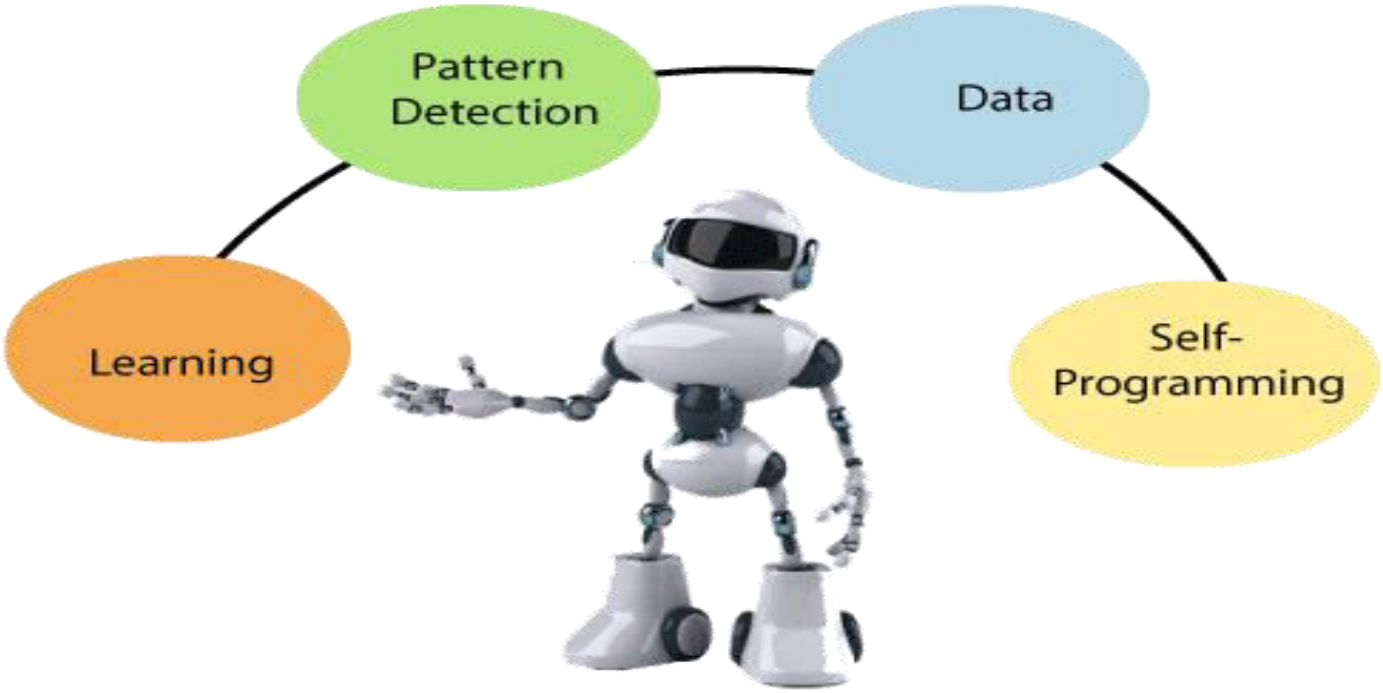
Basics

- Making machine to learn
- Learning through experience
- Statistical analysis of data
- Herbert Simon- “Learning is any process by which a system improves performance from experience”
- Machine Learning is a sub-set of artificial intelligence where computer algorithms are used to autonomously learn from data and information

Machine Concepts



Contd..



Evolution

- 1950 — Alan Turing creates the “Turing Test” to determine if a computer has real intelligence. To pass the test, a computer must be able to fool a human into believing it is also human.

- 1952 — Arthur Samuel wrote the first computer learning program. The program was the game of checkers.

- 1957 — Frank Rosenblatt designed the first neural network for computers (the perceptron), which simulate the thought processes of the human brain.

- 1967 — The “nearest neighbor” algorithm was written, allowing computers to begin using very basic pattern recognition.

Definition by Tom Mitchell (1998):

- Machine Learning is the study of algorithms that
  improve their performance P
  at some task T
  with experience E

- A well-defined learning task is given by <T, P, E>

Samuels Checker Player

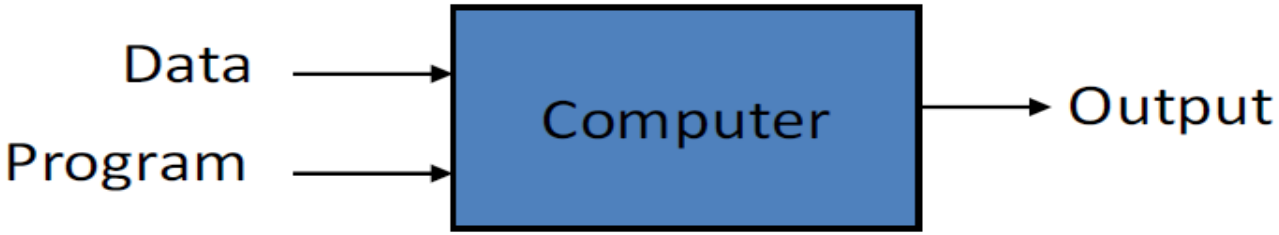
- “Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.” -Arthur Samuel (1959)

<T, P, E>

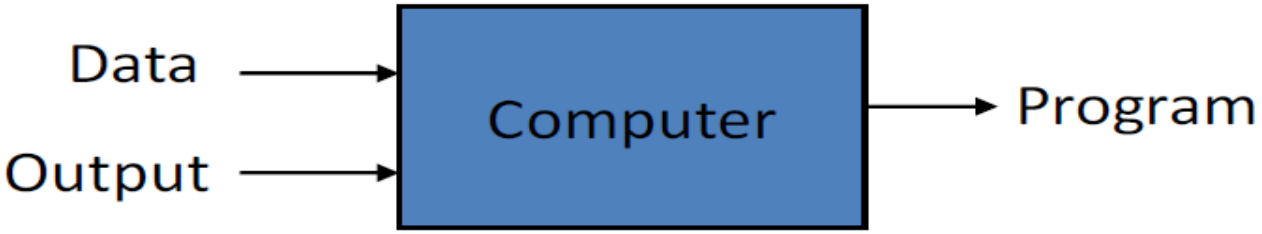
- T: Playing checkers
- P: Percentage of games won against an arbitrary opponent
- E: Playing practice games against itself

Traditional Vs ML system

Traditional Programming



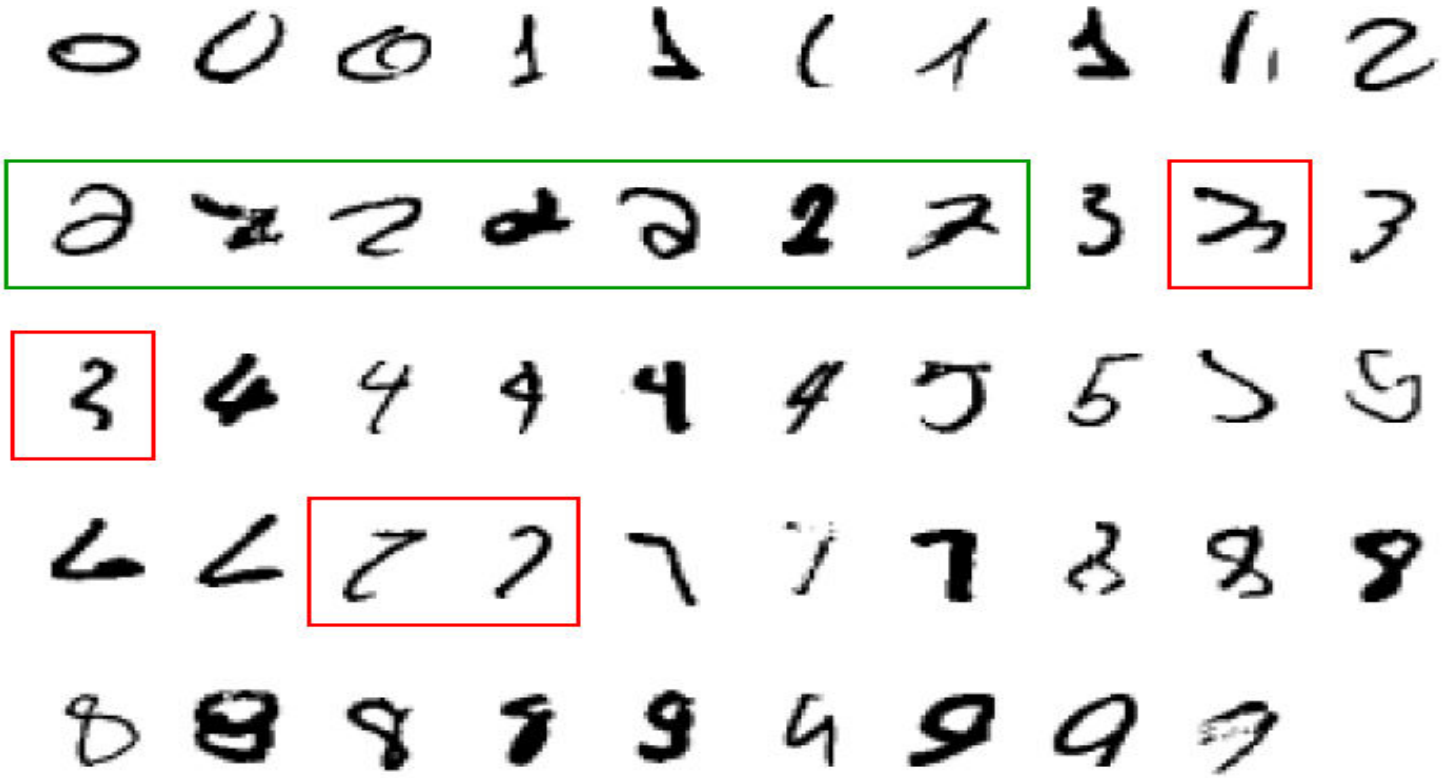
Machine Learning



ML is used when

- Human expertise does not exist (navigating on Mars)
- Humans can't explain their expertise (speech recognition)
- Models must be customized (personalized medicine)
- Models are based on huge amounts of data (genomics)

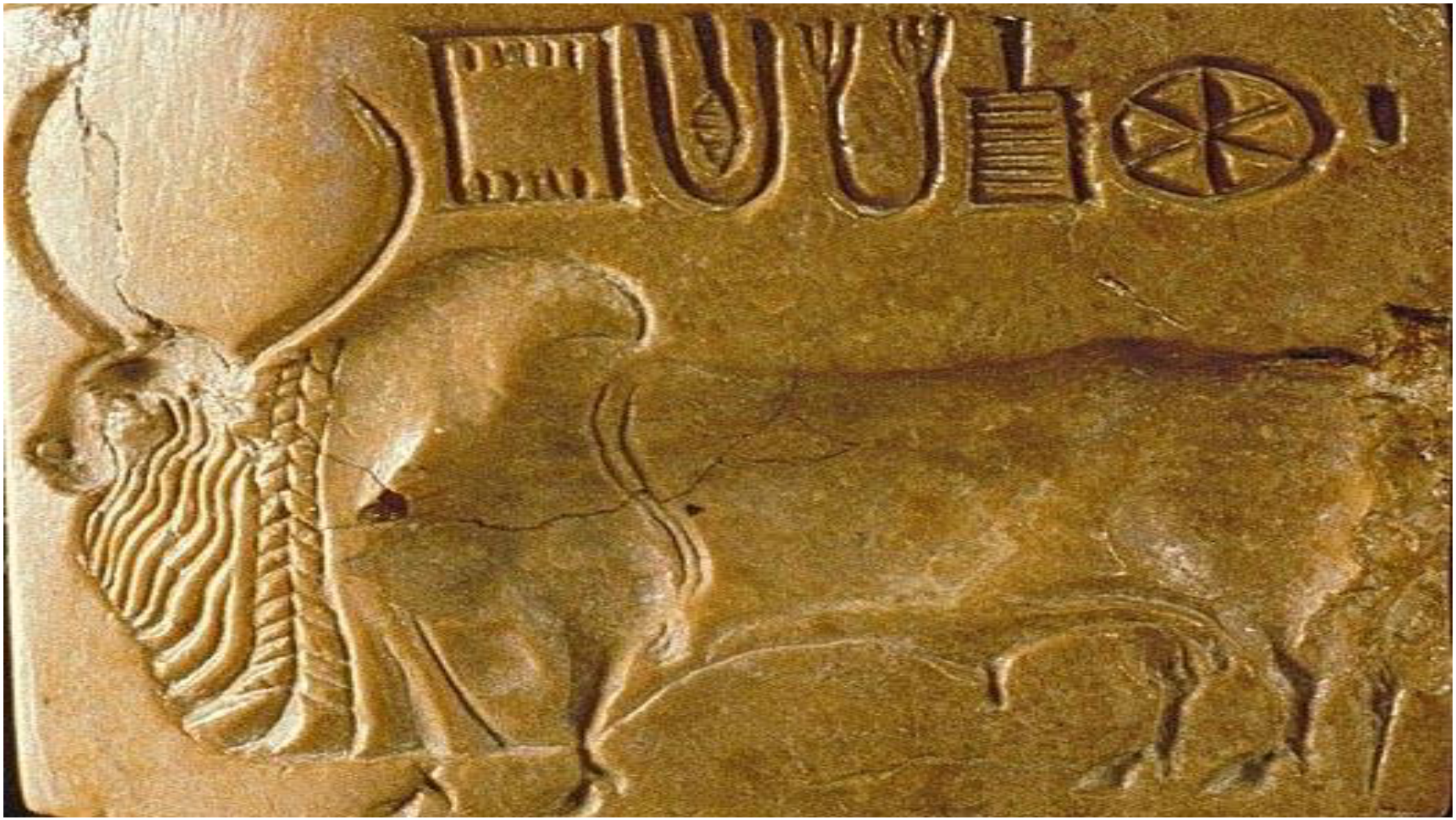
classic example of a task that requires machine learning:
It is very hard to say what makes a 2



Slide credit: Geoffrey Hinton

Contd..

Contd..



Contd..

Contd..

- Indus Valley Civilization

Contd..


na ko ta bae cho no bi ru ma poe mae ngae bo wa
ra uu to cha mwoe choe mwa ro maa ba tae pae fo chuu
bu noe tuu fa nae mwoa rae nuu sa toe wae sae kuu sho
pa ku choo ka ruu rga mwc kae tu ngo ngoa choa shu koe
shoa nguu puu toa shuu su poa nge nu fae mwi so taa mu
chu oe re ha roa ryo noa nma ya yoa yae i wo yoe
fi ki ngi ni mi wi chi pi si yo ti ri u

Some more examples of tasks

- Recognizing patterns:
  - Facial identities or facial expressions
  - Handwritten or spoken words
  - Medical images
- Generating patterns:
  - Generating images or motion sequences
- Recognizing anomalies:
  - Unusual credit card transactions
  - Unusual patterns of sensor readings in a nuclear power plant
- Dfdf

AI vs ML

- AI stands for Artificial intelligence, where intelligence is defined as the ability to acquire and apply knowledge.
- ML stands for Machine Learning which is defined as the acquisition of knowledge or skill.

- AI leads to intelligence or wisdom whereas ML leads to knowledge.

- AI's key uses include- Siri, customer service via chatbots Expert Systems, Google Translate, Intelligent humanoid robots such as Sophia.
- The most common uses of ML- Facebook's automatic friend suggestions, Google's search algorithms, Banking fraud analysis, Stock price forecast

Classification: Applications

Two causal relations:

- Strike of the railway *can* cause Norman to be late
- Strike of the railway *can* cause Martin to be late

IMPORTANT:

These relations are **NOT** absolute!!

Strike of the railway does **NOT** guarantee that Norman and Martin will be late for sure. It **ONLY** increases the probability (chance) of lateness.

What is BBN:

Bayesian Belief Network (**BBN**) is a directed acyclic graph associated with a set of conditional probability distributions.

BBN is a set of nodes connected by directed edges in which:

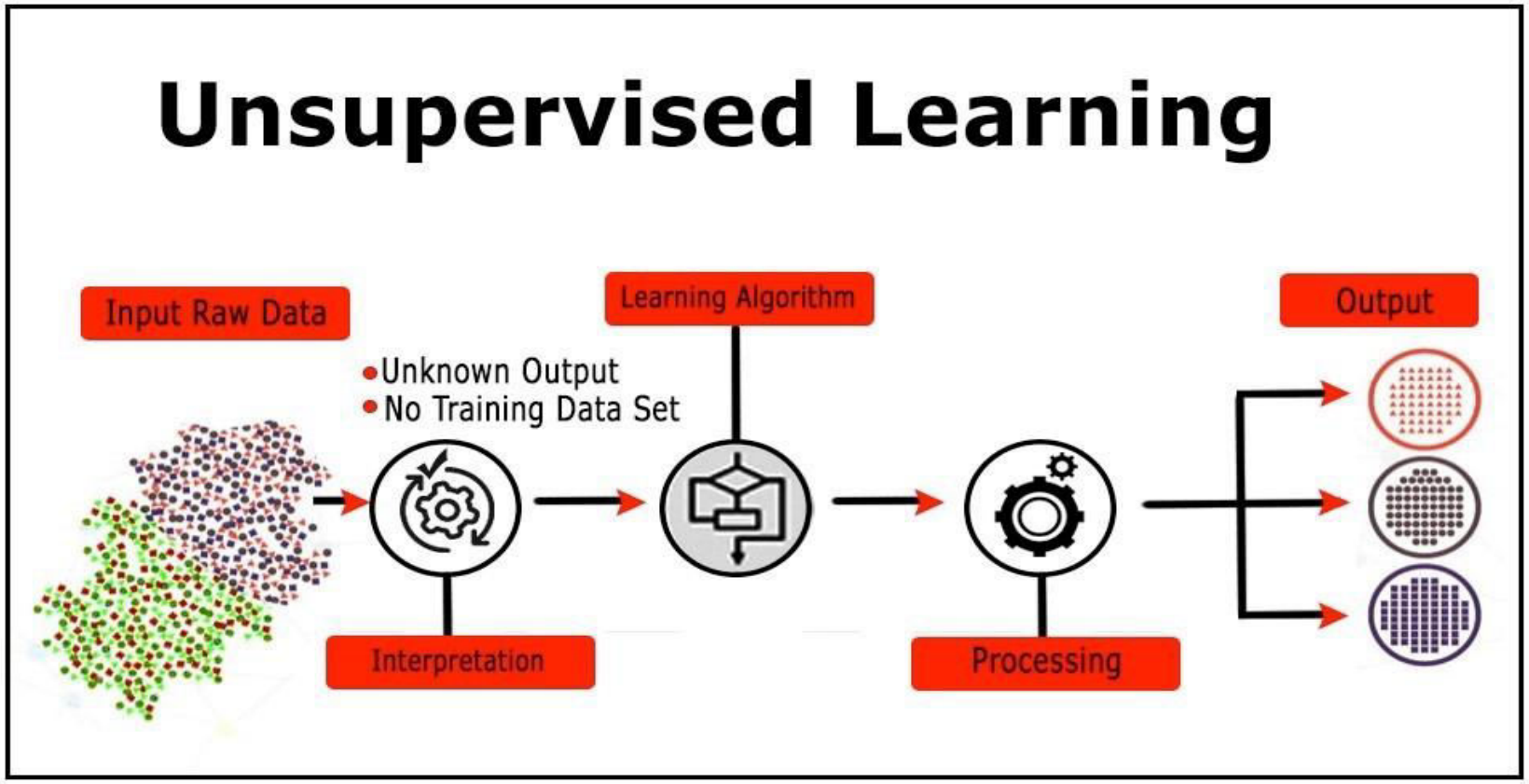
- ***nodes*** represent discrete or continuous random variables in the problem studied,
- ***directed edges*** represent direct or causal relationships between variables and do not form *cycles*,
- each node is associated with a ***conditional probability distribution*** which quantitatively expresses the strength of the relationship between that node and its parents.

So BN = (DAG, CPD)

- DAG: directed acyclic graph (BN's structure)
  - Nodes: random variables (typically binary or discrete, but methods also exist to handle continuous variables)
- Arcs: indicate probabilistic dependencies between nodes (lack of link signifies conditional independence)

- CPD: conditional probability distribution (BN's parameters)
  - Conditional probabilities at each node, usually stored as a table (conditional probability table, or CPT)

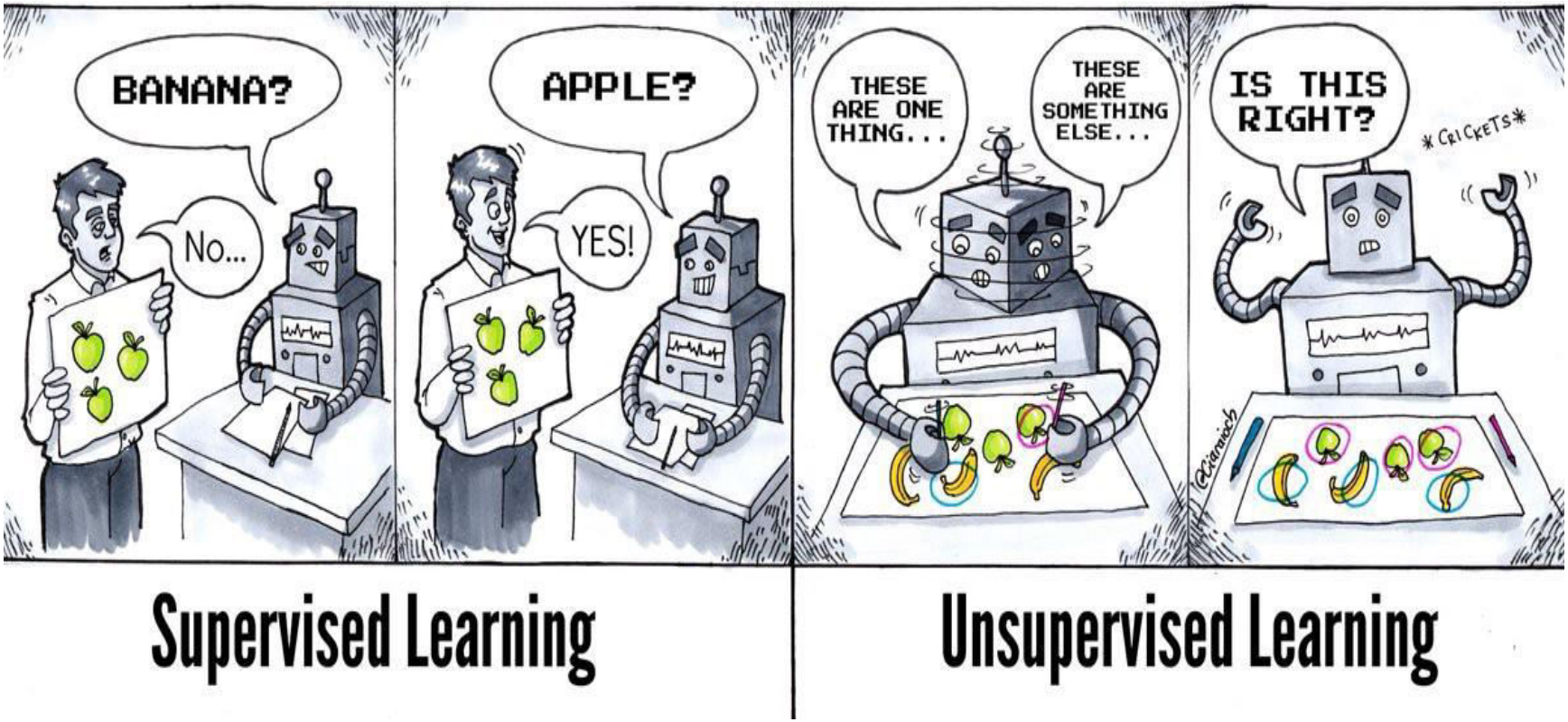
Contd..


Unsupervised Learning
Input Raw Data
Learning Algorithm
Output
Unknown Output
No Training Data Set
Interpretation
Processing

Clustering Models-

- K- means clustering
- Hierarchical clustering
- Single linkage clustering
- Average linkage clustering
- Complete linkage clustering
- DBSCAN clustering etc.

Supervised Vs Unsupervised



Reinforcement Learning

- Topics:
  - Policies: what actions should an agent take in a particular situation
  - Utility estimation: how good is a state (→used by policy)
- No supervised output but delayed reward
- Credit assignment problem (what was responsible for the outcome)
- Applications:
  - Game playing
  - Robot in a maze
  - Multiple agents, partial observability, ...

Resources: Journals

- Journal of Machine Learning Research www.jmlr.org
- Machine Learning
- IEEE Transactions on Neural Networks
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Annals of Statistics
- Journal of the American Statistical Association
- ...

Resources: Conferences

- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Neural Information Processing Systems (NIPS)
- Computational Learning
- International Joint Conference on Artificial Intelligence (IJCAI)
- ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)
- IEEE Int. Conf. on Data Mining (ICDM)

Lecture 2

Bayesian Learning

Shesh N Sahu

Asst. Prof. CSE, UTD

Contents

- Bayesian probability
- Bayes theorem
- MAP
- Bayesian network
- Joint probability
- Construction of BBN
- Conditional Independence
- Concluding remarks

Bayesian probability

- Bayesian probability is the notion of probability which talks about partial beliefs

- Bayesian estimation calculates the validity of a propositions.
- It is calculated based on two attribute
  - Prior estimate
  - New relevant evidence

- Based on above posterior bayes estimation can be calculated.
- Key to these a Bayes theorem can be proposed.

Bayes theorem

- It is a mathematical formula for determining conditional probability.
- Conditional probability is the likelihood of an outcome occurring, based on a previous outcome having occurred in similar circumstances.
- Prior probability, in Bayesian statistical inference, is the probability of an event occurring before new data is collected.
- Posterior probability is the revised probability of an event occurring after taking into consideration the new information.

Contd..

Bayes Rule: $$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- P(h) = prior probability of hypothesis h
- P(D) = prior probability of training data D
- P(h|D) = probability of h given D (posterior density)
- P(D|h) = probability of D given h (likelihood of D given h)

Maximum A posterior hypothesis (MAP)

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

The Goal of Bayesian Learning: the most probable hypothesis given the training data (Maximum A Posteriori hypothesis)

$$h_{MAP} = \arg\max_{h \in H} P(h \mid D)$$

$$= \arg\max_{h \in H} \frac{P(D \mid h)P(h)}{P(D)}$$

$$= \arg\max_{h \in H} P(D \mid h)P(h)$$

Bayesian network

- It is compact and expressive graphical representation of events/variable.
- Represents the dependence/independence among events/variable.
- Node represents events/variable and arc represents dependence.
- The structure of the graph shows the conditional independence.
- The graph is DAG, that directed acyclic graph.

Joint probability distribution

- Bayesian network represent joint probability distribution of variable efficiently.
- A joint probability distribution represents a probability distribution for two or more random variables.

Let,

$X: \{x1,\ x2,\ x3,\ x_n\}$

Then joint probability disributon can be given by

$$p\,(x1,\ x2,\ x3,\ \ldots..x_n) = p\,(x1).\ p(x2/x1).\ p\,(x3/x1\ x2).\ p(x_n/\ x1\ x2\ x_{n-1})$$

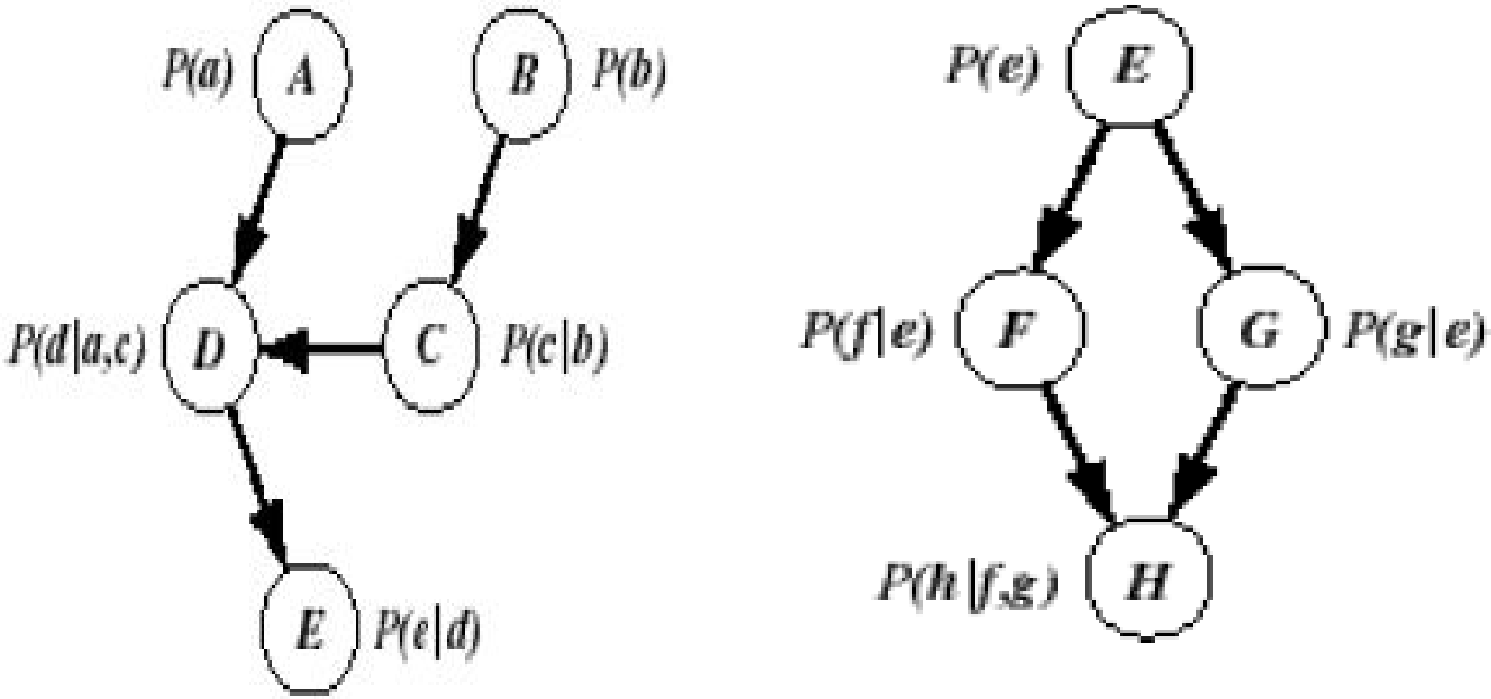
BN: Conditional independence

- Two components in a Bayesian network-
  - Graph components
  - Numerical probabilities (CPT)
  - BBN= (DAG, CPT)
- Bayesian network consider the conditional independence of events/variable.

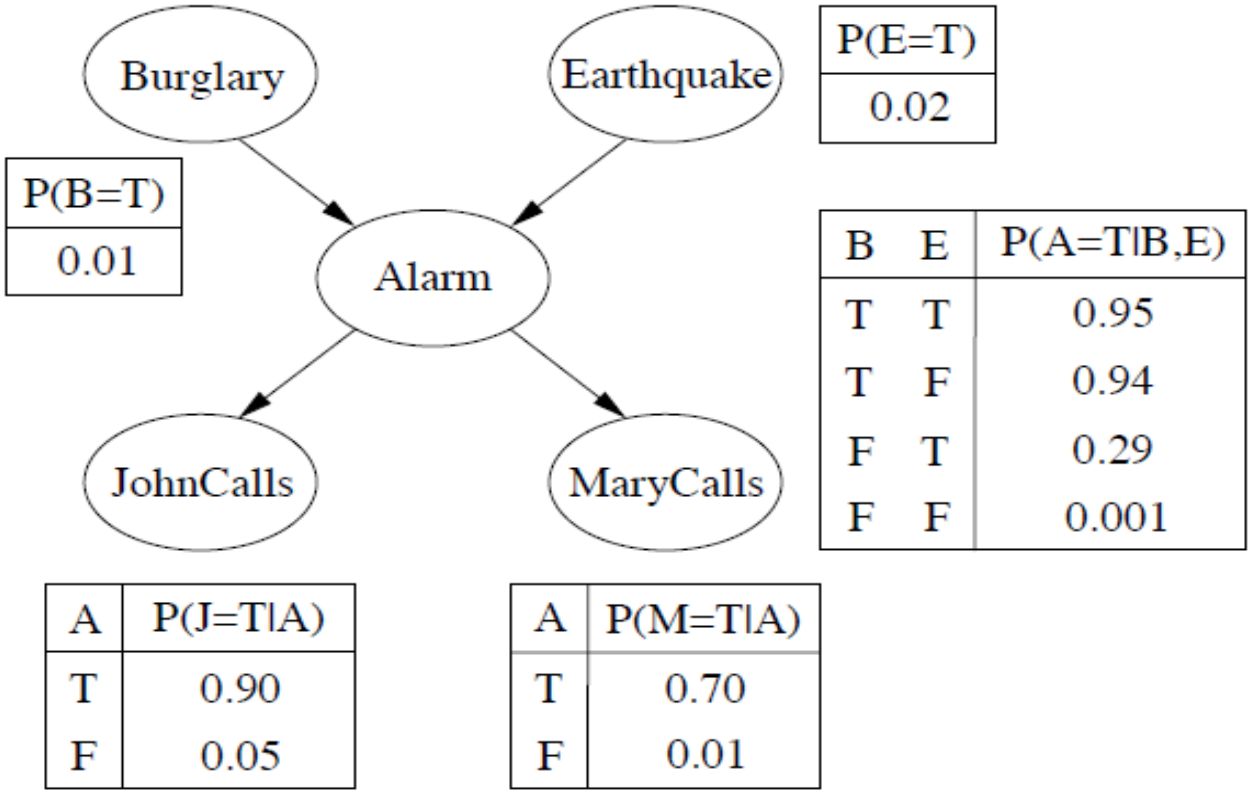
- Each node/variable is conditionally independent of its non descendants given its immediate parents

$$p\,(x1,\ x2,\ x3,\ \ldots\ldots x_n) = \Pi p\left(x_i / parents\left(x_i\right)\right)$$

Find the probability in BN



BBN: Problem 1



P(B=T)
0.01

P(E=T)
0.02

B	E	P(A=T|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J=T|A)
T	0.90
F	0.05

A	P(M=T|A)
T	0.70
F	0.01

Contd..

- What is the probability that the alarm has sounded but neither a burglary nor an earthquake has occurred, and both John and Marry call?

Applications of Bayesian Networks

- Machine learning
- Statistics
- Computer vision
- Natural language processing
- Speech recognition
- Error-control codes
- Bioinformatics
- Medical diagnosis
- Weather forecasting

What is BBN:

Bayesian Belief Network (**BBN**) is a directed acyclic graph associated with a set of conditional probability distributions.

BBN is a set of nodes connected by directed edges in which:

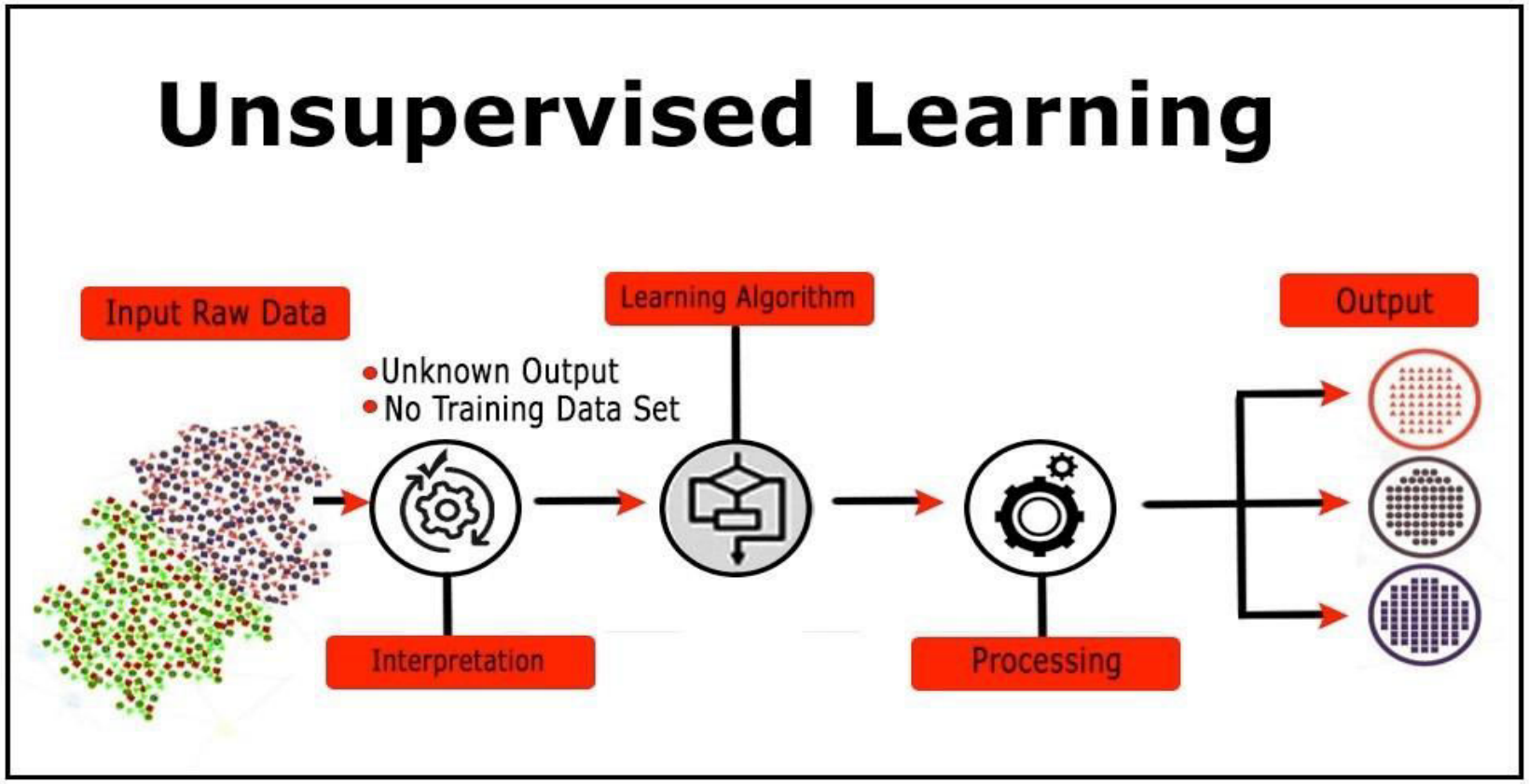
- ***nodes*** represent discrete or continuous random variables in the problem studied,
- ***directed edges*** represent direct or causal relationships between variables and do not form *cycles*,
- each node is associated with a ***conditional probability distribution*** which quantitatively expresses the strength of the relationship between that node and its parents.

So BN = (DAG, CPD)

- DAG: directed acyclic graph (BN's structure)
  - Nodes: random variables (typically binary or discrete, but methods also exist to handle continuous variables)
- Arcs: indicate probabilistic dependencies between nodes (lack of link signifies conditional independence)

- CPD: conditional probability distribution (BN's parameters)
  - Conditional probabilities at each node, usually stored as a table (conditional probability table, or CPT)

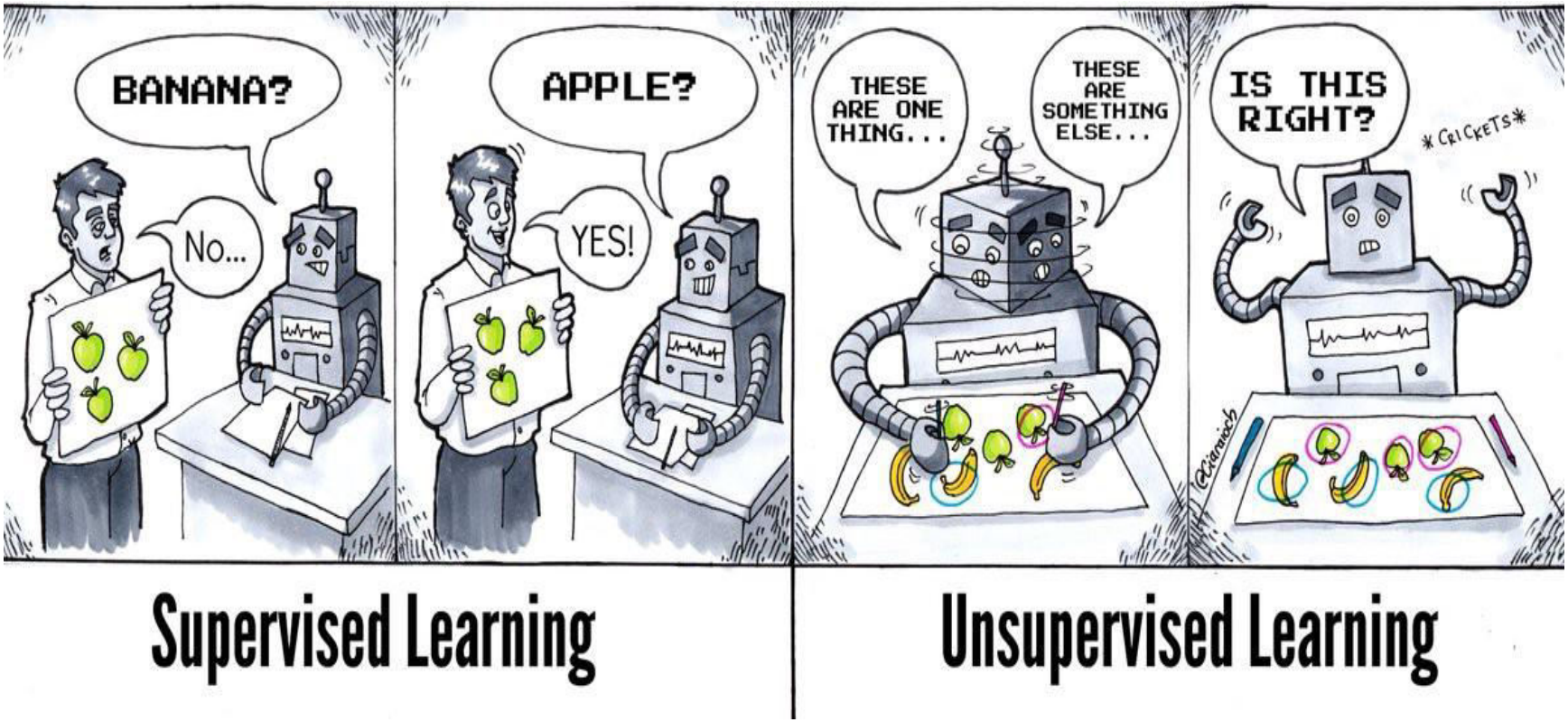
Contd..


Unsupervised Learning
Input Raw Data
Learning Algorithm
Output
Unknown Output
No Training Data Set
Interpretation
Processing

Clustering Models-

- K- means clustering
- Hierarchical clustering
- Single linkage clustering
- Average linkage clustering
- Complete linkage clustering
- DBSCAN clustering etc.

Supervised Vs Unsupervised



Reinforcement Learning

- Topics:
  - Policies: what actions should an agent take in a particular situation
  - Utility estimation: how good is a state (→used by policy)
- No supervised output but delayed reward
- Credit assignment problem (what was responsible for the outcome)
- Applications:
  - Game playing
  - Robot in a maze
  - Multiple agents, partial observability, ...

Resources: Journals

- Journal of Machine Learning Research www.jmlr.org
- Machine Learning
- IEEE Transactions on Neural Networks
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Annals of Statistics
- Journal of the American Statistical Association
- ...

Resources: Conferences

- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Neural Information Processing Systems (NIPS)
- Computational Learning
- International Joint Conference on Artificial Intelligence (IJCAI)
- ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)
- IEEE Int. Conf. on Data Mining (ICDM)

Lecture 3

Bayesian Learning
Naïve Bayes

Shesh N Sahu

Asst. Prof. CSE, UTD

Contents

- Bayesian probability
- Bayes theorem
- MAP
- Bayesian network
- Joint probability
- Construction of BBN
- Conditional Independence
- Concluding remarks

Bayesian probability

- Bayesian probability is the notion of probability which talks about partial beliefs

- Bayesian estimation calculates the validity of a propositions.
- It is calculated based on two attribute
  - Prior estimate
  - New relevant evidence

- Based on above posterior bayes estimation can be calculated.
- Key to these a Bayes theorem can be proposed.

Bayes theorem

- It is a mathematical formula for determining conditional probability.
- Conditional probability is the likelihood of an outcome occurring, based on a previous outcome having occurred in similar circumstances.
- Prior probability, in Bayesian statistical inference, is the probability of an event occurring before new data is collected.
- Posterior probability is the revised probability of an event occurring after taking into consideration the new information.

Contd..

Bayes Rule: $$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- P(h) = prior probability of hypothesis h
- P(D) = prior probability of training data D
- P(h|D) = probability of h given D (posterior density)
- P(D|h) = probability of D given h (likelihood of D given h)

Maximum A posterior hypothesis (MAP)

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

The Goal of Bayesian Learning: the most probable hypothesis given the training data (Maximum A Posteriori hypothesis)

$$h_{MAP} = \arg\max_{h \in H} P(h \mid D)$$

$$= \arg\max_{h \in H} \frac{P(D \mid h)P(h)}{P(D)}$$

$$= \arg\max_{h \in H} P(D \mid h)P(h)$$

Joint probability distribution

- Bayesian network represent joint probability distribution of variable efficiently.
- A joint probability distribution represents a probability distribution for two or more random variables.

Let,

$X: \{x1,\ x2,\ x3,\ x_n\}$

Then joint probability disributon can be given by

$$p\,(x1,\ x2,\ x3,\ \ldots..x_n) = p\,(x1).\ p(x2/x1).\ p\,(x3/x1\ x2).\ p(x_n/\ x1\ x2\ x_{n-1})$$

Naïve Bayes

- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.
- The fundamental Naive Bayes assumption is that each feature makes an: independent and equal.

Data set I

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Contd..

- Given a new instance, predict its label?

I. X= (Outlook= Sunny, Temperature= Cool, Humidity= High, Wind= Strong)

- P(Yes/X) = ?
- P(No/X) = ?

II. Today = (Outlook= Sunny, Temperature= Hot, Humidity= Normal, Wind= False)

Contd..

- Given a new instance, predict its label?

I. X= (Outlook= Sunny, Temperature= Cool, Humidity= High, Wind= Strong)

- P(X/Yes) = 0.0053
- P(X/No) = 0.0206

Applications of Bayesian Networks

- Machine learning
- Statistics
- Computer vision
- Natural language processing
- Speech recognition
- Error-control codes
- Bioinformatics
- Medical diagnosis
- Weather forecasting

What is BBN:

Bayesian Belief Network (**BBN**) is a directed acyclic graph associated with a set of conditional probability distributions.

BBN is a set of nodes connected by directed edges in which:

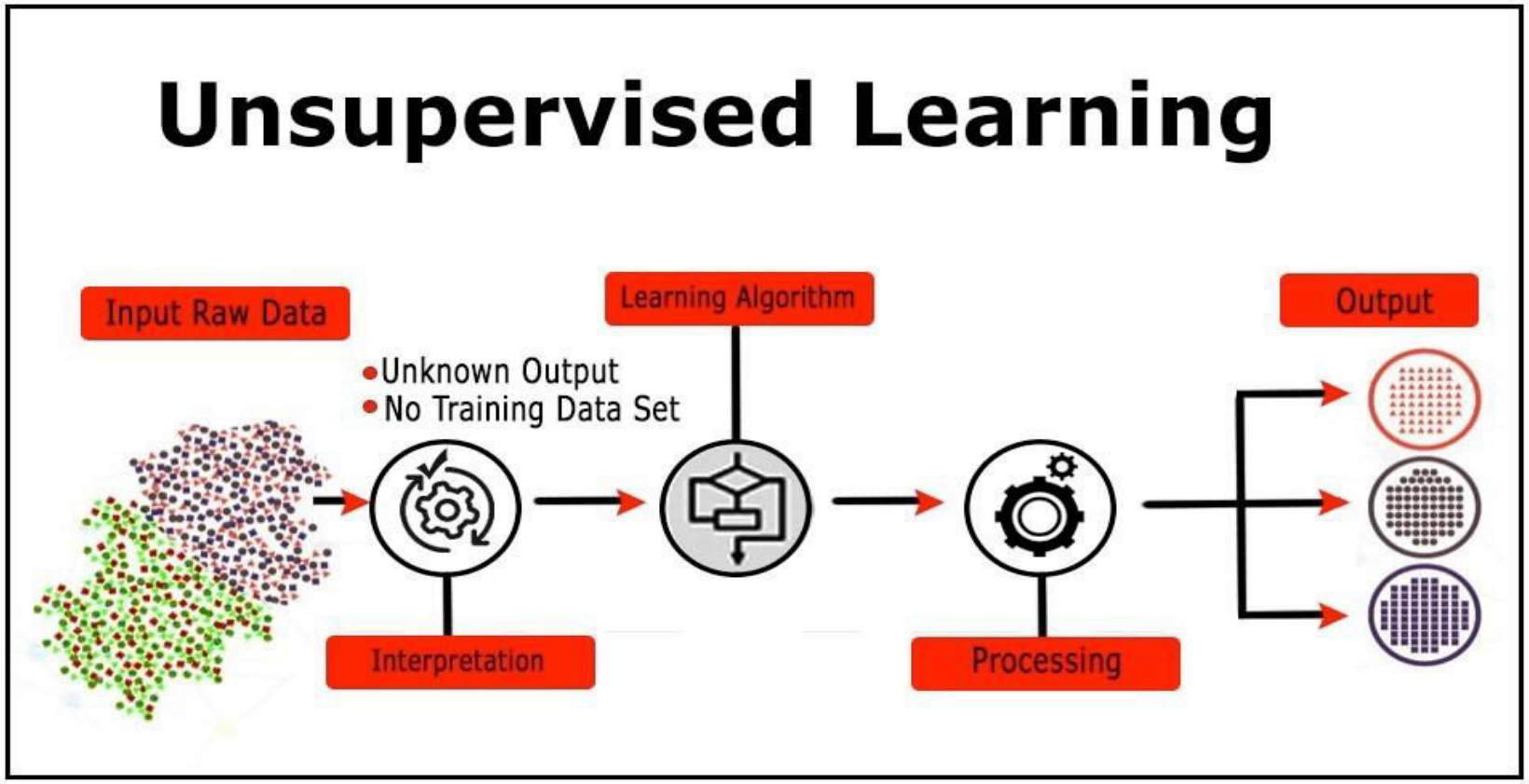
- ***nodes*** represent discrete or continuous random variables in the problem studied,
- ***directed edges*** represent direct or causal relationships between variables and do not form *cycles*,
- each node is associated with a ***conditional probability distribution*** which quantitatively expresses the strength of the relationship between that node and its parents.

So BN = (DAG, CPD)

- DAG: directed acyclic graph (BN's structure)
  - Nodes: random variables (typically binary or discrete, but methods also exist to handle continuous variables)
- Arcs: indicate probabilistic dependencies between nodes (lack of link signifies conditional independence)

- CPD: conditional probability distribution (BN's parameters)
  - Conditional probabilities at each node, usually stored as a table (conditional probability table, or CPT)

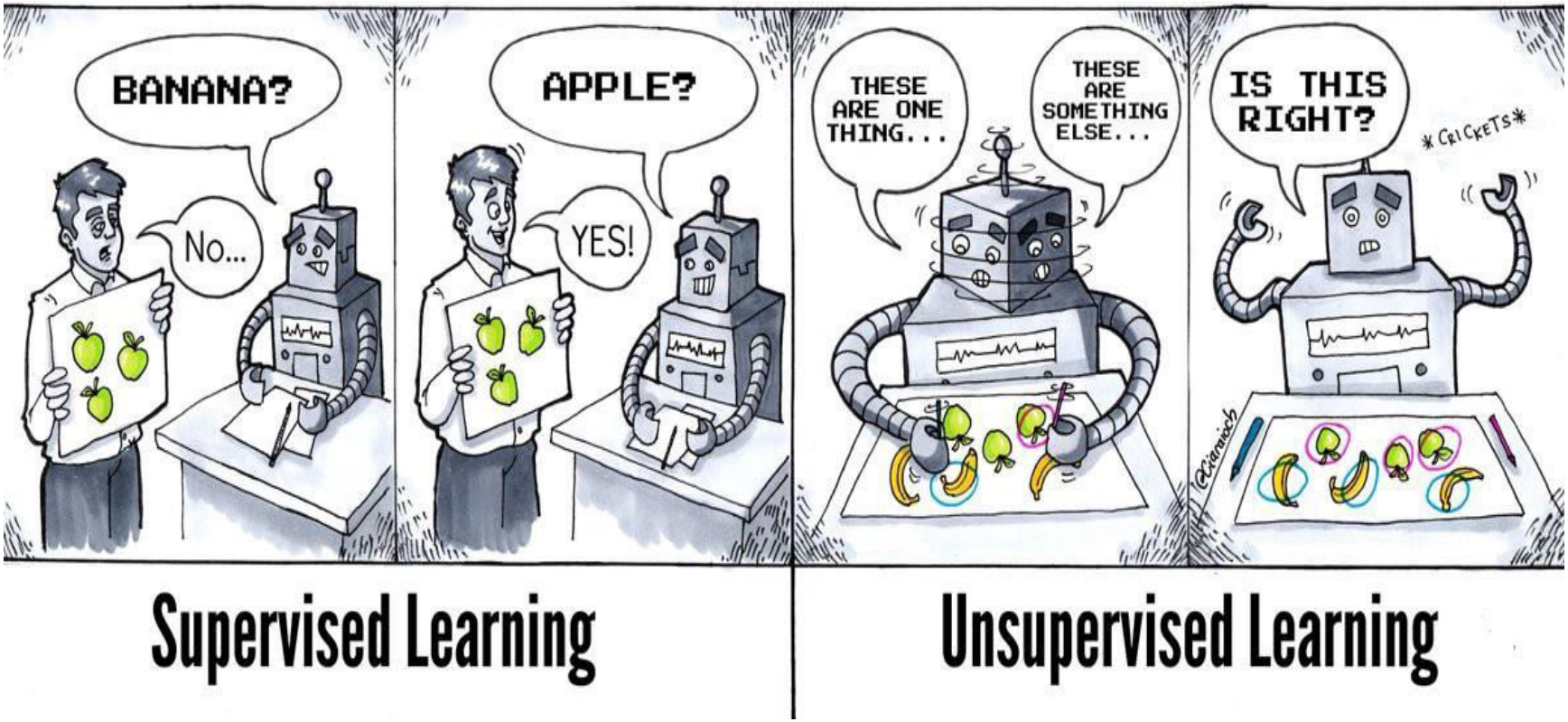
Contd..


Unsupervised Learning
Input Raw Data
Unknown Output
No Training Data Set
Learning Algorithm
Output
Interpretation
Processing

Clustering Models-

- K- means clustering
- Hierarchical clustering
- Single linkage clustering
- Average linkage clustering
- Complete linkage clustering
- DBSCAN clustering etc.

Supervised Vs Unsupervised



Reinforcement Learning

- Topics:
  - Policies: what actions should an agent take in a particular situation
  - Utility estimation: how good is a state (→used by policy)
- No supervised output but delayed reward
- Credit assignment problem (what was responsible for the outcome)
- Applications:
  - Game playing
  - Robot in a maze
  - Multiple agents, partial observability, ...

Resources: Journals

- Journal of Machine Learning Research www.jmlr.org
- Machine Learning
- IEEE Transactions on Neural Networks
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Annals of Statistics
- Journal of the American Statistical Association
- ...

Resources: Conferences

- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Neural Information Processing Systems (NIPS)
- Computational Learning
- International Joint Conference on Artificial Intelligence (IJCAI)
- ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)
- IEEE Int. Conf. on Data Mining (ICDM)

Lecture 4

Statistical Analysis

Shesh N Sahu

Asst. Prof. CSE, UTD

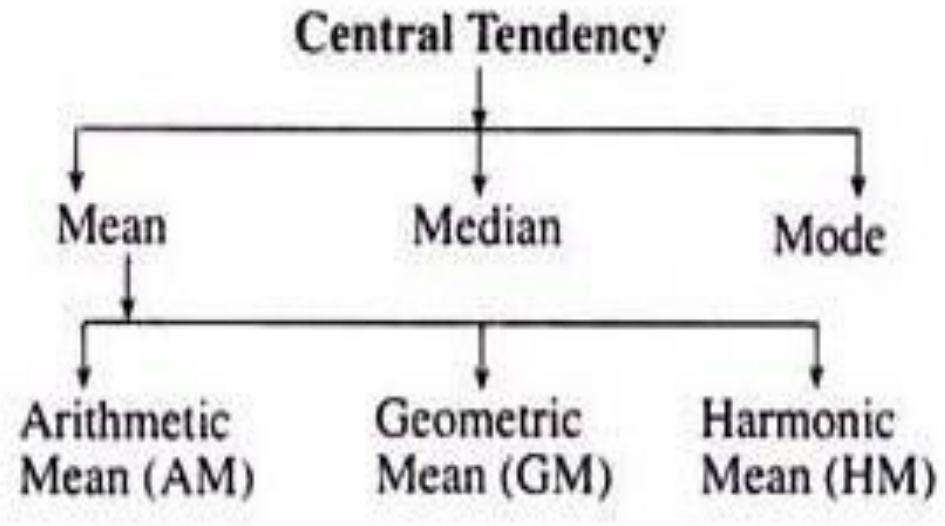
Contents

- Measure of central tendency
- Types
- Measure of dispersion
- Standard Deviation
- SD Problem
- Correlation and Regression
- Conditional Independence
- Concluding remarks

Measure of central tendency

- In any set of observation, an average denotes describes the entire series of observations.
- Since it occupies central position, so observation are larger or smaller than it, averages are also known as measure of central tendency.
- There are 3 measures of central tendency- mean, median and mode.

Types


Central Tendency
Mean
Median
Mode
Arithmetic Mean (AM)
Geometric Mean (GM)
Harmonic Mean (HM)

Contd..

- AM: It is a simplest and easiest to understand and fulfills all the criteria of a satisfactory average.

$$A = \frac{1}{n} \sum_{i=1}^{n} a_i$$

A = arithmetic mean
n = number of values
a_i = data set values

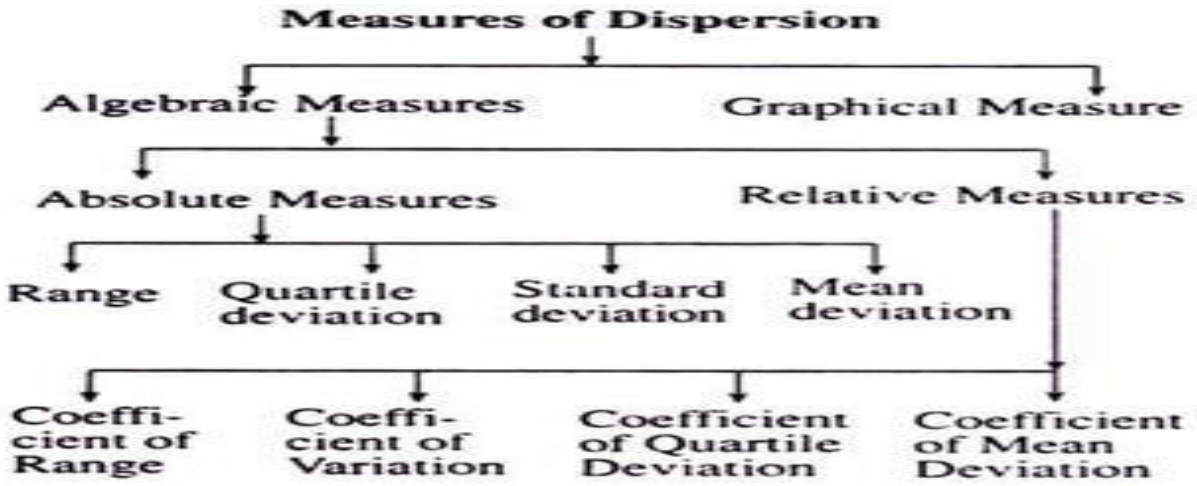
- GM: It is rigidly defined. It is difficult to calculate.

- HM: It is the reciprocal of the AM of reciprocals of observations.

Measure of dispersion

- Measure of dispersion is used to denote the degree of heterogeneity in the data.

- A measure of dispersion is designed to state numerically the extents to which individual observation vary on the average.



Standard deviation (SD)

- Standard deviation (SD): It is rigidly defined, based on all observations, calculated fairy easily.

- It is most suitable among all different measures and also least affected by sampling fluctuation.

$$\sigma = \sqrt{\frac{\sum (x - \text{mean})^2}{n}}$$

x is a set of numbers

mean is the average of the set of numbers

n is the size of the set

σ is the standard deviation

SD: Example

	City A Forecast	City B Forecast
Monday	95	90
Tuesday	93	81
Wednesday	95	95
Thursday	94	91
Friday	96	86
Saturday	94	82
Sunday	95	78

- SD of city A= 0.89
- SD of city B= 5.7

S.D.: Properties

- SD is independent of the change of origin, i.e. if y=x-c, where c is constant then,

  SD of x= SD pf y

This implies that the same SD will be obtained if each of the observations is increased or decreased by constant.

- If two variable x and z are related that z= ax+b for each x= xi, where a and b are constant then

  SD of x= lxl. SD of x

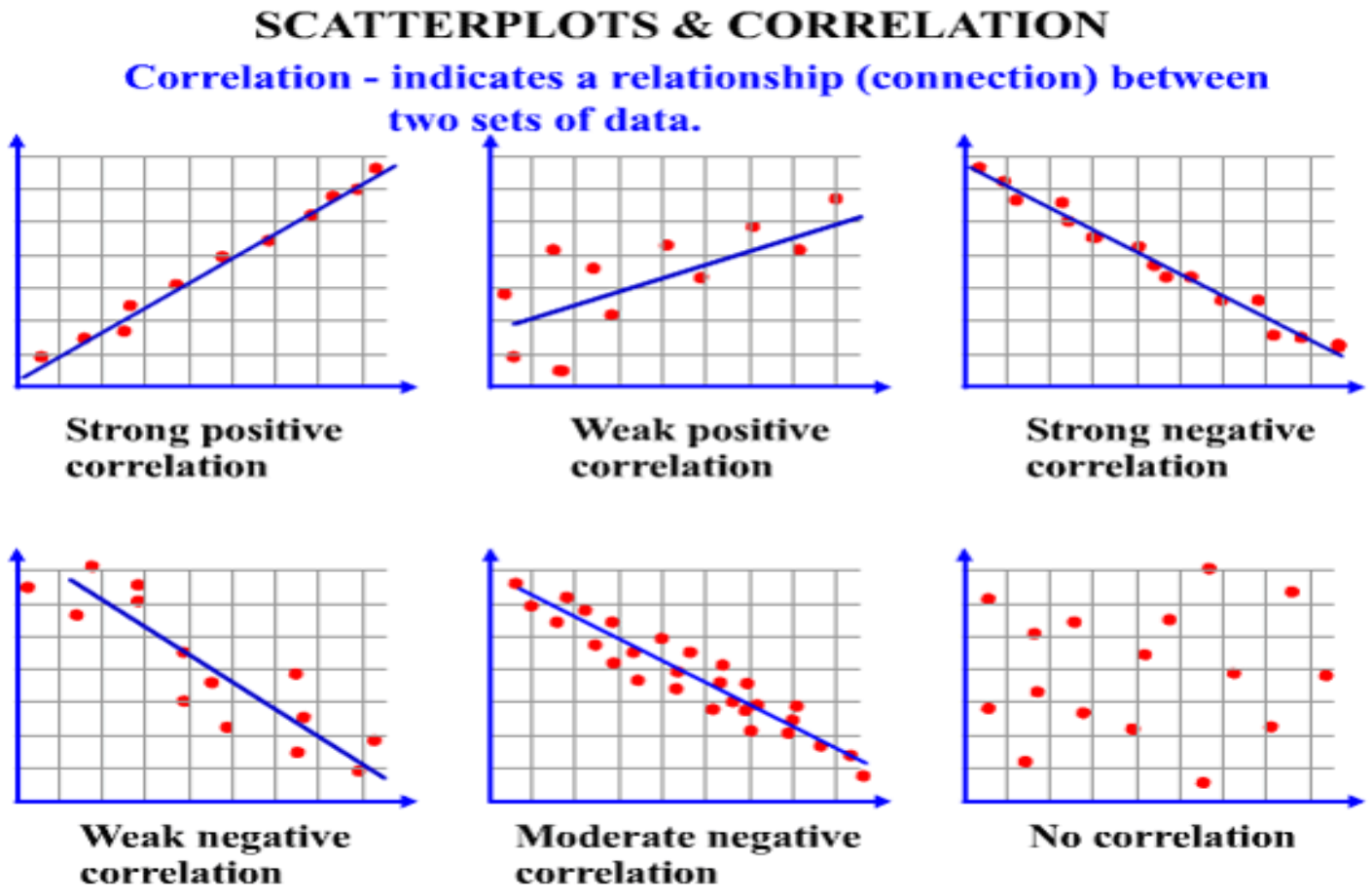
It implies that SD does not dependent on origin but dependent upon scale of measurements.

- Variance: It is represented as Sigma. It is square of SD.

Correlation and Regression

- Correlation is concerned with the measurement of the "strength of association" between variable.
- Regression is concerned with the "prediction" of the most likely value of one variable when the value of other variable is known.
- Scatter diagram indicates the nature of association between the two variable i.e. the types correlation between them.

Contd..


SCATTERPLOTS & CORRELATION
Correlation - indicates a relationship (connection) between two sets of data.
Strong positive correlation
Weak positive correlation
Strong negative correlation
Weak negative correlation
Moderate negative correlation
No correlation

Covariance

- Given a set of pairs of observation (x1, y1), (x2, y2)..(xn, yn)
- The covariance is expressed as

$$\operatorname{cov}(X, Y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}.$$

- This form is generally used for calculation. It has properties similar to those of variance i.e. square of SD.
- Correlation coefficient (r): Linear correlation can be measured by correlation coefficient.

$$Correlation = \frac{Cov(x, y)}{\sigma x * \sigma y}$$

- -1<=r<= +1

Contd..

- Correlation coefficient does not depends on the origin or scale of observation.

Regression

- The word “regression” is used to denote estimation or prediction of the average value of one variable for a specified value of other variable.
- In a linear regression the relationship between the variable is assumed to be linear.

Data set I

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Resources: Conferences

- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Neural Information Processing Systems (NIPS)
- Computational Learning
- International Joint Conference on Artificial Intelligence (IJCAI)
- ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)
- IEEE Int. Conf. on Data Mining (ICDM)

Lecture 5

Statistical Analysis

Shesh N Sahu

Asst. Prof. CSE, UTD

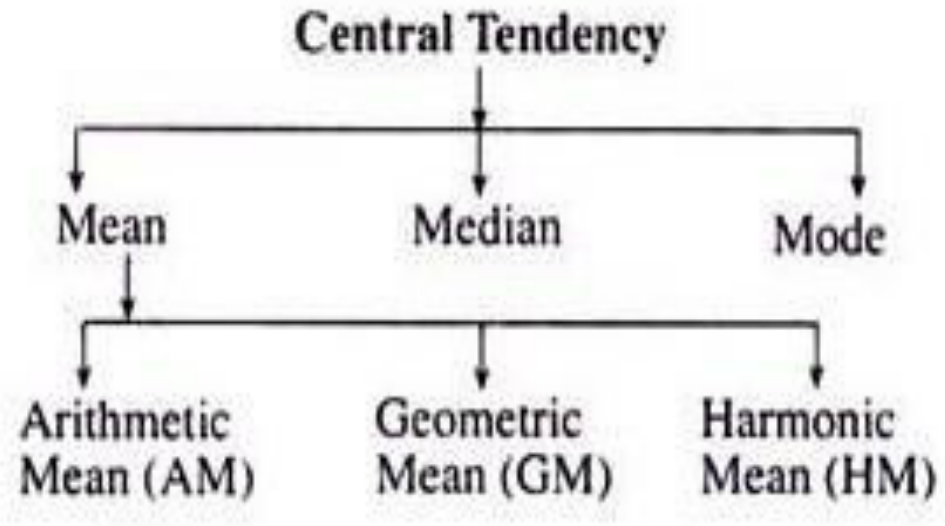
Contents

- Measure of central tendency
- Moment
- Central moment
- First moment
- Second moment
- Third moment
- Fourth moment
- Concluding remarks

Measure of central tendency

- In any set of observation, an average denotes describes the entire series of observations.
- Since it occupies central position, so observation are larger or smaller than it, averages are also known as measure of central tendency.
- There are 3 measures of central tendency- mean, median and mode.

Types


Central Tendency
Mean
Median
Mode
Arithmetic Mean (AM)
Geometric Mean (GM)
Harmonic Mean (HM)

Moments

- In statistics, moments are measures of the shape and variability of a data set.
- hey are used to describe the location and dispersion of the data.
- The four commonly used moments in statistics are-
  1. the mean,
  2. variance,
  3. skewness, and
  4. kurtosis.

Contd..

Let,

Given n observations x1, x2, x3, x_n and an arbitrary contatnt A

Then

$\frac{1}{n}\sum(x - A)$ is called the 1st moment about A

$\frac{1}{n}\sum\Big(x - A\Big)2$ is called the 2nd moment about A

$\frac{1}{n}\sum\Big(x - A\Big)3$ is called the 3rd moment about A

and so on.

Moment about mean (central moment)

- The first central moment is the expected value, known also as mean, or average.
- It measures the location of the central point.
- First moment about mean-

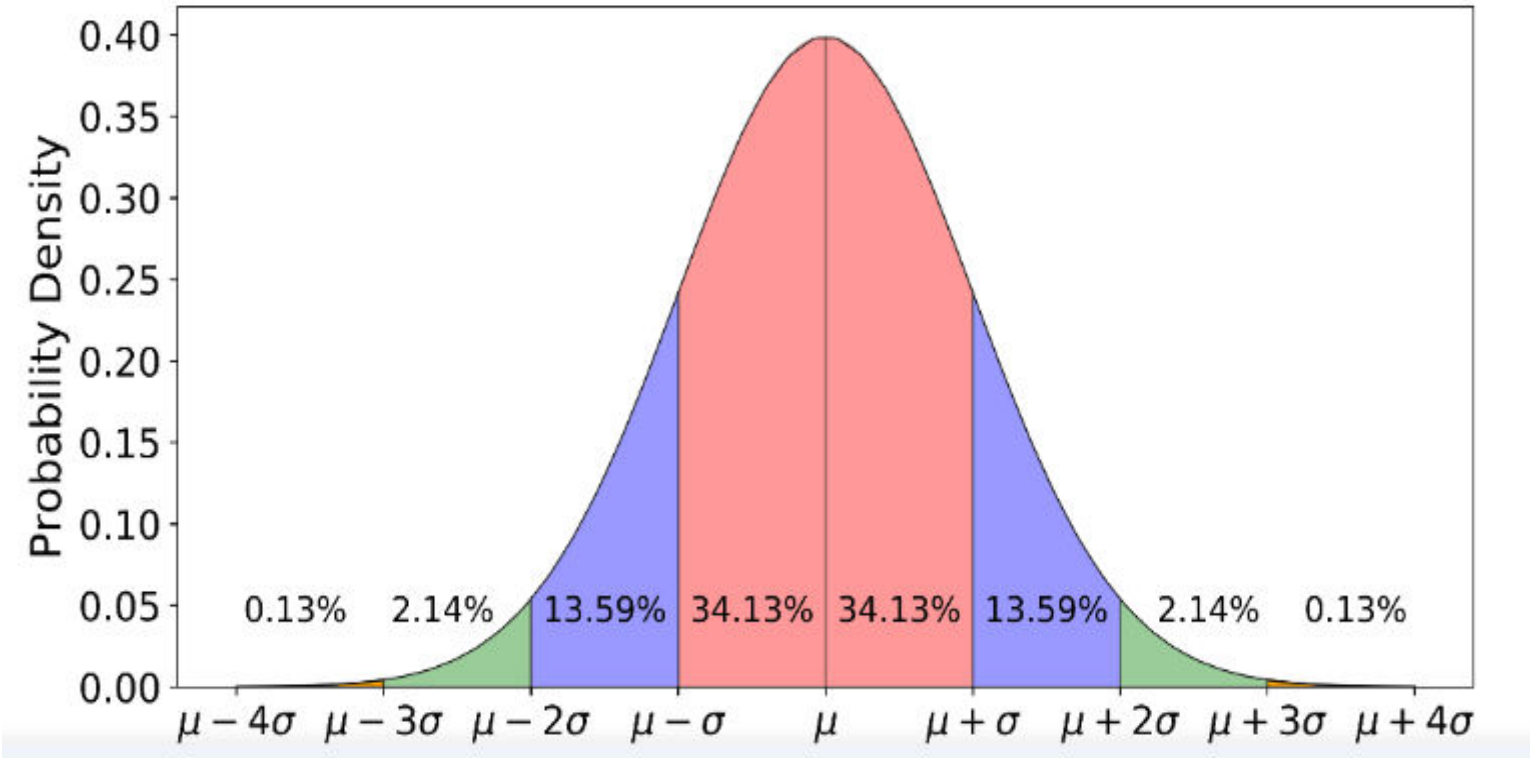
$$\frac{1}{\mathrm{n}}\sum\left(x - x'\right) = 0$$

Cond..
2nd Moment

- The second central moment is “Variance”.
- It measures the spread of values in the distribution OR how far from the normal.
- Variance represents how a set of data points are spread out around their mean value.
- 2nd Moment about mean is-

$$\frac{1}{\mathrm{n}}\sum\left(x-\ x'\right)\dot{2}\ =\sigma 2$$

Contd..



- 1st Standard Deviation: 68.27% of the data points lie
- 2nd Standard Deviation: 95.45% of the data points lie
- 3rd Standard Deviation: 99.73% of the data points lie

3rd Moment: Skewness

- The third statistical moment is “Skewness”.
- It measures how asymmetric the distribution is about its mean.

- 3rd Moment about mean-

$$\frac{1}{\mathrm{n}}\sum\left(x-\ x'\right)\dot{3}$$

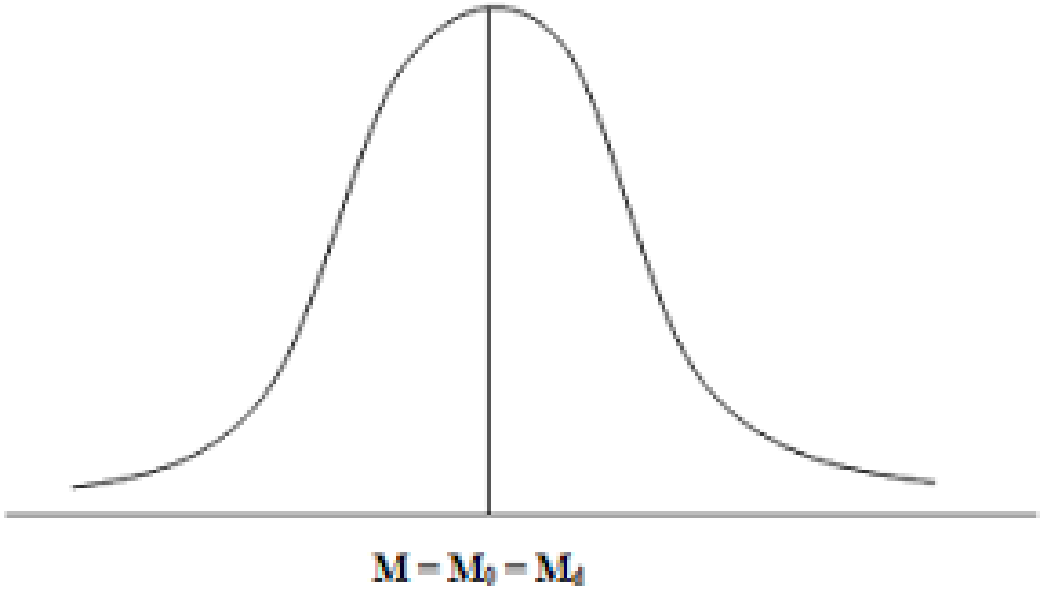
Skewness: Types

- We can differentiate three types of distribution with respect to its skewness:

1. Positive skewness
2. Negative skewness
3. Zero skewness

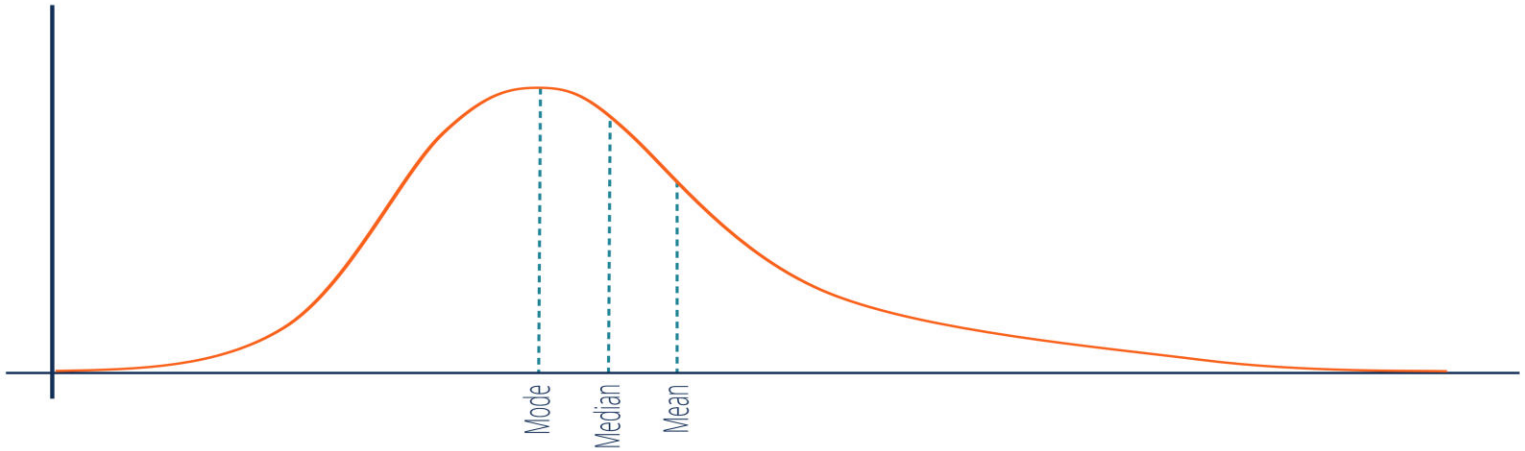
Contd..
Zero skewness (Symmetrical)

- If both tails of a distribution are symmetrical, and the skewness is equal to zero, then that distribution is symmetrical.



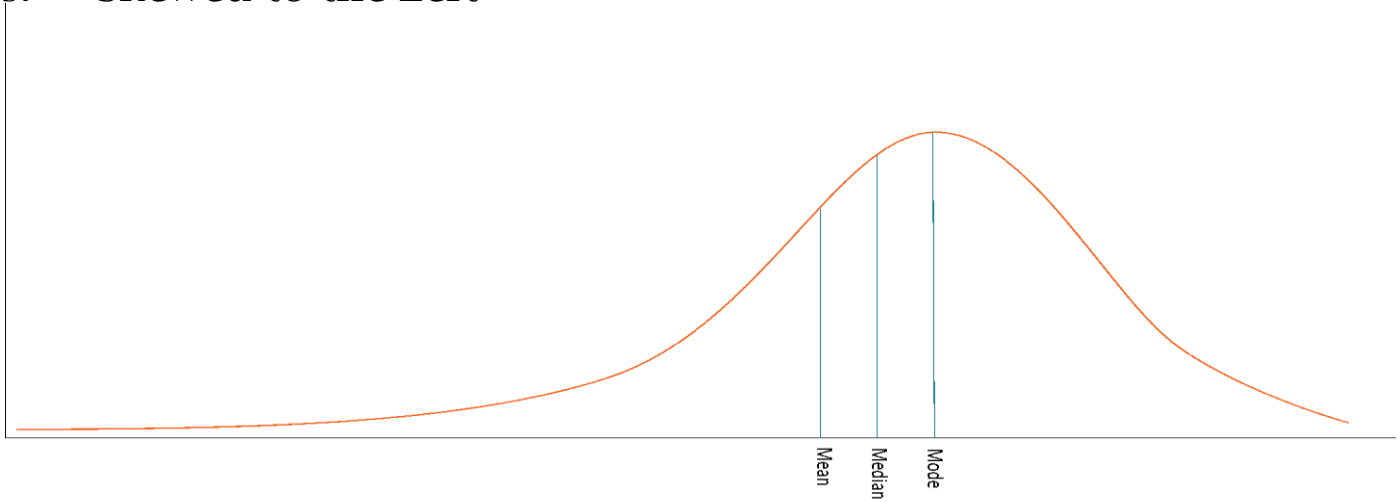
Contd..
Positive skewness

- In these types of distributions, the right tail (with larger values) is longer.
- So, this also tells us about 'outliers' that have values higher than the mean. Sometimes, this is also referred to as:
  1. Right-skewed
  2. Right-tailed
  3. Skewed to the Right



Contd..
Negative skewness (Symmetrical)

- In these types of distributions, the left tail (with small values) is longer. So, this also tells us about 'outliers' that have values lower than the mean. Sometimes, this is also referred to as:
  1. Left-skewed
  2. Left-tailed
  3. Skewed to the Left



Skewness: Observation

- In general, Skewness will impact the relationship of mean, median, and mode in the described manner:

1. For a Symmetrical distribution: Mean = Median = Mode

2. For a positively skewed distribution: Mode < Median < Mean (large tail of high values)

3. For a negatively skewed distribution: Mean < Median < Mode (large tails of small value)

4^{th} Moment: Kurtosis

- Kurtosis refers to the degree of “peakedness” of the frequency curve.
- It measures the amount in the tails and outliers.
- It focuses on the tails of the distribution and explains whether the distribution is flat or rather with a high peak.
- Fourth moment-

$$\frac{1}{\mathrm{n}}\sum\left(x - x'\right)4$$

Contd..

- Kurtosis can be expressed by-

$$Fischer's\ Kurtosis = \sum_{i=1}^{N} \frac{\frac{X_i - \bar{X}}{N}}{S^4} - 3$$

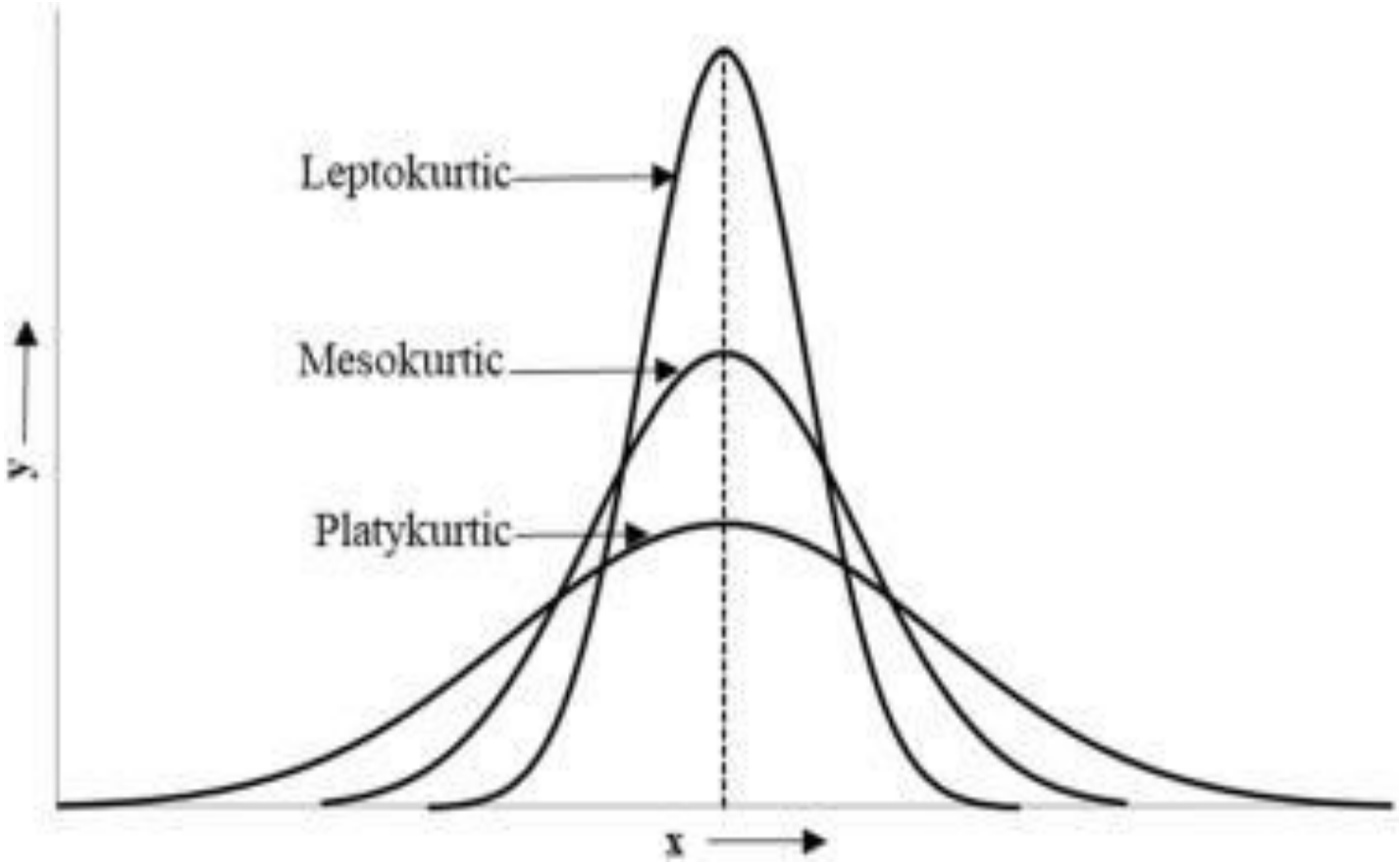
Where,

$\bar{X}$ *is the mean,*

N *is sample size,*

S *is standard deviation*

Contd..



Contd..

- A distribution is said to be platykurtic, when kurtosis value is negative.
- A distribution is said to be mesokurtic, when kurtosis value is zero.
- A distribution is said to be leptokurtic, when kurtosis value is positive.

Kurtosis: Observation

- For a normal distribution, the value of Kurtosis equal to 3
- For Kurtosis not equal to 3, there are the following cases:
- Kurtosis<3 [Lighter tails]: Negative kurtosis indicates a broad flat distribution.
- Kurtosis>3 [Heavier tails]: Positive kurtosis indicates a thin pointed distribution.

Expectation

Let a discrete random variable x assume the values

$x1, x2, x3, x_n$ with probablity $p1, p2, p3, \ldots . pn$ respectively.

Then, the expectation or expected value of x can be represented as $E(x)$

$$E(x) = \sum pi.xi$$

Expectation: Mean and Variance

Then, the expectation or expected value of x *can be represented as* $E(x)$

$$E(x) = \sum pi\,.xi$$

Mean of a probability distribution is expected value of x

$$Mean(\mu) = E(x)$$

Variance is the expected value of $(x-\mu)2$, μ is mean

Then

$$\text{Variance } (\sigma 2) = E(x-\mu)2$$

$$= E(x)2 - \mu 2$$

Resources: Conferences

- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Neural Information Processing Systems (NIPS)
- Computational Learning
- International Joint Conference on Artificial Intelligence (IJCAI)
- ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)
- IEEE Int. Conf. on Data Mining (ICDM)

Lecture 6

Statistical Analysis III

Shesh N Sahu

Asst. Prof. CSE, UTD

Contents

- Law of large number
- Moment
- Central moment
- First moment
- Second moment
- Third moment
- Fourth moment
- Concluding remarks

Law of large number

- The law of large numbers has a very central role in probability and statistics.
- It states that if you repeat an experiment independently a large number of times and average the result, what you obtain should be close to the expected value.
- There are two main versions of the law of large numbers. They are called the weak and strong laws of the large numbers.
- It uses the notion of Independent and identically distributed (IID) random variable.

Definition:

- It says As the number of times the experiments is repeated the empirical probability will approach the theoretical probability.

For i.i.d. random variables $X_1, X_2, \ldots, X_n$, the **sample mean**,

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

- Expectation-

$$\begin{aligned} E[\overline{X}] &= \frac{EX_1 + EX_2 + \ldots + EX_n}{n} \\ &= \frac{nEX}{n} \\ &= EX. \end{aligned}$$

Contd..

The Law of Large Numbers

$$\frac{X1 + X2 + X3 + \ldots\ldots + Xn}{n \text{ (number of variables)}} \longrightarrow E(X) \quad \text{Where} \quad n \longrightarrow \infty$$

WallStreetMojo

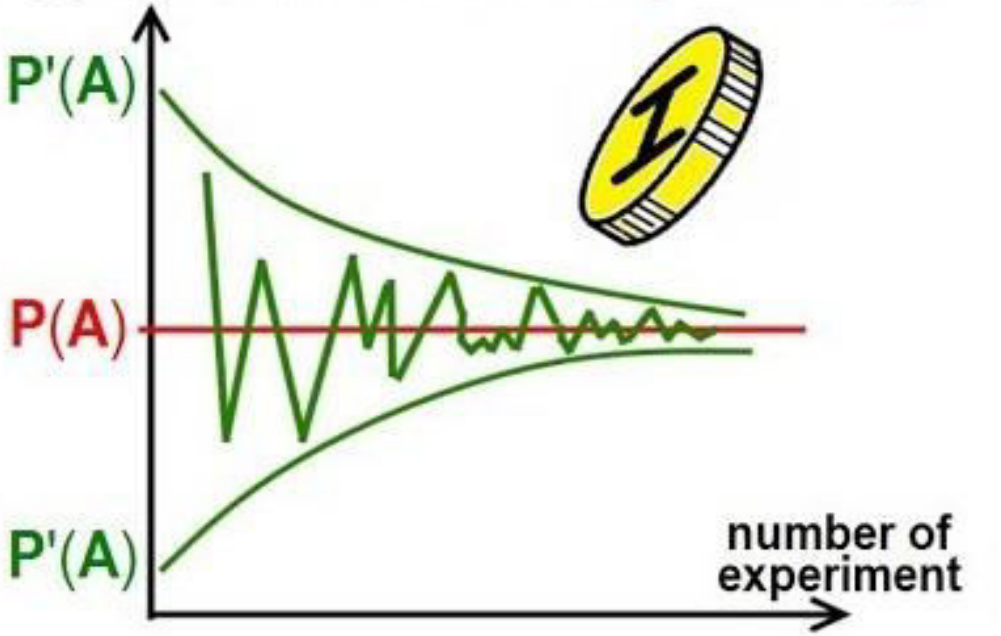
Contd..

Law of Large Numbers

As the number of the times the experiment is related the empirical probability will approach the theoretical probability.

P(A) = theoretical probability

P'(A) = experimental (empirical) probability



Weak of law of large number

- The mean of the sample will converge in probability to the expected value of its respective probability, as the number of sample goes infinity.

$$\lim_{n \to \infty} P(|\overline{X}_n - \mu| \geq \epsilon) = 0$$

Strong law of large number

- As you take more sample up to infinity, the sample mean almost surely equals to the population mean.

- Suppose $X_1, X_2, X_3, \ldots$ are i.i.d with $E(X_i) = \mu < \infty$, then $\bar{X}_n$ converges to μ as $n \to \infty$ with probability 1. That is

$$P\left(\lim_{n\to\infty} \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \mu\right) = 1$$

- This is called convergence almost surely.

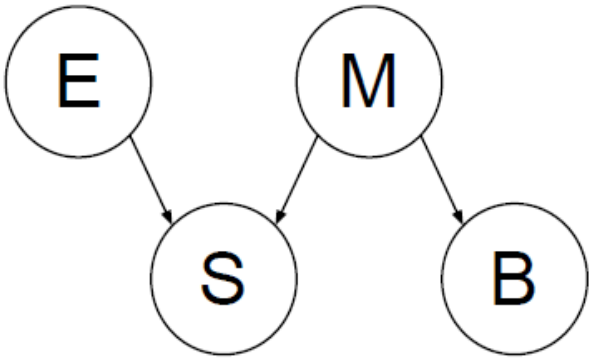
Problem II

A smell of sulphur (S) can be caused either by rotten eggs (E) or as a sign of the doom brought by the Mayan Apocalypse (M). The Mayan Apocalypse also causes the oceans to boil (B). The Bayesian network and corresponding conditional probability tables for this situation are shown below. For each part, you should give either a numerical answer (e.g. 0.81) or an arithmetic expression in terms of numbers from the tables below (e.g. $0.9 \cdot 0.9$).

Note: be careful of doing unnecessary computation here.

$P(E)$	
$+e$	0.4
$-e$	0.6

$P(S\|E,M)$			
$+e$	$+m$	$+s$	1.0
$+e$	$+m$	$-s$	0.0
$+e$	$-m$	$+s$	0.8
$+e$	$-m$	$-s$	0.2
$-e$	$+m$	$+s$	0.3
$-e$	$+m$	$-s$	0.7
$-e$	$-m$	$+s$	0.1
$-e$	$-m$	$-s$	0.9



$P(M)$	
$+m$	0.1
$-m$	0.9

$P(B\|M)$		
$+m$	$+b$	1.0
$+m$	$-b$	0.0
$-m$	$+b$	0.1
$-m$	$-b$	0.9

Questions?

- What is the probability that Ocean boils?
- What is the probability of Mayan Apocalypse is occurring given the oceans are boiling?
- What is the probability Mayan Apocalypse given the ocean are boiling, Sulphur and rotten eggs?

Lecture 7

Multidimensional space

Shesh N Sahu

Asst. Prof. CSE, UTD

Multidimensional Space

- A dimension of a point is defined as minimum number of coordinates that are needed to specify a point in it.
- The ability to understand and think through dimensions is an important part of human intelligence.

- A multi-dimensional space can apply to efficiently managing any types of resources and modeling the interest of various objects including individual user and organization (e.g., company, community and state).

- Multi-dimensional space is space for multi dimensional data that can manage, coordinate, map, and compose methodologies as well as guide applications and predict development.

Application I:

- A smart online library system:

  It provide services for all types of users, including individual users, organizations (research institutions and enterprises), and strategic services through establishing the multi-dimensional users' mental space, the multi-dimensional resource space, and the mapping between the mental space and the resource space.

a. It manages all library resources through a multi-dimensional category space, which can adapt to the increasing of new resources.
b. It records the users' operations (queries, chats, and clicks for browsing and downloading resources) and the downloaded resources.
c. It discovers the dimensions and the interest semantic link networks on the operations and resources.

Application II:

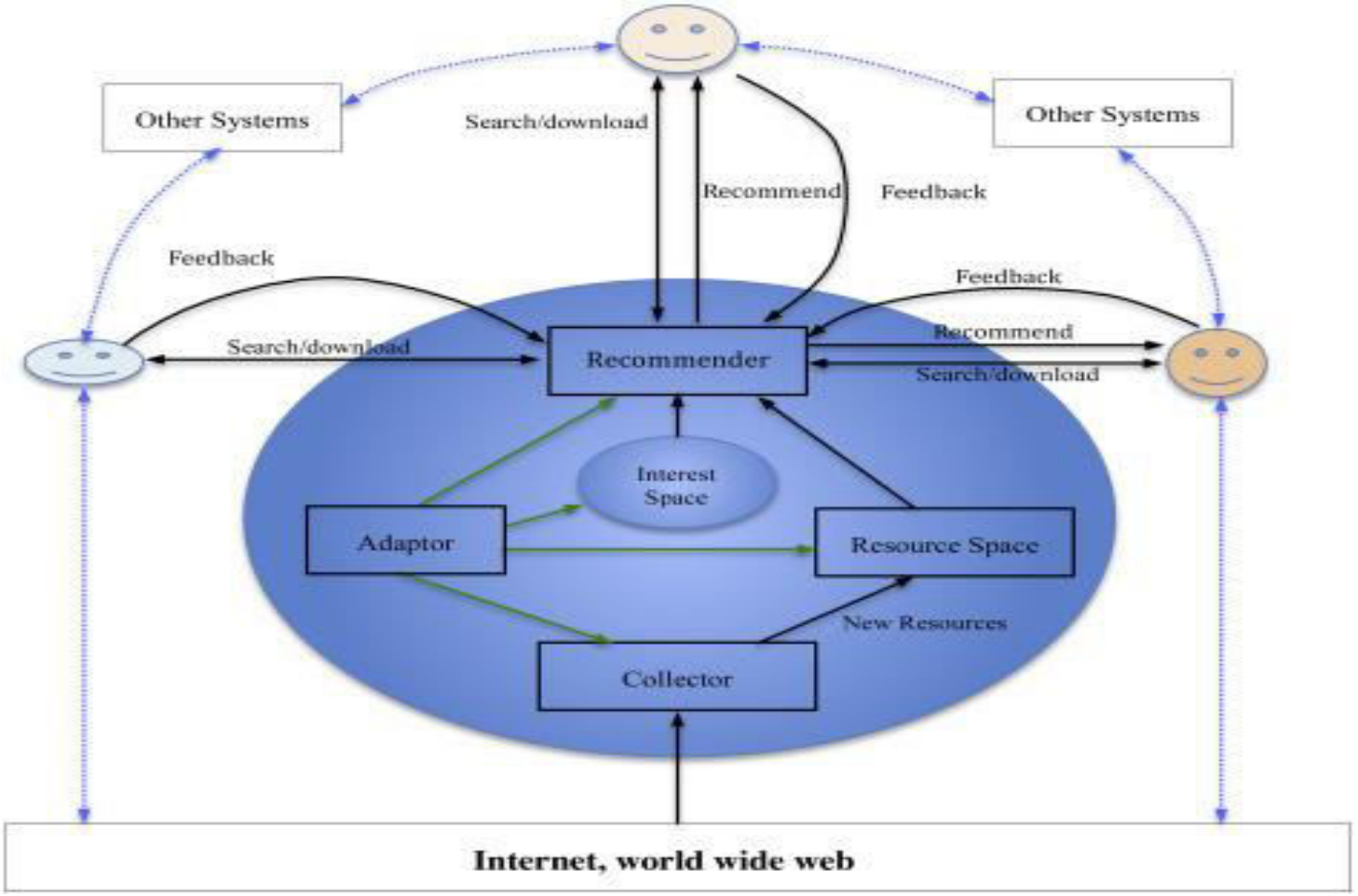
- A smart website can be designed with the following features::

a. It has a multi-dimensional resource space for storing webpages. Webpages will be stored in and retrieved from a point of the space.

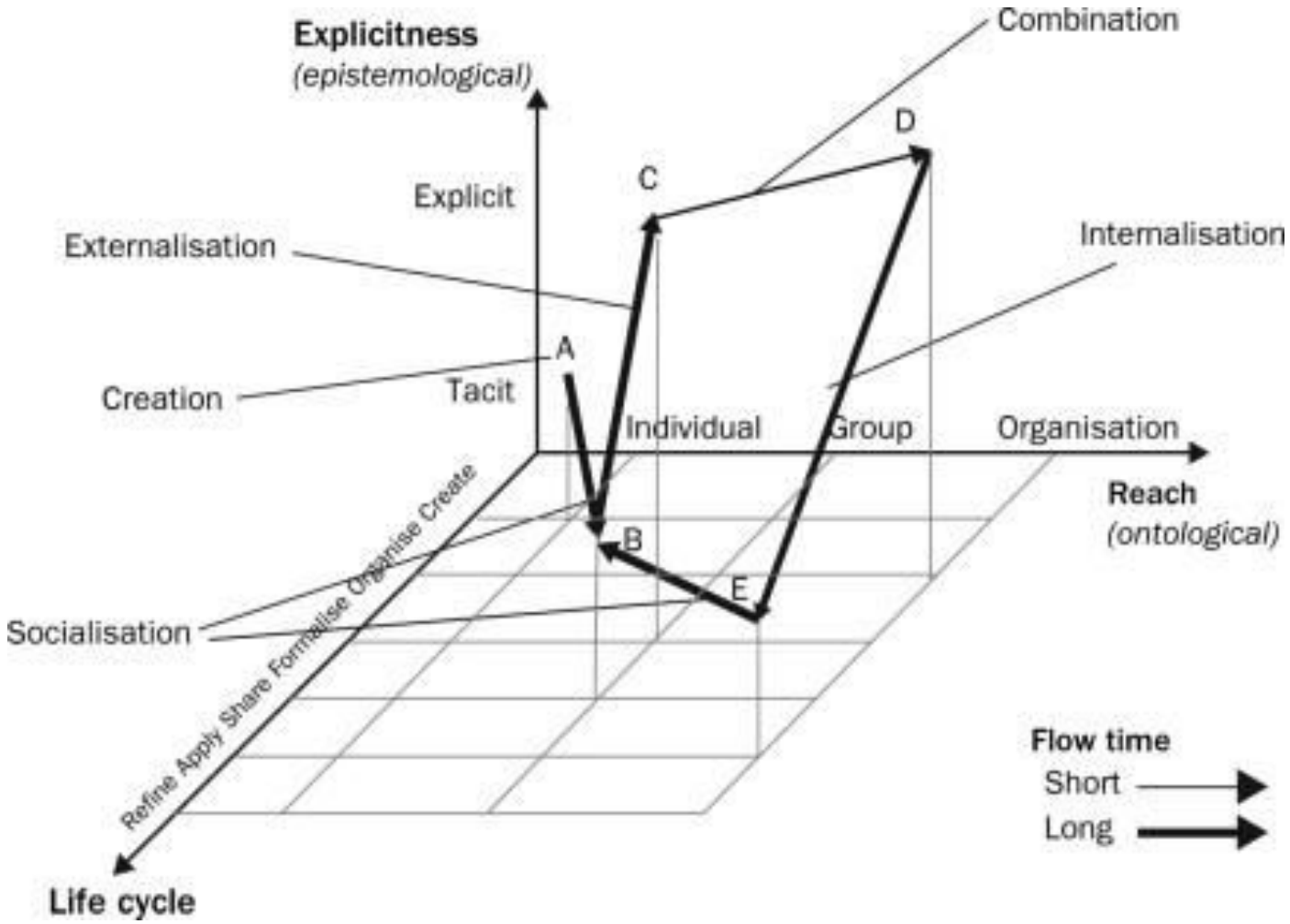
b. It has a collector (crawler) that can collect or create webpages and store them in the multi-dimensional resource space.

c. It generates a multi-dimensional interest space by extracting dimensions from users' operation history and the contents of recommendation.

Contd..


Other Systems
Search/download
Other Systems
Recommend
Feedback
Feedback
Feedback
Search/download
Recommender
Recommend
Search/download
Interest Space
Adaptor
Resource Space
New Resources
Collector
Internet, world wide web

Contd..


Explicitness
(epistemological)
Combination
D
C
Explicit
Externalisation
Internalisation
A
Creation
Tacit
Individual
Group
Organisation
Reach
(ontological)
B
E
Socialisation
Refine Apply Share Formalise Organise Create
Flow time
Short
Long
Life cycle

Contd..

- The vertical axis represents the dimension explicitness, which characterizes the degree to which knowledge has been articulated in explicit form.
- The horizontal axis represents the dimension reach, which characterizes the level of social aggregation associated with knowledge flows.
- The third axis represents the dimension life cycle, which characterizes the kind of activity associated with knowledge flows.

Metric Space

- A metric space is a set X that has a notion of the distance d(x, y) between every pair of points x, y ∈ X.
- A metric on a set is a function that satisfies the minimal properties we might expect of a distance.
- A metric space is made up of a nonempty set and a metric on the set.

Definition: Metric space

- A metric d on a set X is a function $d : X \times X \rightarrow R$ such that for all x, y $\in$ X.

  (1) $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$;

  (2) $d(x, y) = d(y, x)$ (symmetry);

  (3) $d(x, y) \leq d(x, z) + d(z, x)$ (triangle inequality).

A metric space (X, d) is a set X with a metric d defined on X.

Normed Vector Space

- A normed vector space is a vector space in which each vector is associated with a scalar value called a norm.

- Ex: In a standard Euclidean vector spaces, the length of each vector is a norm:

Definition: Normed Vector Space

- A normed vector space is a vector space in which each vector is associated with a scalar value called a norm. In a standard Euclidean vector spaces, the length of each vector is a norm:

- A normed vector space is vector space (V,F) associated with a function $\|.\|:V\rightarrow R$, called a norm, that obeys the following axioms:

- $\forall v \in V,\ \|v\| \geq 0$
- $\forall v \in V, \forall \alpha \in F,\ \|\alpha v\| = |\alpha| \|v\|$
- $\forall v,u \in V,\ \|u+v\| \leq \|u\| + \|v\|$

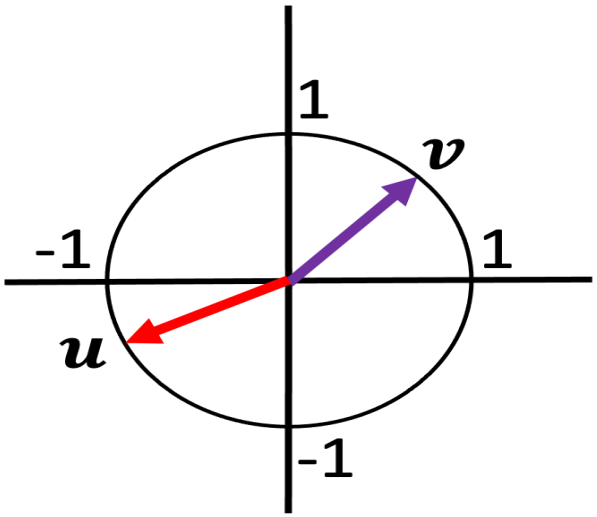
Contd..

- Axiom 1 says that all vectors should have a positive length. This enforces our intuition that a "length" is a positive quantity.

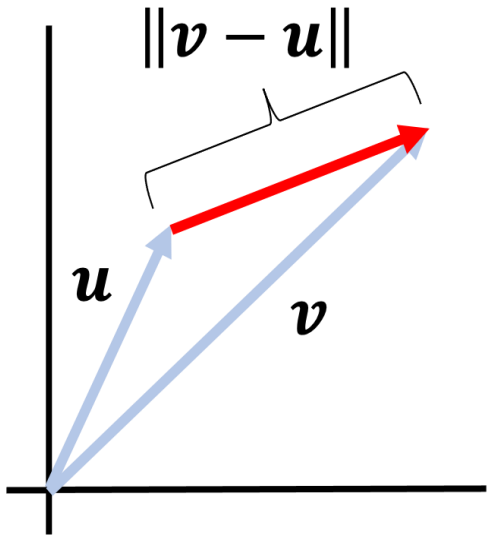
- Axiom 2 says that if we multiply a vector by a scalar, it's length should increase by the magnitude (i.e. the absolute value) of that scalar.

- Axiom 3 says that the length of the sum of two vectors should not exceed the sum of the lengths of each vector.

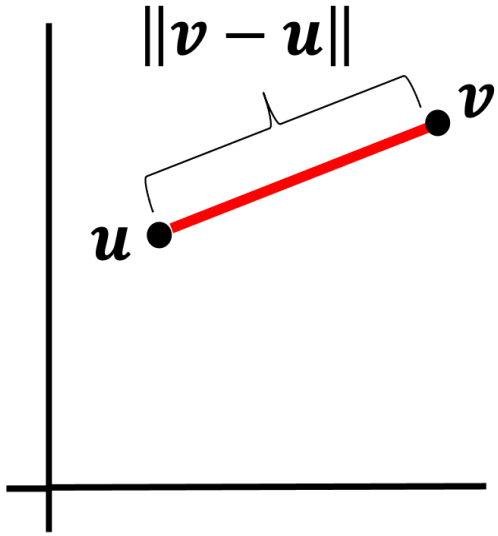
Find vector name?



Cond..



As arrows



As points

Dot Product Space

- An inner product is a generalization of the dot product. In a vector space, it is a way to multiply vectors together, with the result of this multiplication being a scalar.
- More precisely, for a real vector space, an inner product <·,·> satisfies the following four properties. Let u, v, and w be vectors and alpha be a scalar, then:

1. Non Negativity: <v,v> >= 0 and equal if and only if v=0.
2. Conjugate symmetry: <v,w> = <w,v>.
3. Linearity: <u+v, w> = <u,w> + <v,w>.

Lecture on

Types of space

Shesh N Sahu

Asst. Prof. CSE, UTD

Pre-Hilbert Space

- A Pre-Hilbert space is a vector space equipped with an inner product, but it may not be complete. In other words, it is a space where you can calculate inner products (like in a Hilbert space), but it may not have all the nice properties of a Hilbert space, such as completeness.

- Example: Euclidean space ℝ^n, equipped with the usual dot product, is a Pre-Hilbert space. You can calculate inner products of vectors in ℝ^n, but it is not complete because some sequences may not converge within ℝ^n.

Hilbert Space

- A Hilbert space is a complete inner product space, meaning it is a vector space where you can calculate inner products, and it is also complete, which means that all Cauchy sequences (sequences whose elements get arbitrarily close to each other) converge to a point within the space

- Example: The space of square-integrable functions L^2(Ω), where Ω is a measurable set, is a Hilbert space. In this space, you can calculate inner products of functions, and it is complete, meaning any Cauchy sequence of functions in L^2(Ω) converges to a function also in L^2(Ω).

Contd..

- In machine learning, Hilbert spaces are often used to describe feature spaces where learning algorithms operate.
- For example, in kernel methods like Support Vector Machines (SVMs), the data is implicitly mapped into a high-dimensional Hilbert space, where the algorithm performs linear operations to find optimal decision boundaries.
- The terms "Pre-Hilbert" and "Hilbert" are more mathematical in nature and might not be commonly used in everyday machine learning discussions.
- They are important for understanding the mathematical foundations of some machine learning algorithms, particularly those involving kernel methods.

Decision theory

- Decision theory is a field of study that provides a systematic framework for making decisions in the presence of uncertainty.
- It involves identifying possible decision alternatives, assessing the likelihood of various outcomes, and evaluating the consequences or utility associated with each possible decision.
- The objective of decision theory is to select the decision that maximizes expected utility or minimizes expected costs or risks, depending on the context.

Example:

- Choosing a Mode of Transportation:
- Imagine you need to decide on the mode of transportation for your daily commute to work, and you have three options: taking the bus, driving a car, or riding a bicycle. Your decision will depend on various factors, including time, cost, and convenience.
- 1. Identify Decision Alternatives:

  Option 1: Taking the Bus

  Option 2: Driving a Car

  Option 3: Riding a Bicycle
- 2. Assess Probabilities and Outcomes:
- 3. Determine Preferences and Utilities:
- 4. Calculate Expected Utility:
- 5. Make the Decision:

Contd..

- **2. Assess Probabilities and Outcomes:**

  - Probabilities represent the likelihood of various events. In this case, you might consider:
  - The probability of traffic congestion when driving.
  - The probability of a bus delay.
  - The probability of encountering bad weather while riding a bicycle.
  - Outcomes represent the results or consequences associated with each option, such as:
  - Time spent commuting.
  - Cost of transportation (e.g., bus fare, fuel costs, maintenance).
  - Health benefits (e.g., exercise gained from biking).
  - Environmental impact (e.g., carbon emissions from driving).

Contd..

- **3. Determine Preferences and Utilities:**

  - You assign subjective values or utilities to each outcome based on your personal preferences. For example:

  - You might value time saved highly, so the utility of shorter commute times is higher.

  - You might value health and environmental benefits, so biking has positive utility in those aspects.

  - You might dislike traffic and stress associated with driving, so it has a lower utility in those aspects.

- **4. Calculate Expected Utility:**

  - Decision theory involves calculating the expected utility of each option by considering the probabilities and utilities associated with each outcome.

  - For instance, if the expected utility of taking the bus is higher than that of driving or biking, it would be the rational choice.

Contd..

- In this example, decision theory helps you make an informed and rational decision by considering not only the immediate costs and benefits but also the uncertainties associated with each option.

- It's a valuable tool for decision-making in various fields, including business, economics, healthcare, and more, where choices often involve risk and uncertainty.

Minimizing the misclassification rate

- Minimizing the misclassification rate is a fundamental objective in machine learning, particularly in classification tasks.
- It refers to the process of reducing the number of incorrect predictions made by a machine learning model when categorizing data points into different classes or categories.
- The misclassification rate is a key metric used to assess the performance of a classification model and is often quantified as a percentage or fraction of incorrectly classified instances relative to the total number of instances.

How to minimize the misclassification rate

- 1. Understanding Misclassification:

  - In a binary classification problem, you typically have two classes: positive and negative. Misclassification occurs when the model assigns a data point to the wrong class. There are two types of misclassification:

  - False Positives (Type I Error): Occur when the model incorrectly predicts a positive outcome when it should be negative.

  - False Negatives (Type II Error): Occur when the model incorrectly predicts a negative outcome when it should be positive.

Contd..

- 2. Choosing an Appropriate Metric:

  - Before you can minimize misclassification, you should choose an appropriate evaluation metric that aligns with your problem's goals. Common classification metrics include accuracy, precision, recall, F1-score, and the confusion matrix.

- 3. Model Selection and Training:

  - Select a machine learning algorithm that is suitable for your classification task. Train the model using a labeled dataset, ensuring that it learns the underlying patterns in the data.

- 4. Hyperparameter Tuning:

  -Fine-tune the model's hyperparameters to achieve better classification performance. Grid search, random search, or automated hyperparameter optimization techniques can help you find the best combination of hyperparameters.

Contd..

- 5. Feature Engineering:

  - Carefully preprocess and engineer the features used by the model. This step can improve the model's ability to discriminate between different classes.

- 6. Balancing Class Distribution:

  - If your dataset has imbalanced classes (one class significantly outnumbers the other), consider using techniques like oversampling, under sampling, or generating synthetic data to balance the class distribution.

Contd..

- 7. Threshold Adjustment:

  - Adjust the classification threshold if necessary. By changing the threshold for classifying instances as positive or negative, you can control the trade-off between precision and recall, which can help minimize misclassification errors.

- 8. Ensemble Methods:

  - Consider using ensemble methods like Random Forests or Gradient Boosting, which combine the predictions of multiple models to reduce misclassification rates.

Contd..

- 9. Error Analysis:

  - Conduct a thorough error analysis to understand the types of misclassifications that are occurring. This can provide insights into potential improvements, such as collecting more data for specific classes or refining features.

- 10. Continuous Monitoring and Model Maintenance:

  - After deploying the model in a real-world setting, continuously monitor its performance and retrain it as needed to adapt to changing data distributions and minimize misclassification errors over time

Minimizing the expected loss

- It is a key objective in decision theory and statistical decision-making.
- It involves making decisions that minimize the average or expected loss or cost associated with those decisions.
- This concept is often used in situations where decisions involve uncertainty or risk, and it provides a systematic framework for selecting actions that are both rational and informed. Here's a more detailed explanation:

How to minimizes

- 1. Expected Loss or Risk:

  - In decision theory, "loss" refers to the negative consequences or costs associated with a decision. This can include monetary costs, time, resources, or any other relevant measure of loss.

  - The "expected loss" or "expected risk" represents the average loss that would be incurred over a large number of repetitions of a decision-making process, taking into account the probabilities of different outcomes.

- 2. Decision Space:

  - Decision theory considers a set of possible actions or decisions, often referred to as a "decision space." Each decision leads to a different set of outcomes.

Contd..

- 3. Loss Function:

  - A critical component of minimizing expected loss is defining a "loss function" or "cost function." This function quantifies the cost or loss associated with each possible outcome for a given decision.

  - The loss function typically takes two arguments: the actual outcome and the decision made. It assigns a numerical value (the loss or cost) based on how well the decision aligns with the actual outcome.

Contd..

- 4. Probabilistic Outcomes:

  - Many real-world decisions involve uncertainty. Therefore, decision theory considers the probabilities associated with different outcomes. These probabilities represent the likelihood of each outcome occurring.

- 5. Expected Loss Calculation:

  - The expected loss is calculated by taking a weighted average of the losses for each possible decision, with the weights being the probabilities of each outcome. The formula for expected loss (E[Loss]) is often expressed as:

  $E[Loss] = \Sigma$ (Probability of Outcome i) * (Loss associated with Outcome i)

Problem: Loss calculation

- Let's consider a medical diagnosis example. A doctor needs to decide whether to perform an expensive diagnostic test on a patient with certain symptoms. The decision has two possible outcomes: "Positive Diagnosis" or "Negative Diagnosis." The doctor defines a loss function that quantifies the cost associated with each outcome. The cost of a false positive (unnecessary test) is high, and the cost of a false negative (missed diagnosis) is also high due to potential health risks.

- - Probability of Positive Diagnosis = 0.15
- - Probability of Negative Diagnosis = 0.85

- - Loss associated with Positive Diagnosis = $1,000
- - Loss associated with Negative Diagnosis = $5,000

Contd..

- 6. Decision Making:

  - The goal in minimizing expected loss is to select the decision that results in the lowest expected loss. In other words, you aim to choose the action that, on average, minimizes the negative consequences of your decisions.

- 7. Risk Preferences:

  - The specific choice of the loss function and the degree to which you aim to minimize expected loss may depend on your risk preferences. Some individuals or organizations may be risk-averse and prioritize minimizing the worst-case loss, while others may be more risk-tolerant and focus on expected value.

- Example: Medical Diagnosis

Lecture on

Minimization of loss

Shesh N Sahu

Asst. Prof. CSE, UTD

Minimizing the expected loss

- It is a key objective in decision theory and statistical decision-making.
- It involves making decisions that minimize the average or expected loss or cost associated with those decisions.
- This concept is often used in situations where decisions involve uncertainty or risk, and it provides a systematic framework for selecting actions that are both rational and informed. Here's a more detailed explanation:

How to minimizes

- 1. Expected Loss or Risk:

  - In decision theory, "loss" refers to the negative consequences or costs associated with a decision. This can include monetary costs, time, resources, or any other relevant measure of loss.

  - The "expected loss" or "expected risk" represents the average loss that would be incurred over a large number of repetitions of a decision-making process, taking into account the probabilities of different outcomes.

- 2. Decision Space:

  - Decision theory considers a set of possible actions or decisions, often referred to as a "decision space." Each decision leads to a different set of outcomes.

Contd..

- 3. Loss Function:

  - A critical component of minimizing expected loss is defining a "loss function" or "cost function." This function quantifies the cost or loss associated with each possible outcome for a given decision.

  - The loss function typically takes two arguments: the actual outcome and the decision made. It assigns a numerical value (the loss or cost) based on how well the decision aligns with the actual outcome.

Contd..

- 4. Probabilistic Outcomes:

  - Many real-world decisions involve uncertainty. Therefore, decision theory considers the probabilities associated with different outcomes. These probabilities represent the likelihood of each outcome occurring.

- 5. Expected Loss Calculation:

  - The expected loss is calculated by taking a weighted average of the losses for each possible decision, with the weights being the probabilities of each outcome. The formula for expected loss (E[Loss]) is often expressed as:

  $E[Loss] = \Sigma$ (Probability of Outcome i) * (Loss associated with Outcome i)

Problem: Loss calculation

- Let's consider a medical diagnosis example. A doctor needs to decide whether to perform an expensive diagnostic test on a patient with certain symptoms. The decision has two possible outcomes: "Positive Diagnosis" or "Negative Diagnosis." The doctor defines a loss function that quantifies the cost associated with each outcome. The cost of a false positive (unnecessary test) is high, and the cost of a false negative (missed diagnosis) is also high due to potential health risks.

- - Probability of Positive Diagnosis = 0.15
- - Probability of Negative Diagnosis = 0.85

- - Loss associated with Positive Diagnosis = $1,000
- - Loss associated with Negative Diagnosis = $5,000

Contd..

- 6. Decision Making:

  - The goal in minimizing expected loss is to select the decision that results in the lowest expected loss. In other words, you aim to choose the action that, on average, minimizes the negative consequences of your decisions.

- 7. Risk Preferences:

  - The specific choice of the loss function and the degree to which you aim to minimize expected loss may depend on your risk preferences. Some individuals or organizations may be risk-averse and prioritize minimizing the worst-case loss, while others may be more risk-tolerant and focus on expected value.

- Example: Medical Diagnosis

Inference and decision

- Inference and decision are fundamental concepts in the field of statistical decision theory and machine learning, and they play a crucial role in making informed choices based on data and probabilistic reasoning. Let's delve into each concept in detail:

Inference

- Inference involves drawing conclusions or making predictions based on available evidence, data, or observations.
- It's the process of using data to gain insights, uncover patterns, or estimate unknown quantities.
- Inference typically operates in a probabilistic framework, where you use statistical methods to make informed statements about populations or events. Here are key aspects of inference:

- **I) Types of Inference:**
  1. Statistical Inference: Involves making statements about a population based on a sample of data. This includes estimating population parameters (e.g., mean, variance) and conducting hypothesis tests.

  2. Machine Learning Inference: In the context of machine learning, inference often refers to using a trained model to predict or classify new, unseen data points. It can also include interpreting the model's internal representations.
  3. 3. Hypothesis Testing:

  - Hypothesis testing is a common statistical inference technique. It involves setting up null and alternative hypotheses and using sample data to determine whether there's enough evidence to reject the null hypothesis in favor of the alternative.

Contd..

- 2. Bayesian vs. Frequentist Inference:

  - In Bayesian inference, probability is used to quantify uncertainty. It incorporates prior beliefs (prior probabilities) and updates them with observed data using Bayes' theorem to compute posterior probabilities.

  - In frequentist inference, probability is interpreted as long-run frequencies. It involves estimating parameters and making decisions based on data, often using methods like maximum likelihood estimation.

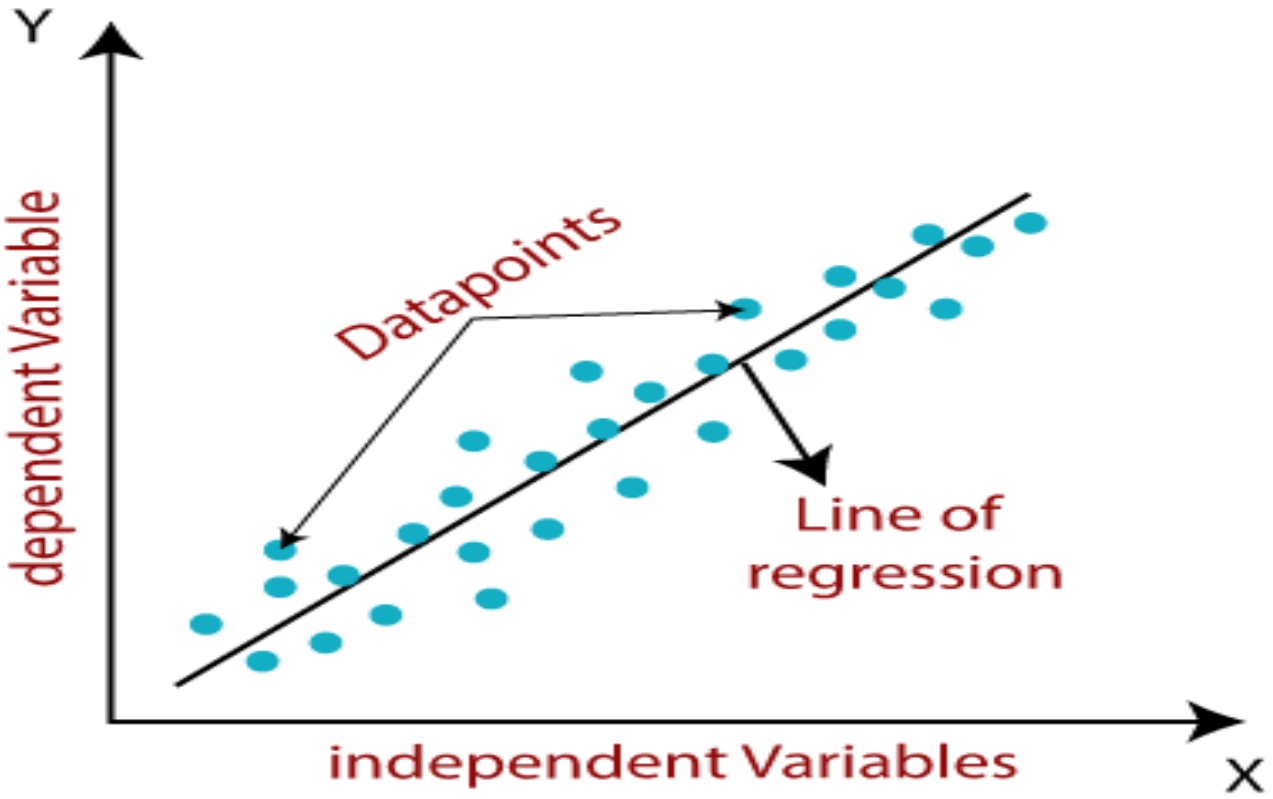
Problem: Predict the glucose level given the age?

S. No	Age (X)	Glucose level (Y)
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

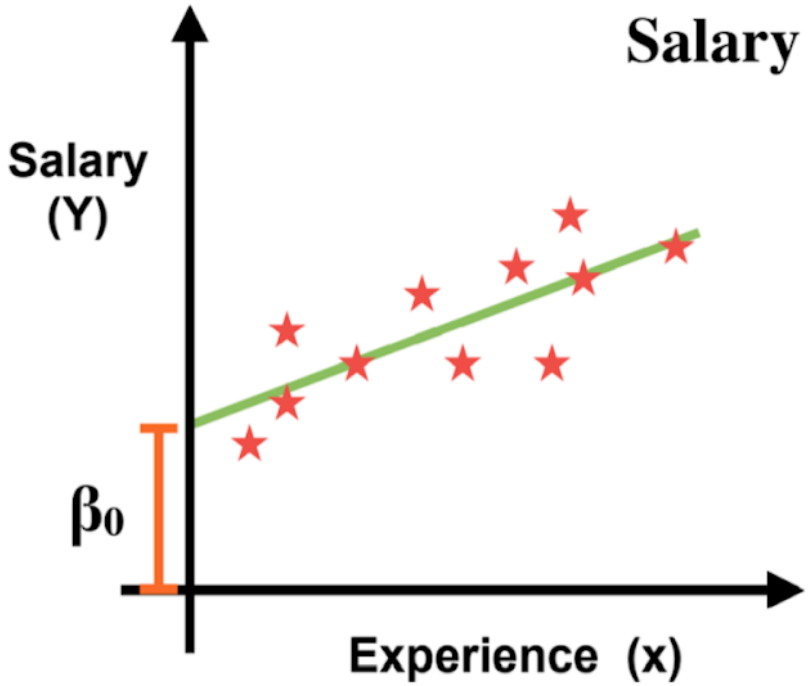
Linear Regression

- Linear regression is one of the easiest and most popular Machine Learning algorithms.
- It is a statistical method that is used for predictive analysis.
- Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.
- Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.

Contd..


Y
dependent Variable
Datapoints
Line of
regression
independent Variables
X

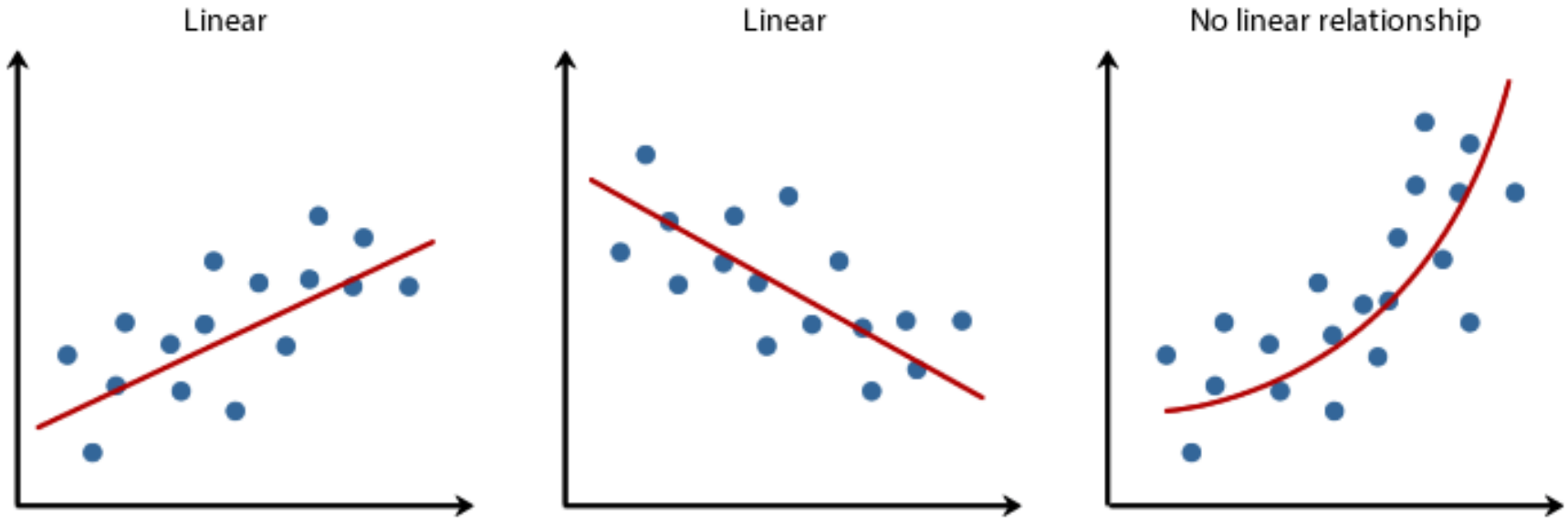
Contd..



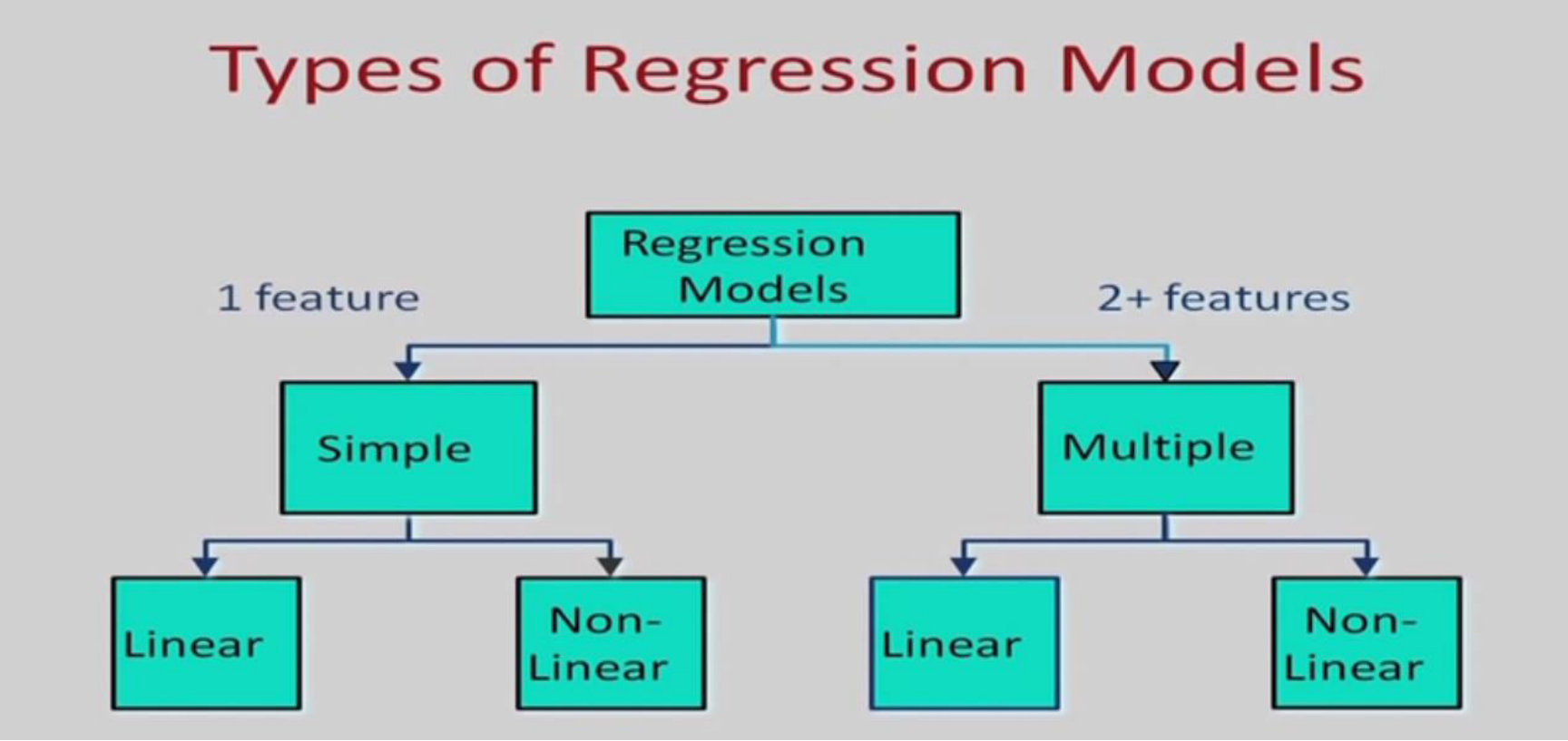
$$\text{Salary} = \beta_0 + \beta_1 * \text{Experience}$$

$$\beta_0 = \$\ 30K$$

Contd..



Types of regression



Linear Regression: Expression

- Mathematically, we can represent a linear regression as:

$$Y = \beta 0 + \beta 1 x + \varepsilon$$

Where,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
β0= intercept of the population
β1 = slope of the population
ε = random error

Contd..

$$\beta_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$\beta_0 = \frac{\sum y - \beta_1 \sum x}{n}$$

Problem: Predict the glucose level given the age?

S. No	Age (X)	Glucose level (Y)
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

Problem?

- Predict the Glucose level of a person whose age is 55 ?

Contd..

- Value of regression coefficient

$\beta 1 = 0.385225$

$\beta 0 = 65.14$

Loss function for regression

- It is a fundamental component of a machine learning or statistical model that is used to quantify the error or discrepancy between predicted values and actual target values (also known as ground truth) in a regression problem.

- The goal of regression is to find a model that minimizes this loss function, which essentially measures how well the model fits the data.

Types of Regression Loss Functions:

1. Mean Squared Error (MSE): The most commonly used loss function for regression. It measures the average of the squared differences between predicted and actual values. It's suitable when the errors should be penalized more for large deviations from the target values.

MSE Loss: $L(y, \hat{y}) = (1/n) * \Sigma(yi - \hat{y}i)^2$

2. Mean Absolute Error (MAE): Measures the average of the absolute differences between predicted and actual values. It's less sensitive to outliers compared to MSE.

MAE Loss: $L(y, \hat{y}) = (1/n) * \Sigma|yi - \hat{y}i|$

Contd..

- - Huber Loss: A hybrid loss function that combines the characteristics of MSE and MAE. It's less sensitive to outliers than MSE and provides a balance between the two.

Lecture on

Regression: Linear Regression

Shesh N Sahu

Asst. Prof. CSE, UTD

Limitation of Naïve Bayes Classifier

- **Assumption of independence**: If the features are correlated, this may result in inaccurate classification result.
- **Lack of flexibility**: Its ability to handle complicated and non-linear relationships between features may be constrained as a result.
- **Limited ability to capture interactions between features**: Naive Bayes may not be able to capture interactions or dependencies between features that are important for classification because it assumes that features are independent of one another.
- **Sensitivity to the choice of prior probabilities**: Naive Bayes requires prior probability specification for each class, which may have an impact on the classification outcomes.

Contd..

- **Limited ability to handle continuous variables:** The Naive Bayes model assumes that the features are discrete or categorical, which prevents it from directly handling continuous variables which may cause information loss and decreased performance.

- **Biased towards features with high frequency**: This could become a problem if some less common but crucial features are missed.

- **Difficulty in handling missing data:** Naive Bayes has trouble handling missing data, and it does so poorly. The entire instance must be discarded or imputed if a feature has a missing value, which can produce biased results.

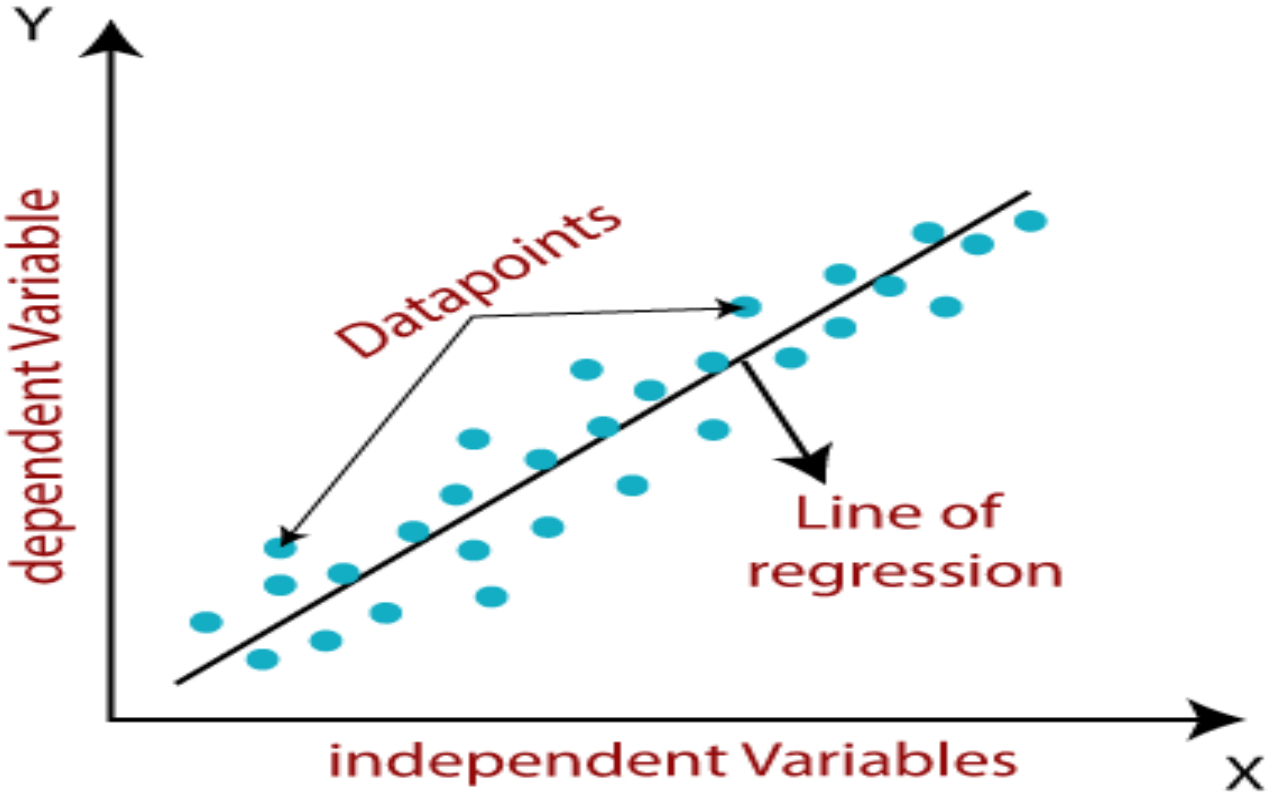
Problem: Predict the glucose level given the age 55 ?

S. No	Age (X)	Glucose level (Y)
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

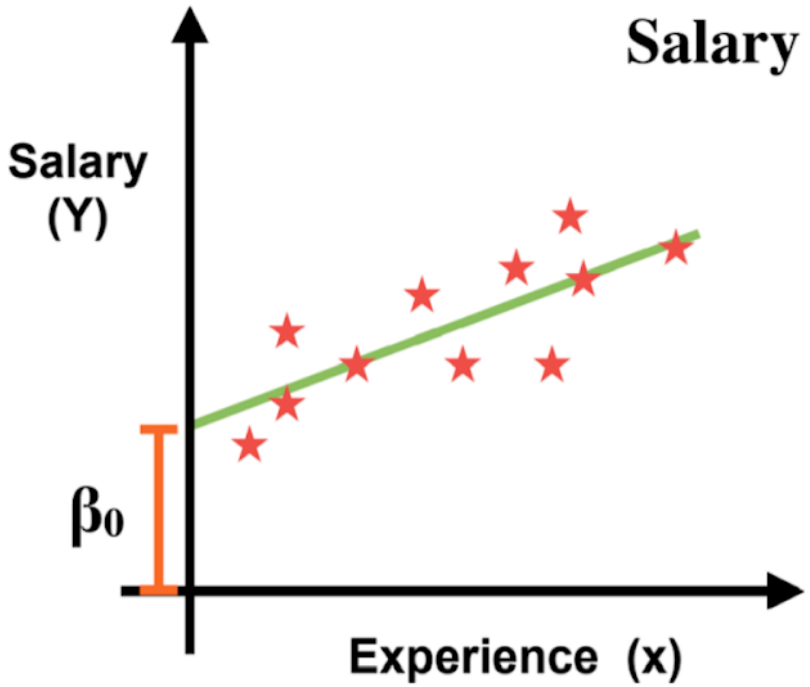
Linear Regression

- Linear regression is one of the easiest and most popular Machine Learning algorithms.
- It is a statistical method that is used for predictive analysis.
- Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.
- Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.

Contd..


Y
dependent Variable
Datapoints
Line of
regression
independent Variables
X

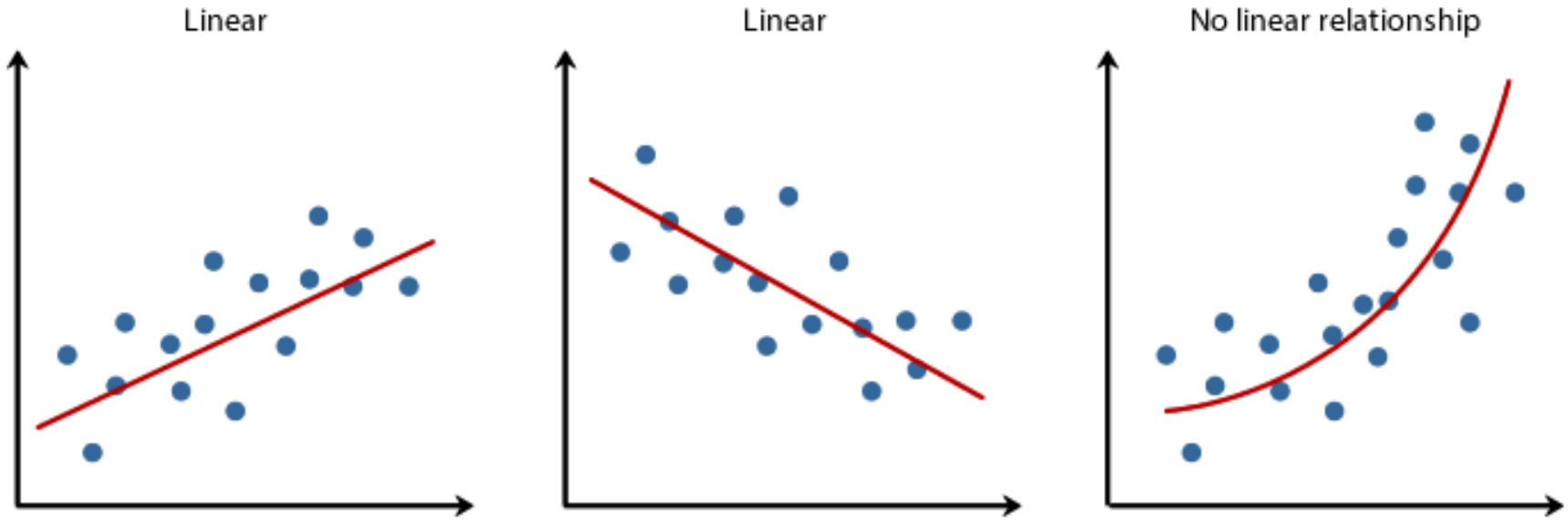
Contd..



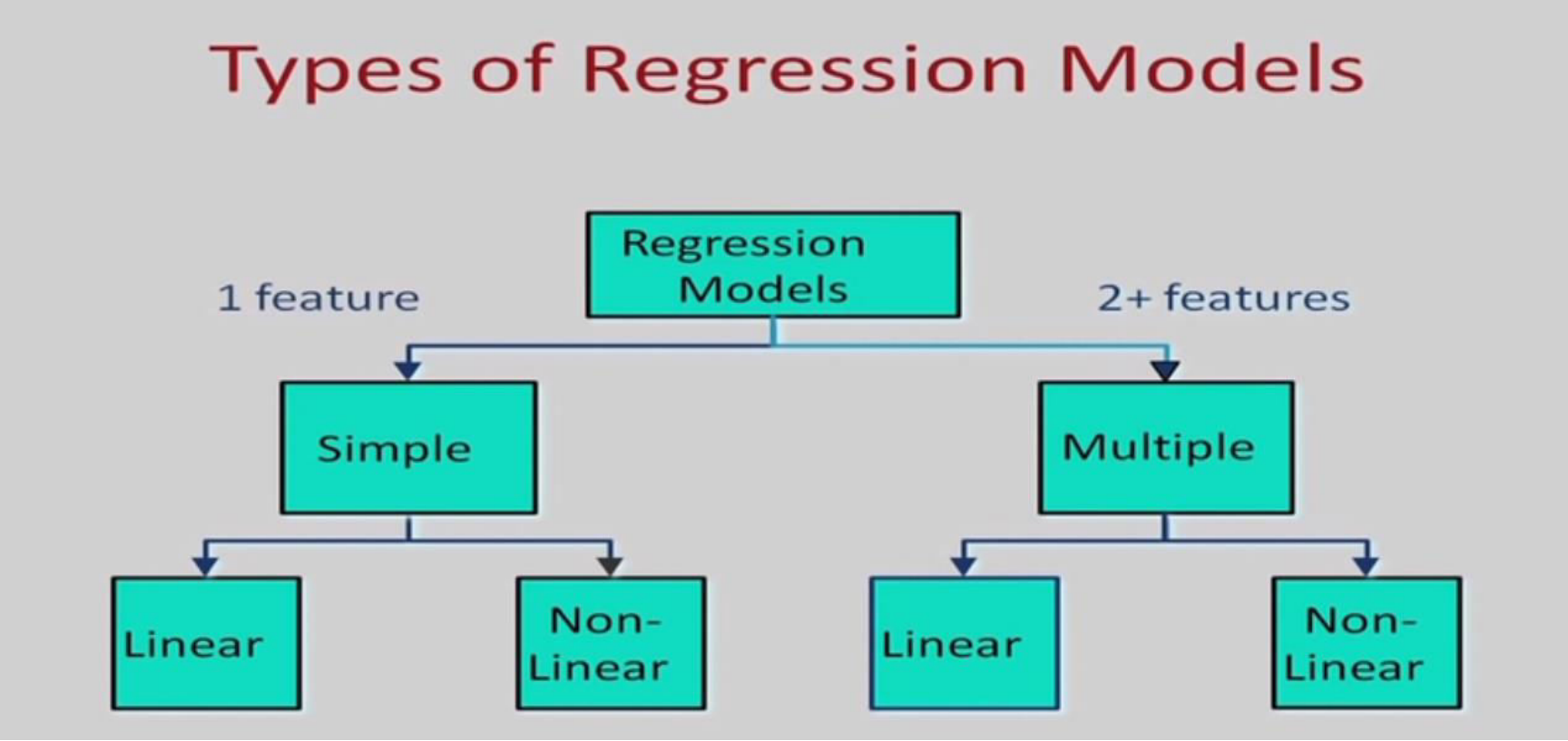
$$\text{Salary} = \beta_0 + \beta_1 * \text{Experience}$$

$$\beta_0 = \$\ 30K$$

Contd..



Types of regression



Linear Regression: Expression

- Mathematically, we can represent a linear regression as:

$$Y = \beta 0 + \beta 1 x + \varepsilon$$

Where,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
β0= intercept of the population
β1 = slope of the population
ε = random error

Contd..

$$\beta_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$\beta_0 = \frac{\sum y - \beta_1 \sum x}{n}$$

Problem: Predict the glucose level given the age?

S. No	Age (X)	Glucose level (Y)
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

Problem?

- Predict the Glucose level of a person whose age is 55 ?

Contd..

- Value of regression coefficient

$\beta 1 = 0.385225$

$\beta 0 = 65.14$

Loss function for regression

- It is a fundamental component of a machine learning or statistical model that is used to quantify the error or discrepancy between predicted values and actual target values (also known as ground truth) in a regression problem.

- The goal of regression is to find a model that minimizes this loss function, which essentially measures how well the model fits the data.

Types of Regression Loss Functions:

1. Mean Squared Error (MSE): The most commonly used loss function for regression. It measures the average of the squared differences between predicted and actual values. It's suitable when the errors should be penalized more for large deviations from the target values.

MSE Loss: $L(y, \hat{y}) = (1/n) * \Sigma(yi - \hat{y}i)^2$

2. Mean Absolute Error (MAE): Measures the average of the absolute differences between predicted and actual values. It's less sensitive to outliers compared to MSE.

MAE Loss: $L(y, \hat{y}) = (1/n) * \Sigma|yi - \hat{y}i|$

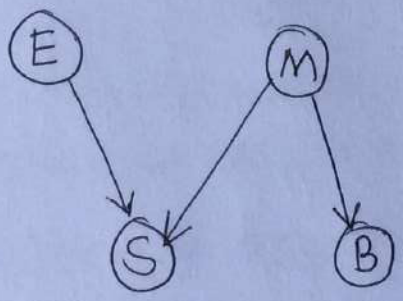
Contd..

- - Huber Loss: A hybrid loss function that combines the characteristics of MSE and MAE. It's less sensitive to outliers than MSE and provides a balance between the two.

4. A smell of Sulphur(S) can be caused by rotten Egg (E) or as a sign of the doom bought by Mayan Apocalypse (M). The Mayan Apocalypse also causes the oceans to boil (B). The BBN and Corresponding Tables are given. Calculate?

P(E)

+e	0.4
-e	0.6



P(M)

+m	0.1
-m	0.9

P(S|E,M)

+e	+m	+s	1.0
+e	+m	-s	0.0
+e	-m	+s	0.8
+e	-m	-s	0.2
-e	+m	+s	0.3
-e	+m	-s	0.7
-e	-m	+s	0.1
-e	-m	-s	0.9

P(B|M)

+m	+b	1.0
+m	-b	0.0
-m	+b	0.1
-m	-b	0.9

(i) what is Probability that Ocean boils?

(i) $P(B) = P(BM) + P(B\bar{M})$

$= P(B|M) * P(M) + P(B|\bar{m}) * P(\bar{m})$

$= 1 * 0.1 + 0.1 * 0.9 = 0.19$

(ii) what is Probability of Mayan Apocalypse is occuring, given the oceans are boiling.

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

$$= \frac{1 * 0.1}{0.19} = 0.52$$

(iii). what is Probability of Mayan Apocalypse given the ocean are boiling, sulphur smelling and rotten Eggs) ?

$$P(M|SBE) = \frac{P(SBME)}{P(SBE)} = ?$$

$P(MSBE) = P(M) * P(E) * P(S|ME) * P(B|M)$

$= 0.1 * 0.4 * 1 * 1 = 0.04$

$P(SBE) = P(SBEM) + P(SBE\bar{M})$

$= 0.04 + P(E) * P(S|\bar{M}E) * P(B|\bar{M})$

$= 0.04 + 0.9 * 0.4 * 0.8 * 0.1$

$= 0.04 + 0.0288 = 0.0688$

$$P(M|SBE) = \frac{0.04}{0.0688} = 0.581$$