



Lecture on

Parameter Estimation: *Theory of estimation*

Shesh N Sahu

Asst. Prof. CSE, UTD



Introduction: Estimation

- **Estimation**, in statistics, any of numerous procedures used to calculate the value of some property of a population from observations of a sample drawn from the population.
- **Estimation** in statistics helps companies, election officials, healthcare professionals, scientists, mathematicians, etc. to determine a trend in data.
- The main **purpose** of estimation in statistics is to be able to measure the behavior of data within a population.
- We can infer something about the population from statistical analysis of the sample which is known as Statistical Inference.

Contd..

- There are two types of problem:
 1. **Firstly** we may have no information at all about some characteristics of the population, especially the values of parameter involved in distribution. **(Problem of estimation)**
 2. **Secondly** some information or hypothetical value may be available, so it must be tested how far this is tenable **(Problem of Test of Hypothesis or Test of significance)**

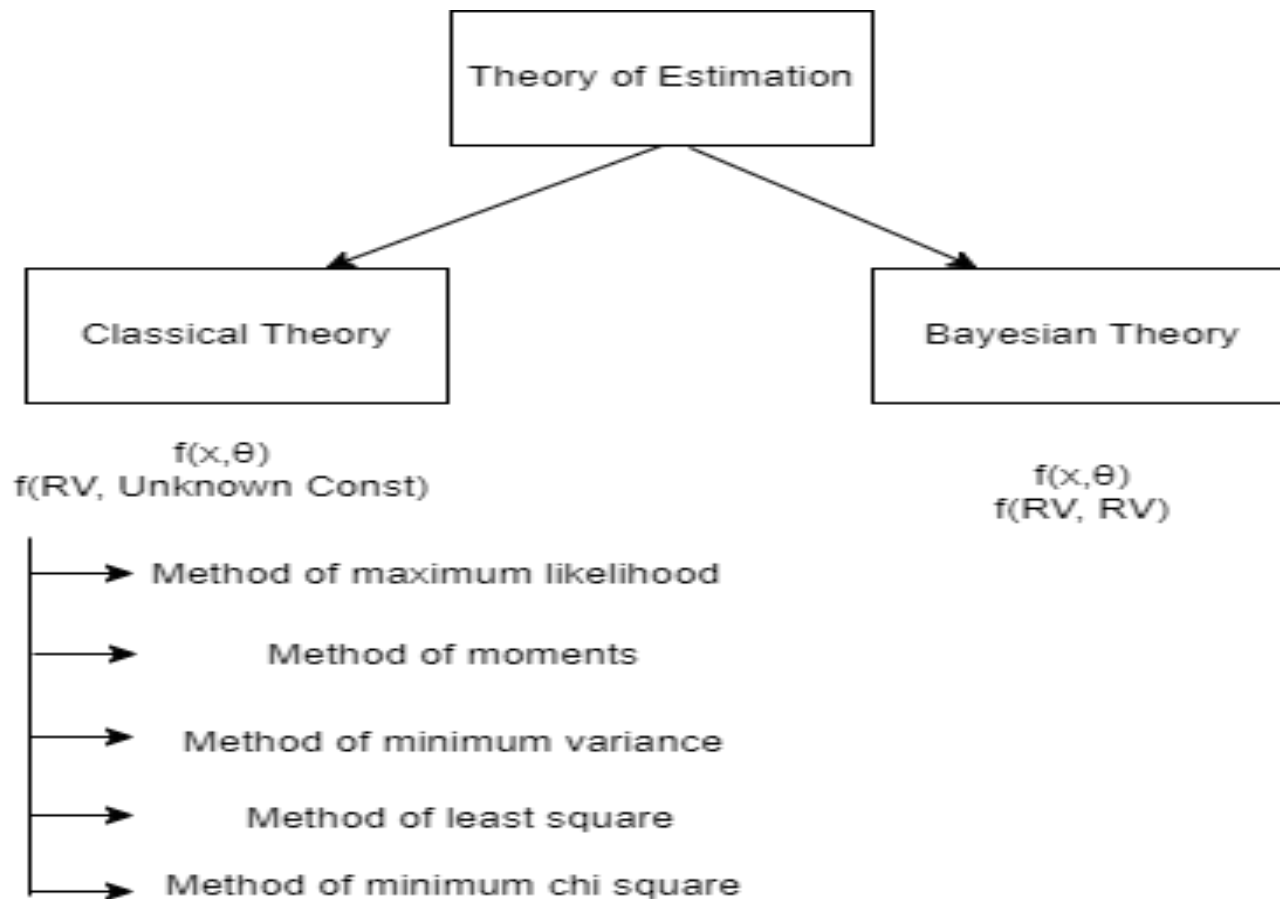
Theory of Estimation

- Suppose we have some random sample $x_1, x_2, x_3, \dots, x_n$ on variable x whose distribution in the population involves an unknown parameter ' θ '.
- It can be written as $f(x, \theta)$
- It is required to find an estimate of ' θ ' on the basis of sample space.
- Estimation can be done in two ways-
 - 1) **Point Estimation**
 - 2) **Interval Estimation**

Contd..

- In **point estimation**, the estimated value is given by a single value, which is a function of sample observation.
- This function is called **Estimator** of the parameter and value in particular sample is called **estimate**.
- In a interval estimation, an interval within which the parameter is expected to lie is given by two entities called **Confidence interval**.
- The two quantities which are used to specify the interval is called **Confidence limit**.

Contd..





Property of Good Estimator

- Unbiasedness
- Consistency
- Efficiency
- Sufficiency

Method of maximum likelihood

- This is a convenient method for finding an estimator which satisfy most of the criteria discussed earlier.
- Let $x_1, x_2, x_3, \dots, x_n$ be the random sample from a population with pdf: $f(\mathbf{x}, \theta)$, where θ is the parameter, the a distribution can be expressed as:

$$L = f(x_1, \theta) f(x_2, \theta), f(x_3, \theta), \dots, \dots, \dots f(x_n, \theta)$$

- This is called likelihood function.

Contd..

- The method of Maximum likelihood consists in choosing as an estimator of ' θ ' that statistic, when submitted for ' θ ', maximize the likelihood function L.
- The goal of MLE is to maximize the likelihood function:

$$L = f(x_1, x_2, x_3, x_4, \dots, x_n / \theta) = \prod_i^n f(x_i / \theta)$$

- For maximization-

$$\frac{\partial L}{\partial \theta} = 0$$

- Since it is difficult to take derivative of joint distribution, so we use log function.

Example:

- A unfair coin is flipped 100 times. 61 head are observed. The coin either has a probability $1/3$, $1/2$ and $2/3$ of flipping a head each time. Find which of the three is MLE?

Contd..

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

P = binomial probability

x = number of times for a specific outcome within n trials

$\binom{n}{x}$ = number of combinations

p = probability of success on a single trial

q = probability of failure on a single trial

n = number of trials



Contd..

$$P\left(H = 61/p = \frac{1}{3}\right) = \binom{100}{61} \left(\frac{1}{3}\right)^{61} \left(\frac{1}{3}\right)^{39}$$

Contd..

$$P\left(H = 61/p = \frac{1}{3}\right) = \binom{100}{61} \left(\frac{1}{3}\right)^{61} \left(\frac{1}{3}\right)^{39}$$

$$P\left(H = 61/p = \frac{1}{2}\right) = \binom{100}{61} \left(\frac{1}{2}\right)^{61} \left(\frac{1}{2}\right)^{39}$$

$$P\left(H = 61/p = \frac{2}{3}\right) = \binom{100}{61} \left(\frac{2}{3}\right)^{61} \left(\frac{1}{3}\right)^{39}$$

Questions:

- Find the MLE of the parameter of Poisson distribution $P(x, m)$?
- Find MLE of the variance of a Normal Distribution $N(\mu, \sigma^2)$

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.

$$\text{MSE Loss: } L(y, \hat{y}) = (1/n) * \sum (y_i - \hat{y}_i)^2$$

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.

$$\text{**MAE Loss:** } L(y, \hat{y}) = (1/n) * \sum |y_i - \hat{y}_i|$$

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

Parameter Estimation: *Theory of estimation*

Shesh N Sahu

Asst. Prof. CSE, UTD



Introduction: Estimation

- **Estimation**, in statistics, any of numerous procedures used to calculate the value of some property of a population from observations of a sample drawn from the population.
- **Estimation** in statistics helps companies, election officials, healthcare professionals, scientists, mathematicians, etc. to determine a trend in data.
- The main **purpose** of estimation in statistics is to be able to measure the behavior of data within a population.
- We can infer something about the population from statistical analysis of the sample which is known as Statistical Inference.

Contd..

- There are two types of problem:
 1. **Firstly** we may have no information at all about some characteristics of the population, especially the values of parameter involved in distribution. **(Problem of estimation)**
 2. **Secondly** some information or hypothetical value may be available, so it must be tested how far this is tenable **(Problem of Test of Hypothesis or Test of significance)**

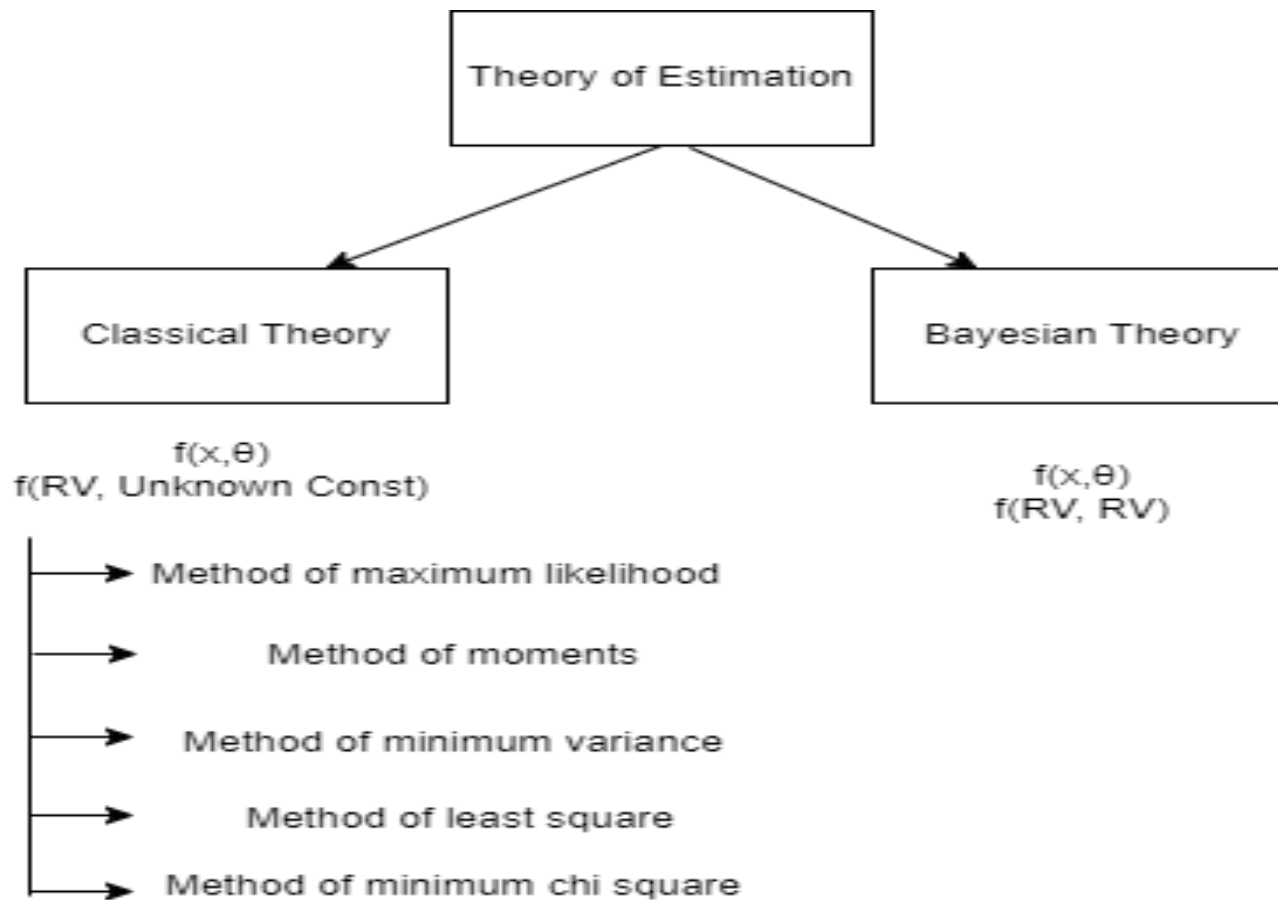
Theory of Estimation

- Suppose we have some random sample $x_1, x_2, x_3, \dots, x_n$ on variable x whose distribution in the population involves an unknown parameter ' θ '.
- It can be written as $f(x, \theta)$
- It is required to find an estimate of ' θ ' on the basis of sample space.
- Estimation can be done in two ways-
 - 1) Point Estimation**
 - 2) Interval Estimation**

Contd..

- In **point estimation**, the estimated value is given by a single value, which is a function of sample observation.
- This function is called **Estimator** of the parameter and value in particular sample is called **estimate**.
- In a interval estimation, an interval within which the parameter is expected to lie is given by two entities called **Confidence interval**.
- The two quantities which are used to specify the interval is called **Confidence limit**.

Contd..





Property of Good Estimator

- Unbiasedness
- Consistency
- Efficiency
- Sufficiency

Method of maximum likelihood

- This is a convenient method for finding an estimator which satisfy most of the criteria discussed earlier.
- Let $x_1, x_2, x_3, \dots, x_n$ be the random sample from a population with pdf: $f(\mathbf{x}, \theta)$, where θ is the parameter, the a distribution can be expressed as:

$$L = f(x_1, \theta) f(x_2, \theta), f(x_3, \theta), \dots, \dots, \dots f(x_n, \theta)$$

- This is called likelihood function.

Contd..

- The method of Maximum likelihood consists in choosing as an estimator of ' θ ' that statistic, when submitted for ' θ ', maximize the likelihood function L.
- The goal of MLE is to maximize the likelihood function:

$$L = f(x_1, x_2, x_3, x_4, \dots, x_n / \theta) = \prod_i^n f(x_i / \theta)$$

- For maximization-

$$\frac{\partial L}{\partial \theta} = 0$$

- Since it is difficult to take derivative of joint distribution, so we use log function.

Example:

- A unfair coin is flipped 100 times. 61 head are observed. The coin either has a probability $1/3$, $1/2$ and $2/3$ of flipping a head each time. Find which of the three is MLE?



Contd..

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

P = binomial probability

x = number of times for a specific outcome within n trials

$\binom{n}{x}$ = number of combinations

p = probability of success on a single trial

q = probability of failure on a single trial

n = number of trials



Contd..

$$P\left(H = 61/p = \frac{1}{3}\right) = \binom{100}{61} \left(\frac{1}{3}\right)^{61} \left(\frac{1}{3}\right)^{39}$$

Contd..

$$P\left(H = 61/p = \frac{1}{3}\right) = \binom{100}{61} \left(\frac{1}{3}\right)^{61} \left(\frac{1}{3}\right)^{39}$$

$$P\left(H = 61/p = \frac{1}{2}\right) = \binom{100}{61} \left(\frac{1}{2}\right)^{61} \left(\frac{1}{2}\right)^{39}$$

$$P\left(H = 61/p = \frac{2}{3}\right) = \binom{100}{61} \left(\frac{2}{3}\right)^{61} \left(\frac{1}{3}\right)^{39}$$

Questions:

- Find the MLE of the parameter of Poisson distribution $P(x, m)$?
- Find MLE of the variance of a Normal Distribution $N(\mu, \sigma^2)$

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.

$$\text{MSE Loss: } L(y, \hat{y}) = (1/n) * \sum (y_i - \hat{y}_i)^2$$

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.

$$\text{**MAE Loss:** } L(y, \hat{y}) = (1/n) * \sum |y_i - \hat{y}_i|$$

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.

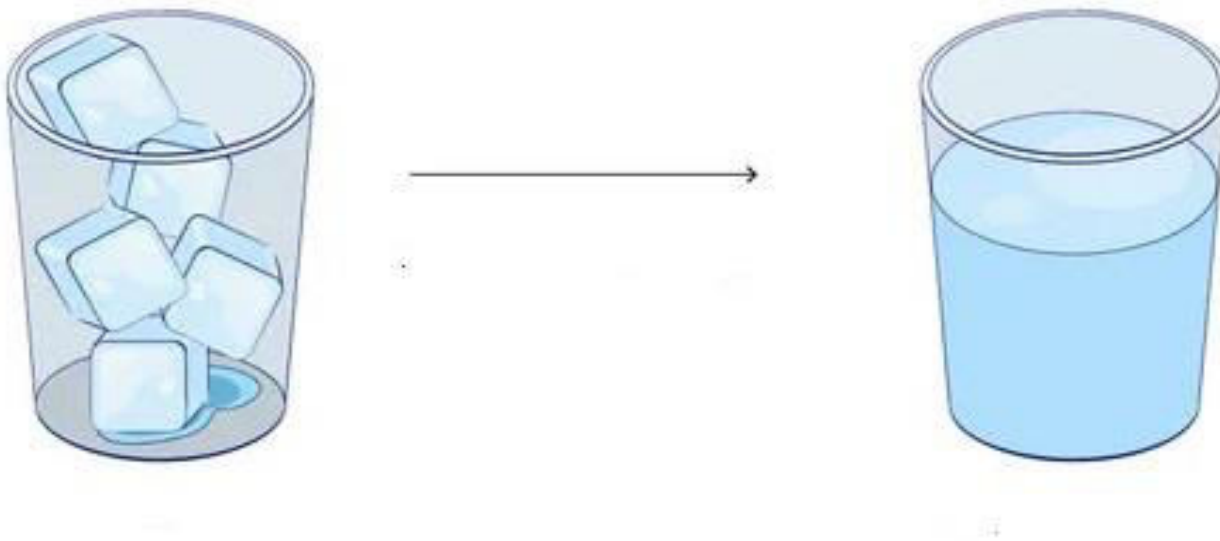
Lecture on

Parameter Estimation: *Maximum Entropy Estimation*

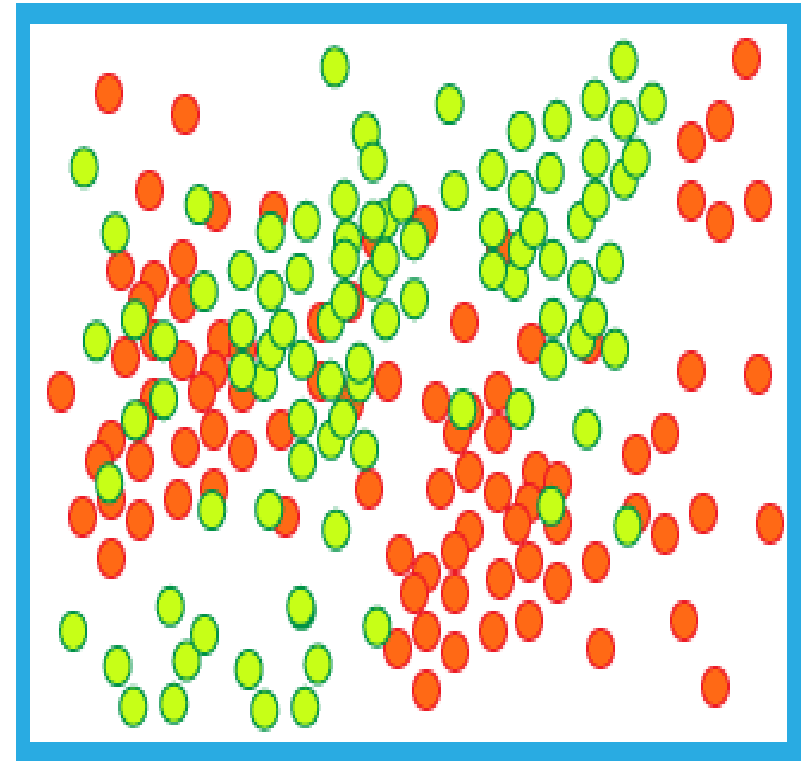
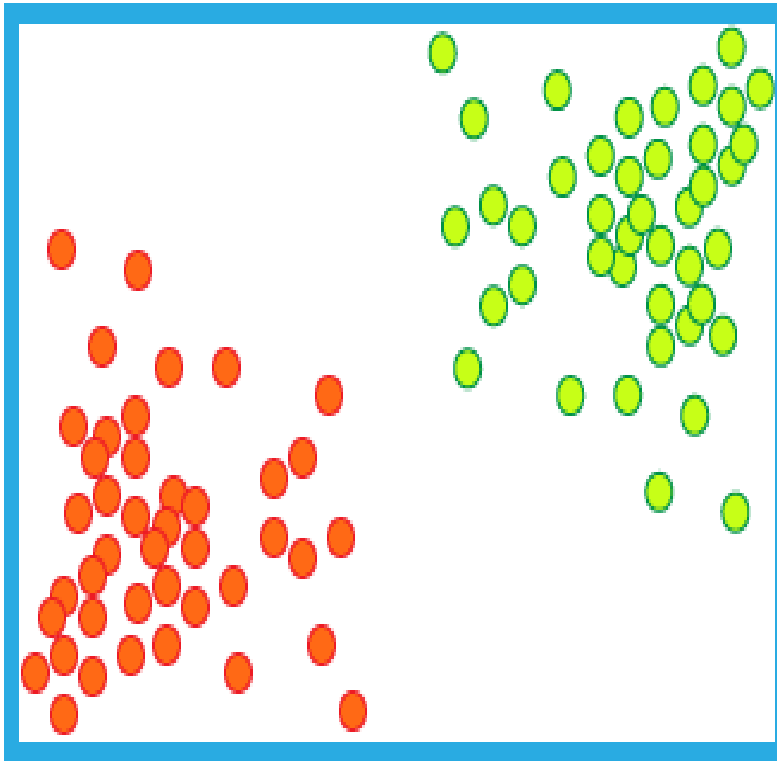
Introduction: Entropy

- Entropy is defined as the randomness or measuring the disorder of the information being processed in Machine Learning.
- We can say that entropy is the machine learning metric that measures the unpredictability or impurity in the system.
- The main **purpose** of entropy in statistics is to be able to measure the behavior of data within a population.

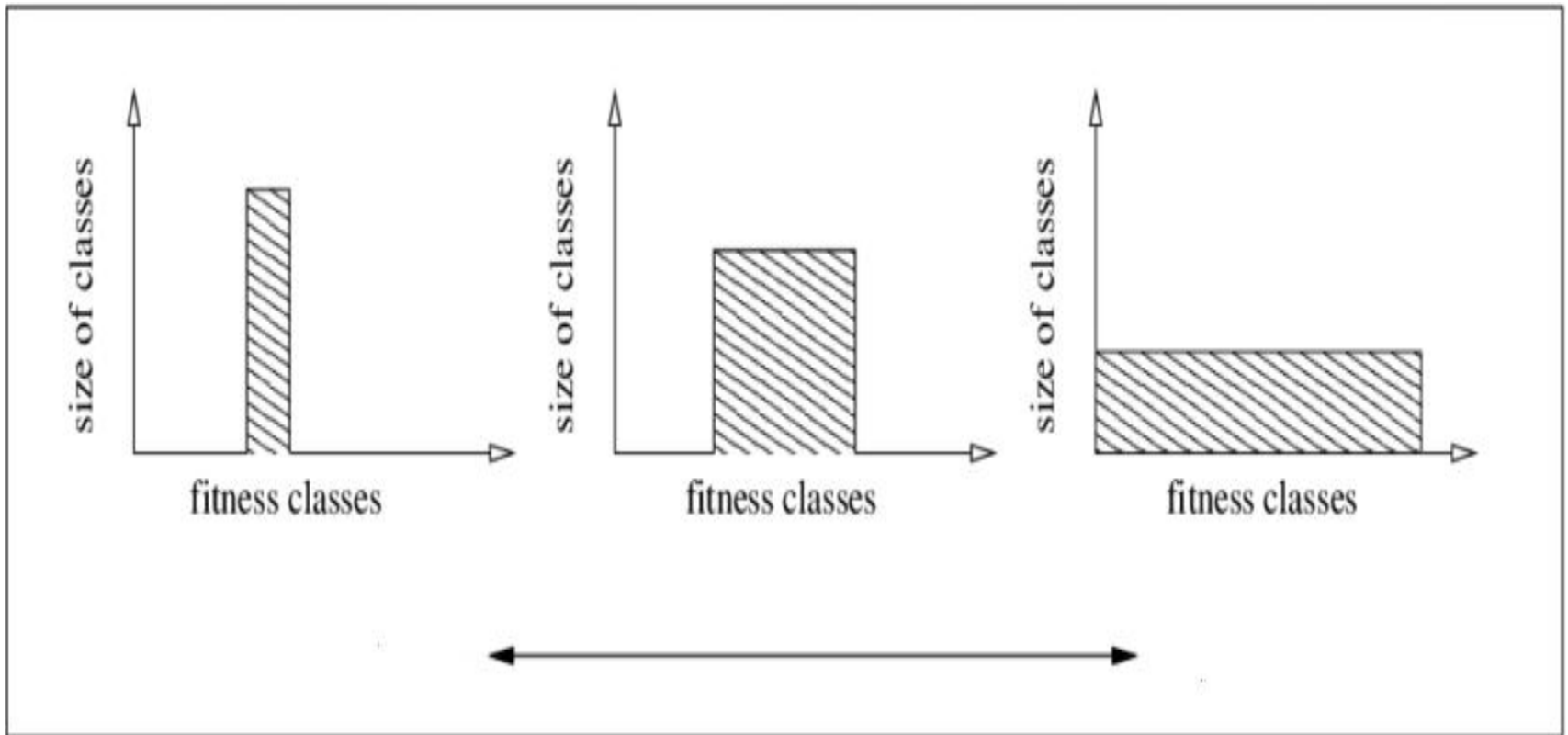
Example



Contd..



Contd..



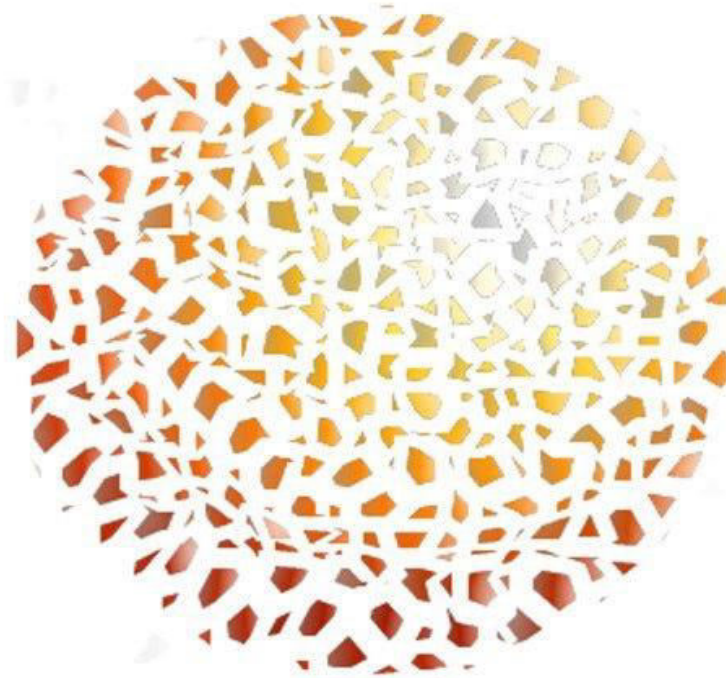
Contd..

ScienceABC



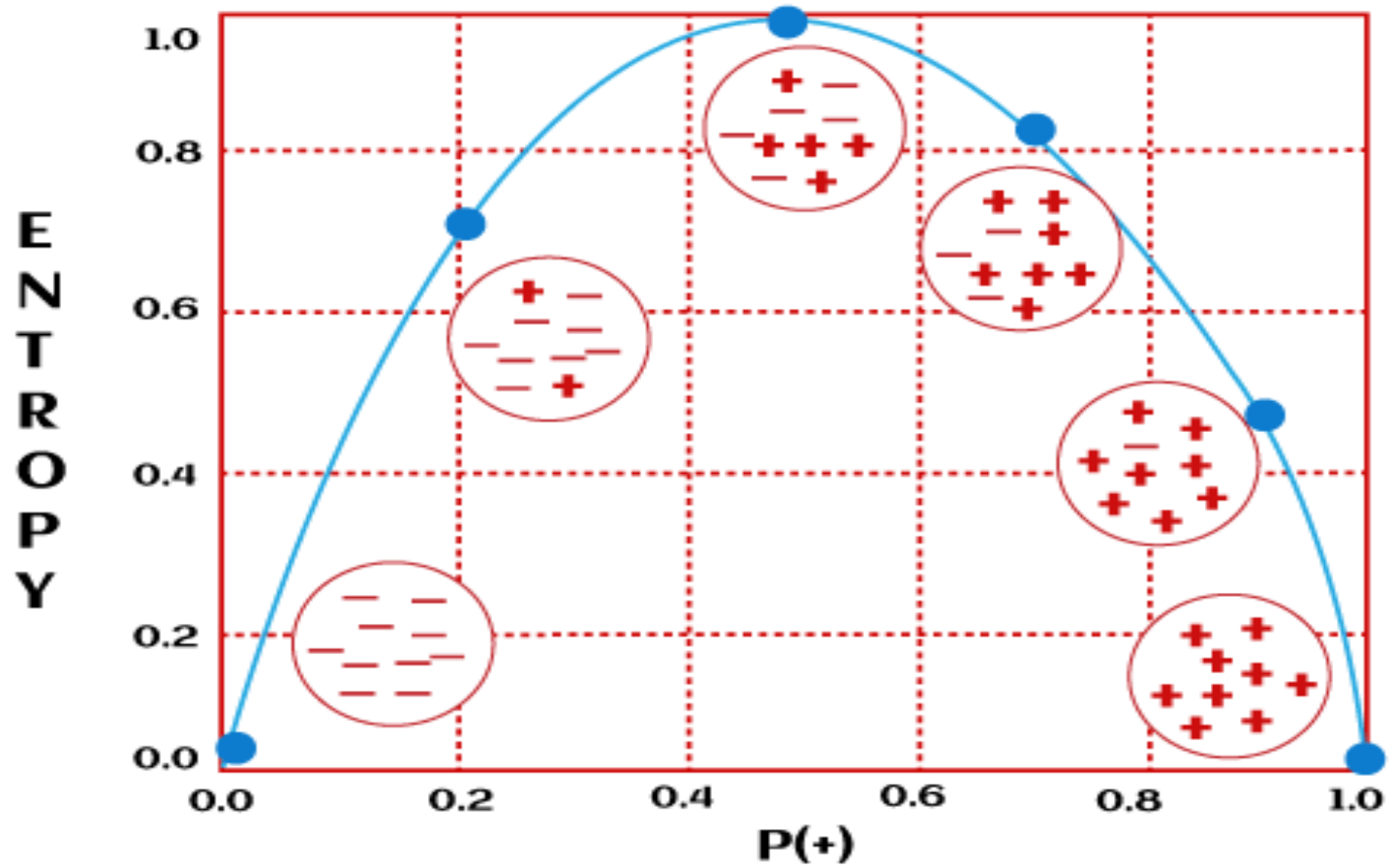
**Highly
Ordered**

**Increase
→
in Entropy**



More disordered

Contd..



Definition: Entropy

1. The entropy of a random variable X with distribution p , denoted by $H(X)$ or sometimes $H(p)$, is a measure of its uncertainty. In particular, for a discrete variable with K states, it is defined by

$$H(X) \triangleq - \sum_{k=1}^K p(X = k) \log_2 p(X = k)$$

Contd..

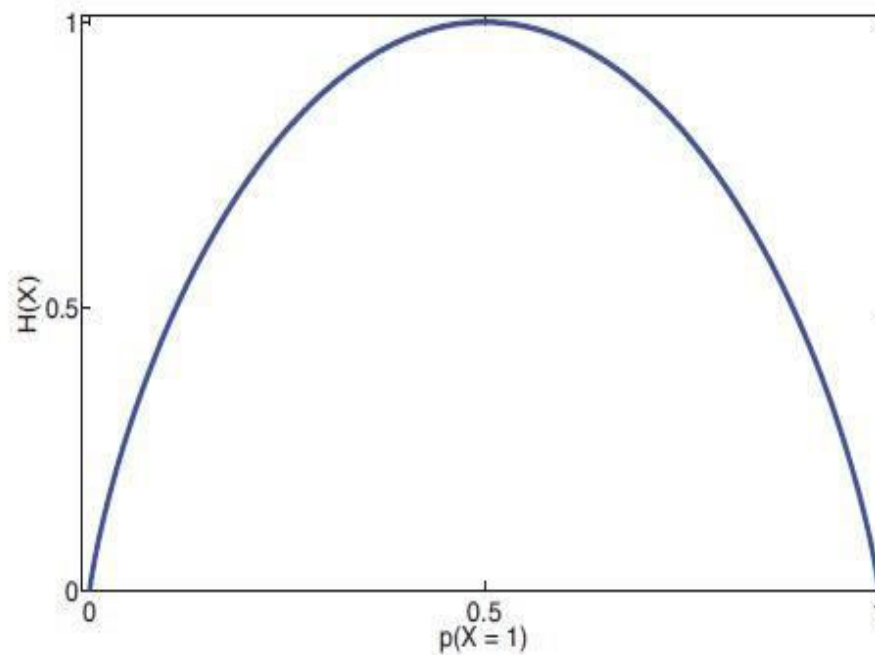


Figure Entropy of a Bernoulli random variable as a function of θ . The maximum entropy is $\log_2 2 = 1$. Figure generated by `bernoulliEntropyFig`.

Binary Entropy

- For the special case of binary random variables, $X \in \{0, 1\}$, we can write $p(X = 1) = \theta$
- and $p(X = 0) = 1 - \theta$. Hence the entropy becomes
- **Find $H(X)$?**



Contd..



Contd..

$$\begin{aligned}\mathbb{H}(X) &= -[p(X=1)\log_2 p(X=1) + p(X=0)\log_2 p(X=0)] \\ &= -[\theta \log_2 \theta + (1-\theta)\log_2(1-\theta)]\end{aligned}$$

Maximum entropy derivation of the exponential family

- Although the exponential family is convenient, is there any deeper justification for its use?
- It turns out that there is: it is the distribution that makes the least number of assumptions about the data, subject to a specific set of user-specified constraints.
- In particular, suppose all we know is the expected values of certain features or functions:

$$\sum f_k(x) p(x) = F_k$$

Principle: MEE

- The principle of maximum entropy or maxent says we should pick the distribution with maximum entropy (closest to uniform).
- To maximize entropy subject to the constraints in Equation 9.71, and the constraints that we need to use Lagrange multipliers.

Lagrange multiplier

- In mathematical optimization, the method of Lagrange multipliers is a strategy for finding the local maxima and minima of a function subject to equation constraints.
- It is named after the mathematician Joseph-Louis Lagrange.
- for functions $f(x), g(x)$, λ is called the Lagrange multiplier.

$$\mathcal{L}(x, \lambda) = f(x) + \langle \lambda, g(x) \rangle$$



- The Lagrangian is given by

$$J(p, \lambda) = - \sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) + \lambda_0 (1 - \sum_{\mathbf{x}} p(\mathbf{x})) + \sum_k \lambda_k (F_k - \sum_{\mathbf{x}} p(\mathbf{x}) f_k(\mathbf{x}))$$

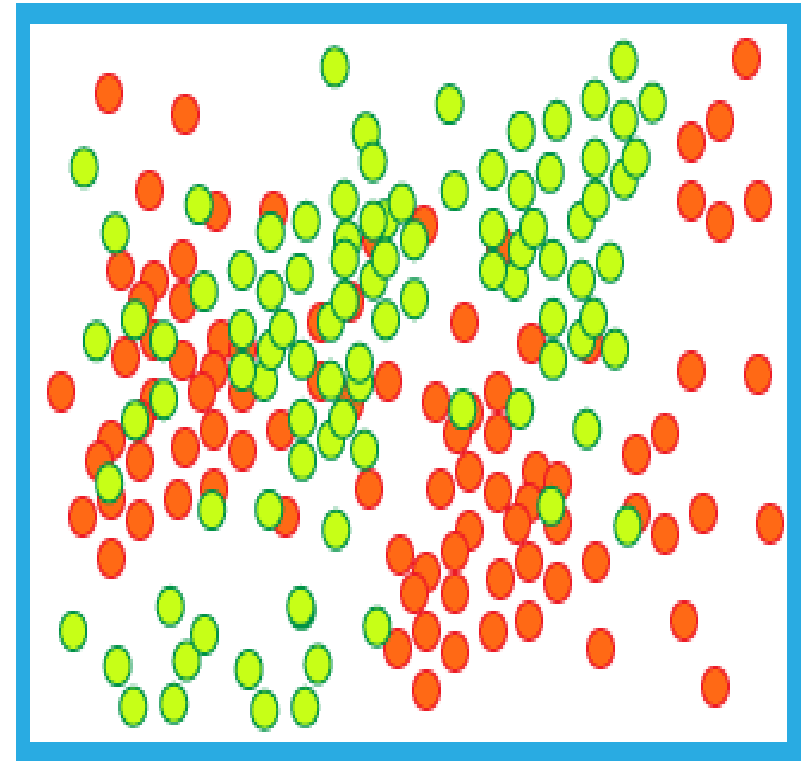
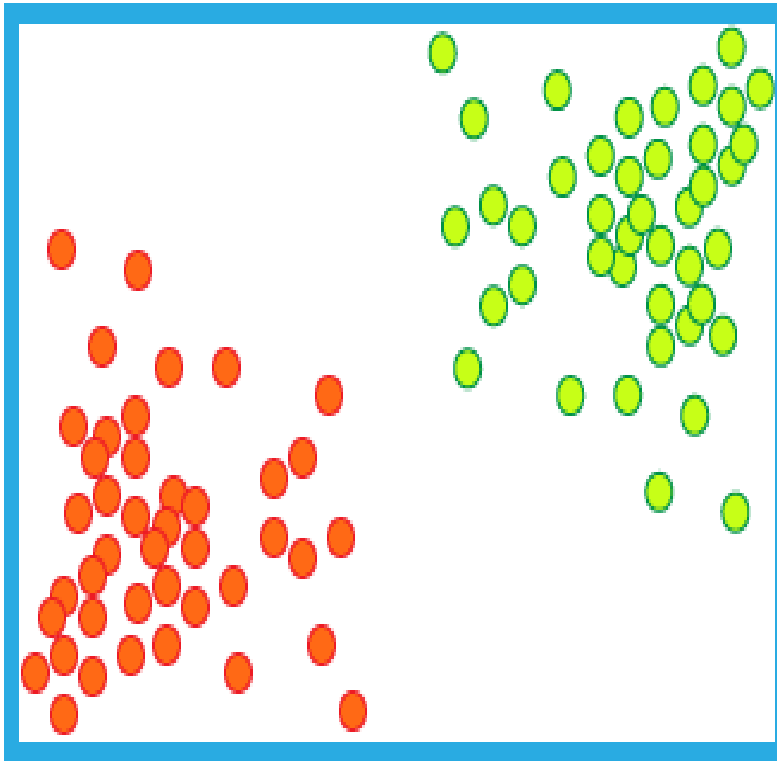


Lecture on Decision Tree

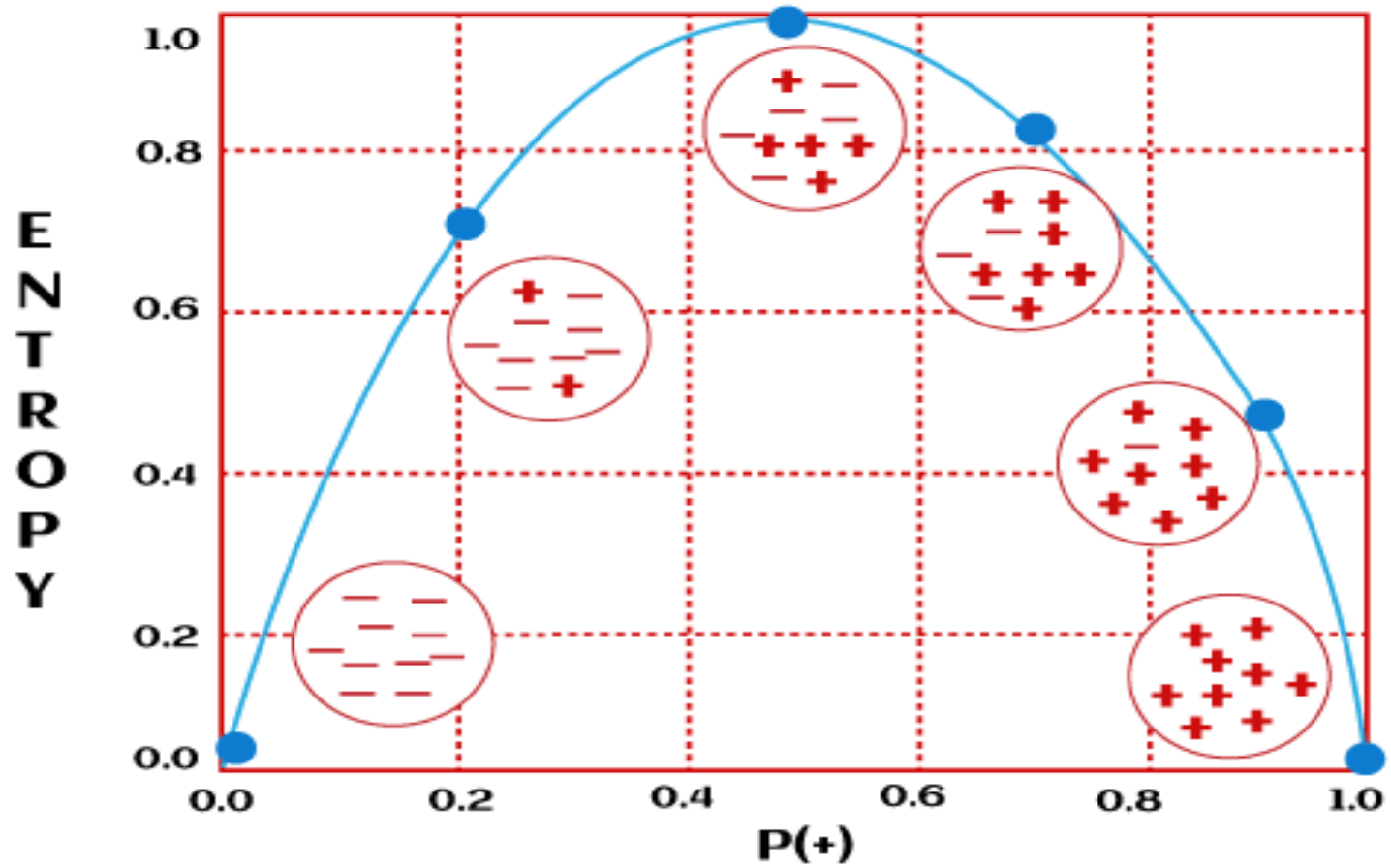
Introduction: Entropy

- Entropy is defined as the randomness or measuring the disorder of the information being processed in Machine Learning.
- We can say that entropy is the machine learning metric that measures the unpredictability or impurity in the system.
- The main **purpose** of entropy in statistics is to be able to measure the behavior of data within a population.

Contd..



Contd..



Definition: Entropy

1. The entropy of a random variable X with distribution p , denoted by $H(X)$ or sometimes $H(p)$, is a measure of its uncertainty. In particular, for a discrete variable with K states, it is defined by

$$H(X) \triangleq - \sum_{k=1}^K p(X = k) \log_2 p(X = k)$$



Decision Tree

- A decision tree is one of the most powerful tools of supervised learning algorithms used for both classification and regression tasks.
- It builds a tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.
- In particular, suppose all we know is the expected values of certain features or functions:



Born

April 30, 1916

Petoskey, Michigan, U.S.

Decision Tree: Evolution

- **Early beginnings (1950s-1960s):** The groundwork for decision tree algorithms can be traced back to the development of information theory by Claude Shannon in the 1940s(1948 paper "A Mathematical Theory of Communication", and is also referred to as Shannon entropy.
- **Iterative Dichotomiser 3 (ID3)** - Ross Quinlan, a computer scientist, introduced the ID3 algorithm in 1986. ID3 uses a greedy top-down approach and selects the best attribute to split the dataset based on information gain.

$$IG(S, A) = H(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \cdot H(S_v) \rightarrow \max(S, A)$$

Where:

$IG(S, A)$ is the Information Gain for splitting the dataset S based on attribute A ,

$H(S)$ is the entropy of the dataset S ,

$\text{Values}(A)$ is the set of all possible values of attribute A ,

S_v is the subset of S where attribute A has the value v ,

$|S_v|$ and $|S|$ denote the cardinality of the sets S_v and S , respectively.

Contd..

- **Classification and Regression Trees (CART)** - Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone introduced the CART algorithm in 1984. CART uses the Gini impurity index to choose the optimal split, and it can handle both classification and regression tasks.

$$Gini(p) = 1 - \sum_{i=1}^C p_i^2$$

Where:

$Gini(p)$ is the Gini impurity index.

C is the number of classes.

p_i is the proportion of samples in class i .



Decision Tree Terminologies

- **Root Node:** It is the topmost node in the tree, which represents the complete dataset. It is the starting point of the decision-making process.
- **Decision/Internal Node:** A node that symbolizes a choice regarding an input feature. Branching off of internal nodes connects them to leaf nodes or other internal nodes.
- **Leaf/Terminal Node:** A node without any child nodes that indicates a class label or a numerical value.

Contd..

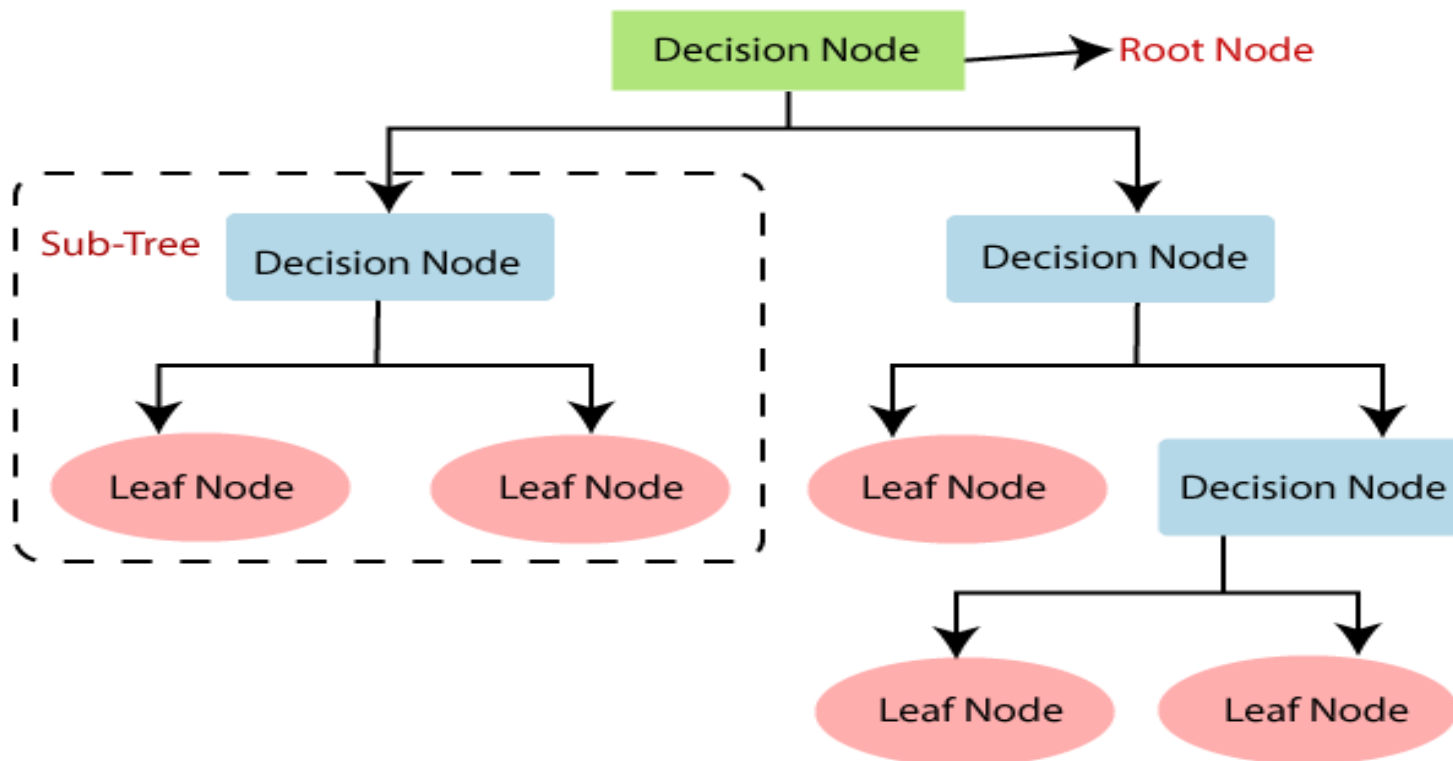
- **Impurity:** A measurement of the target variable's homogeneity in a subset of data. It refers to the degree of randomness or uncertainty in a set of examples. The Gini index and entropy are two commonly used impurity measurements in decision trees for classifications task.
- **Variance:** Variance measures how much the predicted and the target variables vary in different samples of a dataset.
- It is used for regression problems in decision trees. Sum of squared error, Mean Absolute Error, MSE, are used to measure loss function.

Contd..

- **Information Gain:** Information gain is a measure of the reduction in impurity achieved by splitting a dataset on a particular feature in a decision tree.
- The splitting criterion is determined by the feature that offers the greatest information gain.
- It is used to determine the most informative feature to split on at each node of the tree, with the goal of creating pure subsets

Contd..

■ Nodes in Decision Tree



Attribute selection measure

- ASM: By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are:
 1. Information Gain
 2. Gini Index

Information Gain

- Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.
- It calculates how much information a feature provides us about a class.
- According to the value of information gain, we split the node and build the decision tree.
- It can be calculated using the below formula:

$$\text{Information Gain} = \text{Entropy}(S) - [(\text{Weighted Avg}) * \text{Entropy}(\text{each feature})]$$

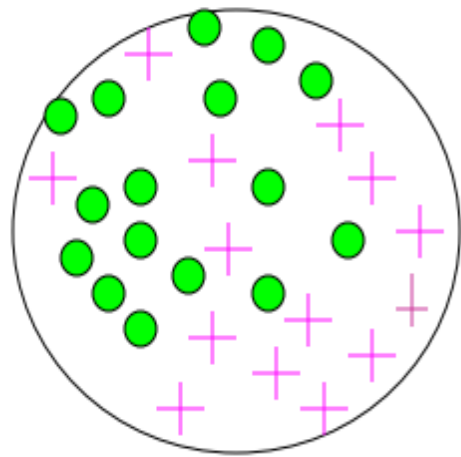
Gini Index:

- Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm.
- An attribute with the low Gini index should be preferred as compared to the high Gini index.
- It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.
- Gini index can be calculated using the below formula:

$$\text{Gini Index} = 1 - \sum_j P_j^2$$

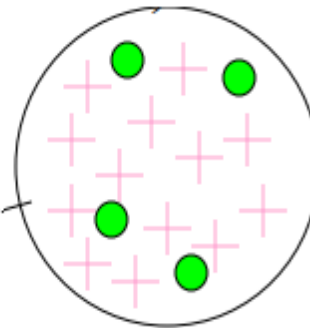
Sample Data: IG calculation

Group of people (Male & Female)



Age > 30

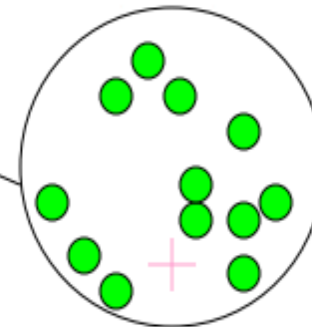
Child 1



17 People

Age < 30

Child 2

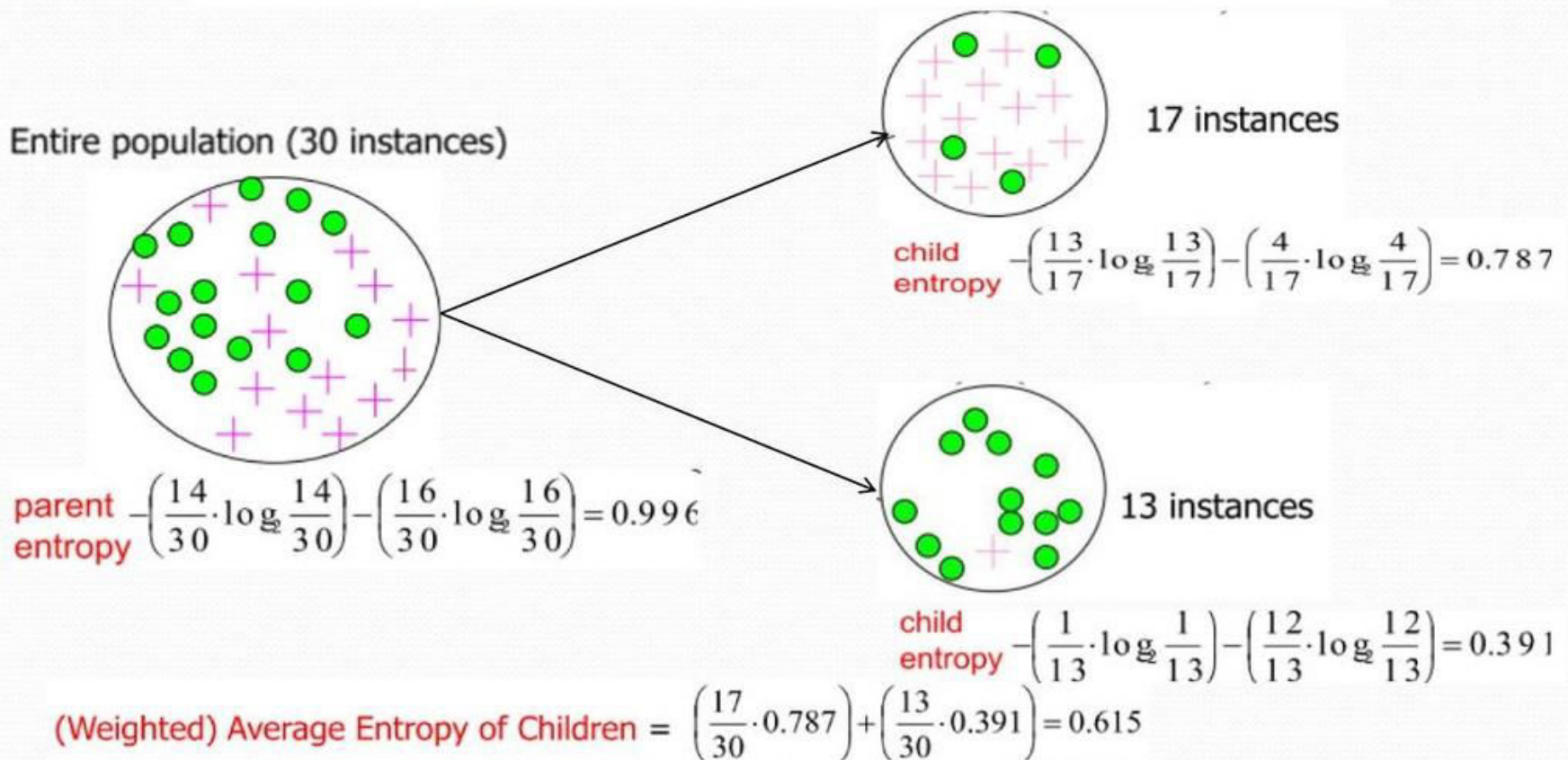


13 People



Contd..

Information Gain = entropy(parent) – [average entropy(children)]




| 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 |






- Entropy(S)= 0.94



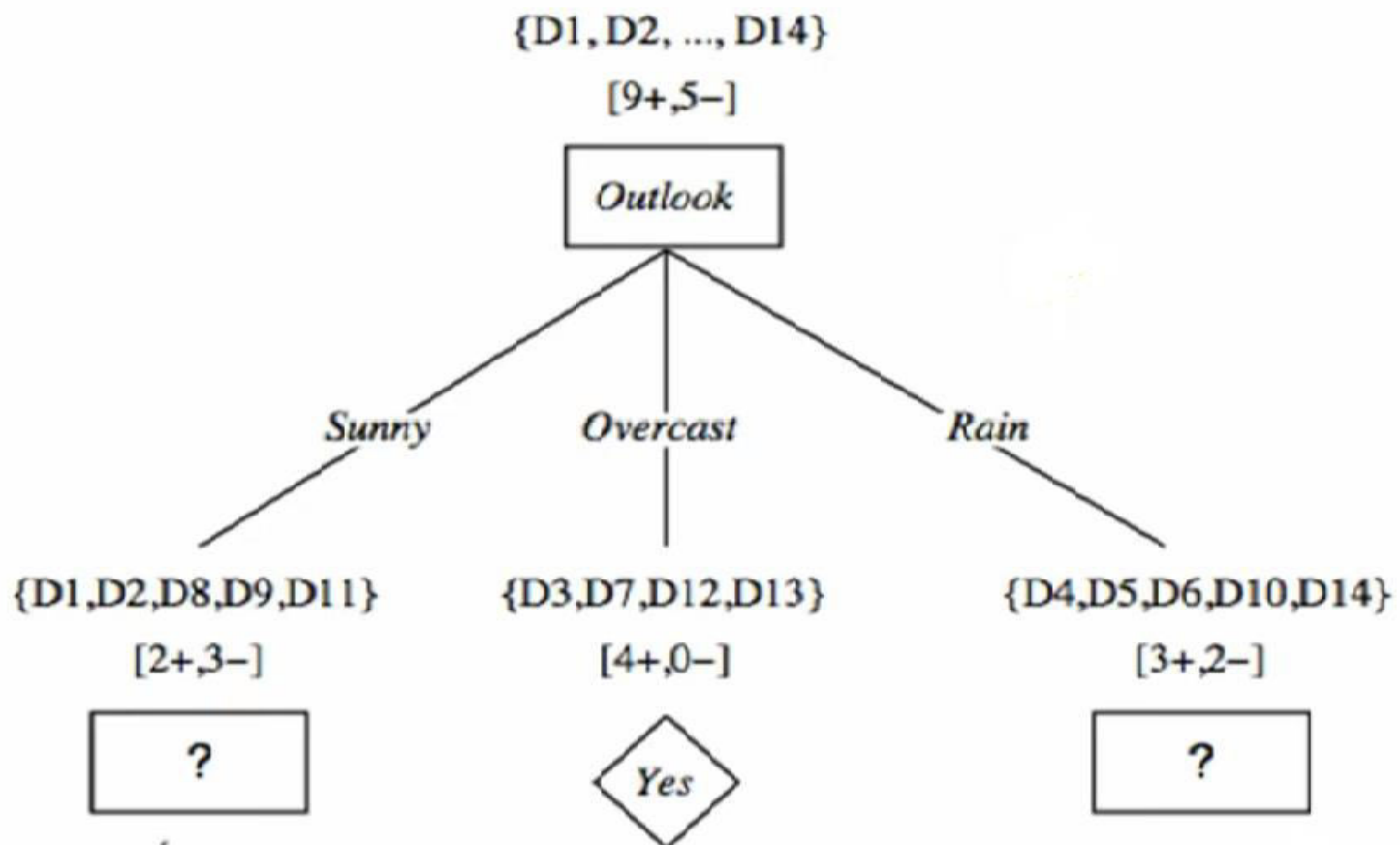
- 
- Entropy (Sunny): 0.971
 - Entropy (Overcast): 0
 - Entropy (Rain): 0.971



- 
- Information Gain (S, Outlook)= 0.2464

IG calculation of each attribute

- Information Gain (S, Outlook)= 0.2464
- Information Gain (S, Temp)= 0.0289
- Information Gain (S, Humidity)= 0.1516
- Information Gain (S, Wind)= 0.0478





Lecture on Decision Tree

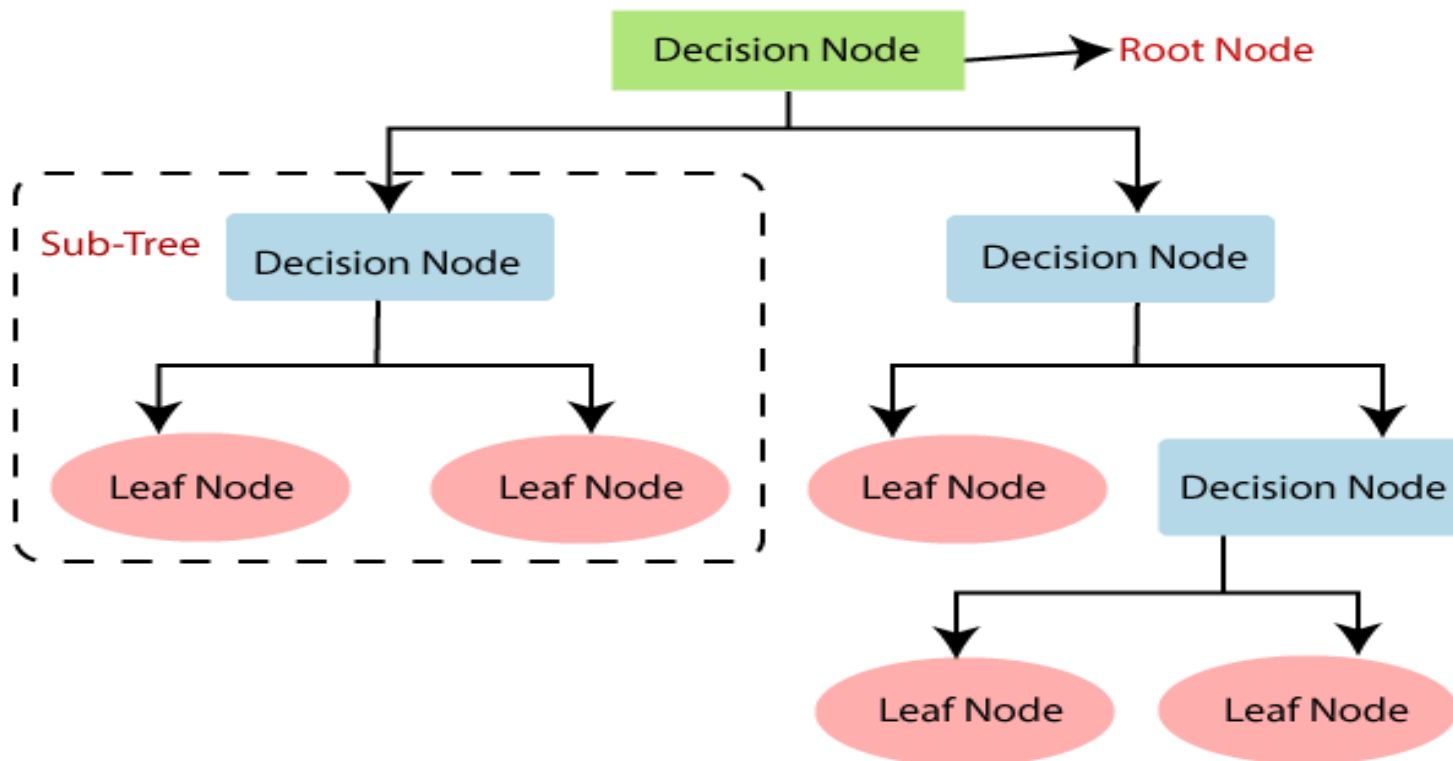


Decision Tree

- A decision tree is one of the most powerful tools of supervised learning algorithms used for both classification and regression tasks.
- It builds a tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.
- In particular, suppose all we know is the expected values of certain features or functions:

Contd..

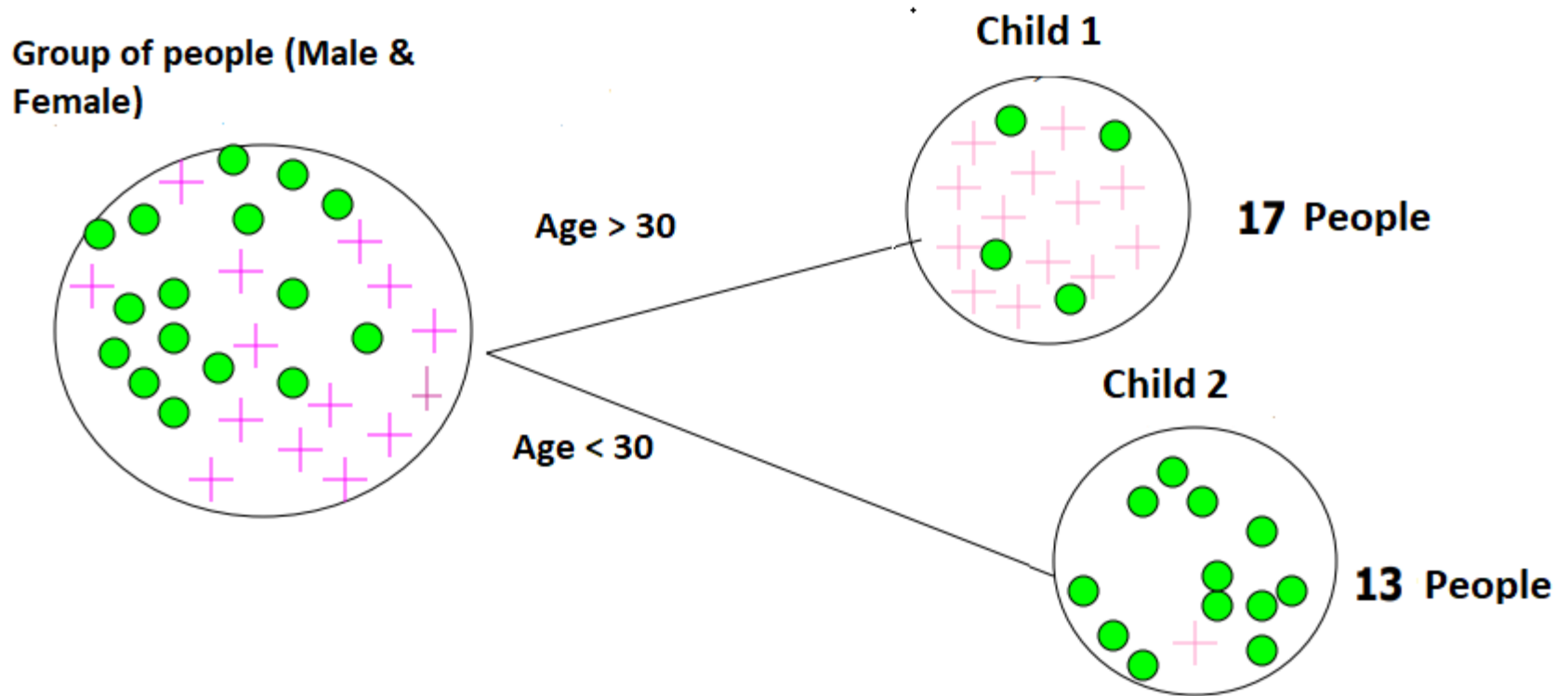
■ Nodes in Decision Tree



Attribute selection measure

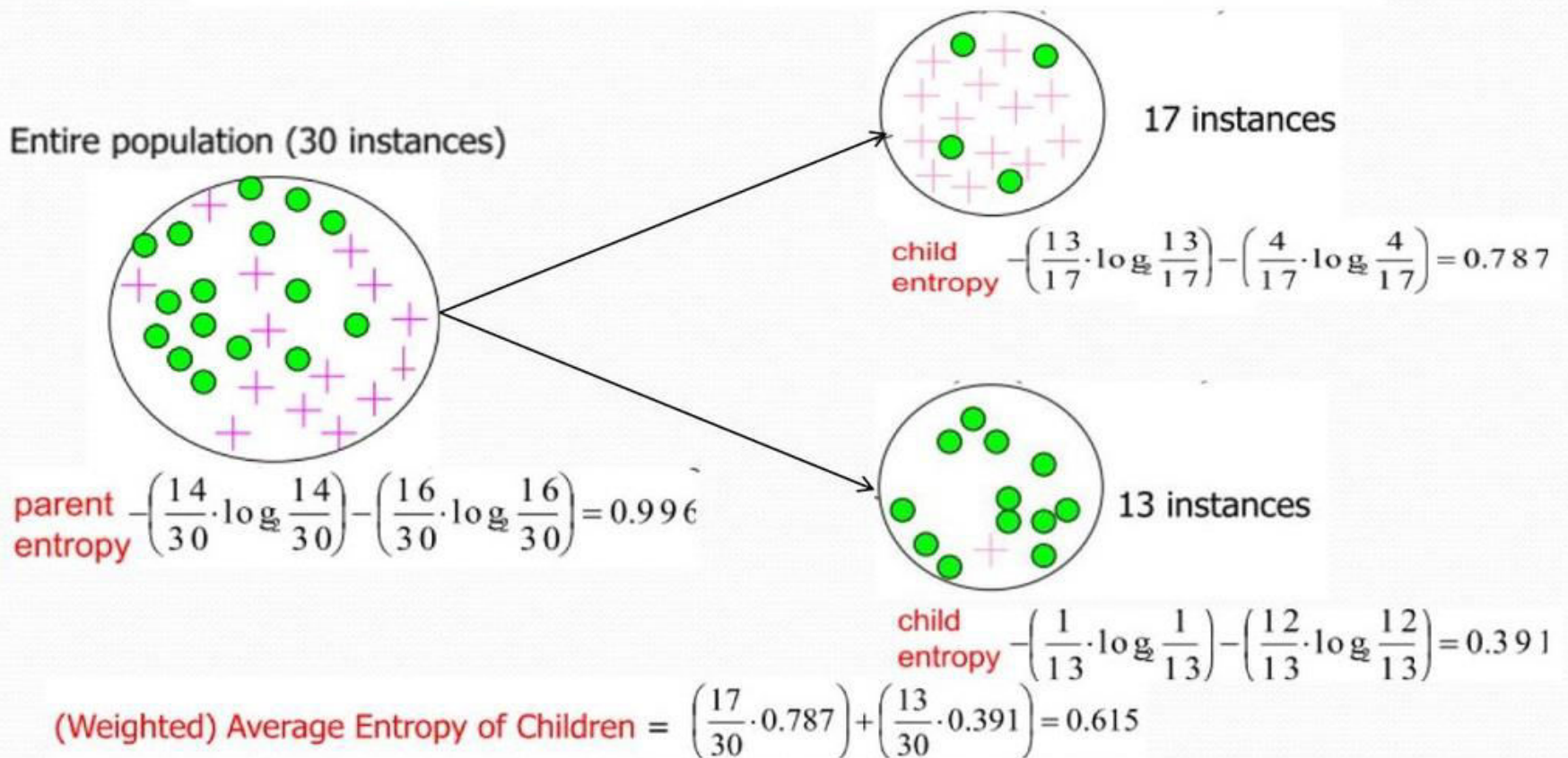
- ASM: By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are:
 1. Information Gain
 2. Gini Index


Sample Data: IG calculation



Contd..

Information Gain = entropy(parent) – [average entropy(children)]





| 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 |





- Entropy(S)= 0.94



Information Gain

- Outlook
- Humidity
- Temperature
- Wind

Attribute: Outlook

- Attribute Value outlook:

- Sunny
- Overcast
- Rain

$$S = [9+, 5-]$$

$$Entropy(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

$$S_{Sunny} \leftarrow [2+, 3-]$$

$$Entropy(S_{Sunny}) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.971$$


$$S_{Overcast} \leftarrow [4+, 0-]$$

$$Entropy(S_{Overcast}) = -\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4} = 0$$


$$S_{Rain} \leftarrow [3+, 2-]$$

$$Entropy(S_{Rain}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.971$$

$$Gain(S, Outlook) = Entropy(S) - \sum_{v \in \{Sunny, Overcast, Rain\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

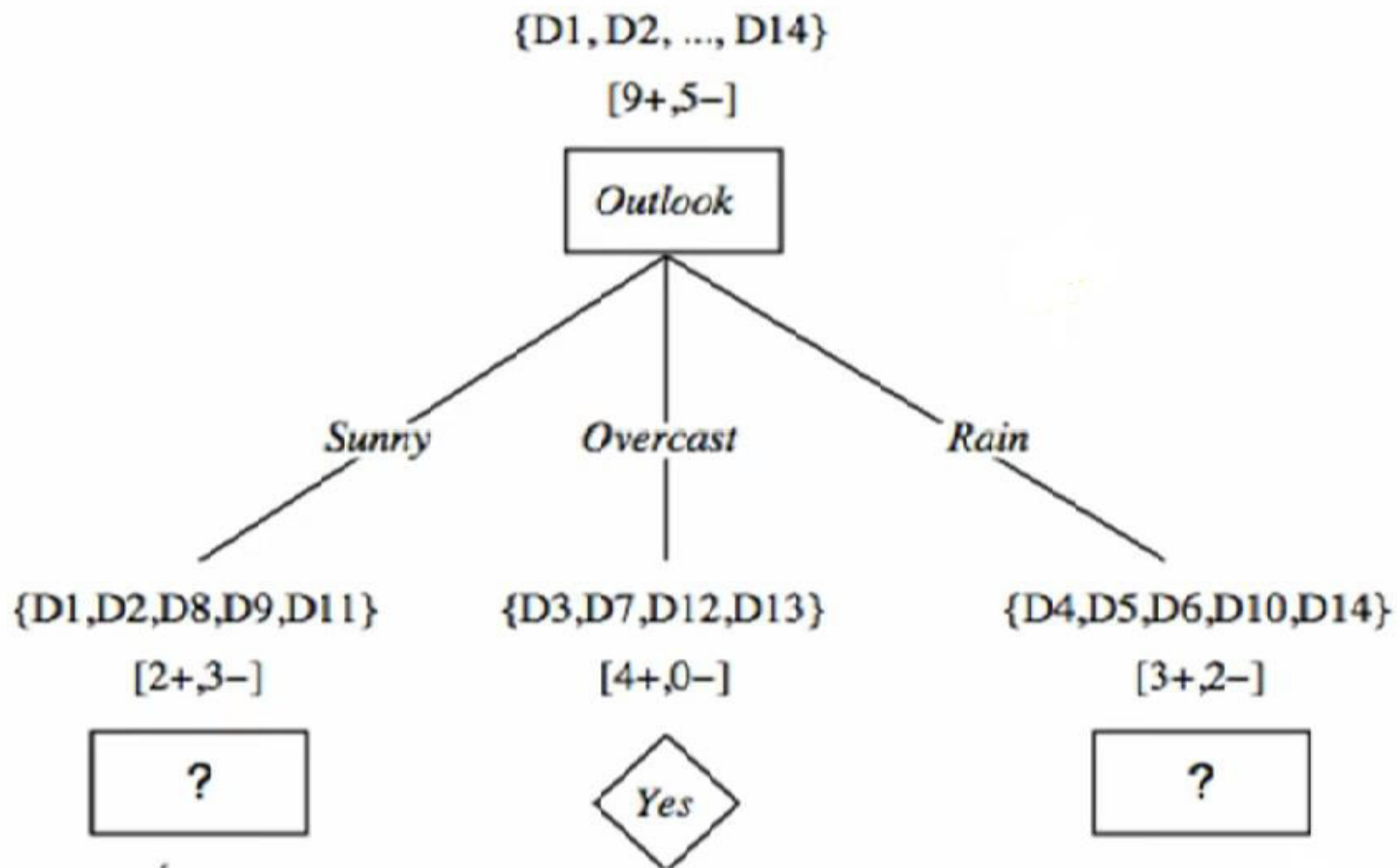
- 
- Entropy (Sunny): 0.971
 - Entropy (Overcast): 0
 - Entropy (Rain): 0.971



- 
- Information Gain (S, Outlook)= 0.2464

IG calculation of each attribute

- Information Gain (S, Outlook)= 0.2464
- Information Gain (S, Temp)= 0.0289
- Information Gain (S, Humidity)= 0.1516
- Information Gain (S, Wind)= 0.0478



Outlook: Sunny

| Day | Temp | Humidity | Wind | Play Tennis |
|-----|------|----------|--------|-------------|
| D1 | Hot | High | Weak | No |
| D2 | Hot | High | Strong | No |
| D8 | Mild | High | Weak | No |
| D9 | Cool | Normal | Weak | Yes |
| D11 | Mild | Normal | Strong | Yes |

IG: Temperature

Values (Temp) = Hot, Mild, Cool

$$S_{Sunny} = [2+, 3-]$$

$$Entropy(S_{Sunny}) = -\frac{2}{5}\log_2 \frac{2}{5} - \frac{3}{5}\log_2 \frac{3}{5} = 0.97$$

$$S_{Hot} \leftarrow [0+, 2-]$$

$$Entropy(S_{Hot}) = 0.0$$

$$S_{Mild} \leftarrow [1+, 1-]$$

$$Entropy(S_{Mild}) = 1.0$$

$$S_{Cool} \leftarrow [1+, 0-]$$

$$Entropy(S_{Cool}) = 0.0$$

$$Gain(S_{Sunny}, Temp) = Entropy(S) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Sunny}, Temp)$$

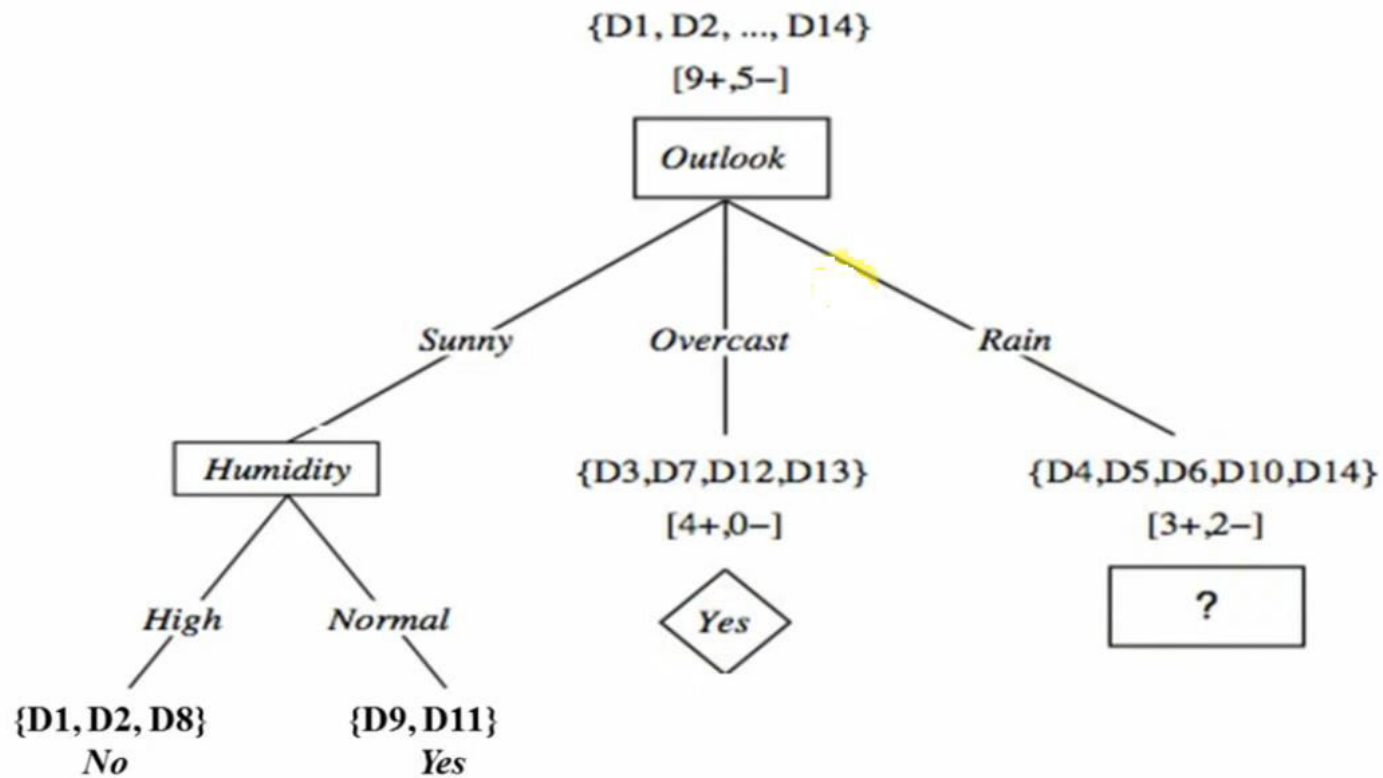
$$= Entropy(S) - \frac{2}{5} Entropy(S_{Hot}) - \frac{2}{5} Entropy(S_{Mild})$$

$$- \frac{1}{5} Entropy(S_{Cool})$$

$$Gain(S_{Sunny}, Temp) = 0.97 - \frac{2}{5} 0.0 - \frac{2}{5} 1 - \frac{1}{5} 0.0 = 0.570$$

- $IG(S_{\text{sunny}}, \text{Humidity}) = 0.570$
- $IG(S_{\text{sunny}}, \text{Humidity}) = 0.97$
- $IG(S_{\text{sunny}}, \text{Wind}) = 0.0192$


Decision Tree: Outlook, Humidity, ?



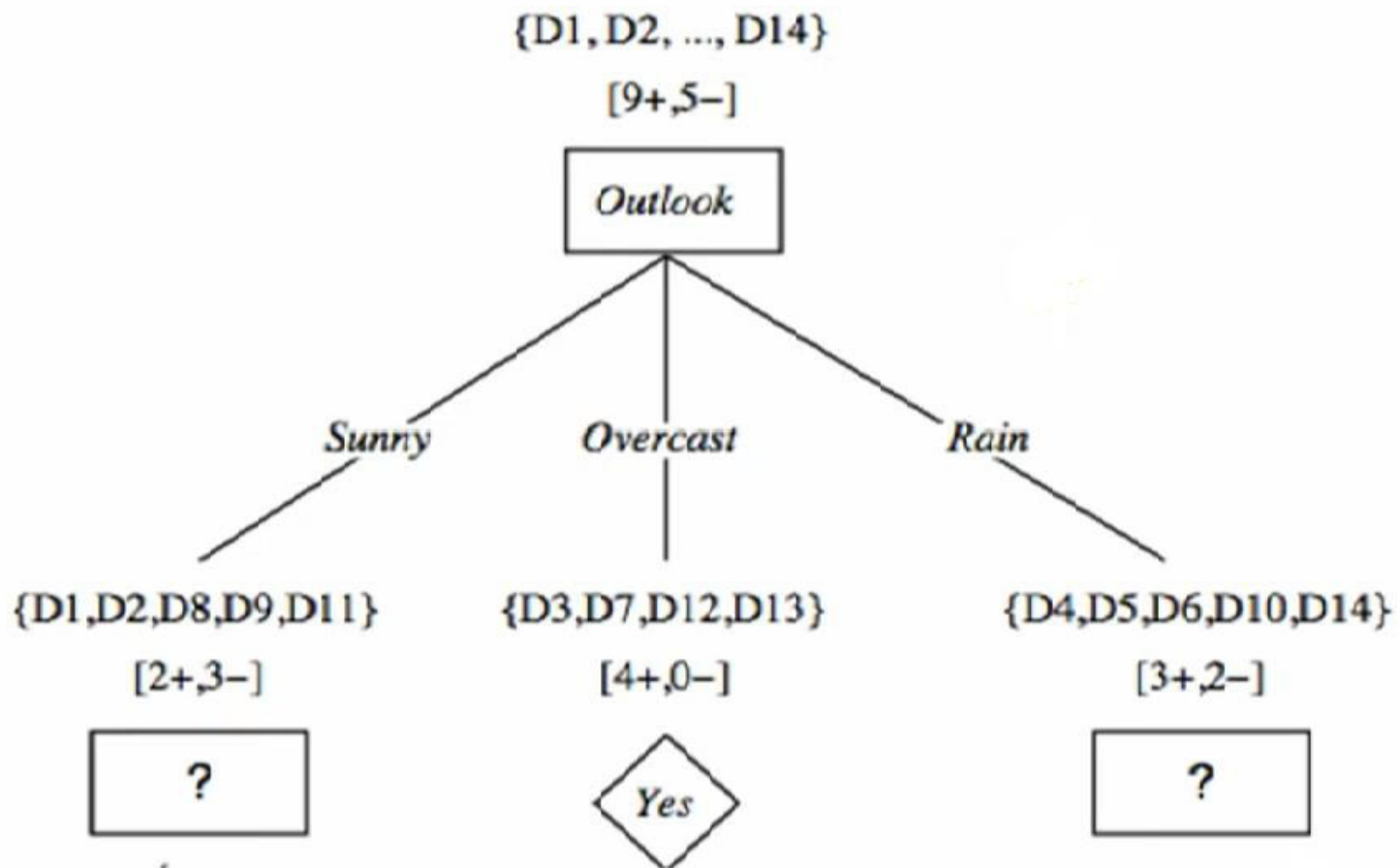


Lecture on

Issue in Decision Tree



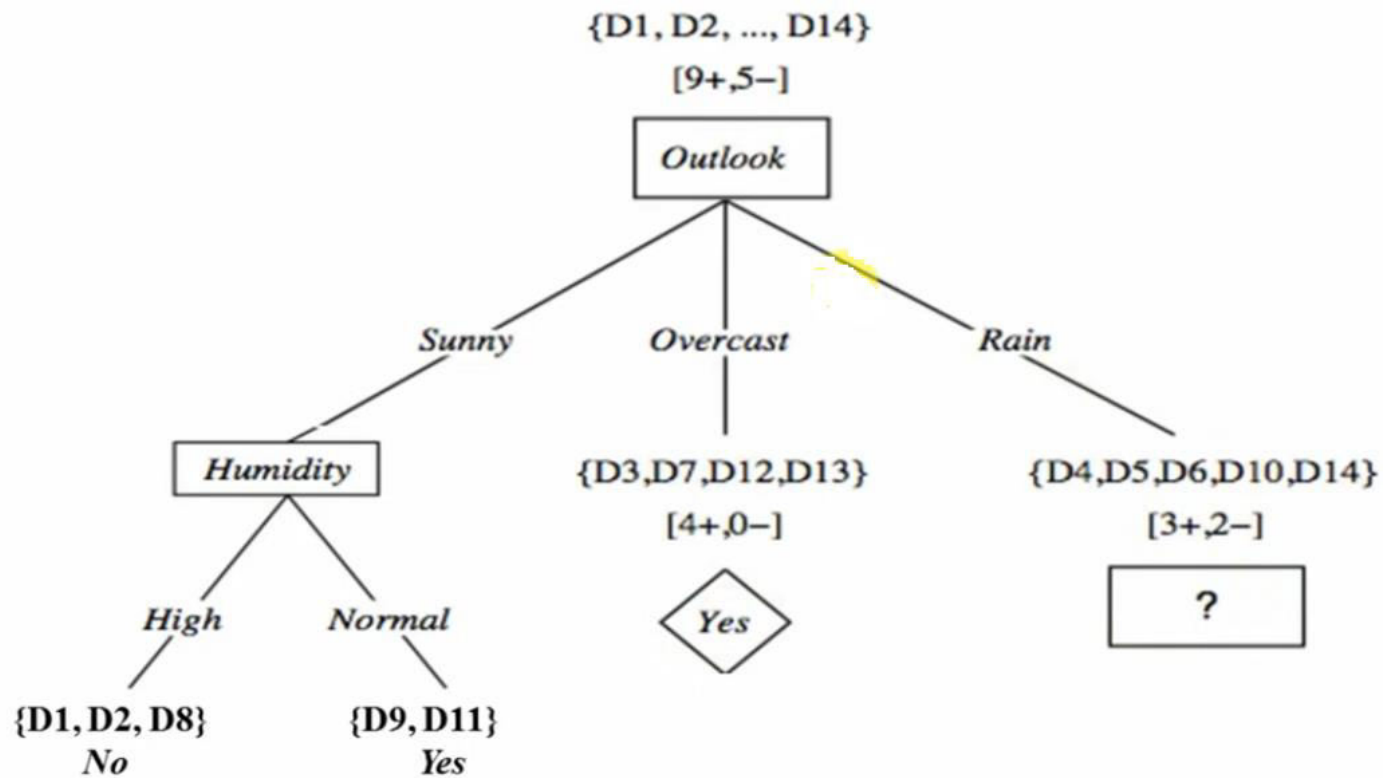
| Day | Outlook | Temperature | Humidity | Wind | Play Tennis |
|-----|----------|-------------|----------|--------|-------------|
| 1 | Sunny | Hot | High | Weak | No |
| 2 | Sunny | Hot | High | Strong | No |
| 3 | Overcast | Hot | High | Weak | Yes |
| 4 | Rain | Mild | High | Weak | Yes |
| 5 | Rain | 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 |



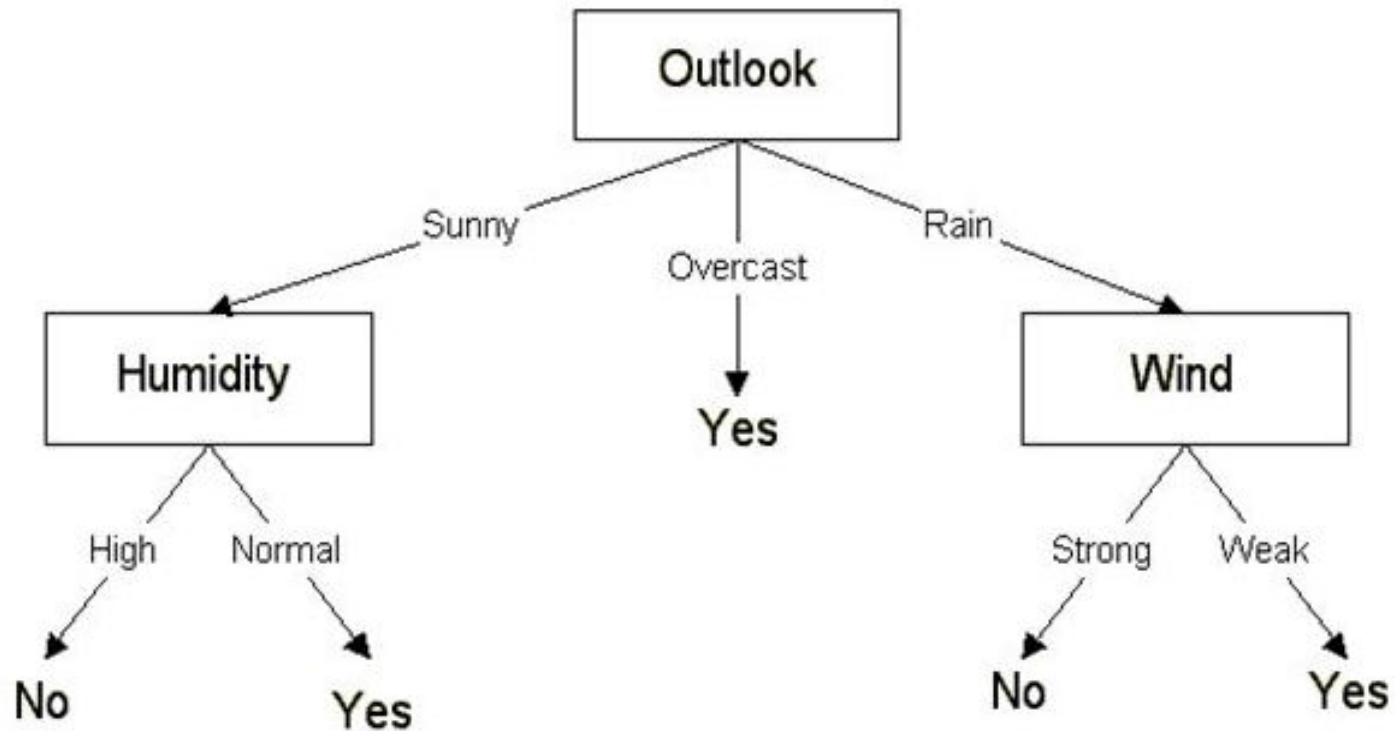
Outlook: Sunny

| Day | Temp | Humidity | Wind | Play Tennis |
|-----|------|----------|--------|-------------|
| D1 | Hot | High | Weak | No |
| D2 | Hot | High | Strong | No |
| D8 | Mild | High | Weak | No |
| D9 | Cool | Normal | Weak | Yes |
| D11 | Mild | Normal | Strong | Yes |

Decision Tree: Outlook, Humidity, ?



Contd..





Issue in DT

1. Avoiding Overfitting the Data
2. Incorporating Continuous-Valued Attributes
3. Alternative Measures for Selecting Attributes
4. Handling Training Examples with Missing Attribute Values
5. Handling Attributes with Differing Costs

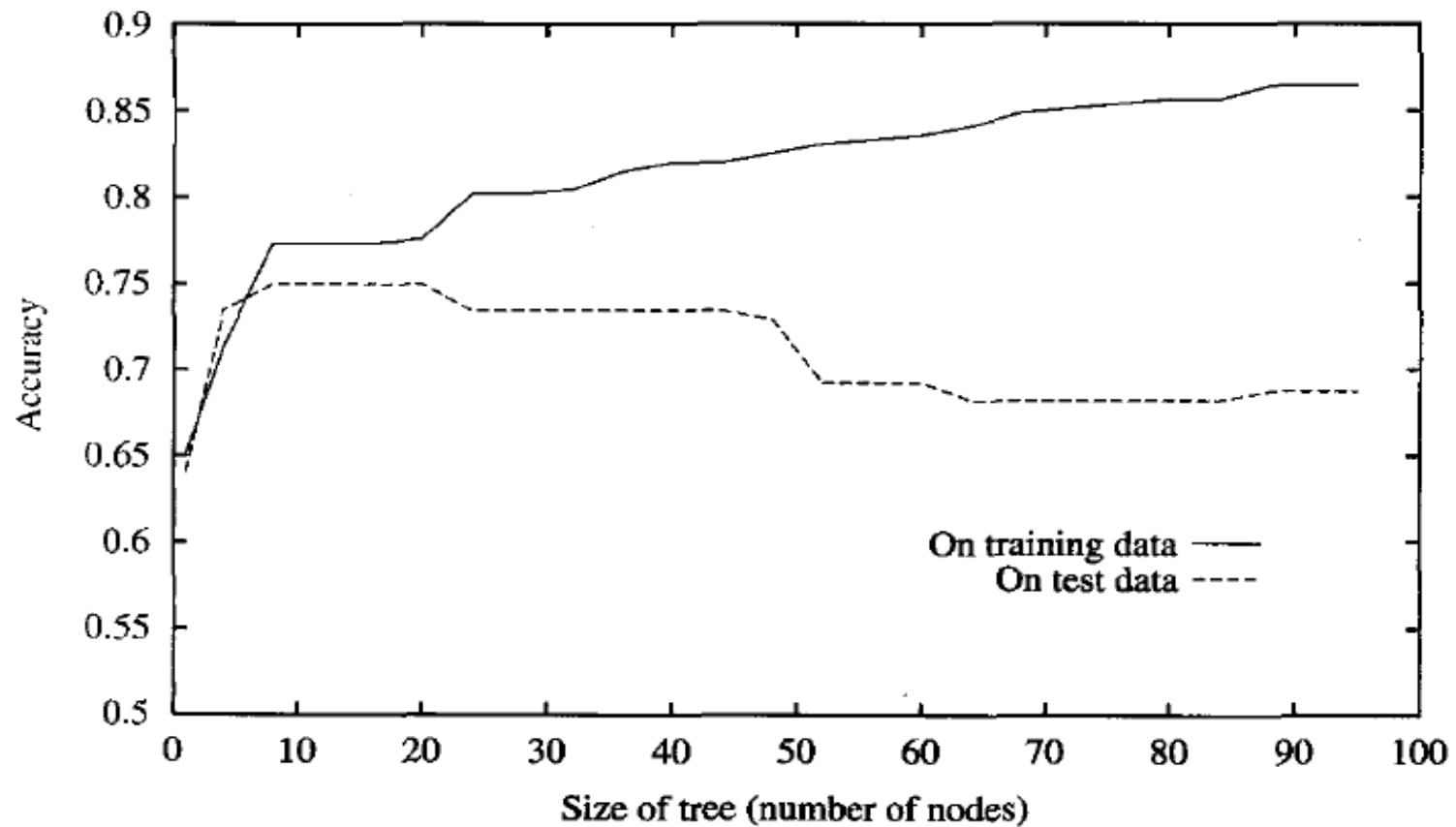
1: Avoiding Overfitting the Data

- We will say that a hypothesis overfits the training examples if some other hypothesis that fits the training examples less well actually performs better over the entire distribution of instances (i.e., including instances beyond the training set).

- **Definition:**

Given a hypothesis space H , a hypothesis $h \in H$ is said to overfit the training data if there exists some alternative hypothesis $h' \in H$, such that h has smaller error than h' over the training examples, but h' has a smaller error than h over the entire distribution of instances.

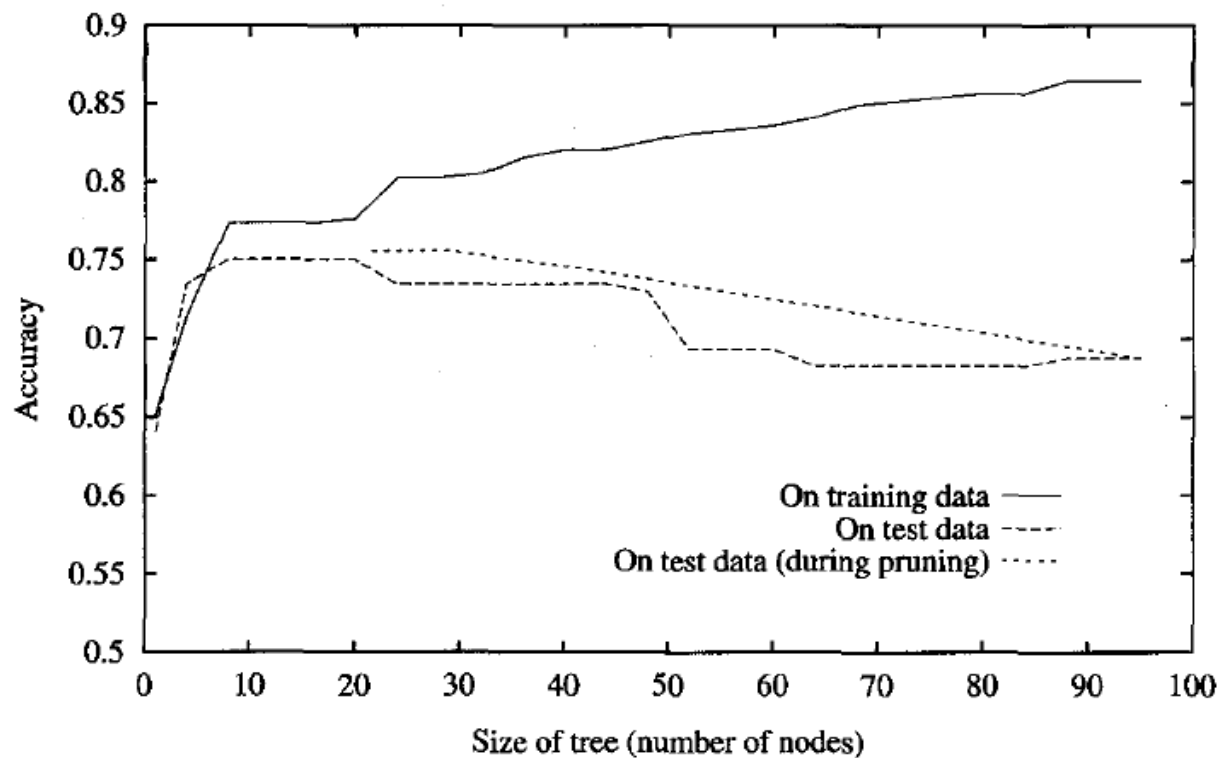
Contd.. (From Tom Mitchel Book)



Contd..

- There are several approaches to avoiding overfitting in decision tree learning. These can be grouped into two classes:
 1. approaches that stop growing the tree earlier, before it reaches the point where it perfectly classifies the training data,
 2. approaches that allow the tree to overfit the data, and then post-prune the tree.
 - a) The first of the above approaches is the most common and is often referred to as a training and validation set approach.

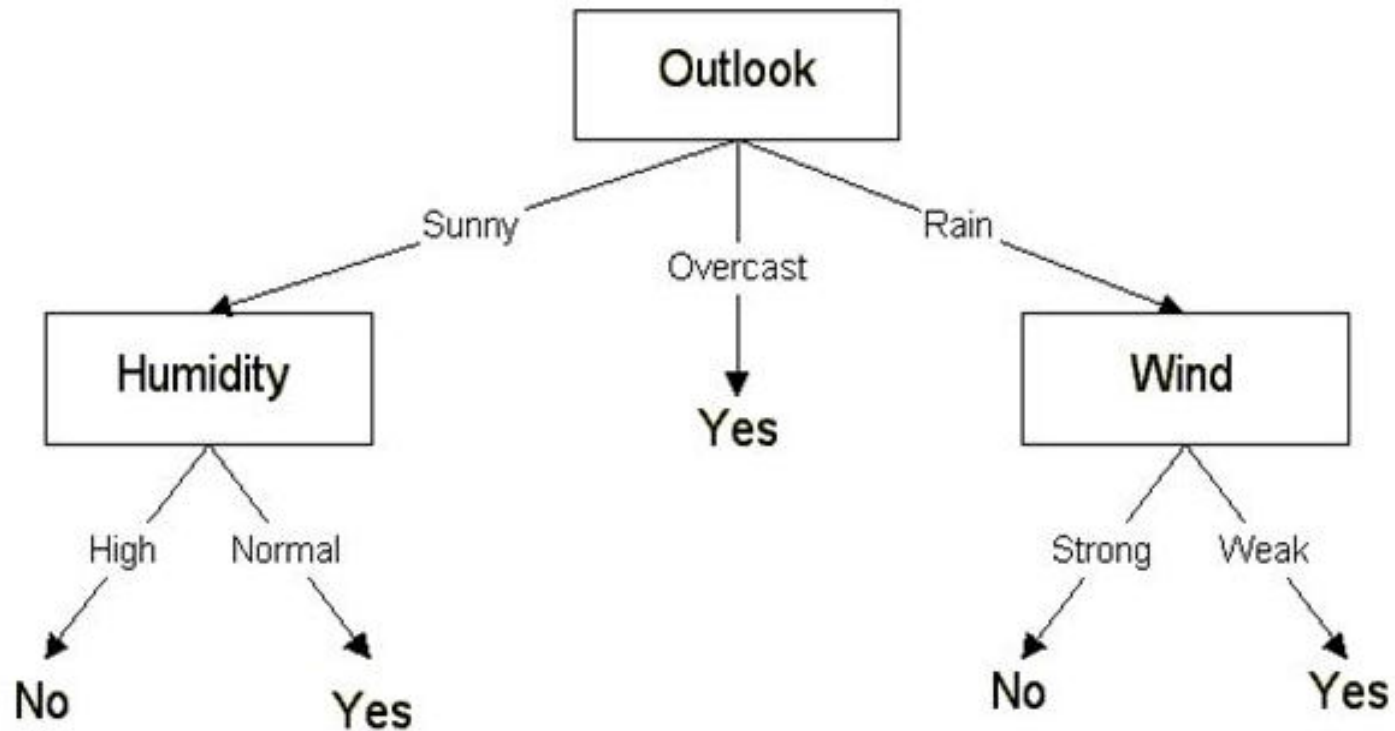
Effect of reduced-error pruning in decision tree learning



Reduced Error Pruning

- Consider each of the decision nodes in the tree to be candidates for pruning.
- Pruning a decision node consists of removing the subtree rooted at that node, making it a leaf node, and assigning it the most common classification of the training examples affiliated with that node.
- Nodes are removed only if the resulting pruned tree performs no worse than the original over the validation set.
- Nodes are pruned iteratively, always choosing the node whose removal most increases the decision tree accuracy over the validation set.
- Pruning of nodes continues until further pruning is harmful (i.e., decreases accuracy of the tree over the validation set).

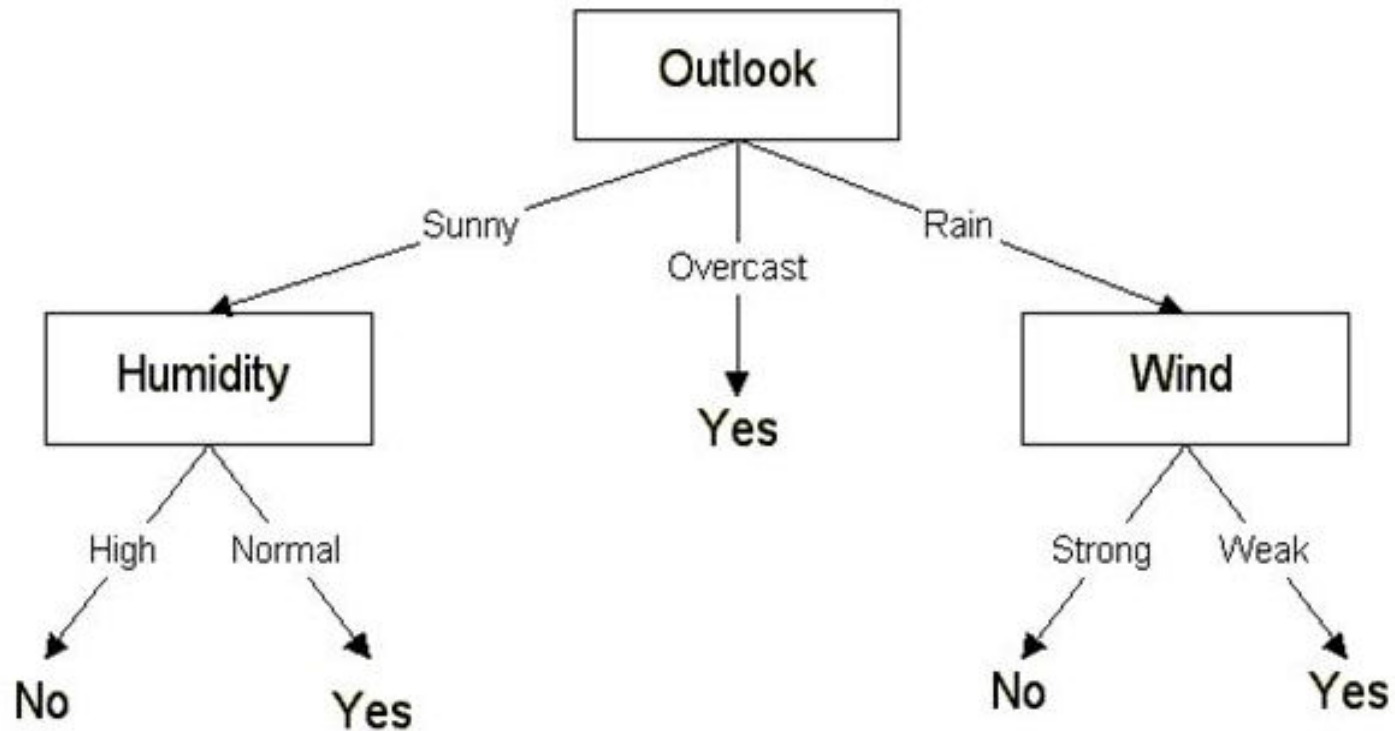
Contd..



Rules based pruning

- In practice, one quite successful method for finding high accuracy hypotheses is a technique we shall call rule post-pruning. A variant of this pruning method is used by C4.5 (Quinlan 1993), which is an outgrowth of the original ID3 algorithm.
- Rule post-pruning involves the following steps:
 1. Infer the decision tree from the training set, growing the tree until the training data is fit as well as possible and allowing overfitting to occur.
 2. Convert the learned tree into an equivalent set of rules by creating one rule for each path from the root node to a leaf node.
 3. Prune (generalize) each rule by removing any preconditions that result in improving its estimated accuracy.
 4. Sort the pruned rules by their estimated accuracy, and consider them in this sequence when classifying subsequent instances.

Contd..



Making Rules

- Each attribute test along the path from the root to the leaf becomes a rule antecedent (precondition) and the classification at the leaf node becomes the rule consequent (postcondition).
- **Rule 1:**
IF (Outlook = Sunny) AND (Humidity = High)
THEN Play Tennis = No



Advantages: Decision tree to rules before pruning

- Converting to rules improves readability. Rules are often easier for to understand.
- Converting to rules allows distinguishing among the different contexts in which a decision node is used.
- Because, each distinct path through the decision tree node produces a distinct rule, the pruning decision regarding that attribute test can be made differently for each path.
- Converting to rules removes the distinction between attribute tests that occur near the root of the tree and those that occur near the leaves



2. Incorporating Continuous-Valued Attributes

- First, the target attribute whose value is predicted by the learned tree must be discrete valued.
- Second, the attributes tested in the decision nodes of the tree must also be discrete valued.
- This second restriction can easily be removed so that continuous-valued decision attributes can be incorporated into the learned tree.
- This can be accomplished by dynamically defining new discrete valued attributes that partition the continuous attribute value into a discrete set of intervals.



Contd..

| | | | | | | |
|---------------------|----|----|-----|-----|-----|----|
| <i>Temperature:</i> | 40 | 48 | 60 | 72 | 80 | 90 |
| <i>PlayTennis:</i> | No | No | Yes | Yes | Yes | No |



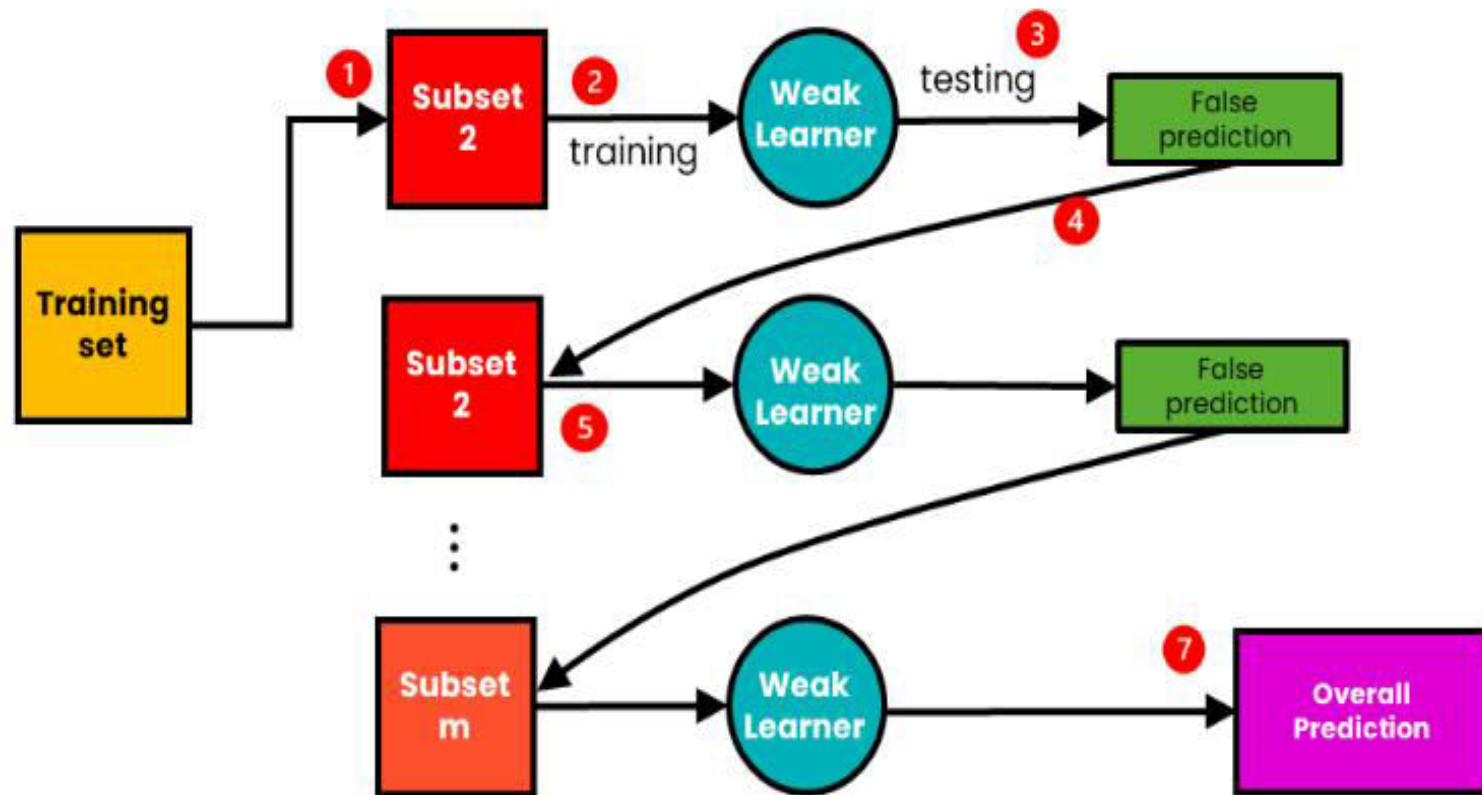
Lecture on

Boosting: Adaboost

Boosting

- Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers.
- It is done by building a model by using weak models in series.
- Firstly, a model is built from the training data.
- Then the second model is built which tries to correct the errors present in the first model.
- This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added.

Contd..





AdaBoost

- AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning.
- It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances.

Algorithm

1. Initialize the dataset and assign equal weight to each of the data point.
2. Provide this as input to the model and identify the wrongly classified data points.
3. Increase the weight of the wrongly classified data points and decrease the weights of correctly classified data points. And then normalize the weights of all data points.
4. if (got required results)
 Goto step 5
else
 Goto step 2
5. End

Step 1: Assigning Weights

First of all data points will be assigned some weights. Initially, all the weights will be equal.

$$w(x_i, y_i) = \frac{1}{N}, \quad i = 1, 2, \dots, n$$

Step 2: Classify the Samples

- a) We start by seeing how well “Chest Pain” classifies the samples and will see how the variables (Patient weight, Blocked Arteries) classify the samples.
- b) We’ll create a decision stump for each of the features and then calculate the Gini Index of each tree.

Contd..

Step 3: Calculate the Influence

We'll now calculate the “Amount of Say” or “Importance” or “Influence” for this classifier in classifying the data points using this formula:

$$\text{Amount of Say} = \frac{1}{2} \log\left(\frac{1 - \text{Total Error}}{\text{Total Error}}\right)$$

Step 4: Calculate TE and Performance

it is necessary to calculate the TE and performance of a stump. So, we need to update the weights because if the same weights are applied to the next model, then the output received will be the same as what was received in the first model.

$$\text{New Sample Weight} = \text{sample weight} \times e^{-\text{amount of say}}$$

Contd..

Step 5: Decrease Errors

Now, we need to make a new dataset to see if the errors decreased or not. For this, we will remove the “sample weights” and “new sample weights” columns and then, based on the “new sample weights,” divide our data points into buckets.

$$\text{New Sample Weight} = \text{sample weight} \times e^{-\text{amount of say}}$$

Step 6: New Dataset

Now, what the algorithm does is selects random numbers from 0-1. Since incorrectly classified records have higher sample weights, the probability of selecting those records is very high.

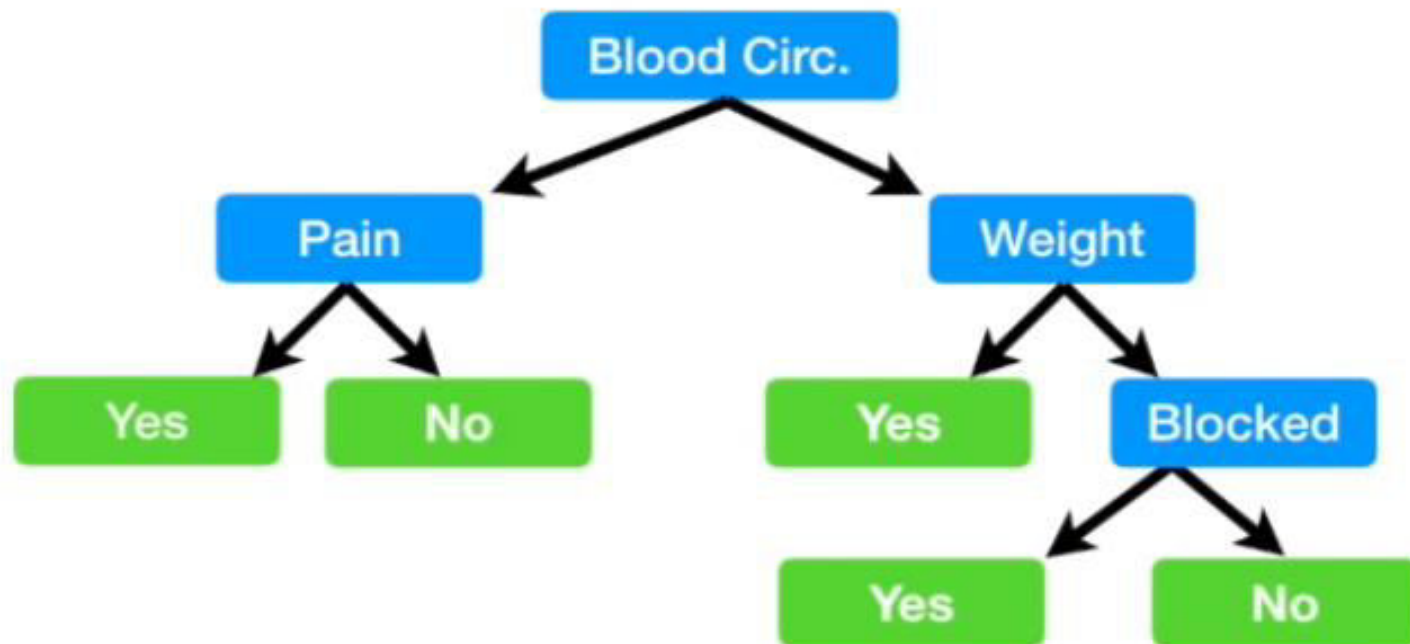
Step 7: Repeat Previous Steps

Now this act as our new dataset, and we need to repeat all the above steps i.e. 1 to 4.

Sample data set

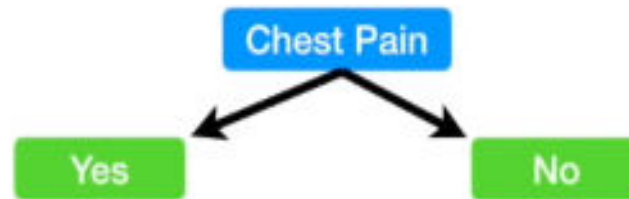
| Chest Pain | Blocked Arteries | Patient Weight | Heart Disease |
|------------|------------------|----------------|---------------|
| Yes | Yes | 205 | Yes |
| No | Yes | 180 | Yes |
| Yes | No | 210 | Yes |
| Yes | Yes | 167 | Yes |
| No | Yes | 156 | No |
| No | Yes | 125 | No |
| Yes | No | 168 | No |
| Yes | Yes | 172 | No |

DT:

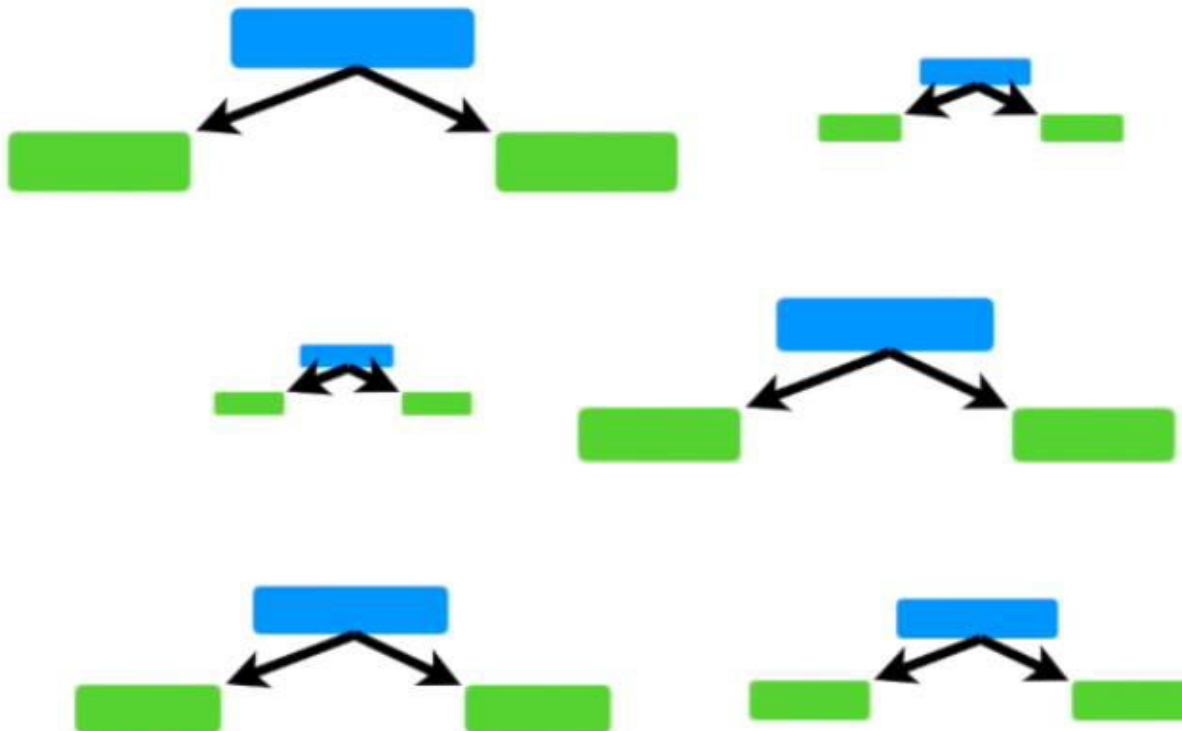


Stumps

- Stumps are weak learners



Contd..



AdaBoost: Basic Idea

- AdaBoost combines a lot of weak learners to make classifications. The weak learners are almost always stumps.
- Some stumps get more say in the classification than others.
- Each stump is made by taking the previous stumps mistakes into accounts.

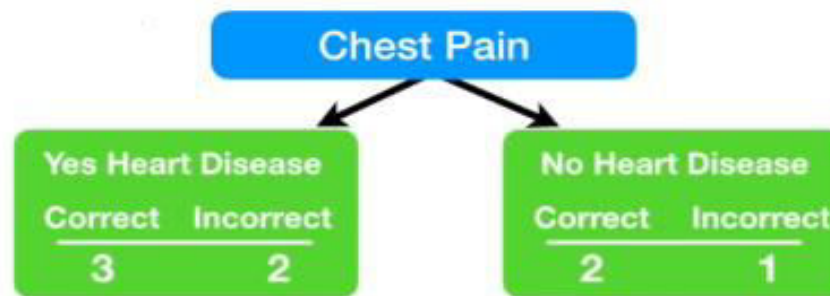
Start AdaBoost

- Add a weight to the sample

| Chest Pain | Blocked Arteries | Patient Weight | Heart Disease | Sample Weight |
|------------|------------------|----------------|---------------|---------------|
| Yes | Yes | 205 | Yes | 1/8 |
| No | Yes | 180 | Yes | 1/8 |
| Yes | No | 210 | Yes | 1/8 |
| Yes | Yes | 167 | Yes | 1/8 |
| No | Yes | 156 | No | 1/8 |
| No | Yes | 125 | No | 1/8 |
| Yes | No | 168 | No | 1/8 |
| Yes | Yes | 172 | No | 1/8 |

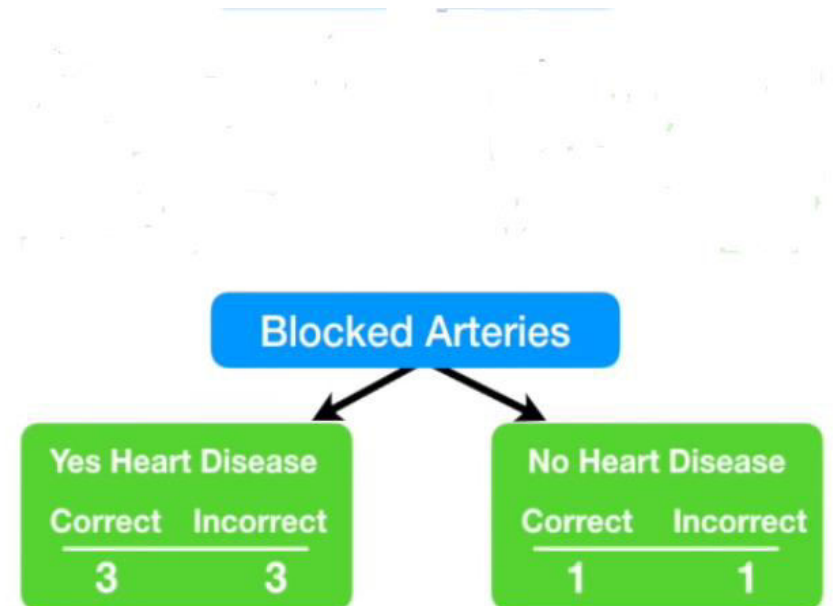
Making stumps

■ Chest Pain



Contd..

| Chest Pain | Blocked Arteries | Patient Weight | Heart Disease | Sample Weight |
|------------|------------------|----------------|---------------|---------------|
| Yes | Yes | 205 | Yes | 1/8 |
| No | Yes | 180 | Yes | 1/8 |
| Yes | No | 210 | Yes | 1/8 |
| Yes | Yes | 167 | Yes | 1/8 |
| No | Yes | 156 | No | 1/8 |
| No | Yes | 125 | No | 1/8 |
| Yes | No | 168 | No | 1/8 |
| Yes | Yes | 172 | No | 1/8 |



Now we do the same thing
for **Blocked Arteries**...

Contd..



Calculate Gini-Index

- Chest pain= ?
- Blocked Arteries =?
- Patient weight= ?

$$Gini = 1 - \sum_j p_j^2$$

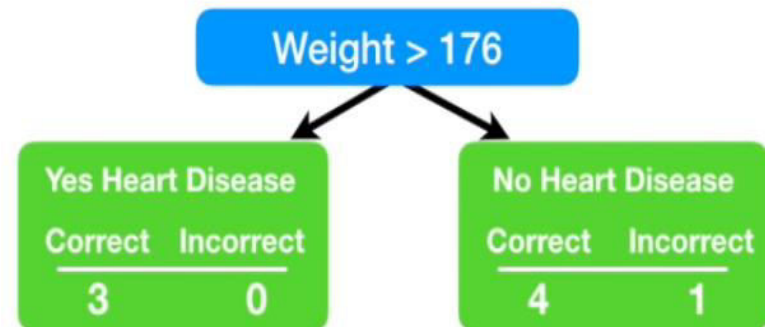
Contd..

- Chest pain= 0.47
- Blocked Arteries = 0.5
- Patient weight= 0.2

Contd..

| Chest Pain | Blocked Arteries | Patient Weight | Heart Disease | Sample Weight |
|------------|------------------|----------------|---------------|---------------|
| Yes | Yes | 205 | Yes | 1/8 |
| No | Yes | 180 | Yes | 1/8 |
| Yes | No | 210 | Yes | 1/8 |
| Yes | Yes | 167 | Yes | 1/8 |
| No | Yes | 156 | No | 1/8 |
| No | Yes | 125 | No | 1/8 |
| Yes | No | 168 | No | 1/8 |
| Yes | Yes | 172 | No | 1/8 |

← This patient, who weighs less than 176, has heart disease, but the stump says they do not.



Contd..

- **Find total error:**

Total error for stumps can be expressed as sum of the weight associated with the incorrectly classified sample.

For Patient weight total error is $1/8$

Amount of say ?

Patient weight = 0.97

$$\text{Amount of Say} = \frac{1}{2} \log\left(\frac{1 - \text{Total Error}}{\text{Total Error}}\right)$$

New Weight ?

New Sample
Weight = sample weight $\times e^{\text{amount of say}}$

Increase weight

New Sample
Weight = sample weight $\times e^{-\text{amount of say}}$

Contd..

| Chest Pain | Blocked Arteries | Patient Weight | Heart Disease | Sample Weight | New Weight |
|------------|------------------|----------------|---------------|---------------|------------|
| Yes | Yes | 205 | Yes | 1/8 | 0.05 |
| No | Yes | 180 | Yes | 1/8 | 0.05 |
| Yes | No | 210 | Yes | 1/8 | 0.05 |
| Yes | Yes | 167 | Yes | 1/8 | 0.33 |
| No | Yes | 156 | No | 1/8 | 0.05 |
| No | Yes | 125 | No | 1/8 | 0.05 |
| Yes | No | 168 | No | 1/8 | 0.05 |
| Yes | Yes | 172 | No | 1/8 | 0.05 |

Contd..

| Chest Pain | Blocked Arteries | Patient Weight | Heart Disease | Sample Weight |
|------------|------------------|----------------|---------------|---------------|
| Yes | Yes | 205 | Yes | 0.07 |
| No | Yes | 180 | Yes | 0.07 |
| Yes | No | 210 | Yes | 0.07 |
| Yes | Yes | 167 | Yes | 0.49 |
| No | Yes | 156 | No | 0.07 |
| No | Yes | 125 | No | 0.07 |
| Yes | No | 168 | No | 0.07 |
| Yes | Yes | 172 | No | 0.07 |



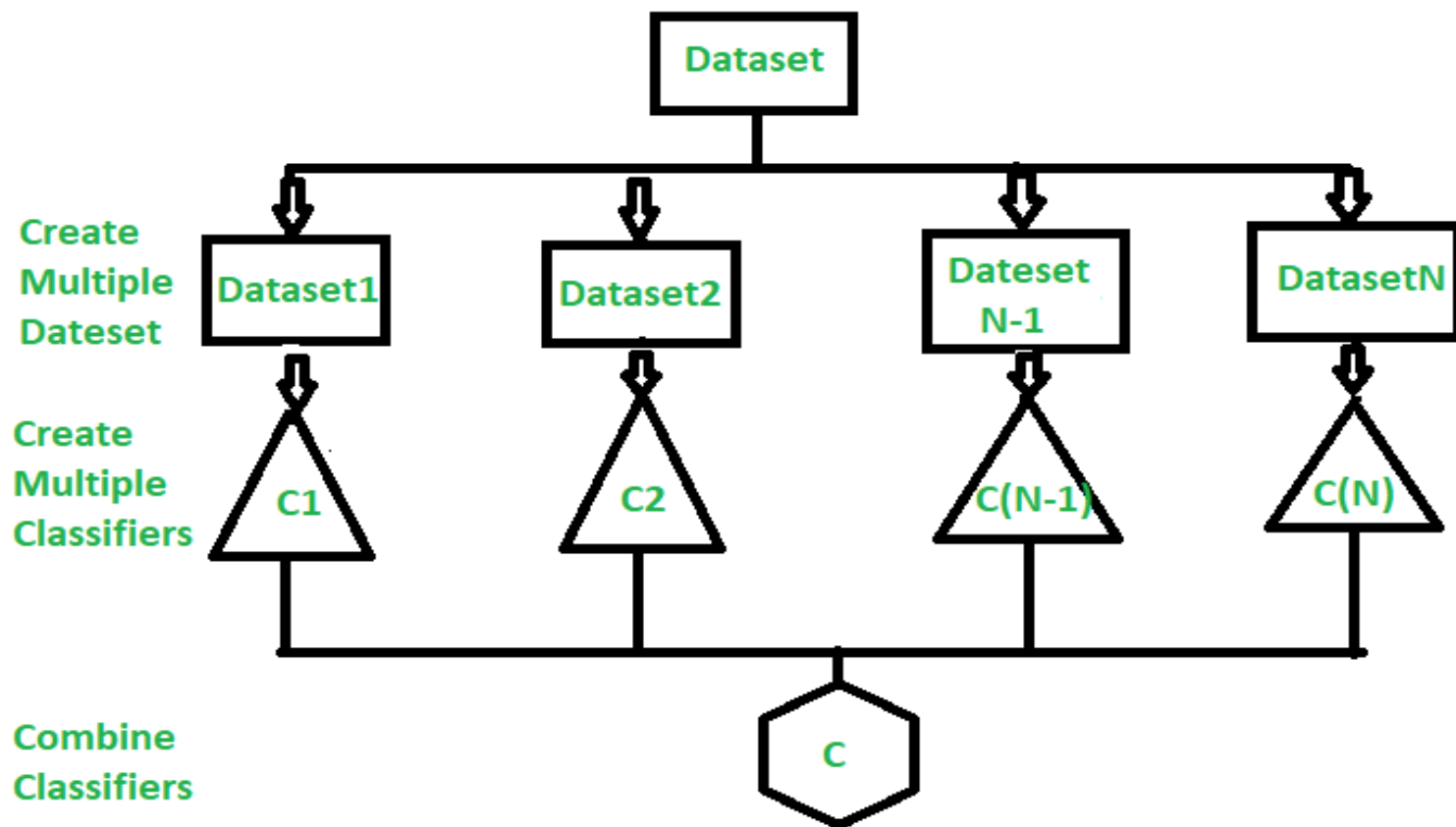
Lecture on

Ensemble Learning

Introduction

- Ensemble learning is a general meta approach to machine learning that seeks better predictive performance by combining the predictions from multiple models.
- This approach allows the production of better predictive performance compared to a single model.
- Basic idea is to learn a set of classifiers (experts) and to allow them to vote.
- Advantage : Improvement in predictive accuracy.
- Disadvantage : It is difficult to understand an ensemble of classifiers.

Contd..



Contd..

- Dietterich (2002) showed that ensembles overcome three problems –
- **Statistical Problem –**
 - The Statistical Problem arises when the hypothesis space is too large for the amount of available data.
 - Hence, there are many hypotheses with the same accuracy on the data and the learning algorithm chooses only one of them!
 - There is a risk that the accuracy of the chosen hypothesis is low on unseen data!
- **Computational Problem –**
 - The Computational Problem arises when the learning algorithm cannot guarantee finding the best hypothesis.
- **Representational Problem –**
 - The Representational Problem arises when the hypothesis space does not contain any good approximation of the target class(es).

Base Learner

- To generate base learner for ensemble learning, there are different methods-
 1. Different algorithms: Decision tree, SVM, Linear regression
 2. Different parameter (hyper-parameter)
 3. Different representations
 4. Different data sample



Bias & Variance

- Low bias and low variance can be achieved using Ensemble learning.
- If we start with a high bias but when we combine it, it may have low bias.
- Because the individual hypothesis may have a new hypothesis after combining these which have low bias.

Weak Learner

- A learner is called weak if it has error less than 0.5.
- Group of weak learner can be combined together to form a strong learner (ensemble model) that achieves better performance than individual weak learners.
- Weak learners are the model which does not perform well by themselves either because of high bias or high variance.
- Underfitting model: High bias and low variance
- Overfitting model: Low bias and high variance

Types of ensemble model

- Based on the choice of weak learners, the ensemble learning model can be classified into two types-

1. Homogeneous model
2. Heterogeneous model

- **Homogeneous model**

A single base learning algorithms is used to make a strong learner.

Eg. Decision tree

- **Heterogeneous model**

Different base learning algorithms are used to make a strong learner.

Eg: Decision tree, SVM, KNN etc.

Types of Ensemble Classifier

- **Bagging**

It considers homogeneous weak learners and focuses on reducing variance.

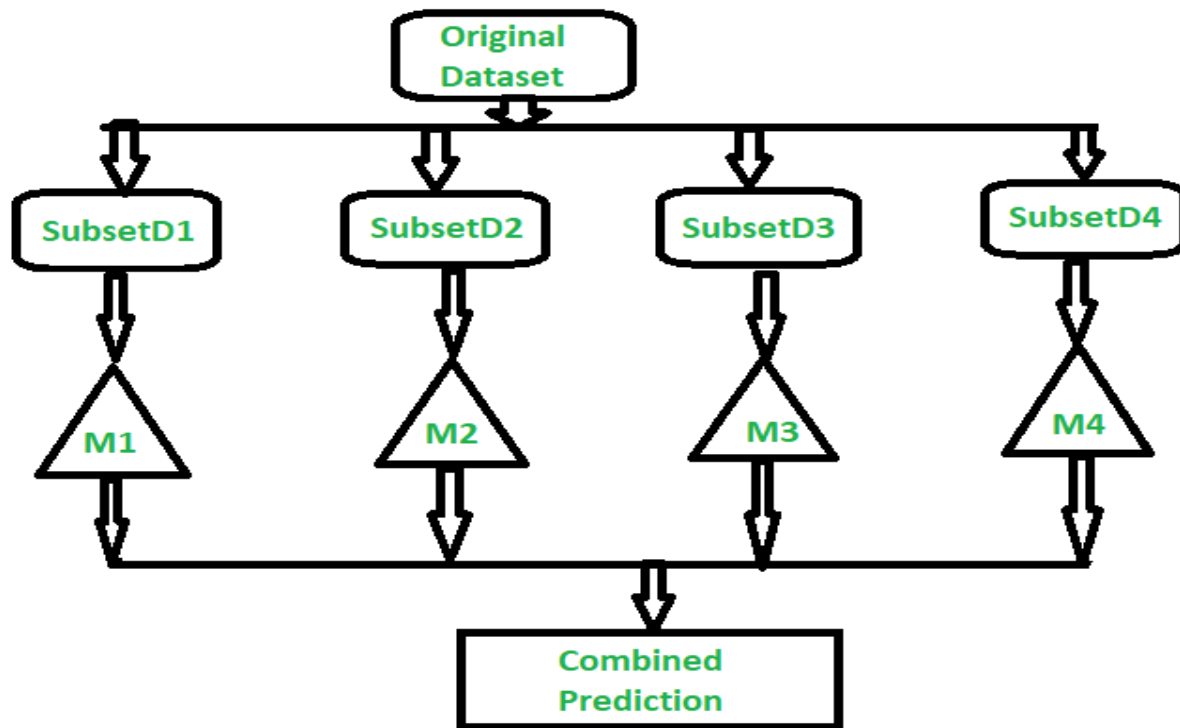
- **Boosting**

It considers homogenous weak learner and focuses on reducing bias.

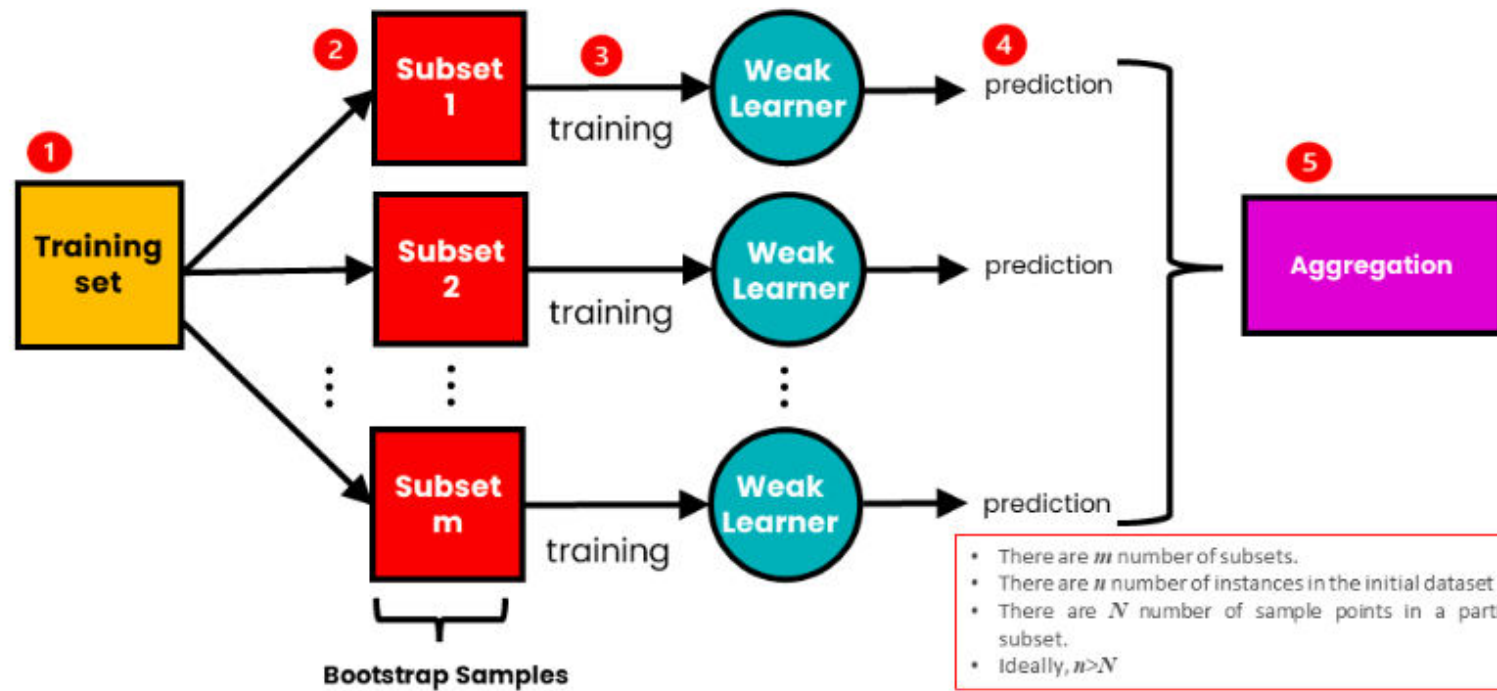
- **Stacking**

It considers heterogeneous weak learners.

Bagging



The Process of Bagging (Bootstrap Aggregation)





Bagging: Steps

1. Multiple subsets are created from the original data set with equal tuples, selecting observations with replacement.
2. A base model is created on each of these subsets.
3. Each model is learned in parallel from each training set and independent of each other.
4. The final predictions are determined by combining the predictions from all the models

Contd..

■ **Random Forest:**

- Random Forest is an extension over bagging.
- Each classifier in the ensemble is a decision tree classifier and is generated using a random selection of attributes at each node to determine the split.
- During classification, each tree votes and the most popular class is returned.



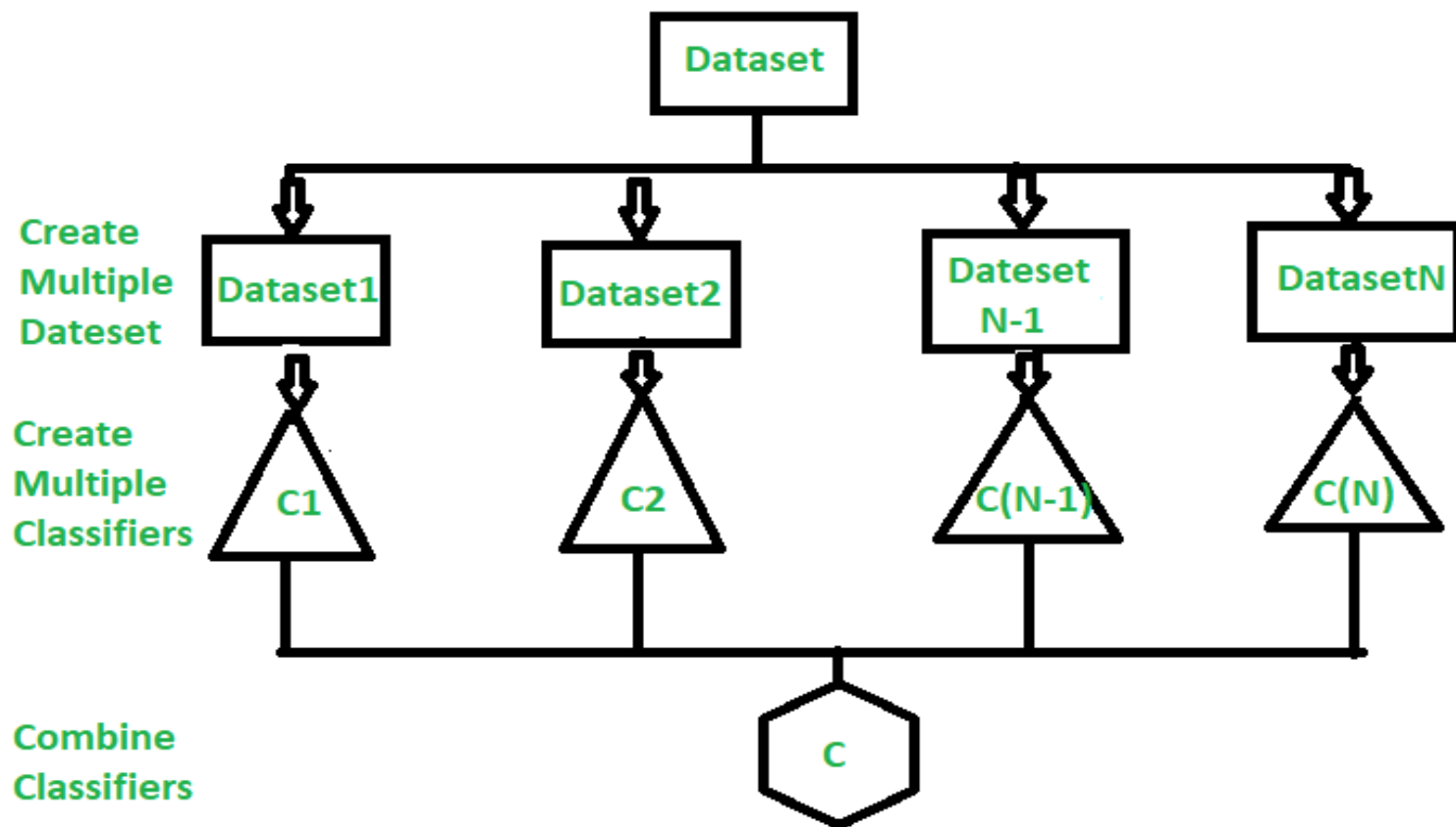
Lecture on

Ensemble Learning

Introduction

- Ensemble learning is a general meta approach to machine learning that seeks better predictive performance by combining the predictions from multiple models.
- This approach allows the production of better predictive performance compared to a single model.
- Basic idea is to learn a set of classifiers (experts) and to allow them to vote.
- Advantage : Improvement in predictive accuracy.
- Disadvantage : It is difficult to understand an ensemble of classifiers.

Contd..



Contd..

- Dietterich (2002) showed that ensembles overcome three problems –
- **Statistical Problem –**
 - The Statistical Problem arises when the hypothesis space is too large for the amount of available data.
 - Hence, there are many hypotheses with the same accuracy on the data and the learning algorithm chooses only one of them!
 - There is a risk that the accuracy of the chosen hypothesis is low on unseen data!
- **Computational Problem –**
 - The Computational Problem arises when the learning algorithm cannot guarantee finding the best hypothesis.
- **Representational Problem –**
 - The Representational Problem arises when the hypothesis space does not contain any good approximation of the target class(es).

Base Learner

- To generate base learner for ensemble learning, there are different methods-
 1. Different algorithms: Decision tree, SVM, Linear regression
 2. Different parameter (hyper-parameter)
 3. Different representations
 4. Different data sample



Bias & Variance

- Low bias and low variance can be achieved using Ensemble learning.
- If we start with a high bias but when we combine it, it may have low bias.
- Because the individual hypothesis may have a new hypothesis after combining these which have low bias.

Weak Learner

- A learner is called weak if it has error less than 0.5.
- Group of weak learner can be combined together to form a strong learner (ensemble model) that achieves better performance than individual weak learners.
- Weak learners are the model which does not perform well by themselves either because of high bias or high variance.
- Underfitting model: High bias and low variance
- Overfitting model: Low bias and high variance

Types of ensemble model

- Based on the choice of weak learners, the ensemble learning model can be classified into two types-

1. Homogeneous model
2. Heterogeneous model

- **Homogeneous model**

A single base learning algorithms is used to make a strong learner.

Eg. Decision tree

- **Heterogeneous model**

Different base learning algorithms are used to make a strong learner.

Eg: Decision tree, SVM, KNN etc.

Types of Ensemble Classifier

- **Bagging**

It considers homogeneous weak learners and focuses on reducing variance.

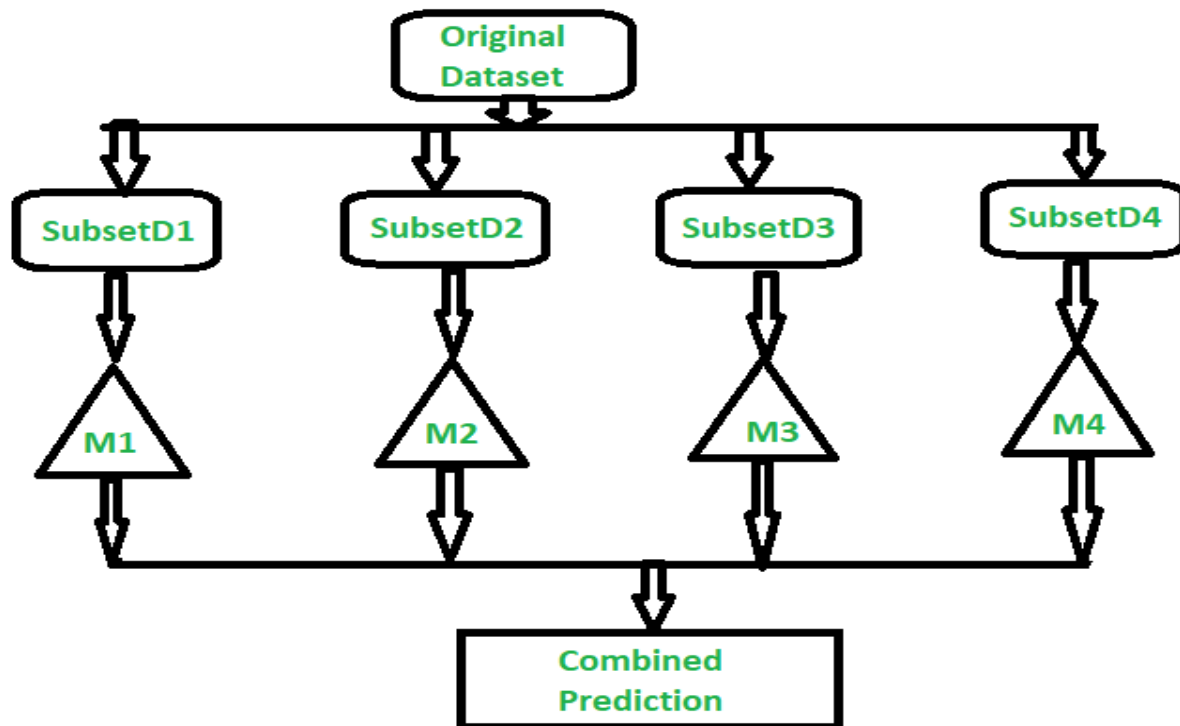
- **Boosting**

It considers homogenous weak learner and focuses on reducing bias.

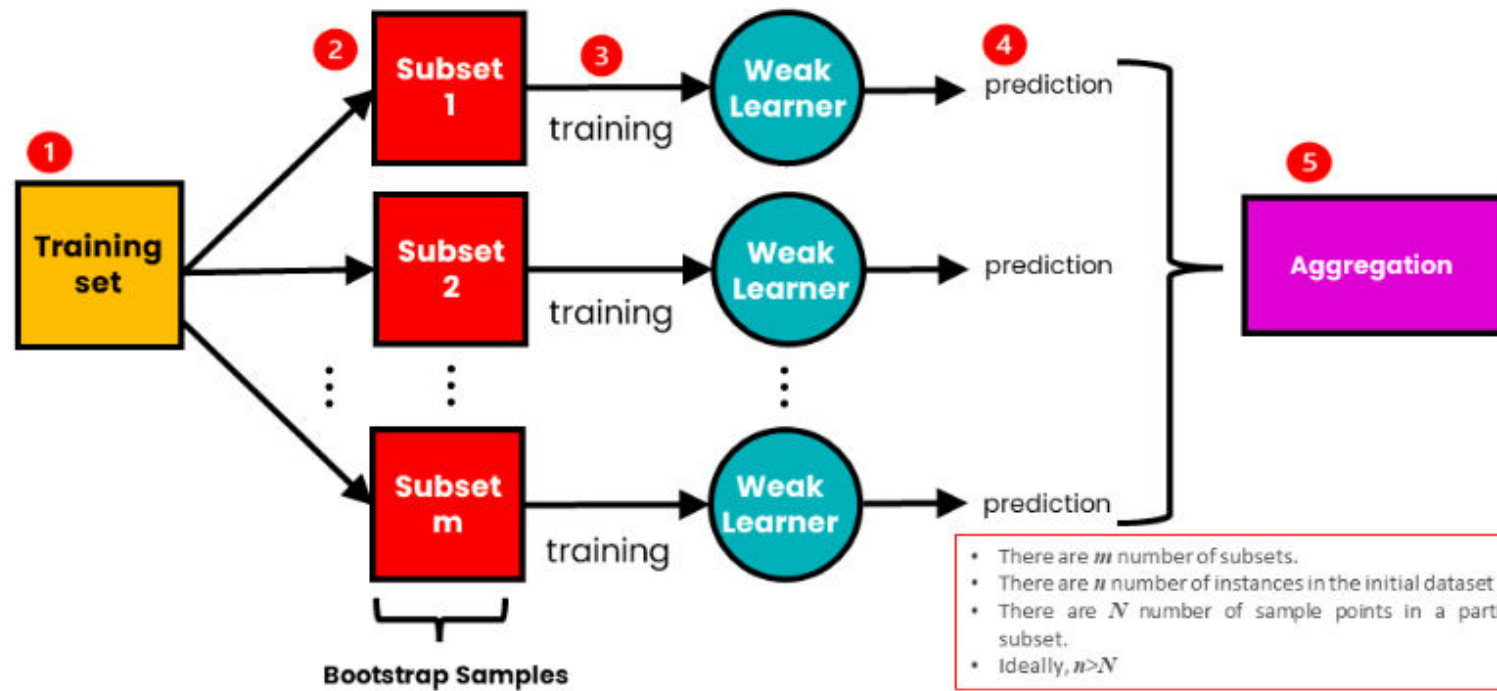
- **Stacking**

It considers heterogeneous weak learners.

Bagging



The Process of Bagging (Bootstrap Aggregation)





Bagging: Steps

1. Multiple subsets are created from the original data set with equal tuples, selecting observations with replacement.
2. A base model is created on each of these subsets.
3. Each model is learned in parallel from each training set and independent of each other.
4. The final predictions are determined by combining the predictions from all the models

Contd..

■ **Random Forest:**

- Random Forest is an extension over bagging.
- Each classifier in the ensemble is a decision tree classifier and is generated using a random selection of attributes at each node to determine the split.
- During classification, each tree votes and the most popular class is returned.

Lecture on

Parameter Estimation: Relative *Entropy Estimation*

Introduction: Relative Entropy

- It is a measure of the dissimilarity between two random quantities.
- The relative entropy is also known as the Kullback–Leibler distance and can be interpreted as measuring the inefficiency of assuming that the distribution is $q(x)$ when the true distribution is $p(x)$.
- This is a type of statistical distance: a measure of how one probability distribution P is different from a second.
- From an information theory point of view, Kullback-Leibler (KL) divergence is a relative entropy measure that estimates how close we have modeled an approximate distribution $q(x)$ with respect to the unknown distribution $p(x)$.

Definition:

- One way to measure the dissimilarity of two probability distributions, p and q , is known as the KL divergence or relative entropy. This is defined as follows:

$$\mathbb{KL}(p||q) \triangleq \sum_{k=1}^K p_k \log \frac{p_k}{q_k}$$





Contd..

$$\text{KL} (p||q) = \sum_k p_k \log p_k - \sum_k p_k \log q_k = -\mathbb{H} (p) + \mathbb{H} (p, q)$$

where $\mathbb{H} (p, q)$ is called the **cross entropy**,

$$\mathbb{H} (p, q) \triangleq - \sum_k p_k \log q_k$$

Relative entropy with Boolean RV

- Let, $X = \{0, 1\}$ be a random variable. Consider two distributions p, q on X . Assume, $p(0) = 1-r, p(1) = r$; $q(0) = 1-s, q(1) = s$;
- Consider a above Boolean random variable with two probability distribution p and q , then find relative entropy?



Contd..



Contd..

$$D(p||q) = (1 - r) \log \frac{1-r}{1-s} + r \log \frac{r}{s}$$

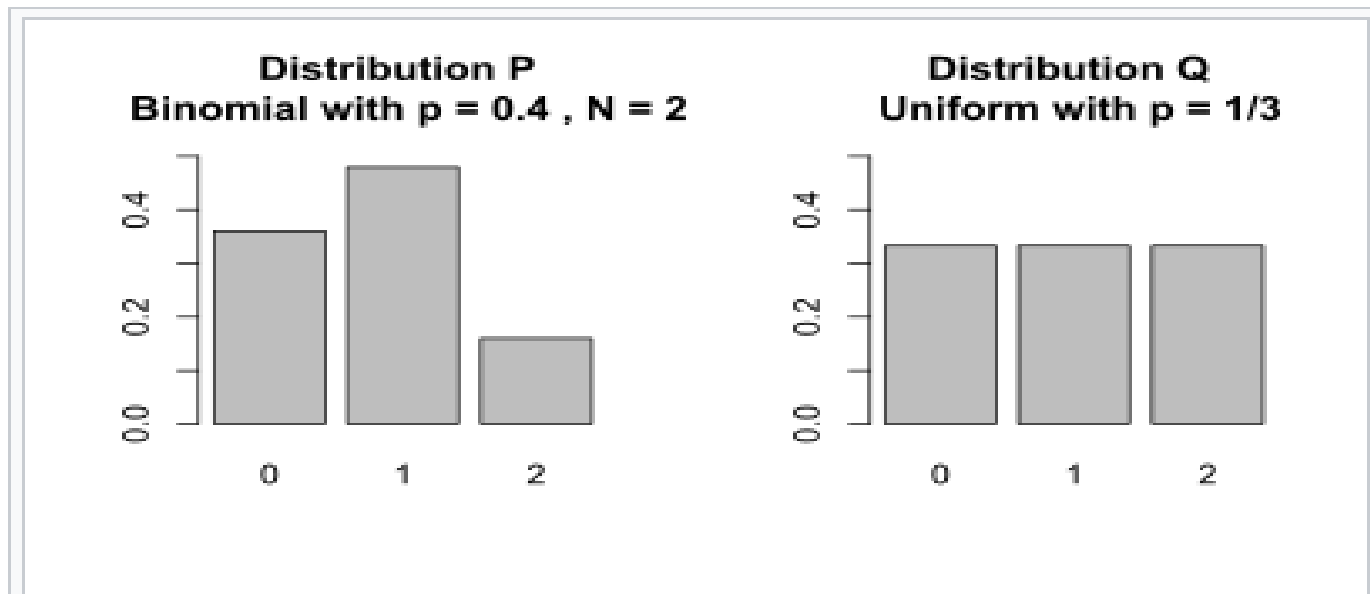
$$D(q||p) = (1 - s) \log \frac{1-s}{1-r} + s \log \frac{s}{r}$$

Observation:

- If, $r=s$, then $D(p||q) = D(q||p) = 0$,
- else for $r \neq s$, $D(p||q) \neq D(q||p)$

Contd..

- Let P and Q be the distributions shown in the table and figure. P is the distribution on the left side of the figure, a binomial distribution with $N=2$ and $p = 0.4$
- Q is the distribution on the right side of the figure, a discrete uniform distribution with the three possible outcomes $X = \{0, 1, 2\}$, with $p = 1/3$





Solution:

| x | 0 | 1 | 2 |
|----------------------------|----------------|-----------------|----------------|
| Distribution $P(x)$ | $\frac{9}{25}$ | $\frac{12}{25}$ | $\frac{4}{25}$ |
| Distribution $Q(x)$ | $\frac{1}{3}$ | $\frac{1}{3}$ | $\frac{1}{3}$ |

$$\begin{aligned}
 D_{\text{KL}}(P \parallel Q) &= \sum_{x \in \mathcal{X}} P(x) \ln \left(\frac{P(x)}{Q(x)} \right) \\
 &= \frac{9}{25} \ln \left(\frac{9/25}{1/3} \right) + \frac{12}{25} \ln \left(\frac{12/25}{1/3} \right) + \frac{4}{25} \ln \left(\frac{4/25}{1/3} \right) \\
 &= \frac{1}{25} (32 \ln(2) + 55 \ln(3) - 50 \ln(5)) \approx 0.0852996,
 \end{aligned}$$

$$\begin{aligned}
 D_{\text{KL}}(Q \parallel P) &= \sum_{x \in \mathcal{X}} Q(x) \ln \left(\frac{Q(x)}{P(x)} \right) \\
 &= \frac{1}{3} \ln \left(\frac{1/3}{9/25} \right) + \frac{1}{3} \ln \left(\frac{1/3}{12/25} \right) + \frac{1}{3} \ln \left(\frac{1/3}{4/25} \right) \\
 &= \frac{1}{3} (-4 \ln(2) - 6 \ln(3) + 6 \ln(5)) \approx 0.097455.
 \end{aligned}$$



Lecture on

Ensemble Learning

Introduction

- Ensemble learning is a general meta approach to machine learning that seeks better predictive performance by combining the predictions from multiple models.
- This approach allows the production of better predictive performance compared to a single model.
- Basic idea is to learn a set of classifiers (experts) and to allow them to vote.
- Advantage : Improvement in predictive accuracy.
- Disadvantage : It is difficult to understand an ensemble of classifiers.

Bagging

- It is a homogeneous weak learners' model that learns from each other independently in parallel and combines them for determining the model average.
- Bootstrap Aggregating, also known as bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy.
- This algorithms are used in statistical classification and regression.
- It decreases the variance and helps to avoid overfitting. It is usually applied to decision tree methods.
- Bagging is a special case of the model averaging approach.

Sampling

- Sampling denotes the selection of a part of the aggregate statistical material with a view to obtain information about the whole.
 1. Simple random sampling
 2. Purposive Sampling
 3. Stratified Sampling
 4. Systematic Sampling



Simple Random Sampling

- Simple Random Sampling with Replacement
- Simple Random Sampling without Replacement

Bagging Techniques

- Suppose a set D of d tuples, at each iteration i , a training set D_i of d tuples is selected via row sampling with a replacement method (i.e., there can be repetitive elements from different d tuples) from D (i.e., bootstrap).
- Then a classifier model M_i is learned for each training set $D < i$. Each classifier M_i returns its class prediction.



Steps of Bagging

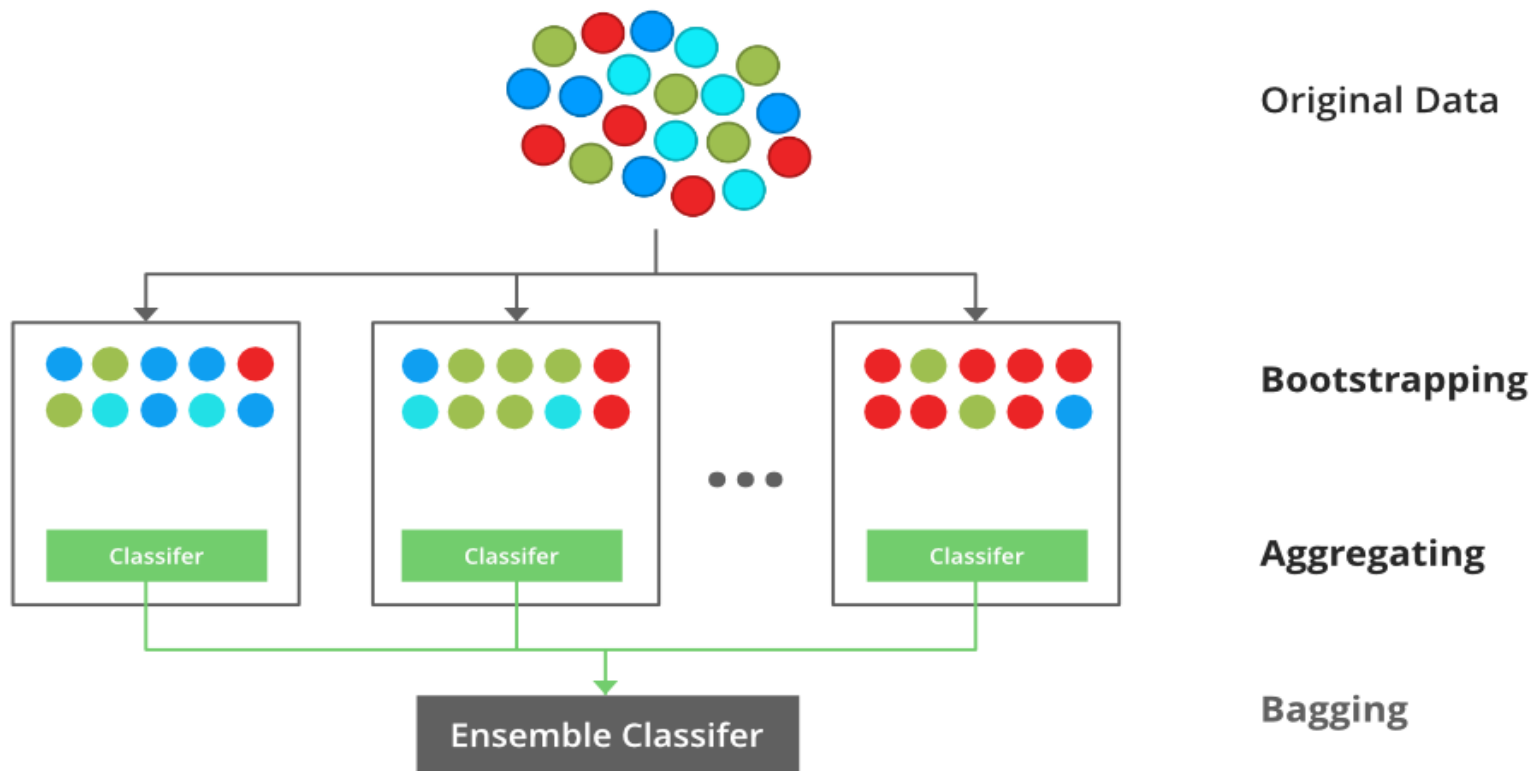
Step 1: Multiple subsets are created from the original data set with equal tuples, selecting observations with replacement.

Step 2: A base model is created on each of these subsets.

Step 3: Each model is learned in parallel with each training set and independent of each other.

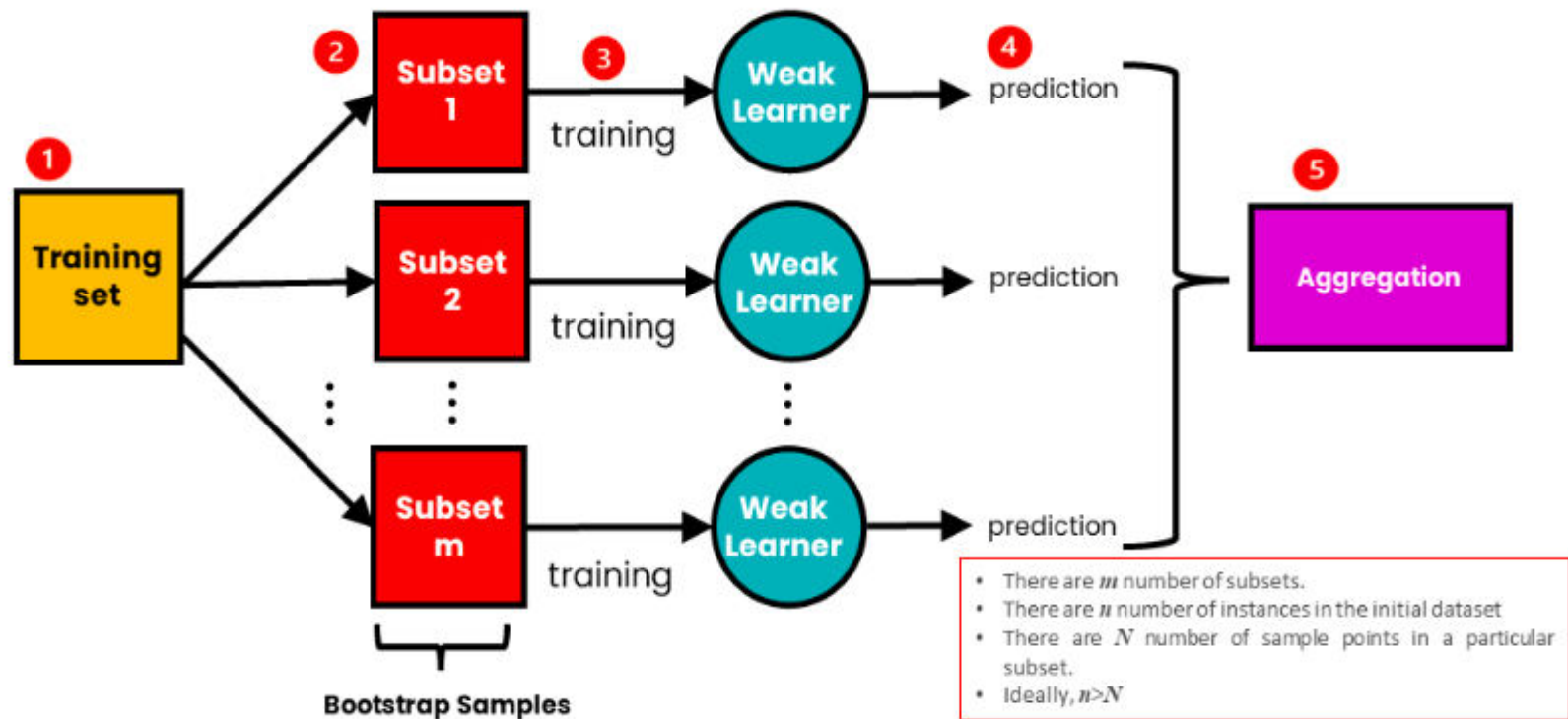
Step 4: The final predictions are determined by combining the predictions from all the models.

Contd..



Contd..

The Process of Bagging (Bootstrap Aggregation)



Contd..

■ **Random Forest:**

- Random Forest is an extension over bagging.
- Each classifier in the ensemble is a decision tree classifier and is generated using a random selection of attributes at each node to determine the split.
- During classification, each tree votes and the most popular class is returned.




Boosting

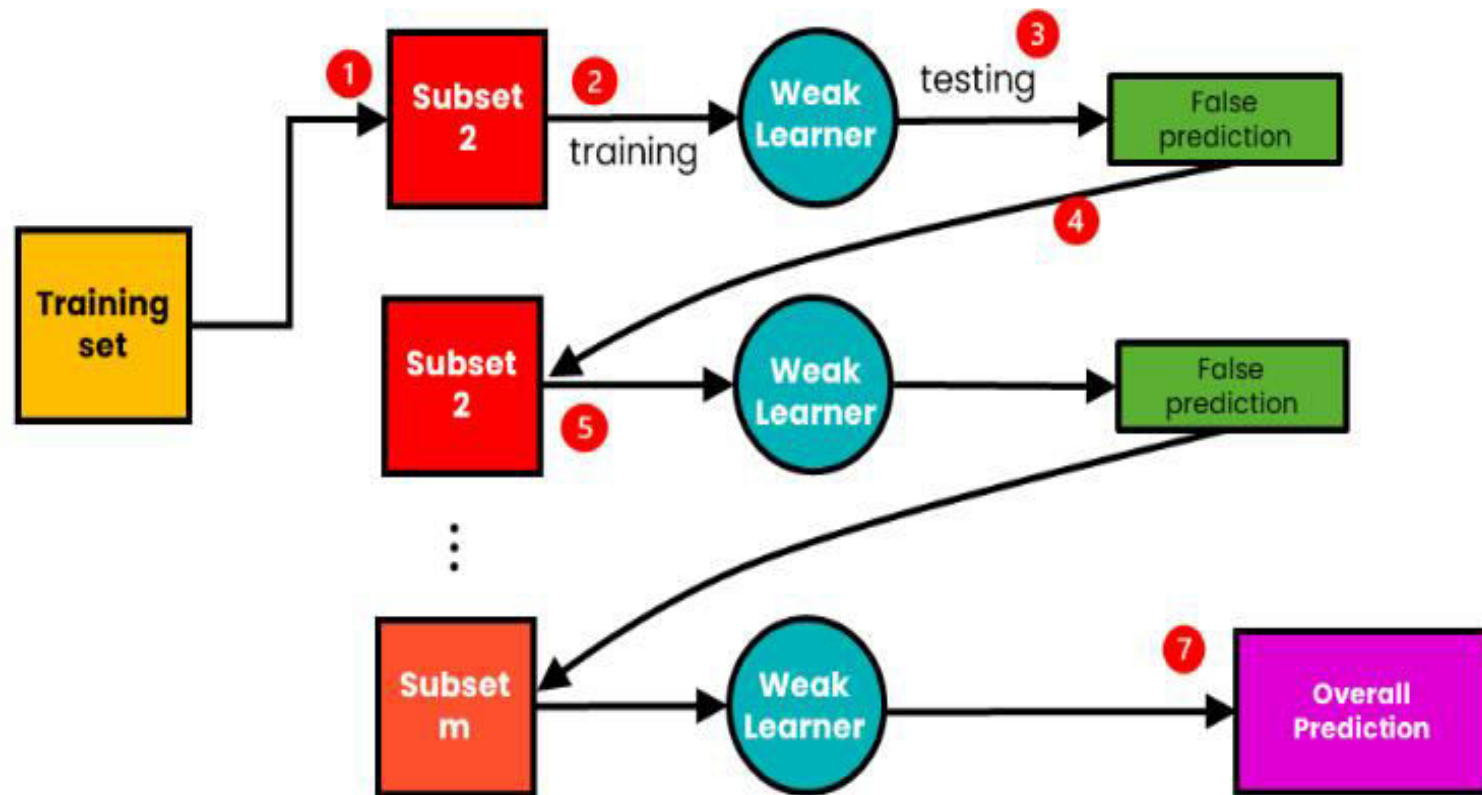
- Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers.
- It is done by building a model by using weak models in series.

Contd..

- Firstly, a model is built from the training data.
- Then the second model is built which tries to correct the errors present in the first model.
- This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models is added.

- 
1. Initialize the dataset and assign equal weight to each of the data point.
 2. Provide this as input to the model and identify the wrongly classified data points.
 3. Increase the weight of the wrongly classified data points and decrease the weights of correctly classified data points. And then normalize the weights of all data points.
 4. if (got required results)
 Goto step 5
else
 Goto step 2
 5. End

Contd..



Lecture on

Parameter Estimation: Mutual Information *Estimation*

Introduction: Relative Entropy

- It is a measure of the dissimilarity between two random quantities.
- The relative entropy is also known as the Kullback–Leibler distance and can be interpreted as measuring the inefficiency of assuming that the distribution is $q(x)$ when the true distribution is $p(x)$.
- This is a type of statistical distance: a measure of how one probability distribution P is different from a second.
- From an information theory point of view, Kullback-Leibler (KL) divergence is a relative entropy measure that estimates how close we have modeled an approximate distribution $q(x)$ with respect to the unknown distribution $p(x)$.

Definition:

- One way to measure the dissimilarity of two probability distributions, p and q , is known as the KL divergence or relative entropy. This is defined as follows:

$$\mathbb{KL}(p||q) \triangleq \sum_{k=1}^K p_k \log \frac{p_k}{q_k}$$





Contd..

$$\text{KL} (p||q) = \sum_k p_k \log p_k - \sum_k p_k \log q_k = -\mathbb{H} (p) + \mathbb{H} (p, q)$$

where $\mathbb{H} (p, q)$ is called the **cross entropy**,

$$\mathbb{H} (p, q) \triangleq - \sum_k p_k \log q_k$$

Relative entropy with Boolean RV

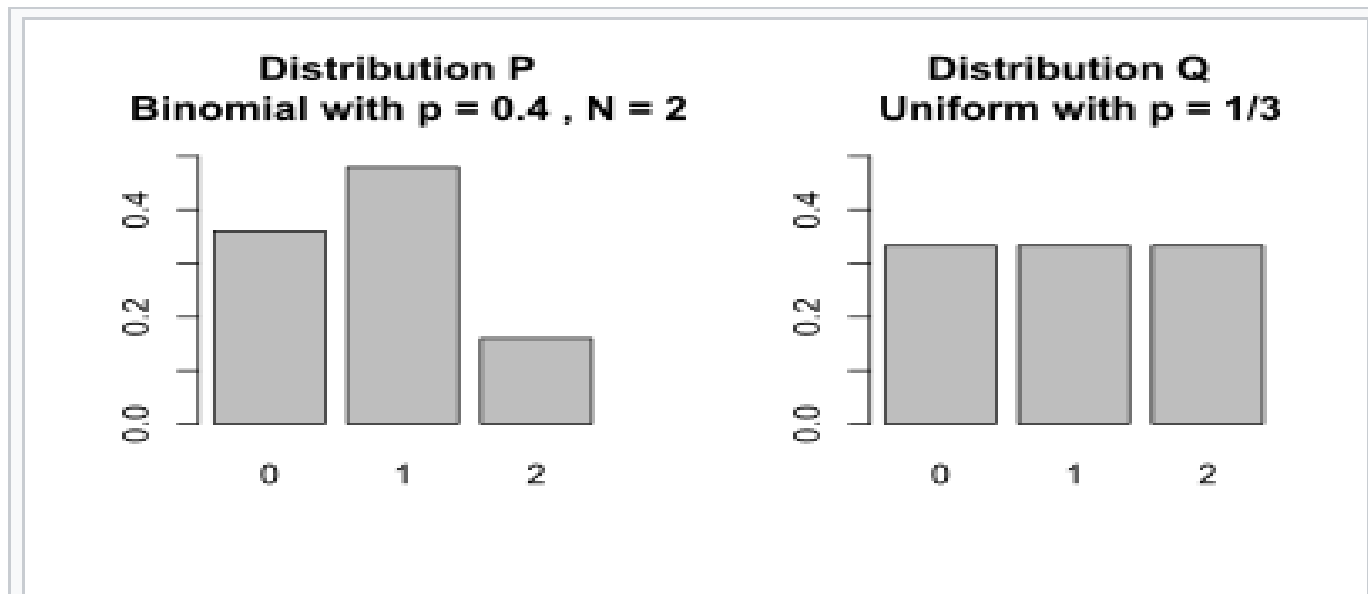
- For the special case of binary random variables, $X \in \{0, 1\}$, we can write $p(X = 1) = \theta$ and $p(X = 0) = 1 - \theta$.
- Consider a Boolean random variable $X = \{0, 1\}$ with two probability distribution p and q , then find relative entropy?



Contd..

Contd..

- Let P and Q be the distributions shown in the table and figure. P is the distribution on the left side of the figure, a binomial distribution with $N=2$ and $p = 0.4$
- Q is the distribution on the right side of the figure, a discrete uniform distribution with the three possible outcomes $X = \{0, 1, 2\}$, with $p = 1/3$





Solution:

| x | 0 | 1 | 2 |
|---------------------------------------|----------------|-----------------|----------------|
| Distribution $P(x)$ | $\frac{9}{25}$ | $\frac{12}{25}$ | $\frac{4}{25}$ |
| Distribution $Q(x)$ | $\frac{1}{3}$ | $\frac{1}{3}$ | $\frac{1}{3}$ |

$$\begin{aligned}
 D_{\text{KL}}(P \parallel Q) &= \sum_{x \in \mathcal{X}} P(x) \ln \left(\frac{P(x)}{Q(x)} \right) \\
 &= \frac{9}{25} \ln \left(\frac{9/25}{1/3} \right) + \frac{12}{25} \ln \left(\frac{12/25}{1/3} \right) + \frac{4}{25} \ln \left(\frac{4/25}{1/3} \right) \\
 &= \frac{1}{25} (32 \ln(2) + 55 \ln(3) - 50 \ln(5)) \approx 0.0852996,
 \end{aligned}$$

$$\begin{aligned}
 D_{\text{KL}}(Q \parallel P) &= \sum_{x \in \mathcal{X}} Q(x) \ln \left(\frac{Q(x)}{P(x)} \right) \\
 &= \frac{1}{3} \ln \left(\frac{1/3}{9/25} \right) + \frac{1}{3} \ln \left(\frac{1/3}{12/25} \right) + \frac{1}{3} \ln \left(\frac{1/3}{4/25} \right) \\
 &= \frac{1}{3} (-4 \ln(2) - 6 \ln(3) + 6 \ln(5)) \approx 0.097455.
 \end{aligned}$$