Introduction to Machine Learning

- Prof. Balaraman Ravindran | IIT Madras

Problem Solving Session (Week-8)

Shreya Bansal

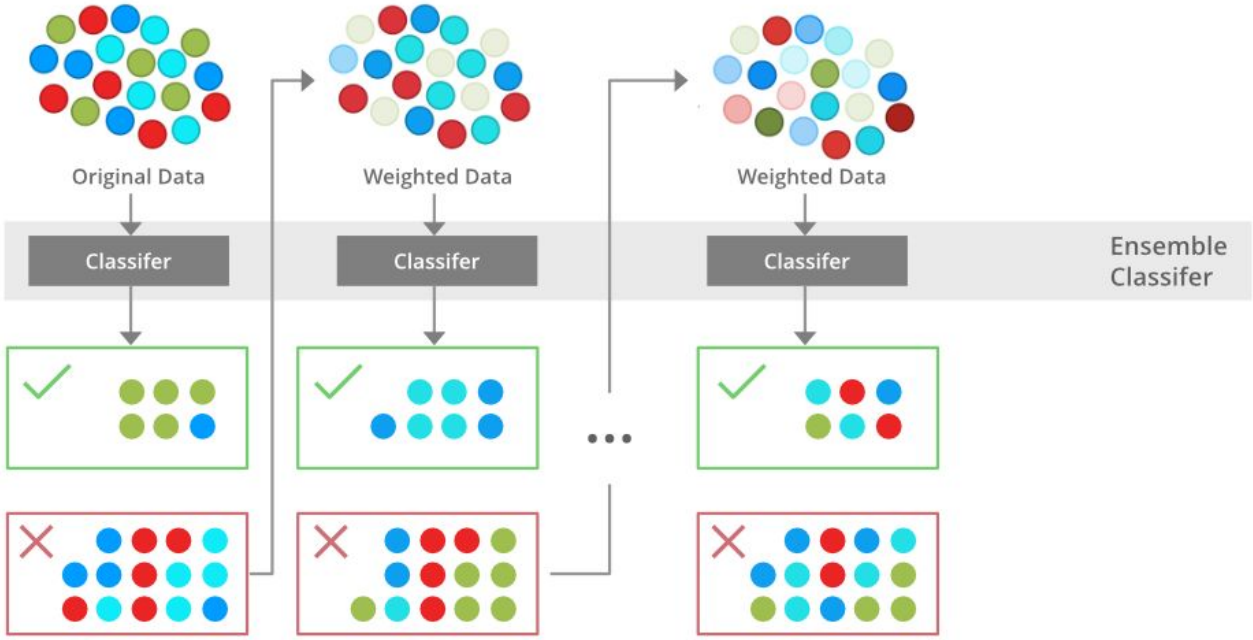
PMRF PhD Scholar
IIT Ropar

Week-8 Contents

---

1. Gradient Boosting
2. Random Forest (Bagging)
3. Naive Bayes
4. Bayesian Network
5. Multiclass Classification

Introduction to Boosting



Introduction to Boosting

- **What is Boosting?**
    - Stage-wise process to improve classifiers.
    - At each stage, errors from the previous stage are reduced.

- **Key Characteristics:**
    - Focuses on minimizing errors iteratively.
    - Examples: AdaBoost, LogitBoost, GradientBoost.

AdaBoost Steps

1. **Initialize Weights**:
   - Assign equal weights to all training examples: $w_i = \frac{1}{N}$.
2. **For each iteration** $t = 1, 2, \ldots, T$:
   - Train a weak learner $h_t(x)$ using weighted data.
   - Compute error: $\epsilon_t = \sum_{i=1}^{N} w_i \cdot \mathbb{I}(h_t(x_i) \neq y_i)$.
   - Compute learner weight: $\alpha_t = \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t} \right)$.
   - Update weights: $w_i \leftarrow w_i \cdot \exp\left(-\alpha_t \cdot y_i \cdot h_t(x_i)\right)$.
   - Normalize weights: $w_i \leftarrow \frac{w_i}{\sum_{j=1}^{N} w_j}$.
3. **Final Classifier**:
   - $H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t \cdot h_t(x)\right)$.

AdaBoost and Exponential Loss

- **AdaBoost:**
  - **Uses exponential loss.**
  - **Related to logistic loss (LogitBoost).**

- **Why AdaBoost is Popular:**
  - **Nice analytical properties.**
  - **Widely used for decades.**

Gradient Boosting Overview

- **What is Gradient Boosting?**
    - **A modern approach to boosting (last decade).**
    - **Often used with decision trees (Gradient Boosted Decision Trees).**

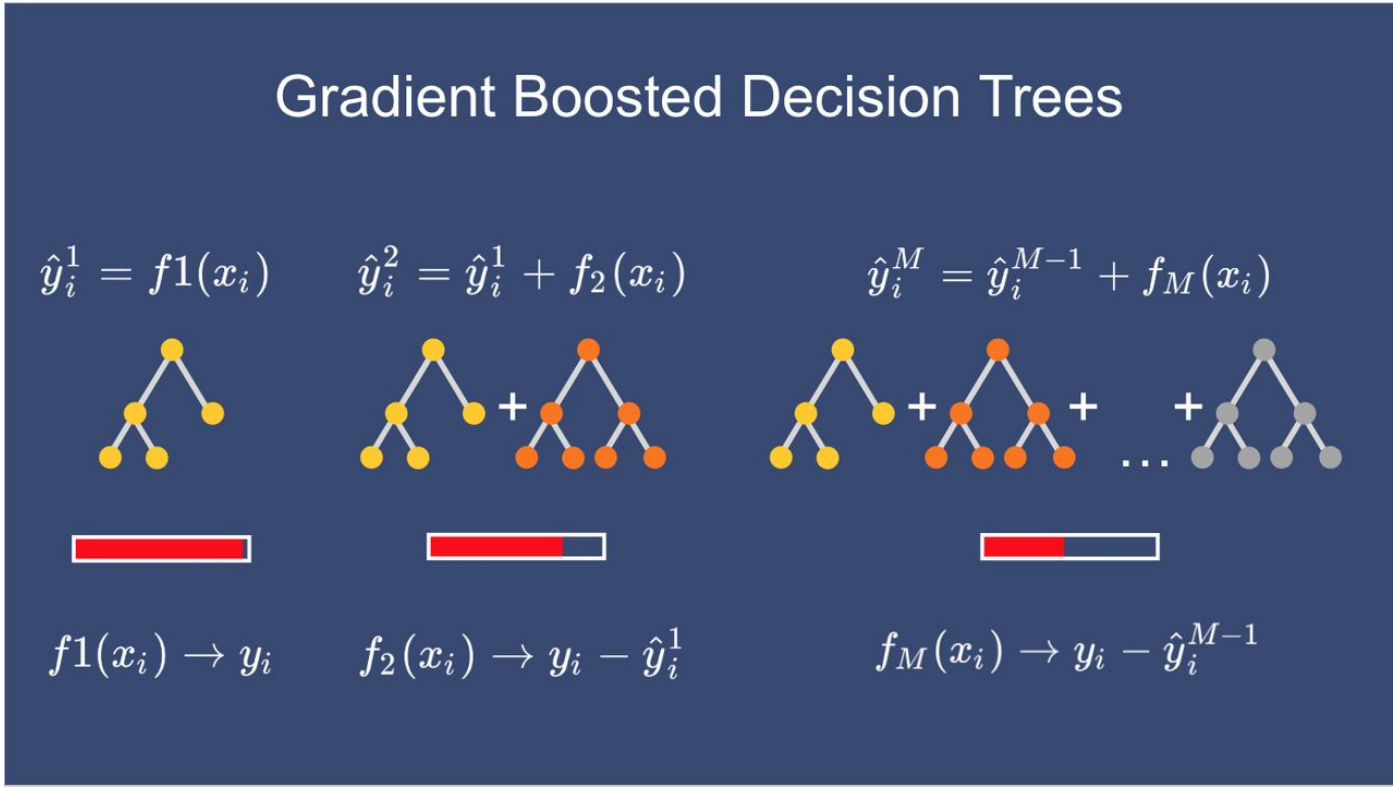
- **Why Gradient Boosting is Popular:**
    - **Hard to beat in many applications.**
    - **Combines the power of boosting and decision trees.**

Boosting with Trees

- **Boosting with Decision Trees:**
  - Combine outputs of multiple trees (forest).
  - Each tree corrects errors from previous trees.

- **Process:**
  - Build a tree to predict output.
  - Compute residuals (errors).
  - Build next tree to predict residuals.
  - Repeat until errors are minimized.

Gradient Boosting



Gradient Boosting

Problem Setup

We have a **regression problem** with 4 data points. The goal is to predict y based on feature x.

Data Point	x	y
1	1	2
2	2	4
3	3	6
4	4	8

Gradient Boosting Steps

Step 1: Initialize the Model

- Start with an initial prediction for all data points. A common choice is the **mean of the target values**:

$$F_0(x) = \text{mean}(y) = \frac{2+4+6+8}{4} = 5.$$

- Initial predictions:

$$F_0(x_1) = 5, \quad F_0(x_2) = 5, \quad F_0(x_3) = 5, \quad F_0(x_4) = 5.$$

Gradient Boosting

Step 2: Compute Residuals

- Residuals are the differences between the actual values (y) and the current predictions ($F_0(x)$):

$$\text{Residual} = y - F_0(x).$$

Data Point	x	y	$F_0(x)$	Residual ($y - F_0(x)$)
1	1	2	5	-3
2	2	4	5	-1
3	3	6	5	1
4	4	8	5	3

Step 3: Train a Weak Learner on Residuals

- Train a **weak learner** (e.g., a decision tree) to predict the residuals.
- Let's assume the weak learner predicts the following:

$$h_1(x_1) = -3, \quad h_1(x_2) = -1, \quad h_1(x_3) = 1, \quad h_1(x_4) = 3.$$

Gradient Boosting

Step 4: Update the Model

- Update the model by adding the weak learner's predictions to the current model:

$$F_1(x) = F_0(x) + h_1(x).$$

- Updated predictions:

$$F_1(x_1) = 5 + (-3) = 2, \quad F_1(x_2) = 5 + (-1) = 4,$$
$$F_1(x_3) = 5 + 1 = 6, \quad F_1(x_4) = 5 + 3 = 8.$$

Step 5: Compute New Residuals

- Compute residuals for the updated model:

$$\text{Residual} = y - F_1(x).$$

Data Point	x	y	$F_1(x)$	Residual ($y - F_1(x)$)
1	1	2	2	0
2	2	4	4	0
3	3	6	6	0
4	4	8	8	0

Gradient Boosting

Step 6: Repeat the Process

- Since the residuals are now zero, the model has perfectly fit the data. In practice, we would repeat the process until the residuals are minimized or a stopping criterion is met.

Final Model

- The final model is the sum of the initial prediction and all weak learners:

$$F(x) = F_0(x) + h_1(x) + h_2(x) + \cdots + h_T(x).$$

- In this example:

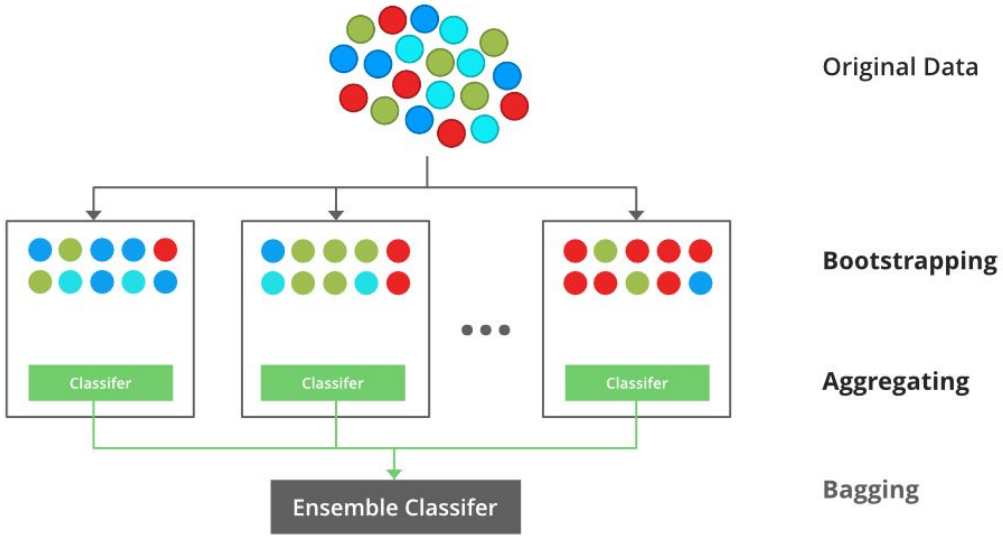
$$F(x) = 5 + h_1(x).$$

Gradient Boosting Mechanics

- **Gradient Descent Analogy:**
  - Start with an initial guess (F0).
  - Compute gradient of loss function.
  - Move in the direction of steepest descent.

- **Additive Model:**
  - Sequentially add corrections to improve predictions.

Bagging (Bootstrap Aggregating)



Introduction to Bagging

- Bagging (Bootstrap Aggregating) is an ensemble method that reduces variance by training multiple models on different subsets of the data.
- It works best when the base models (classifiers) are uncorrelated.

How Bagging Works

1. Take a dataset and sample with replacement to create multiple datasets (bootstrap samples).
2. Train multiple classifiers independently on these samples.
3. Average the predictions (for regression) or use majority voting (for classification).
4. Reduces variance by ensuring that small changes in data do not affect the model much.

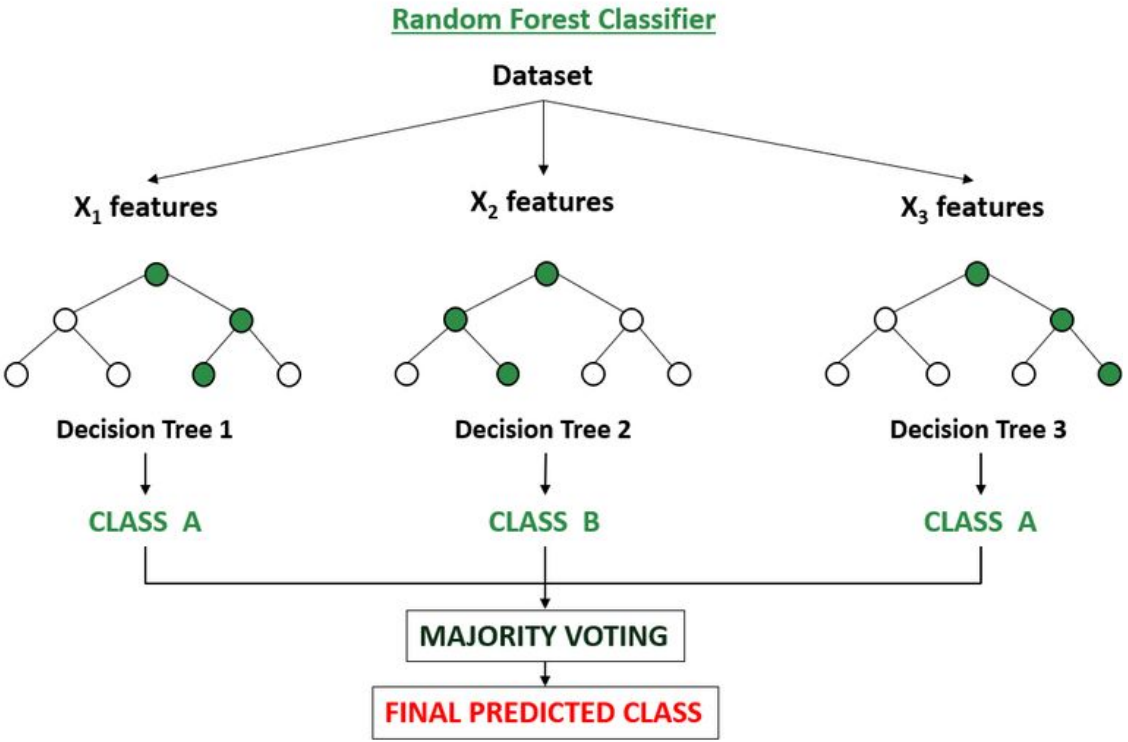
The Role of Correlation in Bagging

- If classifiers are highly correlated, bagging does not help much in variance reduction.
- The goal is to make classifiers as uncorrelated as possible.
- The more uncorrelated the classifiers, the greater the variance reduction.

Introduction to Random Forests

- Random Forest = Bagging + Feature Randomization.
- Standard bagging may lead to correlated trees because the most predictive features are frequently chosen.
- Random forests introduce randomness by selecting a subset of features at each node.

Introduction to Random Forests



How Random Forests Work

1. Sample with replacement to create bootstrap datasets.
2. For each tree:
    a. Select T random features from the total P features at each split.
    b. Find the best split using only these T features.
    c. Repeat recursively to grow the tree.
3. Final prediction: Majority voting (classification) or averaging (regression).

Why Random Forests Work Well

- Reduces correlation between trees.
- Leads to better variance reduction than standard bagging.
- Works well even when some features are noisy or less predictive.

Boosting vs. Random Forests

Random Forests	Gradient Boosting
Reduces variance	Reduces bias
Trees are trained independently	Trees are trained sequentially
Works well with high-variance models	Works well with high-bias models
Good for large datasets	More computationally expensive

Introduction to Bayesian Classification

- **Definition:** Bayesian classification is a probabilistic approach to classification based on Bayes' Theorem. It provides a systematic way to calculate the probability of a hypothesis given evidence.

- **Key Advantages:**
    - Handles uncertainty effectively
    - Provides probabilistic outputs
    - Works well with small datasets

- **Applications: Spam detection, medical diagnosis, image classification, etc.**

Bayes′ Theorem

- **Bayes′ Theorem states:**

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- **Where:**
- **P(H|E)** is the posterior probability (probability of hypothesis H given evidence E)
- **P(E|H)** is the likelihood (probability of evidence E given H)
- **P(H)** is the prior probability (initial probability of hypothesis H)
- **P(E)** is the marginal probability (probability of evidence E across all hypotheses)

Example

- **Suppose a person takes a medical test for a disease.**
- **Find P(Disease|Positive Test)**

- **Applying Bayes' Theorem:**

$$P(\text{Disease}|\text{Positive Test}) = \frac{P(\text{Positive Test}|\text{Disease}) * P(\text{Disease})}{P(\text{Positive Test})}$$

Example

- Prior probability P(Disease) = 0.01
- Likelihood P(Positive Test|Disease) =0.9
- False positive rate P(Positive Test|No Disease)=0.05

- P(No Disease) = 1-P(Disease) = 1-0.01=0.99
- Overall test positivity rate P(Positive Test)=

P(Disease) *P(Positive Test|Disease) + P(No Disease) * P(Positive Test|No Disease)

0.01 * 0.9 + 0.99 * 0.05 =0.0585

Example

- **Applying Bayes' Theorem:**

  (0.9 * 0.01)/0.0585 = 0.154

- **This means there is only a 15.4% chance the person actually has the disease despite testing positive!**

Bayes Optimal Classifier

- **The Bayes Optimal Classifier is the theoretical best classifier that minimizes classification error.**

$$C^* = \arg\max_C P(C|X)$$

- **Challenges:**
  - Requires knowledge of all class probability distributions, which is often impractical
  - Computationally expensive

- **Example: If we have a dataset with three classes (A, B, and C) and the following probabilities: P(A|X)=0.4 ,P(B|X)=0.35 ,P(C|X)=0.25 then the optimal classifier would assign X to class A.**

k-Nearest Neighbors (KNN) and Bayes

- **KNN can be interpreted in a Bayesian context as estimating probabilities based on local density.**
- **Key Idea:** $P(C|X) \approx \frac{k_C}{k}$

Where k_C is the number of neighbors belonging to class C, and is the total number of neighbors.

- **Example: If 3 out of 5 nearest neighbors belong to Class A, then: P(A|X)=%**

Naïve Bayes Classifier

A common application of Bayesian classification is the Naïve Bayes classifier, which assumes that all features are independent given the class.

Formula:

$$P(C|X_1, X_2, ..., X_n) \propto P(C) \prod_{i=1}^{n} P(X_i|C)$$

Example: Consider spam classification based on words in an email. If the presence of words "free," "win," and "prize" contribute to spam likelihood:

$$P(Spam|X) \propto P(Spam)P(free|Spam)P(win|Spam)P(prize|Spam)$$

Example: Spam Detection

- Let's classify an email as spam or not spam based on the presence of words like "free", "win", "money", etc.

Email	Contains "free"	Contains "win"	Contains "money"	Spam (Yes/No)
1	Yes	No	Yes	Yes
2	No	Yes	No	No
3	Yes	Yes	No	Yes
4	No	No	Yes	No

Example: Spam Detection

Step 1: Compute Priors

$$P(\text{Spam}) = \frac{2}{4} = 0.5, \quad P(\text{Not Spam}) = \frac{2}{4} = 0.5$$

Step 2: Compute Likelihoods

$$P(\text{free}|\text{Spam}) = \frac{2}{2} = 1, \quad P(\text{free}|\text{Not Spam}) = \frac{0}{2} = 0$$

$$P(\text{win}|\text{Spam}) = \frac{1}{2} = 0.5, \quad P(\text{win}|\text{Not Spam}) = \frac{1}{2} = 0.5$$

$$P(\text{money}|\text{Spam}) = \frac{1}{2} = 0.5, \quad P(\text{money}|\text{Not Spam}) = \frac{1}{2} = 0.5$$

Example: Spam Detection

Step 3: Classify a New Email

Suppose a new email contains "free" and "money" but not "win". Compute:

$$P(\text{Spam}|\text{free}, \text{money}) \propto P(\text{Spam})P(\text{free}|\text{Spam})P(\text{money}|\text{Spam})$$

$$= 0.5 \times 1 \times 0.5 = 0.25$$

$$P(\text{Not Spam}|\text{free}, \text{money}) \propto P(\text{Not Spam})P(\text{free}|\text{Not Spam})P(\text{money}|\text{Not Spam})$$

$$= 0.5 \times 0 \times 0.5 = 0$$

Since **0.25 > 0**, the email is classified as **Spam**.

Advantages and Disadvantages of Naïve Bayes

- **Pros:**
  - **Works well with high-dimensional data**
  - **Computationally efficient**
  - **Performs well with small datasets**

- **Cons:**
  - **Assumption of feature independence is often unrealistic**
  - **Poor performance when features are highly correlated**

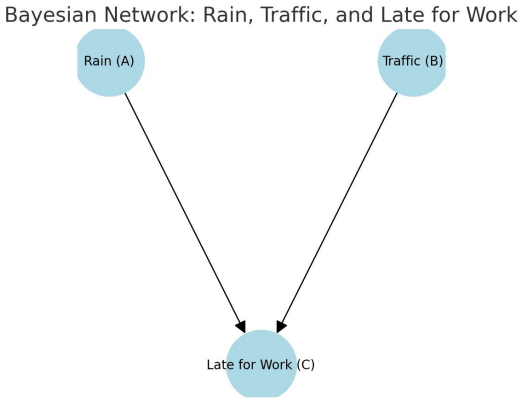
Introduction to Bayesian Belief Networks

- A Bayesian Belief Network is a probabilistic graphical model that represents a set of variables and their conditional dependencies using a directed acyclic graph (DAG).
- **Why is it important?**
    - Helps in reasoning under uncertainty
    - Reduces computational complexity
    - Widely used in AI, diagnostics, and decision-making
- **Example Scenario:**
    - Predicting the probability of a person having a disease based on symptoms and test results.

Components of a Bayesian Belief Network

- **Nodes:** Represent random variables
- **Edges:** Directed links showing dependencies
- **Conditional Probability Table (CPT):** Defines the probability of each variable given its parents

- **Example:**
- Let A be "Rain", B be "Traffic", and C be "Late for Work"
- The probability of being late depends on both rain and traffic.



Joint Probability Distribution in Bayesian Networks

- The joint probability of multiple variables can be written as:
- $P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1, X_2)\ldots P(X_n \mid X_1, \ldots, X_{n-1})$

- Factorization using conditional independence:
- If X_3 depends only on X_1, we simplify the joint probability.
- $P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)\ldots P(X_n \mid X_1, \ldots, X_{n-1})$

Example Bayesian Network with Probabilities

- Consider a simple Bayesian Network:
- Variables: "Cloudy", "Rain", "Sprinkler", "Wet Grass"
- Edges: "Cloudy → Rain", "Cloudy → Sprinkler", "Rain → Wet Grass", "Sprinkler → Wet Grass"
- The joint probability factorization:
- $P(Cloudy, Rain, Sprinkler, WetGrass) = P(Cloudy)P(Rain|Cloudy)P(Sprinkler|Cloudy)P(WetGrass|Rain, Sprinkler)$

Understanding Dependency Relations

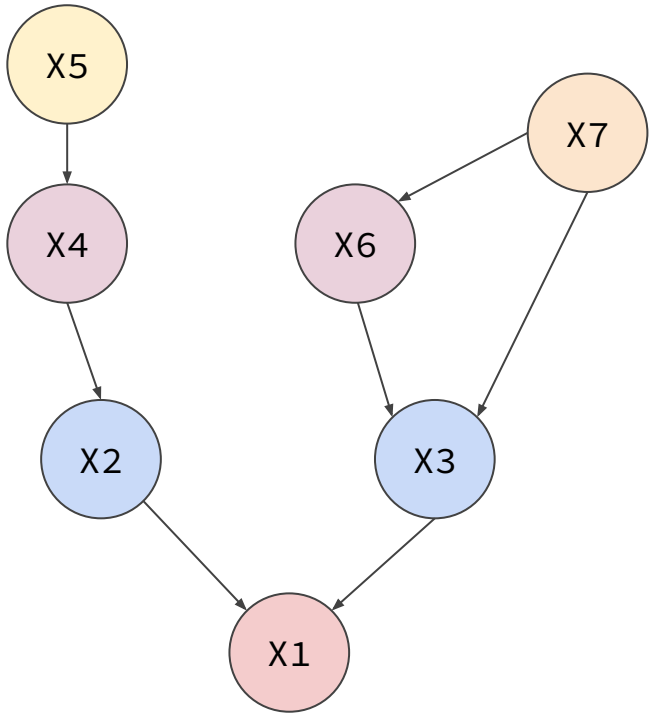
- **Example Dependency Relations:**
  - **X1 depends on X2 and X3**
  - **X3 depends on X6 and X7**
  - **X4 depends on X5**
  - **X6 depends on X7**
  - **X2 depends on X4**
  - **Some variables are completely independent**
- **These relationships can be represented graphically for better clarity.**

Understanding Dependency Relations

- $P(X_1,X_2,X_3,X_4,X_5,X_6,X_7)$

  $= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_1,X_2,X_3)$
  $P(X_5|X_1,X_2,X_3,X_4)P(X_6|X_1,X_2,X_3,X_4,X_5)\ P(X_7|X_1,X_2,X_3,X_4,X_5,X_6)$

  $=P(X_1|X_2,X_3)P(X_2|X_4)P(X_3|X_6, X_7)P(X_4|X_5)\ P(X_5)P(X_6|X_7)\ P(X_7)$



Conditional Independence in Bayesian Networks

- **Definition: Two variables A and B are conditionally independent given C if:**
- $P(A,B|C)=P(A|C)P(B|C)$
- **Example:**
- **Given "Rain", the probability of "Wet Grass" does not depend on "Sprinkler".**

Conditional Independence Example

- **Scenario:**
- If X2 is known, then X4 and X1 become independent.
- Without knowing X2, X4 and X1 remain dependent.

- **Real-world Analogy:**
- Consider the relationship between a cricket match and Dhoni's presence.
- If we know whether a cricket match is happening, Dhoni's presence becomes independent of the general sports context.

D-Separation in Bayesian Networks

- **D-Separation (Directional Separation) is a method used in Bayesian Networks to determine whether two variables are independent given some observed evidence.**
- **Definition**

  Two variables X and Y are d-separated (conditionally independent) given a set of observed variables Z if all paths between X and Y are "blocked" by Z.

D-Separation in Bayesian Networks

A path is blocked in the following cases:

1. **Chain Structure (X → Z → Y) or (X ← Z ← Y)**
    a. **If Z is observed, it blocks the path.**
    b. **If Z is not observed, it does not block the path.**
2. **Fork Structure (X ← Z → Y)**
    a. **If Z is observed, it blocks the path.**
    b. **If Z is not observed, it does not block the path.**
3. **Collider Structure (X → Z ← Y)**
    a. **If Z is NOT observed, the path is blocked.**
    b. **If Z OR any of its descendants are observed, the path is unblocked.**

D-Separation in Bayesian Networks

- A path is blocked in the following cases:
- Chain Structure (X → Z → Y) or (X ← Z ← Y)
    - If Z is observed, it blocks the path.
    - If Z is not observed, it does not block the path.
- Explanation:
    - If Z is observed, the path between X and Y is blocked (X and Y become independent).
    - If Z is not observed, X and Y remain dependent.

Example of D-Separation

- Example: X = Rain, Z = Wet Roads, Y = Traffic Jam
- Graph Representation: Rain→Wet Roads→Traffic Jam
- If we do not observe Wet Roads, Rain can still affect Traffic Jam.
- If we observe Wet Roads, knowing whether it rained or not doesn't change our belief about the Traffic Jam (Rain and Traffic Jam become independent).
- Mathematically: $P(Y|X,Z)=P(Y|Z)$
- (Traffic Jam depends only on Wet Roads once we observe it).

D-Separation in Bayesian Networks

- A path is blocked in the following cases:
- Fork Structure ($X \leftarrow Z \rightarrow Y$)
    - If Z is observed, it blocks the path.
    - If Z is not observed, it does not block the path.
- Explanation:
    - If Z is observed, the path between X and Y is blocked (X and Y become independent).
    - If Z is not observed, X and Y remain dependent.

Example of D-Separation

- Example: X = Sports Popularity, Z = Cricket Match, Y = Stadium Crowd
- Graph Representation: Sports Popularity←Cricket Match→Stadium Crowd
- If we do not observe the Cricket Match, Sports Popularity and Stadium Crowd seem related (both depend on Z).
- If we observe the Cricket Match, Sports Popularity and Stadium Crowd become independent (since Cricket Match fully explains the reason for the crowd).
- Mathematically: $P(X|Y,Z)=P(X|Z)$
- (Sports Popularity depends only on Cricket Match when we observe it).

D-Separation in Bayesian Networks

- **A path is blocked in the following cases:**
- **Collider Structure (X → Z ← Y)**
    - **If Z is NOT observed, the path is blocked.**
    - **If Z OR any of its descendants are observed, the path is unblocked.**
- **Explanation:**
    - **If Z is NOT observed, the path between X and Y is blocked (X and Y are independent).**
    - **If Z or any of its descendants are observed, the path becomes active, and X and Y become dependent.**

Example of D-Separation

- Example: X = Rain, Z = Wet Shoes, Y = Sprinklers
- Graph Representation: Rain→Wet Shoes←Sprinklers
- If we do not observe Wet Shoes, Rain and Sprinklers are independent (knowing about Rain doesn't tell us about Sprinklers).
- If we observe Wet Shoes, then knowing it didn't rain makes Sprinklers more likely to have been on, and vice versa.
- Mathematically: $P(X|Y,Z) \neq P(X|Z)$
- (Rain and Sprinklers become dependent once we observe Wet Shoes).

Summary of D-Separation

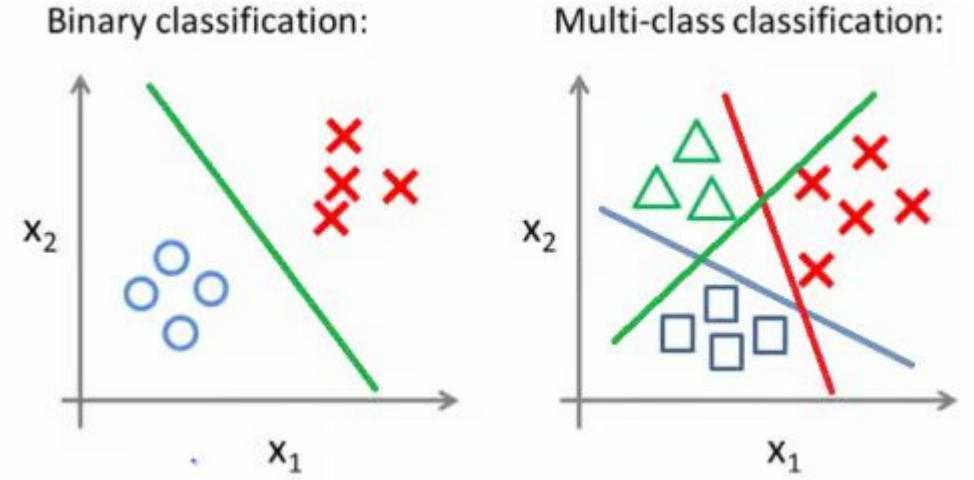
Summary of Blocking Conditions

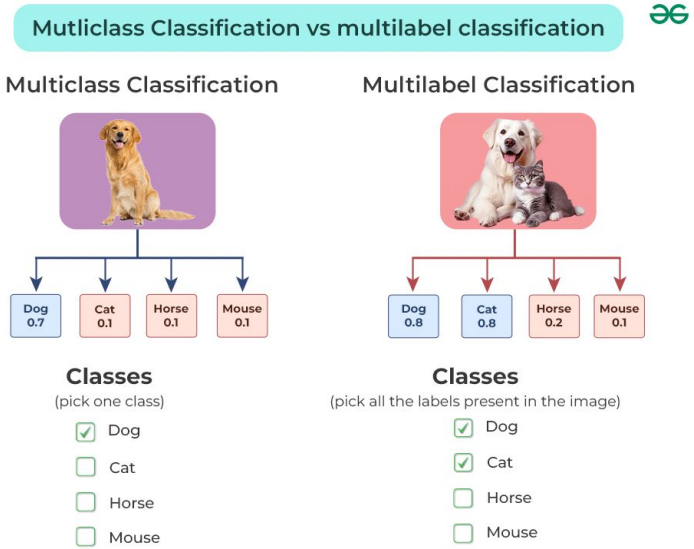
Graph Type	Path Blocked When	Path Open When
Chain $X \to Z \to Y$	Z is **observed**	Z is **not observed**
Fork $X \leftarrow Z \to Y$	Z is **observed**	Z is **not observed**
Collider $X \to Z \leftarrow Y$	Z is **not observed**	Z OR any of its **descendants** are observed

Real-World Applications of Bayesian Networks

- Medical Diagnosis: Predicting diseases based on symptoms.
- Spam Filtering: Probabilistic classification of emails.
- Fault Diagnosis: Identifying failures in complex systems.
- AI and Robotics: Decision-making under uncertainty.

Multi-class Classification





Naturally Multi-Class Classifiers

- Neural Networks
- Decision Trees (handles multiple classes naturally)
- Naïve Bayes & Bayesian Classifiers

Inherently Two-Class Classifiers

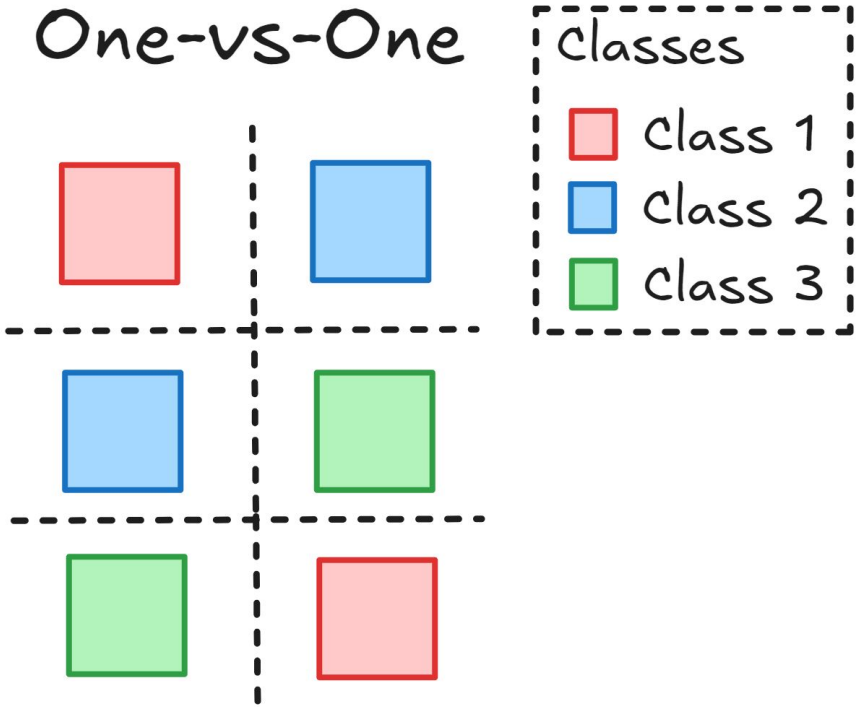
- Support Vector Machines (SVMs)
- Logistic Regression (basic form is binary, but has multi-class extensions)
- Discriminant Function-Based Classifiers

Converting Binary to Multi-Class

- One-vs-One (OvO) Strategy
  - Trains n(n-1)/2 classifiers
  - Balanced, but computationally expensive
- One-vs-All (OvA) Strategy
  - Trains n classifiers
  - Class imbalance issues

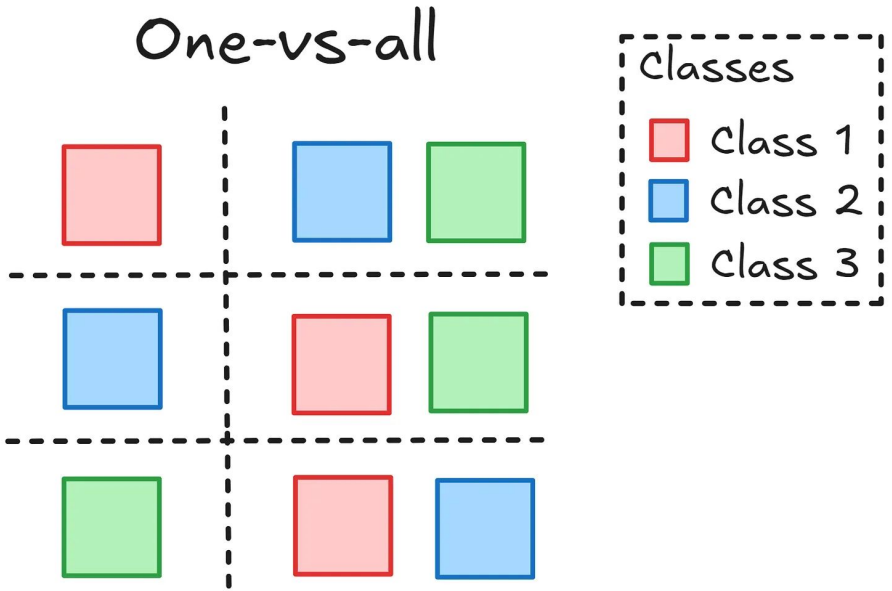
Converting Binary to Multi-Class

- One-vs-One (OvO) Strategy
  - Trains n(n-1)/2 classifiers
  - Balanced, but computationally expensive



Converting Binary to Multi-Class

- One-vs-All (OvA) Strategy
  - Trains n classifiers
  - Class imbalance issues



Converting Binary to Multi-Class

- **Tournament Approach- variant of 1 vs 1**
  a. **Pairwise Classification:**
     i. Train a binary classifier for each possible pair of classes.
     ii. Each classifier determines which of the two classes is more likely for a given sample.
  b. **Tournament Structure:**
     i. Arrange classes in a bracket-like structure.
     ii. In each round, classifiers compare pairs, and one class is eliminated while the other advances.
  c. **Final Decision:**
     i. The last remaining class after all rounds is the predicted class.

Converting Binary to Multi-Class

- Tournament Approach- variant of 1 vs 1
  - Example of 4-Class Classification (A, B, C, D):
  - First Round:
    - A vs. B → Winner advances
    - C vs. D → Winner advances
  - Second Round:
    - Winner (A/B) vs. Winner (C/D) → Final Class Prediction

Challenges in Multi-Class Classification

- **Class Imbalance**
    - **Some classes might have far more data than others**
    - **Example: 1 class with 1M samples, others with 1,000**
- **Ways to Address Class Imbalance**
    - **Over-sampling / Under-sampling**
    - **Class Weighing**
    - **Hierarchical Classification**

Hierarchical Classification

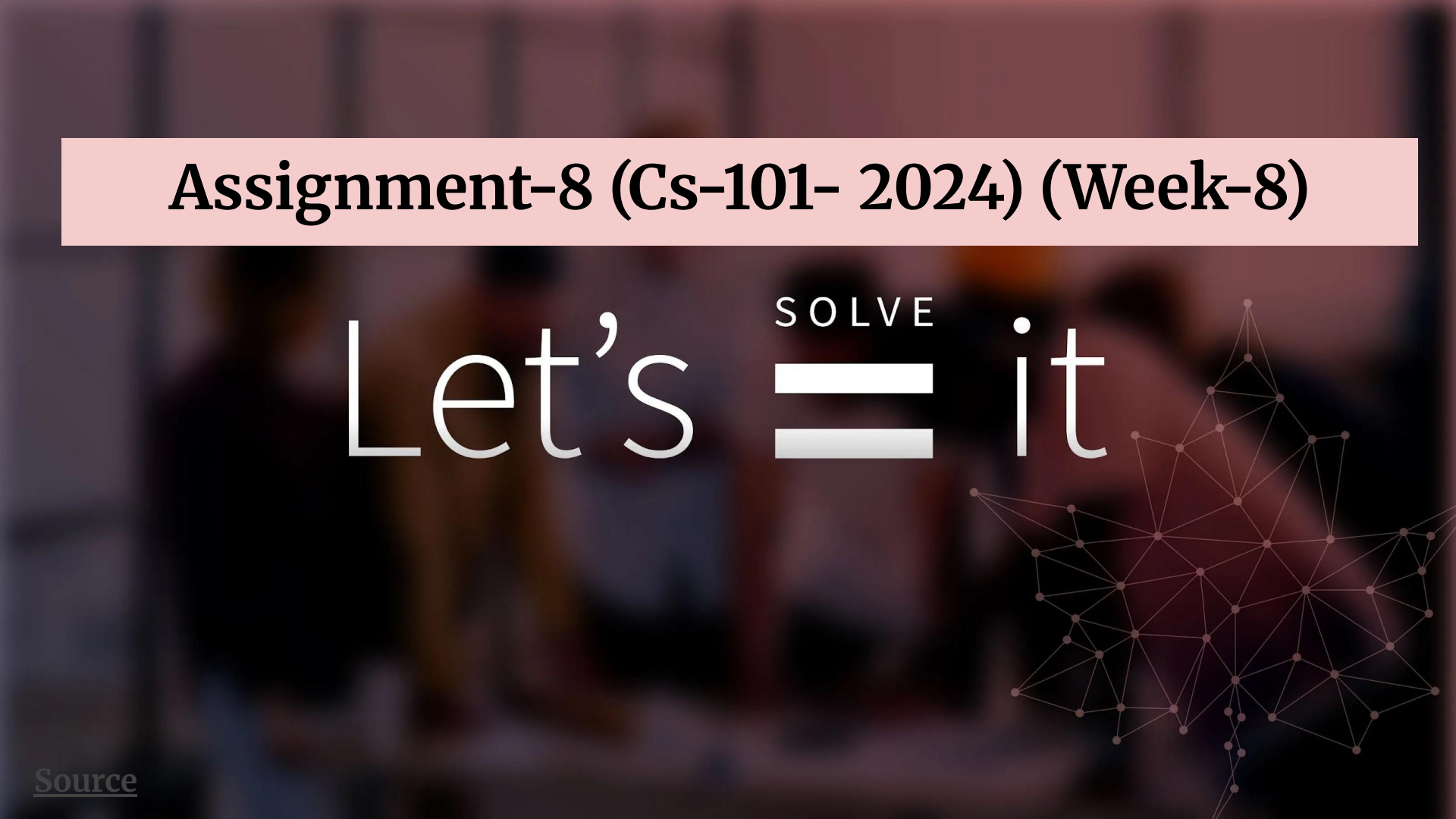
- Group classes into hierarchies
- Step-wise classification (broad to specific)
- Example:
  - First level: Entertainment vs. News
  - Second level: News → Politics, Sports
  - Third level: Politics → National, International

Hierarchical Classification

Clustering for Hierarchical Classification

- Use clustering to group similar classes
- Based on class conditional densities
- Helps manage class imbalance by structuring data

Assignment-8 (Cs-101- 2024) (Week-8)



Source

Question-1

01:00

In Bagging technique, the reduction of variance is maximum if:

a) The correlation between the classifiers is minimum
b) Does not depend on the correlation between the classifiers
c) Similar features are used in all classifiers
d) The number of classifiers in the ensemble is minimized

Question-1- Correct answer

In Bagging technique, the reduction of variance is maximum if:

a) The correlation between the classifiers is minimum
b) Does not depend on the correlation between the classifiers
c) Similar features are used in all classifiers
d) The number of classifiers in the ensemble is minimized

Correct options: (a)-This ensures diverse predictions that effectively average out errors

Question-2

01:00

If using squared error loss in gradient boosting for a regression problem, what does the gradient correspond to?

a) The absolute error
b) The log-likelihood
c) The residual error
d) The exponential loss

Question-2- Correct answer

If using squared error loss in gradient boosting for a regression problem, what does the gradient correspond to?

a) The absolute error

b) The log-likelihood

c) The residual error- $\Delta(y-f(x;w))2 = 2(y - f(x;w))\Delta f(x;w)$

d) The exponential loss

Correct options: (c)

Question-3

01:00

In a random forest, if T (number of features considered at each split) is set equal to P (total number of features), how does this compare to standard bagging with decision trees?

a) It's exactly the same as standard bagging
b) It will always perform better than standard bagging
c) It will always perform worse than standard bagging
d) Can not be determined

Question-3 - Correct answer

In a random forest, if T (number of features considered at each split) is set equal to P (total number of features), how does this compare to standard bagging with decision trees?

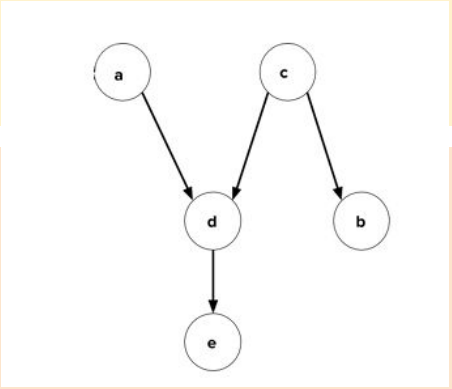
a) It's exactly the same as standard bagging
b) It will always perform better than standard bagging
c) It will always perform worse than standard bagging
d) Can not be determined

Correct options: (a)

Question-4

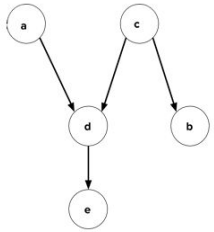
03:00

Multiple Correct: Consider the following graphical model, which of the following are true about the model? (multiple options may be correct)



a) d is independent of b when c is known
b) a is independent of c when e is known
c) a is independent of b when e is known
d) a is independent of b when c is known

Question-4

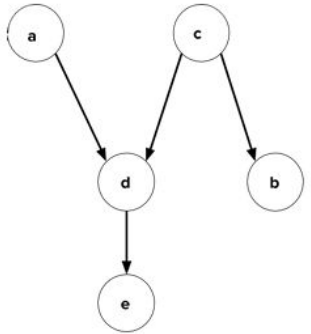


03:00

a) d is independent of b when c is known -**fork (2)(a)**
b) a is independent of c when e is known- **collider (3)(b)**
c) a is independent of b when e is known- **from above, a and c are dependent , so b depend on c → a and b are dependent (e don't separate a,b)**
d) a is independent of b when c is known—> **knowing c, d and b are independent and since d depends on a, so and b are also independent**

Question-4 - Correct answer

Multiple Correct: Consider the following graphical model, which of the following are true about the model? (multiple options may be correct)



a) d is independent of b when c is known
b) a is independent of c when e is known
c) a is independent of b when e is known
d) a is independent of b when c is known

Correct options: (a)(d) - Refer slide for d-separation condition

Question-5

01:00

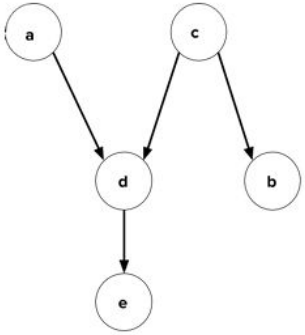
Consider the Bayesian network given in the previous question. Let “a”, “b”, “c”, “d” and “e” denote the random variables shown in the network. Which of the following can be inferred from the network structure?

a) “a” causes “d”

b) “e” causes “d”

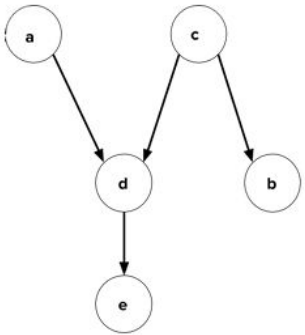
c) Both (a) and (b) are correct

d) None of the above



Question-5 - Correct answer

Consider the Bayesian network given in the previous question. Let “a”, “b”, “c”, “d” and “e” denote the random variables shown in the network. Which of the following can be inferred from the network structure?



a) “a” causes “d”

b) “e” causes “d”

c) Both (a) and (b) are correct

d) None of the above

Correct options: (d)-Node “d” is dependent on both “a” and “c” and “e” can not cause “d”

Question-6

01:00

A single box is randomly selected from a set of three. Two pens are then drawn from this container. These pens happen to be blue and green colored. What is the probability that the chosen box was Box A?

a) 37/18
b) 15/56
c) 18/37
d) 56/15

Box	Green	Blue	Yellow
A	3	2	1
B	2	1	2
C	4	2	3

Question-6- Explanation

01:00

A single box is randomly selected from a set of three. Two pens are then drawn from this container. These pens happen to be blue and green colored. What is the probability that the chosen box was Box A?

Let E be event choosing blue, green

P(Box A | (blue,green))= P(A|E) = [P(E|A) * P(A)]/P(E)

$P(E|A)= {}^{3}C_{1} * {}^{2}C_{1} / {}^{6}C_{2} = 3*2 / 15 = 2/5$ $P(A) = 1/3$

P(E)= P(A)* P(E|A) + P(B)* P(E|B) + P(B)* P(E|B)

1/3 * 2/5 + 1/3 * 1/5 + 1/3 * 2/9 =1/3 * 37/45

Box	Green	Blue	Yellow
A	3	2	1
B	2	1	2
C	4	2	3

P(A|E) = (1/3 * 2/5) / (1/3 * 37/45)

2/5 * 45/37 = 18/37

Question-6 - Correct answer

A single box is randomly selected from a set of three. Two pens are then drawn from this container. These pens happen to be blue and green colored. What is the probability that the chosen box was Box A?

a) 37/18

b) 15/56

c) 18/37

d) 56/15

Box	Green	Blue	Yellow
A	3	2	1
B	2	1	2
C	4	2	3

Correct options: (c)

Question-7

03:00

State True or False: The primary advantage of the tournament approach in multi class classification is its effectiveness even when using weak classifiers.

a) True

b) False

Question-7 - Correct answer

State True or False: The primary advantage of the tournament approach in multi class classification is its effectiveness even when using weak classifiers.

a) True

b) False - Disadvantage: Early elimination risk: A weak classifier in the early rounds may cause misclassification.

Correct options: (b)

Question-8

03:00

A data scientist is using a Naive Bayes classifier to categorize emails as either “spam” or“notspam”. The features used for classification include:

- Number of recipients(To,Cc,Bcc)
- Presence of“spam”keywords(e.g.,”URGENT”,”offer”,”free”)
- Time of day the email was sent
- Length of the email in words

Which of the following scenarios, if true, is most likely to violate the key assumptions of Naive Bayes And Potentially Impact its performance?

a) The Length Of The Email follows non-Gaussian Distribution
b) The Time Of Day Is Discretized Into Categories(morning,afternoon,evening,night)
c) The proportion of spam emails in the training data is lower than in real-world email traffic
d) There's Strong correlation between the presence Of the word”free”and the length of the email

Question-8 - Correct answer

A data scientist is using a Naive Bayes classifier to categorize emails as either “spam” or“notspam”. The features used for classification include:

- Number of recipients(To,Cc,Bcc)
- Presence of“spam”keywords(e.g.,”URGENT”,”offer”,”free”)
- Time of day the email was sent
- Length of the email in words

Which of the following scenarios, if true, is most likely to violate the key assumptions of Naive Bayes And Potentially Impact its performance?

a) The Length Of The Email follows non-Gaussian Distribution
b) The Time Of Day Is Discretized Into Categories(morning,afternoon,evening,night)
c) The proportion of spam emails in the training data is lower than in real-world email traffic
d) There's Strong correlation between the presence Of the word”free”and the length of the email

Correct options: (d)- This scenario violates the Naive Bayes assumption of feature independence,as it the features are dependent on each other.

Question-9

03:00

Consider the two statements:

Statement 1: Bayesian Networks are inherently structured as Directed Acyclic Graphs (DAGs).

Statement 2: Each node in a bayesian network represents a random variable, and each edge represents conditional dependence.

Which of these are true?

a) Both the statements are True.
b) Statement 1 is true, and statement 2 is false.
c) Statement 1 is false, and statement 2 is true.
d) Both the statements are false.

Question-9 - Correct answer

Consider the two statements:

Statement 1: Bayesian Networks are inherently structured as Directed Acyclic Graphs (DAGs).

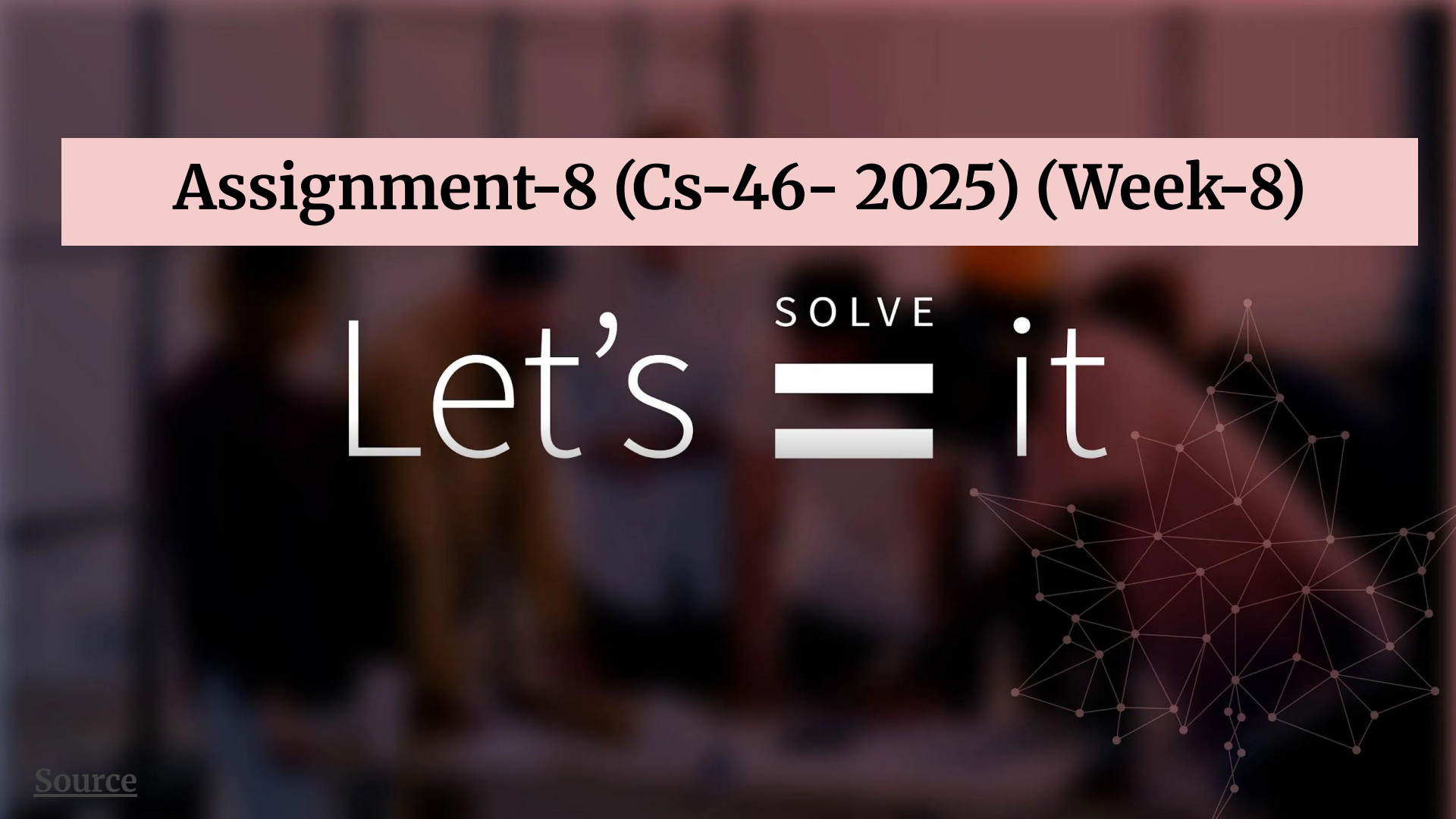
Statement 2: Each node in a bayesian network represents a random variable, and each edge represents conditional dependence.

Which of these are true?

a) Both the statements are True.
b) Statement 1 is true, and statement 2 is false.
c) Statement 1 is false, and statement 2 is true.
d) Both the statements are false.

Correct options: (a)

Assignment-8 (Cs-46- 2025) (Week-8)



Source

Question-1

01:00

Which of these statements is/are True about Random Forests?

a) The goal of random forests is to decrease the correlation between the trees.
b) The goal of random forests is to increase the correlation between the trees.
c) In Random Forests, each decision tree fits the residuals from the previous one; thus, the correlation between the trees won't matter.
d) None of these

Question-1- Correct answer

Which of these statements is/are True about Random Forests?

a) The goal of random forests is to decrease the correlation between the trees.
b) The goal of random forests is to increase the correlation between the trees.
c) In Random Forests, each decision tree fits the residuals from the previous one; thus, the correlation between the trees won't matter.
d) None of these

Correct options: (a)

Question-2

01:00

Consider the two statements:

Statement 1: Gradient Boosted Decision Trees can overfit easily.

Statement 2: It is easy to parallelize Gradient Boosted Decision Trees.

Which of these are true?

a) Both the statements are True.

b) Statement 1 is true, and statement 2 is false.

c) Statement 1 is false, and statement 2 is true.

d) Both the statements are false.

Question-2- Correct answer

Consider the two statements:

Statement 1: Gradient Boosted Decision Trees can overfit easily.

Statement 2: It is easy to parallelize Gradient Boosted Decision Trees.

Which of these are true?

a) Both the statements are True.

b) Statement 1 is true, and statement 2 is false.

c) Statement 1 is false, and statement 2 is true.

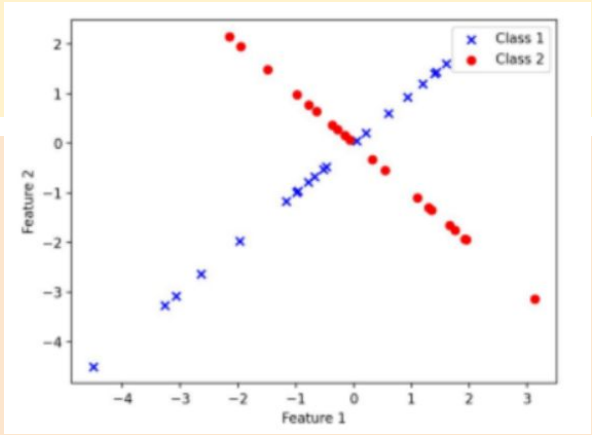
d) Both the statements are false.

Correct options: (b)

Question-3

01:00

A dataset with two classes is plotted below.Does the data satisfy the Naive Bayes assumption?



a) Yes
b) No
c) The given data is insufficient
d) None of these

Question-3 - Correct answer

A dataset with two classes is plotted below. Does the data satisfy the Naive Bayes assumption?

a) Yes
b) No
c) The given data is insufficient
d) None of these

Correct options: (b)

Question-4

x	y
India won the match.	Cricket
The Mercedes car was driven by Lewis Hamilton.	Formula 1
The ball was driven through the covers for a boundary	Cricket
Max Verstappen has a fast car.	Formula 1
Bumrah is a fast bowler.	Cricket
Max Verstappen won the race	Formula 1

Consider the below dataset: Suppose you have to classify a test example “The ball won the race to the boundary” and are asked to compute P(Cricket |“The ball won the race to the boundary”), what is an issue that you will face if you are using Naive Bayes Classifier, and how will you work around it? Assume you are using word frequencies to estimate all the probabilities.

a) There won’t be a problem, and the probability of P(Cricket |“The ball won the race to the boundary”) will be equal to 1.
b) Problem: A few words that appear at test time do not appear in the dataset. Solution: Smoothing.
c) Problem: A few words that appear at test time appear more than once in the dataset. Solution: Remove those words from the dataset.
d) None of these

Question-4 - Correct answer

Suppose you have to classify a test example “The ball won the race to the boundary” and are asked to compute P(Cricket |“The ball won the race to the boundary”), what is an issue that you will face if you are using Naive Bayes Classifier, and how will you work around it? Assume you are using word frequencies to estimate all the probabilities.

a) There won’t be a problem, and the probability of P(Cricket |“The ball won the race to the boundary”) will be equal to 1.
b) Problem: A few words that appear at test time do not appear in the dataset. Solution: Smoothing.
c) Problem: A few words that appear at test time appear more than once in the dataset. Solution: Remove those words from the dataset.
d) None of these

Correct options: (b)

Question-5

01:00

A company hires you to look at their classification system for whether a given customer would potentially buy their product. When you check the existing classifier on different folds of the training set, you find that it manages a low accuracy of usually around 60%. Sometimes, it's barely above 50%.

With this information in mind, and without using additional classifiers, which of the following ensemble methods would you use to increase the classification accuracy effectively?

a) Committee Machine

b) AdaBoost

c) Bagging

d) Stacking

Question-5 - Correct answer

A company hires you to look at their classification system for whether a given customer would potentially buy their product. When you check the existing classifier on different folds of the training set, you find that it manages a low accuracy of usually around 60%. Sometimes, it's barely above 50%.With this information in mind, and without using additional classifiers, which of the following ensemble methods would you use to increase the classification accuracy effectively?

a) Committee Machine

b) AdaBoost

c) Bagging

d) Stacking

Correct options: (b)

Question-6

Type	Single SIM	5G Comaptability	NFC	Total
Budget	15	5	0	20
Mid-Range	20	20	15	30
High End	15	15	15	20

Consider the following data for 20 budget phones, 30 mid-range phones, and 20 high-end phones: Consider a phone with 2 SIM card slots and NFC but no 5G compatibility. Calculate the probabilities of this phone being a budget phone, a mid-range phone, and a high-end phone using the Naive Bayes method. The correct ordering of the phone type from the highest to the lowest probability is?

a) Budget, Mid-Range, High End

b) Budget, High End, Mid-Range

c) Mid-Range, High End, Budget

d) High End, Mid-Range, Budget

Question-6 - Correct answer

Consider the following data for 20 budget phones, 30 mid-range phones, and 20 high-end phones:

Consider a phone with 2 SIM card slots and NFC but no 5G compatibility. Calculate the probabilities of this phone being a budget phone, a mid-range phone, and a high-end phone using the Naive Bayes method. The correct ordering of the phone type from the highest to the lowest probability is?

a) Budget, Mid-Range, High End

b) Budget, High End, Mid-Range

c) Mid-Range, High End, Budget

d) High End, Mid-Range, Budget

Correct options: (c)

Question-7

03:00

Suppose you have a 6 class classification problem with one input variable. You decide to use logistic regression to build a predictive model. What is the minimum number of ($\beta 0$,β) parameter pairs that need to be estimated?

a) 6

b) 12

c) 5

d) 10

Question-7 - Correct answer

Suppose you have a 6 class classification problem with one input variable. You decide to use logistic regression to build a predictive model. What is the minimum number of (β0,β) parameter pairs that need to be estimated?

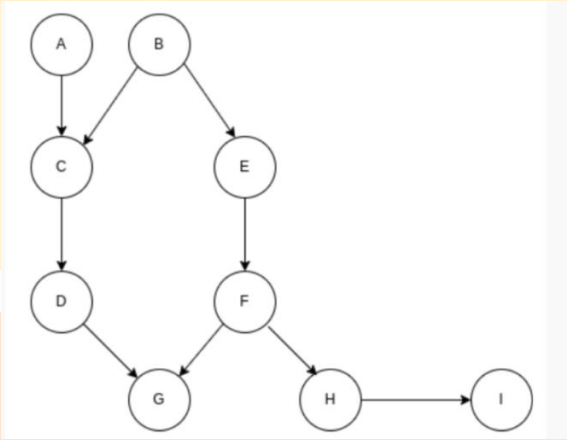
a) 6
b) 12
c) 5
d) 10

Correct options: (c)

Question-8

03:00

The figure below shows a Bayesian Network with 9 variables, all of which are binary. Which of the following is/are always true for the above Bayesian Network?



a) $P(A, B|G) = P(A|G)P(B|G)$
b) $P(A, I) = P(A)P(I)$
c) $P(B, H|E, G) = P(B|E, G)P(H|E, G)$
d) $P(C|B, F) = P(C|F)$

Question-8 - Correct answer

The figure below shows a Bayesian Network with 9 variables, all of which are binary. Which of the following is/are always true for the above Bayesian Network?

a) $P(A, B|G) = P(A|G)P(B|G)$

b) $P(A, I) = P(A)P(I)$

c) $P(B, H|E, G) = P(B|E, G)P(H|E, G)$

d) $P(C|B, F) = P(C|F)$

Correct options: (b)


THANK YOU

Suggestions and Feedback

Next Session:

Tuesday: 21-Sep-2025
3:00 - 5:00 PM