ISHAAN JAIN

1. Challenges in High Dimensional Datasets:
PCA Vs. LDA

High Dimensional datasets like those often seen in genomics, image processing or finance, present key challenges such as the curse of dimensionality, increased computational costs, risk of overfitting, multi-collinearity and reduced model interpretability. Dimensionality reduction is used to mitigate these challenges.

PRINCIPAL COMPONENT ANALYSIS (PCA)

- Objectives - Identify directions (principal components) that capture the maximum variance in the data, disregarding class labels.
- Methodology - Unsupervised transformation into a new set of orthogonal variables (principal components) which are linear combinations of original features.
- Applications - Data visualization, noise reduction, pre-processing for unsupervised learning, exploratory data analysis.

LINEAR DISCRIMINANT ANALYSIS (LDA)

- Objective: find directions that maximize the seperability between multiple known classes by maximizing the ratio of between-class variance to within class-variance
- Methodology: Supervised transformation relying on class labels, primarily used for enhancing classification accuracy.
- Applications: Preprocessing for classification, feature selection in supervised learning, effective where class labels are available.

Aspect	PCA	LDA
Supervision	Unsupervised	Supervised
Objective	Maximize Variance	Maximize class seperability

Methodology	Eigen decomposition of covariance matrix	Eigen decomposition of scatter matrices
Applicability	Any dataset	Classification tasks with labelled data
Limitations	May not aid classification directly	Needs representative labeled data

2. k-Means Vs. Hierarchical Clustering

Clustering is an unsupervised task to group similar items. k-Means and Hierarchical are popular clustering algorithms.

k-Means Clustering -

- How it works - Pre-specifies a number of clusters (k), assigns data points to clusters to minimize within-cluster variance, iteratively refines cluster centroids.
- Advantages - Efficient for large datasets, works well with spherical clusters, requires k in advance.
- Applications - Customer segmentation, image compression, document clustering.

Hierarchical Clustering -

- How it works - Builds a dendrogram (tree of clusters), agglomerative (bottom-up) or divisive (top-down), clusters can be chosen at any level of hierarchy.
- Advantages - No need to specify number of clusters in advance, flexible with cluster shapes, good for smaller datasets.
- Applications - Genomics, market basket analysis, social network analysis.

Criteria	K-Means	Hierarchical
Requires cluster count	Yes	No
Cluster shape	Spherical, uniform size	Varies, flexible
Visualization	Direct cluster labels	Dendrogram, full hierarchy
Scale	large datasets	Small datasets

3. Naive Bayes Algorithm

Derivations from Baye's Theorem

Baye's Theorem defines the probability of a hypothesis given evidence:

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

Naive-Bayes assumes all features are independent given the class label, which simplifies computation to -

$$P(C|X) \text{ or } P(C) \prod_{i=1}^{n} P(X_i|C)$$

Application in Text Classification

- Commonly used for spam detection, sentiment analysis, document organization
- Multinomial and Bernoulli models manage different types of discrete/ categorical data efficiency.

Strengths

- Simple to implement, fast, scales to large datasets, handles high-dimensional data, works well with small training sets.
- Robust to noisy data, less prone to overfitting.

LIMITATIONS

- Assumes feature independence, poor with correlated or interacting features, sensitive to imbalanced class.
- Zero probability [redacted] (for unseen data) addressed by Laplace smoothing.

- May perform poorly when feature distributions differ from model assumptions.

4. Support Vector Machines (SVM)

Correct and optimal Hyperplane

SVM is supervised classification algorithm that identifies a hyperplane in feature space that best seperates data into classes with maximum margin.

- for linearly seperable data, SVM finds a hyperplane maximizing the distance to the nearest points from both classes (support vectors)
- for non-linearly seperable data, kernel functions transform data to higher dimensions to find the optimal seperating hyperplane

Common kernel functions

- Linear kernel - for linearly seperable class
- Polynomial kernel - Can model curved boundaries, flexible for polynomial structures.
- Radial Basis function (RBF) - Most commonly used, handles clusters in complex, high-dimensional spaces
- Sigmoid kernel - Used less often, related to neural network

SVM's are widely used in image classification, bioinformatics, text categorization.

5. Decision Tree Method: Categorical Vs. Continous Splitting

Decision trees use splitting criteria to decide the best way to partition data at each node.

Splitting Criteria

- Information Gain - Based on entropy (impurity of dataset) measures reduction in randomness after a split. Used for categorical and continous attributes, especially in ID3,

C4.5 algorithms

$$IG = Entropy_{before} - \sum_{j=1}^{K} Entropy_{after}^{(j)}$$

- Gini Index - Measures misclassification rate; lower values preferred. Used mainly for classification (binary splits in CART Algorithm)

$$Gini = 1 - \sum_{i=1}^{n} p_i^2$$

Categorical Vs. Continous Splitting

- Categorical: Uses Information to Gain, Gini Index based on Categorical / classes, produces multi-way splits or binary splits.
- Continous: Splits based on threshold values; possible values for splits evaluated using same criteria.

Both these metrics help create effective, interpretable decision trees.