Extra Credit Assignment

# Question 1: Decision Trees

## A

With increase in tree complexity, variance increases and bias decreases which leads to overfitting, while if tree is not complex variance is low and tree can be highly biased towards a few classes.

Diagram

Description automatically generated

## B

As in diagram in (A),

* Train error – bias/variance error
* Test error – total error

Diagram

Description automatically generated

## D

Purpose of tree pruning is to avoid overfitting.

Types of trees pruning techniques:

* Pre-pruning
  + Early stopping of splitting and branching at decision nodes when no significant change in error is found.
  + Faster
* Post-pruning
  + Construct the entire by exhausting all the attributes.
  + Prune the trees that overfit.
  + More accurate.

## E

Diagram

Description automatically generated

Tree 1:

Tree 2:

|  |  |  |
| --- | --- | --- |
| N | Tree 1 Cost | Tree 2 Cost |
| 2 | 22 | 30 |
| 4 | 30 | 34 |
| 8 | 38 | 38 |
| 16 | 46 | 42 |
| 32 | 54 | 46 |

According to the MDL principle,

If N < 16:

Tree 1 is better.

If N > 16:

Tree 2 is better.

## F

A picture containing necklet, accessory

Description automatically generated

# Question 2: Density Estimation

## A

## B

To find the MLE of L we maximize the function by taking partial derivative of Pg and set it to 0.

## C

# Question 3: Clustering

## A

Let us consider +30 & -30 as an added point to set for 2 different cases.

For an outlier to be in its own cluster,

For outlier to be in its own cluster the point must be far enough from A, i.e., distance of point B from the centroid of cluster A must be large enough such that it is not in proximity to the points in A.

A picture containing text, sky, line

Description automatically generated

If the outlier is far enough from the centroids of other clusters the centroid of that point will be the point itself given the avg distance of kth neighbors from their respective cluster.

For cluster A,

Therefore, B is an outlier, and k-means will always form a separate cluster for that point. This shows that K-means clustering is not robust with outliers.

# Question 4: Dimension Reduction

## A

λi represents the variance captured by the ith component.

## C

* ISOMAP and Laplacian Eigenmaps are manifold learning algorithms used to discover low-dimensional embeddings of high-dimensional data.
* ISOMAP uses geodesic distances to measure the pairwise distances between data points by computing the shortest path along the manifold that connects two points.
* Laplacian Eigenmaps uses diffusion or similarity distances to measure pairwise distances by measuring the similarity between probability distributions of random walks starting from each data point.
* Both algorithms construct a weighted graph that connects nearby points, use spectral methods to compute low-dimensional embeddings of the data, and preserve local structure in the data.

# Question 5: Naïve Bayes Classifier

## A

Table

Description automatically generated

Using bag of words -> counting the frequency of each word in the set

Therefore, “Credit Card Deal” is NOT SPAM

## B

Promotion is not present in the dataset so considering as Laplacian estimator where

Therefore, “Credit Card Promotion” is NOT SPAM

# Question 6: Association Rules

Table

Description automatically generated

## B

## C

… (a)

This confidence is high, but the words “Credit Interest” is not frequent.

… (b)

Given the superset {card, credit, interest} having low support and its subsets (credit interest -> card) & (interest -> card credit) will have low confidence. Hence, this rule can be pruned.