Extra Credit Assignment

# Question 1: Decision Trees

## A

As the complexity increases, variance increase, bias decreases and higher variance leads to overfitting.

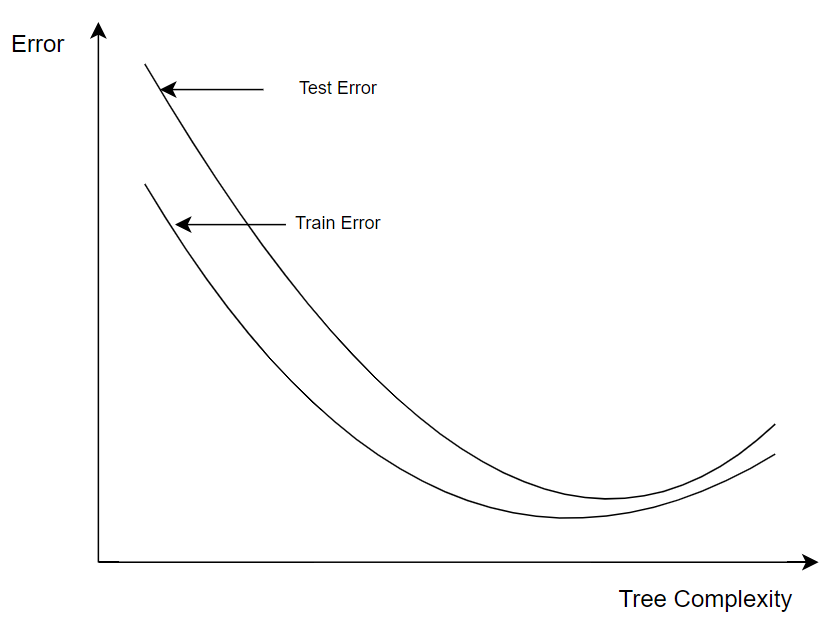
Diagram

Description automatically generated

## B

As in diagram in (A),

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



## 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.

## 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

A picture containing text, sky, line

Description automatically generated

For cluster A,

# 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