Md Hassan

Group 6

CS 4375.004

Similarity and Ensemble

March 25 2023

KNN and Decision Trees

KNN and Decision Trees are different machine learning algorithms that are used in order to train artificial intelligence systems. The process includes training the computer on a given dataset and outputting predictions once trained.

KNN is one of the most popular machine learning algorithms used for classification and regression. After training the data, the user provides an input expecting a prediction based on the input. The algorithm takes the input and compares it to the most similar data point in the dataset. Afterwards, the algorithm computes a prediction based on the label of the nearest neighbors. On the other hand, decision trees separate data into classes. This is done by dividing the dataset into smaller subsets based on the many if-else statements. Moreover, the subsets are divided into smaller subsets which then leads to many different classes. Decision trees can be used for both classification and regression. However, KNN is more suited towards classification problems.

The 3 clustering methods are KMeans clustering, hierarchical clustering and model-based clustering. K Means clustering separates the data into many clusters based on the similarity in the dataset. This is done when the algorithm creates many

centroids and assigns data points that closely relate to a particular centroid. The next clustering method is hierarchical clustering which separates the data by creating a hierarchy. In this method, all the data points start in the same cluster, but it gets divided as it moves on. Model-based clustering uses probability distributions to identify groups in the dataset. It assumes that the data was generated by a model and it tries to fit the different data points in order to get the model.

PCA finds the most significant components in data. After that, it transforms the data to a coordinate system. PCS can be very useful because it is used when operating and visualizing high dimensional data. LDA maximizes the separation between classes in the database by finding the best linear combination of the features of the dataset. LDA is useful when we are trying to reduce the dimensionality of a database while keeping the information needed for classification.