Alekhya Pinnamaneni and Aloksai Choudari

Dr. Mazidi

CS 4375.003

Searching for Similarity

Narrative

Both the knn and decision tree algorithms can be used for regression and classification data sets. Knn algorithms involve predicting values based on the closest neighbors to a data point. The knn algorithm for regression works by predicting the value of a data point based on the values of its closest neighbors. This involves taking the average of the *k* closest training data points to each new data point. The knn algorithm for classification works slightly differently since it is not possible to calculate the average of categories, unlike you can calculate the average of continuous numbers. Instead, knn for classification involves assigning each new data point to the category that contains the majority of the *k* closest training data points. The decision tree algorithm is a greedy algorithm, and therefore is not as accurate as some other algorithms. It involves splitting the data set into smaller and smaller groups until the traits in a group are similar. The decision tree algorithm for regression involves splitting the data into left and right branches at a value *z*, and trying all values in the data to find the optimal solution. The decision tree algorithm for classification uses the number of classes in each region instead to find the optimal solution.

There are three types of clustering: kMeans clustering, hierarchical clustering, and model-based clustering. kMeans clustering works by iteratively assigning each data

point to the closest cluster of the *k* clusters and recalculating the clusters after each assignment. Hierarchical clustering involves a bottom up approach, combining the two closest clusters until all the data points are in one big cluster, forming a dendogram. Model-based clustering works by implementing maximum likelihood and Bayes methods to find the most optimal solution with the most likely model with the most likely number of clusters.

PCA and LDA are dimensionality reduction techniques that aid in cutting down the columns that are used for sampling. PCA, principal components analysis, helps reduce the dimensions of data by reducing the number of axes and creating a new coordinate space for the data. LDA, linear discriminant analysis, is a technique that maximizes class separation, while minimizing in-class deviation. Moreover, LDA is a supervised technique, while PCA is an unsupervised technique. PCA and LDA are used in machine learning because they give data sets more opportunities to be manageable and interpretable. It allows data to not have much risk in losing information or valuable meaning. Overall, both methods make information from data sets more meaningful and give easier interpretation for referability.