Ally de Vera

AMD190012

25 March 2023

CS 4375 Intro to Machine Learning

Searching For Similarity

**A**

To begin a kNN (k Nearest Neighbors) classification, we begin by selecting a k. The most optimal k value is one where an increase in k wouldn’t have a dramatic effect on overall variance, but a decrease would, this value can be seen at the elbow of a graph of variance. Once a k is determined, we randomly assign k centroid cluster points. For each data point in the set, we assign it to a cluster based on which centroid is its nearest point. We repeat this for every data point in the set. Once all data points are assigned to a cluster, we find the mean of each of these clusters and set that as the new centroid point relative to each cluster. We repeat all of this for the new centroids and continue to repeat this until all cluster means are equal to their centroids, meaning there is no shift in the final centroids. After this we have our clusters for this set of initial k’s. We repeat all steps many with k new centroids and whichever set has the least total variance across all clusters is our final classification.

With decision tree classification we go through a series of true/false questions to further refine a grouping for a given an observation. The questions in a decision tree can ask about multiple different attributes across a the observation. Taking an example observation ‘person’ with attributes AGE, GENDER, HEIGHT, WEIGHT. In this, lets say we wanted to classify which weight range (underweight, lean, average, overweight, obese) each person fell into. We begin by refining by gender (i.e. ‘Male? T/F’), from there we go to age (broken down into ~5 year age groups, so ‘<45?’ and further more into greater than ‘<50’ and ‘>40’ – so on and so forth for ages 20-80). After we would follow the same process for height and finally do the same for weight and determine their weight range. For regression we start at our head node and determine what true false split would result in the least variance. For each leaf down to the bottom, we repeat this so that in the end result in a classification with the least variance overall.

**B**

In k-Means clustering we begin by taking k centroids (each centroid being randomly assigned). . For each data point in the set, we assign it to a cluster based on which centroid is its nearest point. We repeat this for every data point in the set. Once all data points are assigned to a cluster, we find the mean of each of these clusters and set that as the new centroid point relative to each cluster. We repeat all of this for the new centroids and continue to repeat this until all cluster means are equal to their centroids, meaning there is no shift in the final centroids. After this we have our clusters for this set of initial k’s. We repeat all steps many with k new centroids and whichever set has the least total variance across all clusters is our final classification. This is very similar to kNN classification but rather than using it for classification of individual points, we use it to cluster multiple points together to pull data from as a cluster.

In hierarchical clustering is different from k-Means in that there is no specified number of clusters prior to running. We first sort each observation into its own cluster (variance = 0, # of clusters = # of observations), then we combine the two closest clusters and repeat until we have one large cluster. After this, we re split our clusters until we are left with k clusters and following our assigned clusters from before. This is not optimal for large data sets, once a point is assigned to a cluster, it remains there. There is no re-assignment as the model learns.

**C**

Principal Component Analysis is a method of reducing dimensionality in a given data set. Rather than making a multi-dimensional graph to represent all attributes, we reduce it down to a 2D plot where points are clustered based on how strongly correlated, they are with each other across their attributes.

Linear Discriminant Analysis also reduces dimensionality by finding a linear combination of attributes that separates the classes while maintaining a low standard deviation.

These is useful in the visualization of data because without this reduction, it would be multiple graphs with different correlations, it reduces its dimensionality to one understandable graph.