3/25/2023

CS 4375.004

Dr. Karen Mazidi

Narrative Document

By:Preston Zacha

KNN or K-Nearest Neighbor can be used on classification and regression. KNN is used for regression by using feature similarity. For example, if you have a data set of the size of a house and you want to find the amount of bathrooms of a certain house. You would look at houses similar to it. So you would say the house in question has the same number of rooms and garages as these other houses. How many bathrooms do these similar houses have? Then you would use that value as your prediction. This is a very simple yet typically very effective way to predict values on a data set. The way KNN works for classification is also by using feature similarity. For classification you would take the values of the similar objects of that column and then the predicted value would be the mode of these values. If there are no repeated values, you would just take the average. Now for finding the similar objects to use you would need to calculate the distance between points, there are a couple of ways to do this. Overall, this is a simple high-level summary of KNN. A couple of things to keep in mind when using this, is that for smaller K values, the bias will go down and variance will go up. For larger K values it's vice versa. Decision trees can also be used for both classification and regression. Decision trees are a greedy algorithm. Overall, they break and continue to break observations down until they have groups of observations that are similar. The main pro and con for decision trees is that they are very interpretable but can be inaccurate at times compared to other algorithms. More specifically for classification each node of the tree is a way to break the data set down into smaller groups. Bringing back the house example, a node of the tree could be “There are more than 3 bathrooms”. This would break down the data set into more similar groups. You would continue to do this until all observations are in similar groups. Now one of the most important things to

consider is what the node should actually be separating the groups by. This is done by using entropy which is essentially the randomness in the data set. Then after each node you can look at the information gain which is how much the entropy decreased or how much the randomness decreased. Decision trees for regression work extremely similarly, they just are able to predict for continuous values instead of discrete values which is what classification is used for.

There is a K-means algorithm for clustering. In this algorithm you would first randomly choose k observations or randomly choose observations to a k group. Then, you would set each observation to the closest centroid. Now, recalculate the centroids. This is the first pass of the algorithm. The second and third step will loop until you can not assign the observation to a closer centroid because of convergence. There is also a hierarchical clustering algorithm. In this algorithm you would first create N clusters where N is equal to the number of observations. Then each observation would be placed in a separate cluster. Next, you need to find the distance between the clusters. Whichever clusters are the closest, you will combine the observations in these clusters into a cluster. This process of finding the closest clusters and merging their observations will loop until there exists only one cluster.

PCA or Principle Component Analysis works in 5 steps. First normalize the variables. Next, create a matrix of size predictors by predictors. Third, calculate eigenvalues and vectors from the covariance matrix. Use these values for finding PC's. Fourth, make a new vector of the features of the observations. Lastly, using your feature vector pick PC's and recast them on the information on the axes. LDA or Linear Discriminant Analysis also works in 5 steps. First, calculate the mean vectors for the classes. Next, calculate in between and within class matrices. Then, calculate eigen values and vectors. Now, use a sort to arrange the vectors by not increasing eigenvalues. Also, select k eigen vectors with the largest values. Lastly, recast the information into a new space. PCA is useful because it greatly increases the interoperability and minimizes information loss. LDA is useful because it reduces dimensionality.