Marcela Pantoja CS 4375.004 3/22/2023

Similarity and Ensemble – Part 5

This document will detail kNN, Decision Trees, kMeans Clustering, Hierarchical Clustering, Model-based clustering, PCA, and LDA.

kNN is a supervised learning algorithm that works by finding the ‘k’ nearest neighbors of an instance. This algorithm groups observations together based on their nearest neighbors, but it does not form a model of the input data. The kNN algorithm also works for both Regression and Classification. kNN works for classification by clustering the data into different classes based on the observations nearest to one another. Due to this, kNN works well on multi-class data sets as it can predict class memberships. However, ‘nearness’ is relative, so it is best to scale the data before performing classification. kNN works for Regression by predicting the data/values of new data points through how close/near the data point is to the points in the training data set. It assigns values based on the values of the data points closest neighbors, and due to this, it is important to scale your data and choose the correct k (number of neighbors) as they will heavily impact your results.

Decision Trees is a supervised, greedy algorithm that runs through the input data and recursively splits it into smaller partitions until the observations are uniform. Like kNN, the Decision Trees algorithm can also be used for both classification or regression, although it is highly interpretable and is overall not the most accurate algorithm. Decision Trees perform regression through seeking to minimize RSS in each separate region, much like how linear regression minimizes RSS overall. This goal is why the top-down greedy approach is chosen, where each split will be made though determining the numeric value for each predictor. The split will then be made and divide the data into 2 regions where the 1st region contains observations less than the split value and the 2nd region contain the remaining values; this process is then repeated recursively until the stopping threshold has been reached. Decision Trees perform classification by following similar steps to the regression method, but instead of measuring RSS per region it instead measures the counts of classes in each region in order to measure the node purity. Splitting on quantitative features will be determined through finding the optimal split point, while qualitative features will either have a natural split, for binary classes, or they will be split in a way where one or more values would go to the left branch and the leftovers will go to the right branch.

There are many different types of clustering algorithms that can be performed, but in this section I will go over three: kMeans, Hierarchical clustering, and model-based clustering. Clustering algorithms in general are unsupervised with the goal of learning more about the input data and finding ‘clusters’ of similar examples/observations.

kMeans clustering is an iterative algorithm where the first step is a random assignment. The manner in which this first step is done is important as there are many different ways to perform it which can cause variations. Two ways to perform a random assignment are: randomly choosing k observations to be the centroids or randomly assigning each of the observations to one of k groups. After random assignment, the 2nd step k-means will perform is assigning every observation to its closest centroid. Step 3 would be recalculating the centroids with regards to the new observation assignments before finally repeating this process (step 2 and 3) until convergence. Hierarchical Clustering is a greedy algorithm that can be performed in a variety of ways, but the bottom-up algorithm works by 1st placing each of the observations into its own cluster. 2nd, calculating the distance between each cluster and every other cluster. The 3rd step is finding which clusters have the shortest distance and combining the two before repeating step 2 and 3 until all the clusters have been merged into one cluster. In the end, this will form a dendrogram which displays the hierarchical nature of the data. Hierarchical clustering tends to work best on small data sets and will bog down with larger sets. Model based clustering works by assuming a variety of different data models and applying both maximum likelihood estimation and Bayes criteria to identify the likeliest model and number of clusters.

Principle Components Analysis (PCA) is an unsupervised feature engineering technique, specifically a dimensionality-reduction technique. PCA is applied to data without any regard for class and works by transforming data into a new coordinate space while reducing the number of axes (principle components). The specific steps by which PCA works are: 1st standardizing the range of initial variables. 2nd, computing the covariance (# of predictor * # of predictors) matrix. 3rd, identifying the principal components by computing the eigenvectors and eigenvalues of the step 2 covariance matrix. 4th, creating a feature vector used to decide which principal components to keep. 5th, recasting the data along the principal component axes. The first principal component is the dimension with the greatest variance and the following principal components are dimensions with steadily decreasing variance. Since PCA is a data-reduction technique, there may be a correspondence between losing data and losing accuracy in a model. Linear Discriminant Analysis (LDA) is a supervised feature engineering (and dimensionality-reduction) technique that, unlike PCA, does have a regard for class when applied. LDA works by finding a linear combination of the different data predictors that will maximize the separation of each of the classes while also minimizing the standard deviation within each class. The specific steps LDA uses are: 1st, it computes the mean vectors for each of the different classes. 2nd, it computes the in-between and within-class scatter matrices. 3rd, it computes the eigenvectors and the eigenvalues. 4th, it sorts the eigenvectors by the decreasing eigenvalues before choosing a k number of eigenvectors with the largest eigenvalues. 5th and lastly, it recasts the data into the new space. After finishing, the LDA can then output the means of all variables by class and the LDA predictions by class can be plotted using the linear discriminants provided by the model. Both PCA and LDA are dimensionality-reducing techniques and as such may be useful for machine learning as they help reduce the number of features in a data set as well as allow us to find the most significant features in a data set. PCA and LDA may also be used for ML as reduction of dimensionality can also lead to a reduction in cost and resources.

References

Mazidi, Karen. “. The Craft of Machine Learning.” *Machine Learning Handbook Using R and Python*, 2nd ed.