Machine Learning Assignment: Similarities

Zachary Canoot, Aarushi Pandey, Brandon Runyon, and Gray Simpson

Intro

In this assignment, our group set out to understand models that look at data based on similarity to training observations. Most commonly, these are things such as K Nearest Neighbors and Decision Trees, but we also have clustering methods. Clustering is one that instead of predicting, tries to find the relationships among unspecified data. We also will be investigating dimensionality reduction and its effect on accuracy.

KNN and Decision Trees: Classification and Regression

Regression:

In kNN, the model looks at the nearest observations and makes predictions based on them, which is called instance-based learning. In the case of regression, the predicted value of a specific observation/row is an average of the neighbors' target value. In this model, the choice of k is important to determine the bias-variance tradeoff. (If k is very small, the classifier will have low bias but high variance.) As k grows, the algorithm becomes less flexible and bias increases while variance decreases. As seen in the Regression notebook, the optimal value for k is often found by cross validation. Before doing this, normalization (applying transformations to the data so that it follows a normal distribution) and scaling (some linear transformation of the data that may or may not result in a normal distribution) are done to ensure optimal performance.

In linear regression, our goal was to minimize RSE over all the data. In decision trees, we want to minimize RSE within each region of the data. In this algorithm, a top-down, greedy approach is used to partition the data. To start, all predictors are examined to see if they make good splits in the data, and for each predictor the numerical value at which to split must be determined. This splitting process is repeated until a stopping threshold is reached.

Classification:

When using kNN for classification, it works very similarly to regression. It still compares the values to the nearest observations to make its prediction, but the way it compares slightly differs. In regression, it would find the mean of the neighbors, but in classification it picks the most common of the neighbors. The classification version of the kNN model retains the same effects of bias and variance as the regression version.

For decision trees in classification, it attempts to split the data into groups that are more uniform. This means that within a group of the resulting tree, it has a higher chance of being the specified class than it would have within the original data set as a whole. This "purity" is often measured by entropy.

Clustering

All types of clustering are trying to take in data, without any given target, and find ways to divide that data into meaningful clusters. There are 3 types that we examined divide a superclass into subclasses in different ways: kMeans Clustering, Hierarchical Clustering, and Model-Based.

kMeans works by creating a k number of points known as centroids. These points correspond to k different clusters that the data may be divided into. Their location in the data is then adjusted after creation until they are in the center of clusters they divide. Each point was used to calculated which data was closest by some given distance calculation, like euclidian distance. The end result was two clusters with tight borders

Hierarchical Clustering works by either dividing the data from the top down or from the bottom up. In one case, the data starts out with each datapoint being a cluster, and then combines every 2 clusters that are the closest together. The end result is a tree that goes up, with greater and greater divisions into how the clusters might divide. On the other hand, with top down hierarchical clustering the data starts out as one cluster and is broken down by which clusters are the farthest apart. The resulting trees can be used and cut at different levels for more and more granular clustering of the data, allowing the analyst to see all the different ways the data can divide.

Model-based clustering works by assuming a various different models that may fit the data, and clusters the data based on the likely hood that each cluster may fit under a probability curve. By minimizing the BIC score, the clusters can be placed under a curve that is most likely to accurately describe that part of the data. The end result are clusters that aren’t bound by border parameters, but able to strongly fit the data.

Principal Component and Linear Discriminant Analysis

Next, we investigated dimensionality reduction on the same data. Since our data was good for clustering, we decided that it would also be a good candidate for dimension reduction and kNN. We were incorrect. Even seeing a poor correlation between data attributes and how popular a given song was, we decided to press on.

PCA, or Principal Component Analysis, works by sorting attributes into ones that can be combined to improve the time it takes for a model to run. The first principal component will be the one of greatest variance, and it decreases from there. LDA, or Linear Discriminant Analysis, is both a dimension reduction technique as well as a machine learning algorithm on its own. LDA also reduces the number of dimensions by trying to find the best way to separate classes via a linear combination of predictors. LDA then focuses on discriminating classes by Bayesian probabilities.

With less data to keep in mind, processes will be faster. PCA aims to make the most effective choices that will minimize accuracy loss. With large amounts of data in a competitive modern age, faster times is often worth a small loss of accuracy so that machine learning is more accessible.

When it came down to the dimension reduction on our data, PCA+kNN and LDA effectively made a coin flip then rated a song as 'bad' or 'horrible'. While the PCA attempt was able to occasionally succeed for the smaller classes, LDA may well have been more accurate due to the fact that it stuck to the larger classes and did not try to sort anything into the smaller classes. PCA nor LDA is able to find or create correlation where there is none.