OCCIDENTAL COLLEGE
COMPUTER SCIENCE DEPARTMENT

Computer Science 349: Machine Learning
Project No. 1: Data Clustering

Dr. Kathryn Leonard

Shasta Clokey

Los, Angeles, CA, September 16, 2017

Description of Data and Research Question

The focus of this project was to inspect variances in multiple distinct features of real and counterfeit bank notes. The two data sets we were supplied consisted of measurements pertaining to banknote authentication and to physical attributes of Swiss banknotes. The data set containing information for banknote authentication was not labeled. It contained four columns of 1372 elements of data used for verifying the authenticity of notes, and one column of 1372 elements stating the true label as zero for real and one for counterfeit. The data set storing information about the physical attributes of swiss banknotes contained six labeled columns each containing two hundred elements; the first hundred of which were real and the second hundred counterfeit. The columns, in order, represented the length, left width, right width, bottom margin, top margin, and diagonal length of the bank notes. Our ultimate goal in this project was to identify if there were specific features in our data sets which could reliably differentiate between real and counterfeit bank notes. We employed the K-Means, K-Nearest-Neighbors, and Naive-Bayes data clustering methods to accomplish this goal.

Feature Selection Rationale

Given the two datasets, it was important to determine which features could potentially differentiate between real and fake data. To make this determination, I scatter plotted every feature of each dataset against every other feature. This allowed me to identify which features provided the clearest distinction between clumps of data. Although I did not label the data at this stage as real or counterfeit, I hypothesised that the features with the most distinction would be the best identifiers of counterfeits when classified by k means, k nearest neighbors, and naive bayes.

Description of Classification Techniques

Each of the three methods we used to determine classification labels involved separating the data into training and test data sets by randomly taking ten percent of the total data as a test set, using the remaining ninety percent of the data as the training data to generate our decision models, then using these classification decision spaces to label our test data as real or counterfeit.

The K-Means label assignment algorithm functions by initializing the locations of k centroids randomly, determining all of the data points closest to each centroid, relocating the centroids to the center of the data points closest to it, and iterating this process until it converges. The generation of these centroids allows for the classification of new “test” data by comparing

the euclidean distances of each centroid from the new data point and assigning the data to whichever centroid is closest.

The K-Nearest-Neighbors label assignment algorithm functions by generating a prediction model using the training data, a set of classifications on that training data, and an integer number of "nearest neighbors". The locations and assignments of the training data points are stored in the model. When the model is used to predict the labels of new data, each data point is identified by the majority vote label of its k nearest data points. This process allows us to vary k in order to find a classification radius which minimizes classification error.

The Naive Bayes algorithm classifies new data by calculating the conditional probability that a data point has a label given the values of the data point's features which are assumed to be mutually exclusive. This conditional probability is calculated by multiplying the probability that each of the features signifies the given label. In this way, we can use the training data to construct our label prediction model, then use this model to classify the rest of our data.

Results

After testing each of the three classification algorithms, it was found that k nearest neighbors worked the best(k = 18), followed by naive bayes, followed by k means(k = 2) in terms of minimizing classification error. I will first discuss the results from the swiss banknote data set, then I will discuss the results pertaining to the banknote authentication data set.

When using the k means classification scheme on the swiss banknote data and specifying k = 2 clusters, the (test, training) errors were (0, 0) when using all of the data in the classification algorithm and (0.0167, 0.05) when using only the bottom margin and diagonal length to generate a classification scheme. When using the k nearest neighbors classification scheme on the swiss banknote data, the (test, training) errors were (0.0056, 0) when using all of the data in the classification algorithm and (0.0111, 0) when using only the bottom margin and diagonal length to generate a classification scheme.When using the naive bayes classification scheme on the swiss banknote data, the (test, training) errors were (0.0056, 0) when using all of the data in the classification algorithm and (0.0056, 0) when using only the bottom margin and diagonal length to generate a classification scheme.

When using the k means classification scheme on the banknote authentication data and specifying k = 2 clusters, the (test, training) errors were (0.3877, 0.3816) when using all of the data in the classification algorithm and (0.3508, 0.3355) when using only the bottom margin and diagonal length to generate a classification scheme. When using the k nearest neighbors classification scheme on the banknote authentication data, the (test, training) errors were (0, 0) when using all of the data in the classification algorithm and (0.0598, 0.0987) when using only the bottom margin and diagonal length to generate a classification scheme. When using the naive bayes classification scheme on the swiss banknote data, the (test, training) errors were

(0.1664, 0.1579) when using all of the data in the classification algorithm and (0.1303, 0.1184) when using only the bottom margin and diagonal length to generate a classification scheme.

Analysis

Given that the probability of error when using each of the three techniques to classify the swiss banknote data are within one percent, I am determining that the swiss banknote dataset is not a good metric to use when determining the effectiveness of k means, k nearest neighbors, and naive bayes. The reason for this indeterminability is that there were simply too few entries in the swiss banknote dataset to represent the typical spread of outliers in a population. Stated simply, it was too easy to differentiate the counterfeit from the real banknotes regardless of the technique. Alternatively, because the banknote authentication dataset included such a large number of entries with a sufficiently small distinction between the two classifications, the difference in efficiency of k means, k nearest neighbors, and naive bayes is easily identifiable. When using the banknote authentication dataset as a metric, the k nearest neighbors classification method was the most effective, followed by naive bayes, making k means the least effective. Given that the error probabilities for k means and naive bayes are within five percent of each other, the two methods should be regarded as comparable to each other with k nearest neighbors only holding a slight advantage at identifying counterfeits. Given that the probability of error is thirty percent higher for the k means classification method when compared to the other two techniques, it is clear that k means is an inferior approach for this dataset.

Future Work

It would be interesting to test these three classification techniques on five to ten more datasets which contained sufficient amounts of data in order to make an accurate determination of which technique works best in different situations. If a counterfeit protection agency was counting on me to give them a determination on real vs fake banknotes, I would research more clustering algorithms, test them on my banknote authentication data, and use the method with the lowest error to make determinations on new data sets of bank notes. Hopefully, this extra level of research and testing would give me me accurate determinations and a bit of job security.