Marielle Billig

Project Progress Report

April 11, 2017

## LINK TO CODE REPO: [**https://github.com/mbillig**](https://github.com/mbillig)

## Introduction:

The primary research question addressed during this segment of the project has been determining if and how the harmonic and timbral features of a song can be used to predict which decade the song is from. To review, this project involves data from the US Billboard Top 100, including song information, year published, and popularity information, as well as data from a dataset describing the harmonic and timbral characteristics of these songs. There are eight harmonic variables and eight timbral variables. Data cleaning and pre-processing was accomplished in the previous progress report period, allowing for the focus to turn to using the harmonic and timbral characteristics to predict the decade.

Several methods were attempted to accomplish this goal. This paper will discuss the details and results of the implementations of Kmeans, Logistic Regression, and SVM.

## KMeans:

Clustering with KMeans was initially very unsuccessful. However, in order to compare different variants on clustering, a silhouette score function was implemented. Silhouette scores range from [-1, 1] with scores near 1 indicating that the tuple is far away from the other clusters, and score near -1 indicates that the tuple may have been classified incorrectly. A silhouette plot can help visualize how big each cluster and the distribution of that cluster’s silhouette score.

When all 16 harmonic and timbral features were included in the model, the average silhouette score was an abysmal 0.147 for 5 clusters. The number of clusters was chosen to be five since there are decades included in the dataset, and it seemed reasonable that there might be 5 clusters. Below, in Figure 1, is the silhouette plot showing these results:

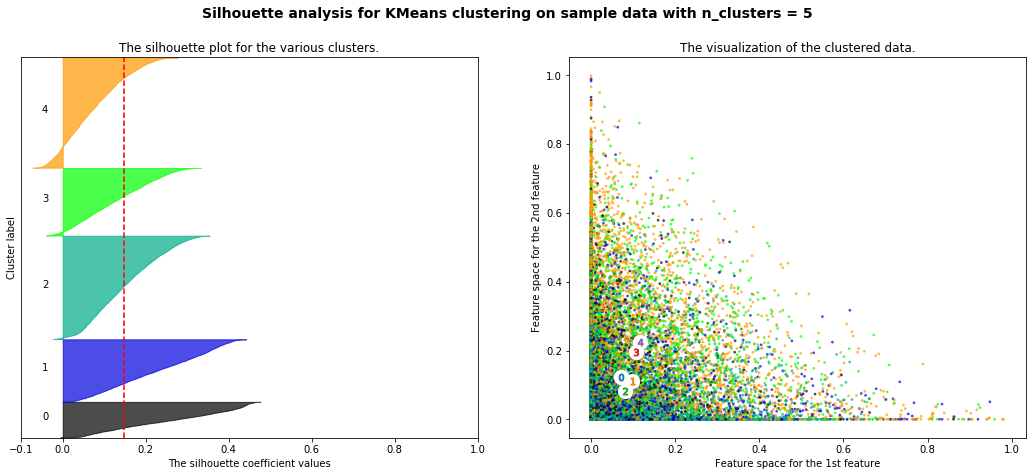


Figure 1 Silhouette Plots for 5 clusters

Since KMeans requires the number of clusters to be given prior, KMeans was also run with between 2 and 7 clusters, but the highest silhouette score of these was only 0.151. Sixteen features seemed like too many, so Principle Component Analysis (PCA) was then run on the data to reduce the data to its 2 principle components. Figure 2 shows the harmonic and timbral data plotted according to the first and second principle components. The distribution is colored based on its true decade. While this plot does not show distinct clusters, it does show an area in the upper left corner that is predominantly 1990 music (yellow).

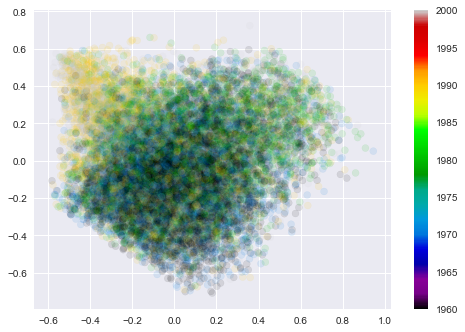


Figure 2 PCA Decomposition

Because of this result, the Kmeans cluster analysis was rerun using the transformed data. This yielded slightly better results with an average silhouette score of 0.355. However, this score is still very low and the visualization of the clustered data does not look particularly promising, so other methods of prediction are being pursued.

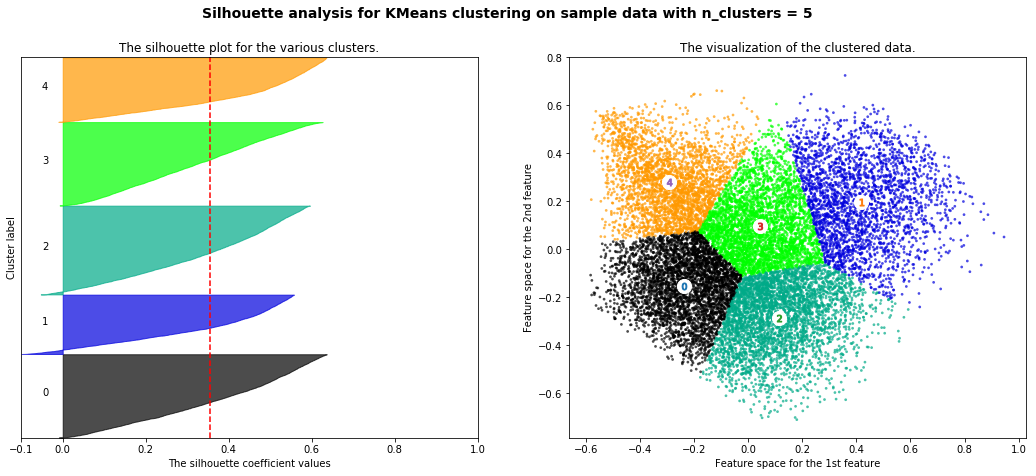


Figure 3 Silhouette Plots for PCA

## Logistic Regression:

For this method, the decade from which a song was from was considered to be a categorical variable. While, an ordinal value might be more accurate, this is considered later.

After partitioning the data into testing and training sets, several different combinations of multi-class approaches and solvers were tested as the inputs for the Logistic Regression function. Logistic Regression models were trained on two of these combinations, but the results were so similar, only the results for the multinomial and newton-cg combination will be discussed specifically. The confusion matrix for the prediction on the training set is shown below in Figure 4. The total number of tuples in the testing partition is 3419, which means that the accuracy of this model is 42.64%

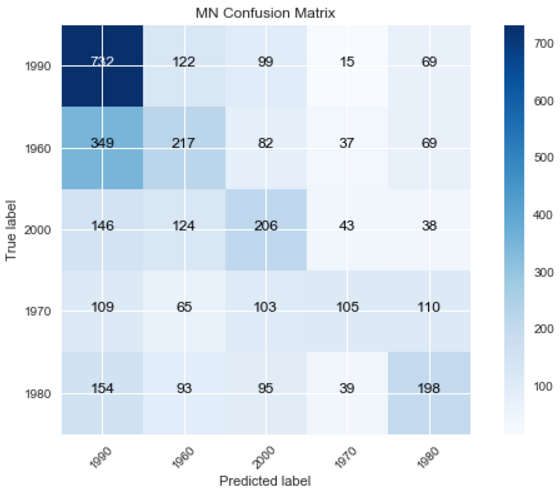
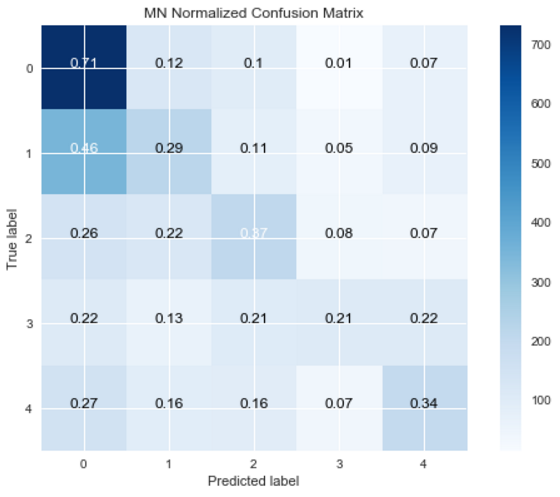
 

Figure 4 Count Confusion Matrix Figure 5 Normalized Confusion Matrix

The confusion matrix, however, shows that a single accuracy measure does not capture the situation well. In fact, the true positive rate for 1990 was 71% whereas the TPR for the other decades was closer to 30%. Figures 4 and 5 indicate that this model is only good at predicting which songs in 1990s. This result may correspond with the results of the PCA plot (Figure 2) since the 1990s was the only decade that had a distinct area on the plot. More research is research to determine what is going on.

## Support Vector Machine:

This model was set up similarly to the logistic regression models. This was just a preliminary investigation so only the linear kernel was used. Below are the unnormalized and normalized confusion matricies for the model.

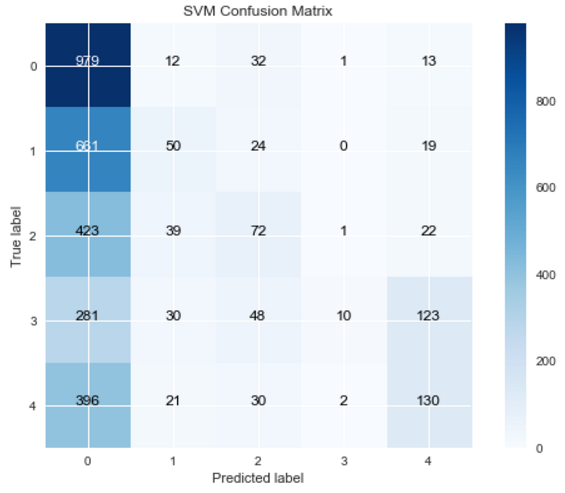
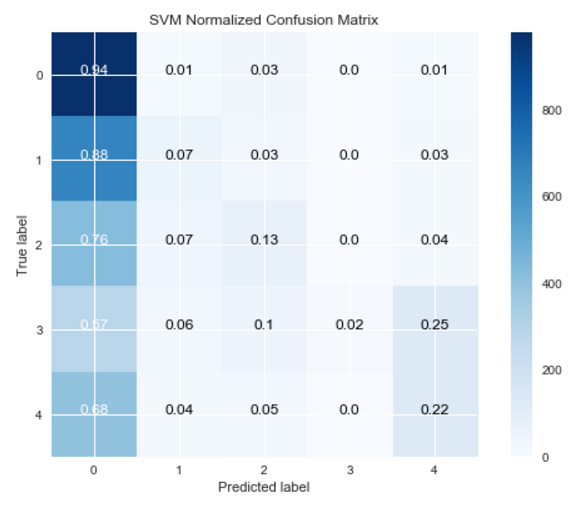
 

Figure 6 Count Confusion Matrix Figure 7 Normalized Confusion Matrix

These confusion matrices show that the SVM model is not very useful because it predominantly just predicts 1990 as the decade, almost regardless of the actual tuple.

## Other Progress:

Besides work on the afore mentioned prediction techniques, progress was also made in the organization of the code. Instead of keep of the code in one notebook, the pre-processing and each of the methods were separated into their own notebooks. This allowed for faster development because the cleaning/pre-processing steps take a long time, and only doing that once and then reading the clean data into the other notebook was more convenient.

Additionally, ordinal logistic regression and linear regression techniques are also in development, but I am still working on how to interpret the results and compare them to the other methods.

## Future Work:

Much future work is still required for this project. So far, only limited results have been seen, so it possible that by fiddling with the inputs better results could be gained. Furthermore, the PCA seems promising, but it would be interesting to do more investigation into which variables it was weighting heavily.

Since there are several variables (there are actually more than the 16 being considered here) a random forest might be another good approach to consider.