Luca Wang

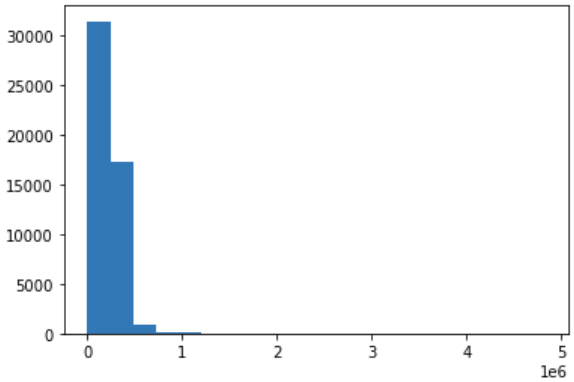
Fundamental Machine Learning

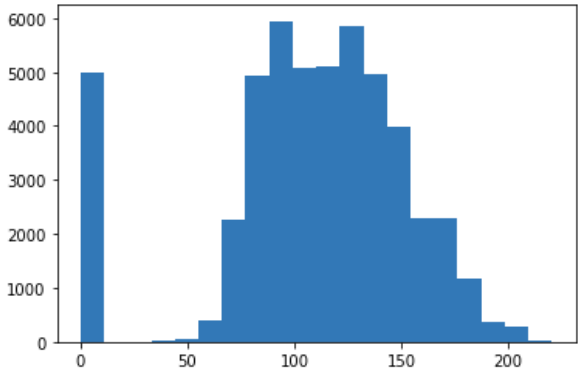
Prof. Pascal Wallisch

Capstone Project Report (Classification)

Data Preprocessing:

First of all, there are 5 rows with nan values in every column. Those lines were directly removed from the dataset. Secondly, the ‘duration_ms’ column has 4939 rows with a value of -1, which should be considered as nan values as well. I plotted the distribution of the column (on the left) and decided to impute nan values with the median value. In addition, the ‘tempo’ column has 4980 ‘?’ values, which should be considered as nan values as well. I plotted the distribution of the column (on the right) and decided to impute nan values with the mean value.



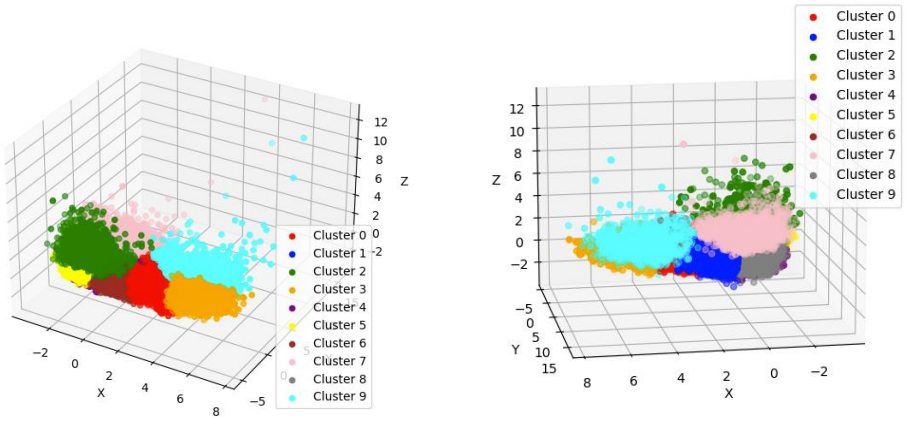


There are also categorical columns in the dataset, including the ‘key’ and the ‘mode’ column. These are columns were hot encoded into columns containing 0 or 1 values.

The 'instance_id', 'artist_name', 'track_name', and 'obtained_date' columns were excluded as they were somehow irrelevant to our problem. Two hot encoded columns, 'key_G#', 'mode_Minor', were also excluded to prevent redundant data. Furthermore, the numerical data were normalized by standard scaler. Finally, the testing dataset (X_test) was established by selecting 500 random samples from each genre, and the rest of the data became the training dataset (X_train).

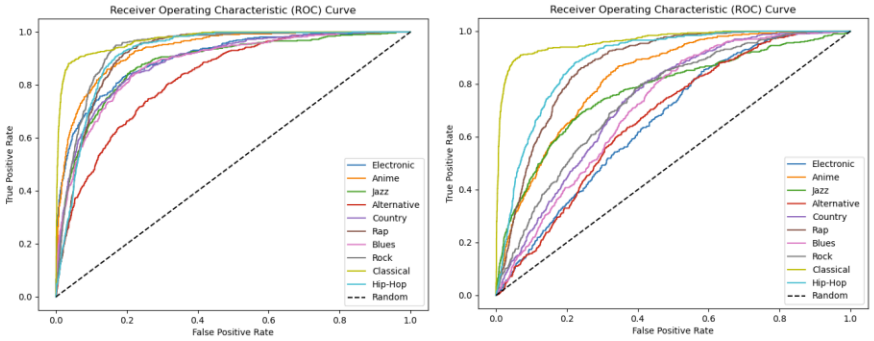
Dimension Reduction and Clustering:

A PCA with n_component of 3 was conducted on both X_train and X_test to reduce the multi-dimensional dataset into its 3 major components. Then I utilized k-means clustering with n_clusters of 10 to group the data into 10 clusters. The 3D visualization of the clustering result is shown below (from 2 different angles). From the plot, we may find out that the clustering done by k-means is clear-cut, and we may observe 10 genre clusters unambiguously.



Classification:

My first approach was a logistic regression model with the original dataset. The final AUC of this model was 0.9102, and the ROC curve of 10 genres is shown below on the left. I also trained the logistic regression model with the dataset after dimension reduction. The final AUC of the second model was 0.7971 and the ROC curve is shown below on the right. In conclusion, the logistic regression on the original dataset does a good job with a relatively high AUC.



Extra Credit:

I trained a two-layer neural network classifier on the original data with ReLU as activation function and CrossEntropyLoss as loss function. The learning rate was 0.001 and epoch size was 20. The AUC is 0.8075. In this case, neural network is doing a worse job than logistic regression.