CS5033 - SL Project Checkpoint

Airi Shimamura Khoi Trinh

1. Introduction

For our Supervised Learning project this semester, we want to create a few models to predict a user's preference for a song on the Spotify streaming platform.

Each song on Spotify has their own set of 11 features. These features are: danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, and tempo.

A user can request their streaming history from Spotify via the P rivacy section of their account. For the purpose of this project, Khoi's one-year streaming history from February 2022 to February 2023 will be used as the 1 class, for songs we like; and about 4200 random songs which Khoi has never heard will be used as the 0 class, for songs we do not like.

The dataset has 56,210 songs in class 1, and 4,170 songs in class 0. However, the songs in class 1 have duplicates, due to the nature of streaming a song multiple times. We expect this number will be much lower once the duplicates are removed.

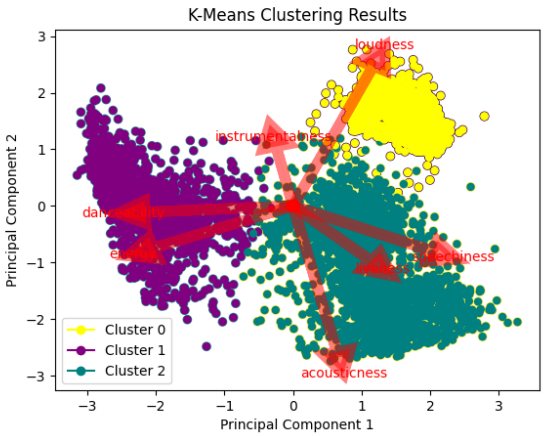

Figure 1. Clustering Results

2. Brief Overview

In this section, some brief overview will be given for the data, as well as some pre-processing and analysis.

One thing we wanted to explore from the data, is how the points are related to one another, through principal component analysis, and clustering. Using k-means clustering, with $k = 3$, we obtained the following figure.

Looking at the figure, we can see that Principal Component 1 separate songs with higher liveness, speechiness, and loudness, and those with lower danceability and energy (these mostly belong to Cluster 0 and 2), and Principal Component 2 separate songs with higher instrumentalness, loudness, and those with lower acousticness (these mostly belong to Cluster 2). Overall, we can see that Cluster 0 has songs that are higher in loudness, speechiness, and liveness. Cluster 1 has songs that are lower in danceability and energy. Cluster 2 has songs high in speechiness, liveness, acousticness, but lower in instrumentalness.

3. Hypotheses

Our first hypothesis is that our supervised learning models will be able to classified these songs as likes or dislikes with at least 90% accuracy.

Our second hypothesis is that out of the four chosen algorithms; Random Forest will give us the best performance; followed by Decision Tree, then Logistic Regression, and finally Naive Bayes. We will evaluate each algorithm's performance by comparing the value of Precision, Recall, and produce ROC curves for each algorithm.

4. Future Work

Khoi will continue refining his logistic regression model, as well as work on the decision tree model. The decision tree model will be useful for random forest, as well.

To enhance the performance of the random forest and Naive Bayes models, Airi will continue re-evaluating them and obtain ROC curves. Moreover, she will expand the random forest model by increasing the number of trees and maximum depth for each tree. In addition, she will experiment with two other criteria, namely entropy and chi-squared, to assess their impact on the model's performance.

Finally, as mentioned in the previous section, the Logistic Regression, Naive Bayes, and Random Forest models will need to be redone, to ensure accurate performance.