COGS 118B: Final Project

Analyzing the Relationship between Song Features and Genre Using Unsupervised Machine Learning

By Bryan, Darren, Rahul, Manan, Nour, and Zakaria

Project Overview

- **Goal**: To determine to what extent the features of a song decide its genre.
- **Motivation**: For artists who create music, knowing this information will allow them to continue creating music that's in a specific genre
- **Related Work**: The most common use of Machine Learning with music revolves around the process of Specific Song Recommendation.
- **Technique**: We reach our goal using three unsupervised Machine Learning algorithms: Principal Component Analysis (PCA), K-Means Clustering, and Gaussian Mixture Model (GMM)

Exploring and Preparing the Data

- *Spotify - All Time Top 2000s Mega Dataset*
- 1994 songs with 14 of their features
  - We focused on 8 of them to help us with clustering
- 149 genres. This is too much....
  - We reduced the number of genres by grouping sub-genres together and only keeping the 10 most popular ones.

Principal Component Analysis (PCA)

- Principal Component Analysis (PCA) focuses on dimensionality reduction.
- In our project, we used PCA once, and used the results from PCA in a few different ways.
- Implementation of PCA.
- Results of PCA.
- We used the results from PCA in 2 different ways.

K-Means

- K-Means is an algorithm that iterates between two steps.
- In the context of our project, we executed K-Means two times: Pre-PCA and Post-PCA.
- Implementation process of K-Means.
- K – Means results (# in clusters)
- K-Means Pre-PCA and Post-PCA results are inaccurate.

K – Means Results

- Analysis of genre percentage in each cluster.

Gaussian Mixture Model (GMM)

- GMM was used to cluster the unlabeled data.
- We opted to execute GMM two times, Pre-PCA and Post-PCA.
- Implementation of GMM.
- GMM results (# in clusters)

Gaussian Mixture Model Results

- Results using the best three features obtained from PCA were inaccurate.
- Results using all eight features looked promising, but…
  - Even though the numbers matched up to the real clustering, looking at the percentage of genres in each clustered showed that no real clustering occurred.

Further Analysis

- One genre per artist

Report and Presentation

- Github and Discord used to facilitate team collaboration and communication.
- Consistent meetings and communication facilitated the working of all team members together throughout all modules of the project.
- Jupyter Notebook and Google Slides used to create Report and Presentation.
- If we had more time with this project, we would have definitely added an extension, or even modified our current project.
- With regards to the goal of our project, we learnt that based on our dataset, the features of a song cannot really control the genre