Project Proposal: Spotify Music Analysis

Alankrit Joshi JOSHI.A@HUSKY.NEU.EDU
Harshdeep Singh SINGH.H@HUSKY.NEU.EDU

1. Problem Description

What problem are we solving? This project aims to mine music attributes data and apply learning algorithms to answer a variety of complex questions like predicting whether a target song was performed live in presence of audience or in a studio with no audience, classifying whether someone would like a song and, lastly, cluster songs based on their similarity. To solve these challenges, we will use supervised learning models to predict and classify and use unsupervised learning techniques to group the data into clusters to determine music similarity.

What are the inputs? The dataset contains processed audio features, for example, Speechiness (A measure from 0.0 to 1.0 that detects the presence of spoken words in a track), Liveness (A measure from 0.0 to 1.0 that detects the presence of an audience in the recording), Acousticness (A confidence measure from 0.0 to 1.0 of whether the track is acoustic) and Valence (A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track). There are a total of 11 such features.

What are the outputs? For continous value predictions using regression, we will use 'Danceability' as target with value from the decimal range [0.000, 1.000]. This is chosen for it is correlatable with a combination of musical elements including tempo, rhythm stability, beat strength and overall regularity, something that our remaining feature set describe in detail. The classification target variable will be primarily 'Likeability' with binary values from the set of $\{0, 1\}$ signifying dislike and like. Finally, we will use the feature set to perform clustering and output the clusters consisting of similar songs.

Why is it interesting? Digitization of every aspect of human activity has led to generation of large amounts of data. Every year millions of users log on to the digital music platform of their choice and contribute to this data, either driven by their own exploration or guided by the recommendation systems of the respective platforms. Music is one of the prominent fields to explore and analyze. Therefore, it is critical to understand the nature of music itself. This would allow deeper understanding of music in analytical terms and act as a precursor exercise to building a great music recommendation system.

2. Data

Which dataset? We will mine data using Spotify's Audio Features API to construct the dataset from around 2000 music records and perform a 60:20:20 split to create training, validation and testing data respectively. A sample record has been shown in Table 1. You can find a bigger sample of this dataset in the references. As mentioned above, the dataset has 11 features.

Why is it appropriate? Spotify is the leading digital music streaming service that gives access to around 40 million songs and this has allowed them to have extensive experience in translating technical audio qualities into normalized human interpretable attributes.

Moreover, the data that they provide is mostly normalized in the range of [0, 1] and can be aggregated in flexible ways, for example, biased distribution of rap and country music.

We are limiting music records from 2017 which should have better Spotify metrics than the recently added records to achieve a consistent and reliable preliminary analysis and categorical segregation.

Song	Acousticness	Liveness	...	Likeability
Stay	0.9	0.1	...	1

Table 1. Sample row with Target variable of Likeability

3. Algorithms

3.1. Exploratory Analysis

We will perform exploratory analysis by cleaning the data, removing outliers and using the music attributes to generate heatmaps and distribution visualizations to analyze correlations. This will help us with feature scaling and feature selection for different models.

3.2. What algorithms to use and how are they appropriate?

3.2.1. Regression

After having a clearer idea about the nature of our features and target, we will add prediction functions using Linear

and Polynomial regression models. This will help us to predict a target value of, for example, Danceability for a given input data point containing other features. We will then begin our first classification task using Logistic Regression on the target variable of Likeability. These techniques are simplistic yet powerful and will aide us in getting approximations of how future models should optimally perform.

3.2.2. SVMs

We will move on to an advanced classification model involving Support Vector Machines (SVMs). We will make models with SVMs using no kernel and then with polynomial and gaussian kernels. We will report the misclassification errors and draw visualizations for the decision boundary and margins in the contour plots. We suspect we need this to better understand the hyperplane separation of data in higher dimensions.

3.2.3. Random Forests

Random forests will help us test out ensemble learning process. We will plot gini indexes in the process and then create our classifier. In the process, we will get a better grip on what features dominate the classification and then use that to contrast against earlier classifications.

3.2.4. k-means Clustering and PCA

To begin to understand the best grouping of similarity based on the music attributes, we will perform clustering analysis. We will utilize k-means to cluster unlabeled data. We will perform this analysis for several values of k, plotting the objective function value for each value of k. We will select the value for k corresponding to the "kink" in the objective function and use this k value for our clustering.

To better understand what are the principal features in the dataset and how they are contributing to similarity of music, we then plan to perform dimensionality reduction to achieve the same. We will start with Principal Component Analysis (PCA) and use the first few principal components to make the plot of the music attributes. These plots will be color-coded based on the similarity type identified through previous k-means cluster analysis. We will then experiment with clustering of other features for varied values of k and plot the results.

Similarly, we will use dimensionality reduction analysis to produce the best human-readable visualization of music similarity using `D3.js`.

3.3. How are these algorithms typically used and how are we using them?

We are using these algorithms with the textbook approach defined in Murphy and with the optimizations we have learnt from assignments. We have described additional details in the respective sections above.

3.4. Have other people used similar algorithms to solve your problem before?

As far as our research goes, we do know Spotify utilizes a mixtures of Collaborative filtering, Audio models and, perhaps, deep learning to work their recommendation systems catered to each user although we could not find any research papers centered around applying these standard machine learning models to this particular dataset. We have attached the resources we were able to find in the references below.

4. Results

What results do you expect to show? The primary results that we are aiming for are to predict certain attributes, classify target labels and correlate different music attributes and find if they can significantly determine music similarity.

We are going to show some visualizations to present the state of the dataset in our exploratory analysis.

For each of the regression techniques, we will report the appropriate SSE/misclassification errors and draw visualizations for them as well as the models' decision boundaries on the datasets.

We aim to do the same for SVMs and contrast performance of different kernels. Experiment results of random forests should provide us information about dominant features and an optimal tree. We are going to plot the same and finally compare all the classification techniques.

The last result involves using the feature set to map out observed patterns or music similarity using unsupervised learning techniques. This should allow us to generate error plots for SSEs vs k and cluster visualization. This visualization should demonstrate the music similarity and trends based on density of clustering

These results will serve to conclude the outcome of the project to predict, classify and determine music similarity.

5. Distribution of Work

- Alankrit Joshi - Linear Regression, SVMs without kernel, Random Forests Analysis, `matplotlib` Visualizations

- Harshdeep Singh - Logistic Regression, SVMs with kernel, Dimensionality Reduction Analysis, k-means Clustering, `D3.js` Visualizations

References

Alankrit Joshi, Harshdeep Singh. Sample spotify analysis dataset. https://gist.github.com/alankritjoshi/118ad8a3b2ac83d10ba44e885b8355c7, 2018. Accessed: 2018-02-26.

Essentia. Essentia documentation. http://essentia.upf.edu/documentation/. Accessed: 2018-02-26.

George McIntire, ODSC. A machine learning deep dive into my spotify data. https://opendatascience.com/blog/a-machine-learning-deep-dive-into-my-spotify-data, 2017. Accessed: 2018-02-26.

Nikhil Sonnad, Quartz. The magic that makes spotifys discover weekly playlists so damn good. https://qz.com/571007/the-magic-that-makes-spotifys-discover-weekly-playlists-so-damn-good, 2015. Accessed: 2018-02-26.

Sander Dieleman. Recommending music on spotify with deep learning. http://benanne.github.io/2014/08/05/spotify-cnns.html, 2014. Accessed: 2018-02-26.

Sophia Ciocca, Hacker Noon. Spotify discover weekly: How machine learning finds your new music. https://hackernoon.com/spotifys-discover-weekly-how-machine-learning-finds-your-new-music-19a41ab76efe, 2017. Accessed: 2018-02-26.

Spotify Developer. Get audio features for a track. https://developer.spotify.com/web-api/get-audio-features/, 2018. Accessed: 2018-02-26.

Tristan Jehan, David DesRoches, The Echo Nest. Analyzer documentation. http://docs.echonest.com.s3-website-us-east-1.amazonaws.com/_static/AnalyzeDocumentation.pdf, 2014. Accessed: 2018-02-26.