Project Preliminary Report

Group Members: Sanat Mouli, Vijay Viswan, Vishaal Bommena

Introduction to Dataset

The problem we are attempting to solve is being able to predict the year a song was releases based on the range of the timbre of the song. The definition of timbre is “the character or quality of a musical sound or voice as distinct from its pitch and intensity” and hence can be used as a good feature for classification. The songs are from years ranging from 1922 - 2011 with differing timbre averages and covariances. Since timbre on its own is a perceived quality of a note rather than a discrete quantity, we will not be able to use it for classification on its own. So we will be using the average perceived value of timbre and the covariance in the value of timbre.

The link to the dataset is: https://www.kaggle.com/uciml/msd-audio-features/version/1#_=_

Using this dataset we will attempt to find the solution to the following questions:

1. The effect on varying values of K and B on the linear svm model
2. Comparison between radial svm and linear svm for the given model
3. Effect of tuning the error term on accuracy.

Preprocessing

The dataset has 1 million rows with 91 columns, but for the scope of this project we will be considering 5 different timbre averages and 5 different timbre covariances. To reduce the amount of data considered by the model, we are only taking 1% of the database which 5153 rows to help narrow the scope of our project. We will be considering this the entire dataset for the scope of this project. We have attempted to use K fold cross validation with a K value of 5 and Bootstrapping with a B value of 30. This will help us narrow the dataset to find the training data that would be beneficial for classification using linear regression. We will be converting the data into a tabular format to feed into the linear SVM model

Model

The models we will be considering are linear SVM and radial SVM. Our decision for selecting these two specific models is influenced by the ease of implementation and compatibility of these models with what we are attempting to accomplish with our data. LinearSVC implements a one-vs-the-rest multi-class support vector machine strategy that scales well with millions of samples and/or features in a linear fashion, with additional flexibility in the choice of penalties and loss functions.

What is left?

While we have made some headway into our project, there are a few tasks that remain.

- We need to test running the model after the preprocessed data from the bootstrapping and k fold cross validation.
- We need to check the effect of hyper parameter tuning on the accuracy of the model.
- We need to make a comparison between linear and radial svm for preprocessed data.