MA678: Midterm Project Proposal

Yudi (Kenny) Mao

Problem Background

Music industry has been transforming from traditional record distribution to modern digital distribution in recent decades. As digital records make success, people are much easier to get, listen and comment on different types, genres of music. Meanwhile, music publishers like Apple, Spotify and Pandora are able to recommend music to their subscribers, based on their appetite and habit, as well as music trend. Thus, how to measure popularity of different music effectively and correctly remains a big problem to music producers and publishers.

Dataset and Method

The Million Song Dataset (MSD), which this analysis project bases on, was created under a grant from the National Science Foundation, project IIS-0713334. It is a free-available collection of audio features and metadata from a million contemporary popular music tracks, provided by The Echo Nest.

From an example track description (*track_id: TRAXLZU12903D05F94*), a list of all the fields associated with each track in the database are introduced. It contains a large amount of numeric data from 46 variables. Several variables will be introduced in the following paragraphs.

This project initially bases on a subset of 10,000 songs of MSD, to explore the relationship between popularity (*song_hotttnesss*) and different metrics including artist, beats, danceability, duration, release year, etc. (Names of variables are omitted here). With the fitted model, which is part of the result of this project, I want to get a deep insight of those most predictive metrics for music popularity.

Furthermore, since music publishers are trying to build a recommendation system not only based on hottest songs, but also on those that share similarity with listening habit of consumers, this project also wants to focus on this topic. Metrics of this topic will be chosen from artist, similar artists, mode, release year, etc.

Analysis Plan

Both the MSD and its subset are available on the labrosa website of Columbia.edu (*https://labrosa.ee.columbia.edu/millionsong/*). Raw data are in HDF5 format, and for convenience of R programming, it should be transformed into csv file format. For data wrangling, basically numerical variables would be paid more attention than strings, and the data would be filtered to different datasets for exploration. Meanwhile, breaking dataset up will decrease file size, thus the analysis will be much easier to handle. After that, it is planned to check several appropriate models, and if possible, I will try to learn some machine learning theory in case some models beyond MA678 content could fit better. A final report should be produced in December, that will include all these procedures, as well as results and discussion part.

Why MSD dataset for midterm project?

There are several key reasons why I choose MSD dataset for my MA678 midterm project, and the first one should always be the strong interests in music and music industry. Music is one of my hobbies for ten years or even more, and the coolest thing is that music apps nowadays can provide music recommendation, identifying songs using machine learning, and distribute diversified new songs conveniently, which was beyond imagination ten years ago. However, the problem is that these functions provided by music publishers are not well developed. I am hoping that the analysis project will at least enhance my understanding of music industry, and if possible, provide some help.