ISyE 6416 – Basic Statistical Methods - Fall 2015
Bonus Project: "Big" Data Analytics
Proposal

Team Member Names: Chen Feng, Mina Georgieva, Tony Yaacoub

Project Title: Music Genre Classification

Problem Statement

Classifying the genre of a song, although an inherently subjective task, comes quite easily to the human ear. Within seconds of hearing a new song one can easily recognize the timbre, distinct instruments, beat, chord progression, lyrics, and genre of the song. For machines on the other hand this is quite a complex and daunting task as the whole "human" experience of listening to a song is transformed into a vector of features about the song. Historically, machines haven't been able to reliably detect many of these musical characteristics that humans recognize in music. Currently, machine learning algorithms haven't been able to surpass the 70% testing accuracy.

The aim of this project is to improve upon the accuracy of genre classification. We are considering a 10-genre classification problem with the following categories: classic pop and rock; classical; dance and electronics; folk; hip-hop; jazz and blues; metal, pop; punk; soul and reggae. The features we will use for classification are timbre, tempo and loudness information.

YouTube, Spotify and similar websites lie behind the motivation for this project. Streaming or broadcasting websites rely on metadata to organize their musical content for easier search and access by the users. A metadata is simply information about the song – album name, artist name, song name, year of publication, genre, etc. While most of the information can easily be extracted from the title of the song, the genre is one of the important features that cannot be easily determined. A lot of the online musical content though lacks this important piece of information. Some websites like Spotify use manual (human) classification of the songs on their website. With the explosion of the musical content online categorizing songs manually can soon become unrealistic. Automatic genre classification would make this process much easier and faster, and it would also improve the quality of the music recommendations. Finally, it will allow for local artists to reach to a greater audience on the web.

Data Source

Our study is based on the Million Song Dataset, which is a freely-available collection of audio features and metadata for a million contemporary popular music tracks. The core of the dataset is the feature analysis and metadata for one million songs, provided by The Echo Nest, a music intelligence and data platform for developers and media companies. The dataset does not include any audio, only the derived features.

A simplified genre dataset is derived from the Million Song Dataset for analysis purposes. The size of this dataset is 59600 by 34 with each row signifying a single song and each column denoting one feature.

The idea is to classify genres based on features. The Echo Nest platform provides too intricate features and genre categories, which makes the task of categorizing songs overly complicated. For example, 'us pop', 'pop', 'indie pop', and 'American pop' could all be merged into one single category, 'pop', which is what we would like to use. In addition to The Echo Nest, there is another open content music database, musicbrainz, which uses features, or tags, that are more appropriate for our purpose. The final genre dataset includes the following 10 different genres: classic pop and rock, folk, dance and electronica, jazz and blues, soul and reggae, punk, metal, classical, pop, hip-hop.

As for features, we use the simple ones from The Echo Nest: loudness, tempo, time signature, key, mode, duration, average and variance of timbre vectors. The genre dataset also provides the artist name and title for each of the songs.

One of the main problems of the Million Song dataset is that the data is unbalanced. For example, the 'classic pop and rock' category has 23,895 tracks, while the 'hip-hop' category only has 434 tracks. For the genres, we rely solely on musicbrainz tags. However, those tags can be wrong or incomplete as they were assigned by humans. These problems can affect testing accuracy. Therefore, we need to pay extra attention during the statistical analysis.

Methodology

Both supervised learning and unsupervised learning algorithms will be applied to our problem. First, we randomly choose 20% of the data points as our testing set; the remaining data will be our training set. We develop our approaches on the training set. For supervised learning, we will use CART (Classification and Regression Tree), random forest, and LDA (Linear Discriminant Analysis) methods; for unsupervised learning, we will use K-means clustering algorithm on the whole dataset instead of the training set only. The performance of each model will evaluated by the following test error statistic:

$$TE = \frac{1}{n}\sum_{i=1}^{n} I\left(Y_i^{test} \neq \hat{f}\left(X_i^{test}\right)\right)$$

where $I\left(Y_i^{test} \neq \hat{f}\left(X_i^{test}\right)\right) = 1$ if $Y_i^{test} \neq \hat{f}\left(X_i^{test}\right)$, and $I\left(Y_i^{test} \neq \hat{f}\left(X_i^{test}\right)\right) = 0$ otherwise, and n is the number of data points in the testing dataset.

Expected Results

We expect to achieve an accuracy of 60-70%, but not much higher. We have to keep in mind that the boundaries separating genre categories are blurry and subjective, and even humans cannot achieve a perfect accuracy rate. Therefore, we expect that clustering algorithms won't be able to fully separate the data due to the overlap. Intuitively, we expect that similar genres are more likely to be conflated with one another. Lastly, some of the algorithms might be prone to overfitting which puts them at a disadvantage. Hopefully, we would also be able to indicate the specific features of each genre category.

One challenge that cannot be tackled in this project is signal processing techniques used to capture the musical features of songs the way humans can hear and recognize them. More advanced signal processing methods will help for a more reliable feature detection and organization.