Classification of Music Genres with eXtreme Gradient Boosting

Jesper Muren*

June 2019

Abstract

The constant growth of the market for music streaming services such as Spotify, Pandora or iTunes is accompanied by a growth in the science of *Music Information Retrieval* or MIR. One big part of MIR is the classification of music, a topic that this thesis will cover. More specifically the classification of songs according to genres will be performed using various characteristics for songs obtained through Spotify. For this thesis the genres we will try to classify are Electro, Hip Hop, Jazz, Pop, R&B and Rock. The purpose of the thesis is to see if we can create a satisfactory classifier for our 6 genres as well as gain insight into how the classifier performs on different genres and with different features. The large quantities of data available when it comes to music has made various machine learning algorithms the natural choice for this task. For this thesis we will limit ourselves to using the relatively new and popular eXtreme Gradient Boosting or XGboosting. XGboosting is a version of gradient boosting which uses an ensemble of decision trees to turn many weak learners into one strong learner. It can be used for classification as well as regression problems and has shown promising performance for both. After tuning the parameters for our XGboost model an overall accuracy of 73.43% was obtained for our classifier, with significantly better performance on Electro, Jazz and Rock compared to R&B and Pop. Important features include for example speechiness for Hip Hop, acousticness for Jazz and danceability for Rock. The result is considered satisfactory when taking into account inherent difficulties in the task at hand as well as the somewhat lacking data set. The largest flaws of the data set is the high percentage of missing data for the lyrics as well as the limited amount of observations.

---

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: jespermuren@gmail.com. Supervisor: Ola Hössjer.