EECS 349: Project Proposal

Eric Hao, Dylan Ong, Max Schuman, Siyu Zhang

April 14, 2016

Task

Our task is to predict the winner of the GRAMMY award for Record Of The Year. There are many factors involved, ranging from popularity of an artist to the genre of the song. A method for predicting whether a record is likely to win this award would not only help artists and producers understand how to work toward winning the award, but also yield insights to subtle differences between popular expectations and actual winners, especially in upset years.

Data

There are numerous song and music databases that are freely available to use, including the Million Song Dataset, which features a vast selection of attributes for each song. Dbopm is a free online database of popular music. We can use these to construct a relevant data set for each GRAMMY season. Also, all previous winners of the award since 1959 are publicly available.

Features

We hope to extract the following features from the web about each song to use in our model, among others: performer(s); year released; gender, age, ethnicity, location, and other relevant demographics of the artist(s); the highest ranking of and duration spent on the Billboard 100 by the record; the number of listens each record received on Spotify, Pandora, and other streaming platforms; the number of downloads of the record on iTunes, Google Music, and other online music stores; the key of the record, and whether it is major or minor; the danceability of the record; the hotttnesss of the artist(s) (a rating from the Million Song Dataset); the duration of the recod; the tempo and time signature of the record; the loudness of the record; and information parsed from the song lyrics of the record.

Initial Approach

We plan to apply Kernel-weighted Regression to deal with our dataset. Since most attributes of music, such as style, genre, may vary a lot after several years, we plan to give different weight to data in different years when we predict the winner of the GRAMMY award 2017. To make accurate prediction, we will weigh recent years results more than the results of years ago. Therefore, instead of using K-Nearest Neighbor, we will use Kernel-weighted Regression to assign different weights to all neighbors. Closer neighbors (recent results) will receive higher weight. And we will use Gaussian-type kernel to convert a distance $d(*,*)$ to a kernel $K(*,*)$.

$$K(\vec{x}, \vec{x_i}) = \exp\left(\frac{-d^2(\vec{x}, \vec{x_i})}{\sigma^2}\right)$$

In kernel regression/classification, nearby points contribute much more to the prediction. A key parameter in defining the Gaussian kernel is σ, also called the width, which determines how quickly the influence of neighbors falls off with distance. We will decide the value of σ by trying different values to make the prediction smoother, as more neighbors weigh in.