PROPOSAL FOR FINAL PROJECT

Introduction to Data Science - esc403spring2019

Spotify Data Analysis

Julian Schmocker

Introduction

The main goal of this project is to find out what properties a song needs to have to be popular. I will try to approach this question from two different angles and work with the two datasets:

Dataset A: UK Number 1 vs. random songs (ca. 2,600 observations of 16 variables)
It contains data from approximately 1,300 songs that were Number 1 songs in the UK (from playlist "Every Official UK Number 1 Ever" from user "Official Charts") and 1,300 random songs.

For Dataset A the aim is to detect the main differences between the two groups.

Dataset B: 10,000 random songs (ca. 10,000 observations of ca. 200 variables)

Here I will focus on the response variable *popularity*, which describes the popularity of the song on Spotify. The dataset B is used to find associations between *popularity* and other variables. Additionally, I will try to predict for given song-attributes the value for *popularity*.

Data

The data used in this project is acquired from the Spotify API by the python library *Spotipy.* The following requests are applied to gather data:

audio_features()
Returns for one or multiple tracks values for variables like *danceability, energy* and *valence*. For details see: https://developer.spotify.com/documentation/web-api/reference/tracks/
Here there is no additional processing necessary. This request is used to acquire dataset A and B.

audio_analysis()
Returns the advanced audio analysis of a track. The Audio Analysis describes the track's structure and musical content, including rhythm, pitch, and timbre. One variable is associated to several values (one value for each section / each segment). As the number of segments and sections vary for each song, I need to summarize these values (by mean, variance or covariance between two variables). I will compute the values:

- Percentage of key changes in a song
- Means, variances and covariances of the pitches of a song
  (I still have to check if I have to shift the pitches in a way that the pitch 0 represents the fundamental tone (not C as per default) to compare the values of the tone pitches)
- Means and variances of timbres of a song

This request is used only for dataset B.

It is not possible to pick completely random songs from the Spotify catalogue. To get almost random tracks I first generate a random word (from a dictionary https://github.com/dwyl/english-words). Then I search with spotipy.search() for tracks which have this word in their name. Then one of these songs gets chosen by a random offset.

Methods

To get an overview, I will first conduct an explanatory data analysis on both of the datasets A and B. I will compute summaries, create scatterplots, histograms, boxplots and a PCA.

To detect differences between the songs which were Number 1 and the random songs (dataset A) I will apply hypothesis testing (T-test, Chi-square-test), a logistic regression and decision trees.

Then I will try to find the best classifier for the two groups in dataset A. Possible methods are:

- Logistic regression
- K-nearest-neighbors
- Support Vector Machines
- Boosting

I will conduct a linear regression to examine if there is an association between *popularity* and other variables in dataset B. Finally, I will try to find the predictor which predicts best the popularity of a song (with cross validation). Possible methods are:

- Linear Regression
- Lasso
- Neural Networks