Machine Learning Engineer Nanodegree
Capstone Proposal

João Pedro Megid Carrilho
February 17st, 2018

Proposal

Domain Background

Machine Learning is one of the most popular subjects being talked these days, supervised learning, unsupervised learning, deep learning, between others are going viral around the globe and from this comes the interest from people to start entering this world of algorithms.

This proposal seeks to analyze Kaggle's "Spotify Song Attributes" dataset using supervised learning methods to create the model that can predict the best if a particular user may or not like a specific song, gathering the results of each and comparing them. This models are being used in various areas, this book mentions this concept of the proposal at section 2.8. Companies suggest other products/services a user may be interested in, increasing user's immersion and profits.

Problem Statement

Selling a service or a product (in this case music subscription) and keeping up with the competition is harder each day but, when you can make good use of each data and feedback a user gives you, the use of good tools like machine learning can help you win the clients' money and fidelity. With the use of machine learning methods, it's possible to find patterns in songs attributes that the user hears and likes (or not), from this patterns new songs can be suggested to the user and, if he keeps liking them, he will probably keep using your service. Thus, our interest here is finding a model that can predict, with a good mean accuracy, the songs a user may or not like.

Datasets and Inputs

This dataset contains 2017 musics/samples (rows) from a single user, each of this samples has 16 columns of which thirteen are music attributes, one column for the songs name (a string), one for the artist name (also a string) and the last, as our target column, if the user likes or not this song, a binary classification. The attributes of each song are the following: acousticness, danceability, duration_ms, energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, time_signature and valence, all these are continuous numerical values. This dataset contains 1020 songs with the "like" label and 997 songs with "don't like" (totalizing the 2017 samples).

For a deeper look into what each attribute means, this link at the Spotify's website.

Solution Statement

The data is going to be explored, analyzed and pre-processed, train and test set will be defined and the supervised learning models LinearSVC, KNeighbors, Ensemble Methods and SVC will be fitted, tuned and the best estimator (the one with the highest mean accuracy of predictions), will be stored.

Benchmark Model

For use of supervised learning model, on the same dataset, this Kaggle submit used Decision Tree and Random Forest algorithms to make predictions, resulting in 0.6998 and 0.7295 of accuracy (the number of predictions made right divided by the total number of predictions) respectively. Also, this blog page, made by the user who uploaded the dataset (GeorgeMcIntire), talks about his own predictive model to this problem and his step-by-step to solve it.

Evaluation Metrics

The evaluation metric used in this project is the mean of the accuracy of predictions made by the Grid Search CV best estimator:

$$Mean\ Accuracy\ = \frac{1}{n} * \sum_{0}^{i} \frac{True\ Positives_{\ i} + True\ Negatives_{\ i}}{Total\ fold\ samples\ tested}$$

Where “n” is the number of cross-validation folds, $true\ positives_{\ i}$ are the music that was correctly predicted, in the “i” fold, as 1(one) or “being liked” and $true\ negatives_{\ i}$ are the music correctly, also predicted in the “i” fold, as 0(zero) or “not liked”.

This metric can be used for evaluation since we have a balanced category of target labels.

Project Design

This project is mainly divided into three parts, data exploration and preparation, training and evaluating the models, choosing the best model. report the final conclusion.

In the first part, the data will be explored to obtain some main statistical description of the data set, feature relevance and distribution will be analyzed, feature scaling and outliers detections are used to provide a more normal distribution. PCA is applied to this altered data and, if possible, dimensionality reduction. Train and test data are going to be defined.

Supervised learning models LinearSVC, KNeighbors, Ensemble Methods and SVC are going to be trained and fitted, if possible, Grid search cross-validation will be applied to the models for choosing the hyper-parameters that give the best accuracy and providing more confidence in the results.The model with the best accuracy will be selected.

The final conclusion will show why the selected model works and should be used, a visual and numeric analysis will confirm this statement.