SpotyPY

SAPIENZA
UNIVERSITÀ DI ROMA

Presentation of our work for the STAT4ACSAI 21-22 course's project

Yusupha Yuwara
Benjamin Barda
Paco Danese

SAPIENZA
UNIVERSITÀ DI ROMA

Overview

- For this project we decided to dive deep into some of the many models and tools seen during the duration of the course.
- We decided that, given our shared passion for music, focusing on this field would be the perfect match of pleasure and “duty”.
- During this brief presentation we present both the challenges and the result we had, in addition to future additions.
- For the full repository of this project visit our public github repository.

SAPIENZA
UNIVERSITÀ DI ROMA

The process

- We could divide the journey of this project into 3 main phases :
  - Data gathering
  - Implementation from scratch
  - Exploring the results

Data gathering

- It was always clear to us that trying to analyse the whole spectrum of existing genres was going to be too big of a task for our limited resources and we were fearing that nothing meaningful would stem out of this approach.

- For the reason mentioned above we decided to limit ourselves to four genres, that in our opinion have both similarities and differences with respect to one another :
  - Jazz
  - Metal
  - Rap
  - Classical

SAPIENZA UNIVERSITÀ DI ROMA

Data gathering

- Once defined the scope of the project we started gathering data. Our first idea was to use the Librosa package in order to extract features from audio files. This proved to be unfeasible indeed given that downloading huge amount of audio-files, and label them, was an effort out of reach for us.

- As an alternative we used the Spotify API, which provide among many things, precomputed features for each song, making the data gathering process faster. Of course we had accept a trade-off given the fact that the Librosa analysis is much more exhaustive than the features provided by Spotify.

Data gathering

- On top of being simpler to store and retrieve, the Spotify precomputed features are “human understandable” and a more detailed explanation is provided in the Spotify developer guide. The ones we considered were :

  - Mood related : Danceability, Valence, Energy, Tempo
  - Loudness, Speechiness, Instrumentalness
  - Context related: Liveness, Acousticness

Data gathering

- In order to build the dataset we downloaded the features appearing in the top Playlists for each one of the genres allowing us to quickly have labelled data.

- After grouping them and cleaning the data frames from missing values and duplicates, we found ourselves with around 1300 datapoints. Which we considered a size good enough to be able to process it with ease on our modest machines.

SAPIENZA
UNIVERSITÀ DI ROMA

Implementation

- Once data has been gathered we decided that it was going to be academically beneficial to code our own models from scratch, trying to avoid as much as possible to use pre-computed modules.

- To make it a fair mix of parametric vs non-parametric, classifications vs clustering algorithm we decided to implement the following.
  - Random forests
  - K-means (++)
  - MoG (EM)
  - GNB
  - Linear and logistic regression models.

SAPIENZA
UNIVERSITÀ DI ROMA

Implementation

- We will not dive deep into the math of each model since this is presented both in the notebooks as well as in the report.

- We think that would be more interesting for the audience to look at the results instead and provide a comparison on the performance of the different models

SAPIENZA
UNIVERSITÀ DI ROMA

Implementation

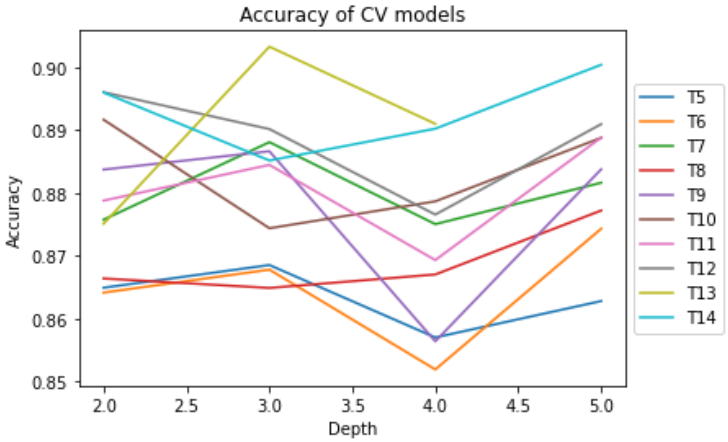
- Decision trees are a well knows tool used for classification as well as for regression.
- Our implementation of the decision Tree uses the Information gain as the objective function.
- We trained the Tree on a 75/25 split for training and evaluation giving us on average an accuracy of 86%.
- The main purpose was for this implantation to be the base of our random forest.

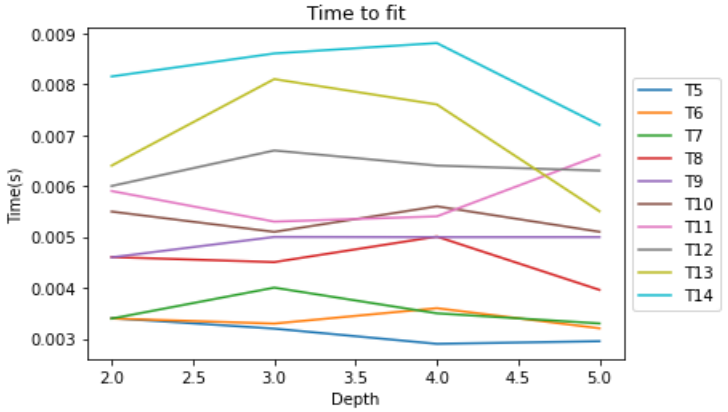
SAPIENZA
UNIVERSITÀ DI ROMA

Implementation

- Having implemented the random forest we decided that it was the perfect opportunity to apply some model selection techniques.

- We performed a 10-fold cross validation tuning both the maximum depth of the trees as well as their number in grid search approach.





Scuola di Specializzazione in Malattie dell'Apparato Cardiovascolare
Direttore Prof. Massimo Volpe
Facoltà di Medicina e Psicologia, Università di Roma Sapienza
Anno Accademico 2013-2014

Dr.ssa/Dr. Nome e Cognome

Grazie per la Vostra Attenzione!

SAPIENZA
UNIVERSITÀ DI ROMA

Progetto Formazione Avanzata in Cardiologia nel Web 2014
Scuola di Specializzazione in Malattie dell'Apparato Cardiovascolare

Direttore: Prof. Massimo Volpe
E-mail: massimo.volpe@uniroma1.it

Coordinatore: Dr. Giuliano Tocci
E-mail: giuliano.tocci@uniroma1.it