SDATA PROJECT

18/01/2021

Supervised classification, regression and data analysis

In this project, we will use a dataset generated using the Spotify API, it contains the following extracted and gathered features for many songs:

- Primary:
  - id (Id of track generated by Spotify)
- Numerical:
  - acousticness (Ranges from 0 to 1)
  - danceability (Ranges from 0 to 1)
  - energy (Ranges from 0 to 1)
  - duration_ms (Integer typically ranging from 200k to 300k)
  - instrumentalness (Ranges from 0 to 1)
  - valence (Ranges from 0 to 1)
  - popularity (Ranges from 0 to 100)
  - tempo (Float typically ranging from 50 to 150)
  - liveness (Ranges from 0 to 1)
  - loudness (Float typically ranging from -60 to 0)
  - speechiness (Ranges from 0 to 1)
  - year (Ranges from 1921 to 2020)
- Binary:
  - mode (0 = Minor, 1 = Major)
  - explicit (0 = No explicit content, 1 = Explicit content)
- Categorical:
  - key (All keys on octave encoded as values ranging from 0 to 11, starting on C as 0, C# as 1 and so on)
  - artists (List of artists mentioned)
  - release_date (Date of release mostly in yyyy-mm-dd format, however precision of date may vary)
  - name (Name of the song)
  - genres (Musical genres)

Exercice 1 : Classification challenge

In this part use : *Spotify_train_dataset.csv* and *Spotify_test_dataset.csv*.
The csv file, for training, contains a dataset of 31728 songs belonging to one of those 15 musical genre :
[Dark Trap, Underground Rap, Trap Metal, Emo, Rap, RnB, Pop, Hiphop, techhouse, techno, trance, psytrance, trap, dnb, hardstyle].

1. Analyze the data.
2. Train some classifiers, analyse results and make conclusion.
3. Using your best classifier predict the musical genre of the 10577 songs of test dataset in order to compete for the challenge!

Exercice 2 : Data analysis

In practice the musical genre in spotify is not that well filled. We are therefore going to work on another dataset in *Spotify_exo2.csv*.

1. Try to predict the "popularity" of a song.
2. How can you handle the class "genre"?
3. What can you show from these data? Patterns, visualization, interpretation...

Instructions : You have to drop 3 files (file names including your names separated with "_"):

- One **.csv** file containing your prediction for the challenge, such as:
  - the *Spotify_test_dataset.csv* completed with a new column named "genre" or a new file containing only one column named "genre",
- One **.pdf** file containing your report,
- One **.zip** compressed archived with your code (*.py or *.ipynb).