Final Capstone Proposal – 2021-05-06

Stu Alden, Thinkful Data Science

I propose to refine machine-learning algorithms for assigning “genre” to musical recordings. This will build nicely on the tools and techniques I’ve learned in the three preliminary capstones.

Rationale, Purpose:

* I have a large music library built from CDs I have purchased over the years. Genres (e.g., jazz, blues, folk) are needed to classify and organize the music so it is conveniently retrievable.
* I (or anyone with a large digital music collection) would benefit from a tool that could assign genres to new music and checks existing genre assignments for validity and consistency.

Dataset:

* I plan to use music from my personal collection of MP3 files, which, over the years, I have built by ripping from my CDs using EAC and encoded with LAME.
* Initially, I plan to work with all music in my library, with the exception of the music currently classified as Rock or Classical (both of which present unique problems which I will try to solve later, after working on the other genres).
* Unlike freely-available datasets on the web, this music dataset will be very large and it will contain entire tracks, not just fragments.

Feature Development:

* I plan to use the **LibROSA** library of audio analysis functions for Python to do most of my feature extraction from the music waveforms.
* Single-thread processing time is roughly 2 songs per minute, so I will need parallelize this process.

Analysis & Tools:

* Starting with the “raw” features, I will undertake an iterative process of EDA, feature selection, and dimensionality reduction, along with a variety of supervised learning algorithms that will be tuned using GridSearchCV.
  + Supervised learning at this stage will allow me to see which features are most important, as well as provide me with a model that can be used to “predict” the genre of new music as I acquire it.
* After I’ve settled on an optimal set of features and an optimized supervised algorithm, I will proceed to investigate several clustering methods on the feature set.
  + Since I have the “ground truth” genres for all tracks, I will be able evaluate the clustering using extrinsic measures (e.g., ARI) as well as intrinsic measures (e.g., Silhouette Coefficient).
  + From this investigation, I will determine the optimal clustering method for my genre-setting task.

Deliverables:

* I will deliver the project as a small set of Jupyter Notebooks. The summary notebook will be organized (using the RISE notebook extension) so that it can be used directly for my presentation as well.
* I will include all code and visualizations used in the project, along with the documentation of any papers or other persons’ code that I may utilize in my work.