Machine Learning with Networks ECEN765.600, Fall 2017

Class Project: Initial Proposal

"On my honor, as an Aggie, I have neither given nor received unauthorized aid on this academic work."

Signature: Joseph Riad

Name: Joseph Riad

1 Introduction and Motivation

Over the years, the music we've been making has proliferated into a very wide variety of genres and styles. One website [1] lists about 380 different sub-genres of music. The task of classifying a given song into one or more of these sub-genres may be daunting, mainly due to the inevitably subjective nature of such a classification, but it can be beneficial in a number of applications. It can be used by individuals to organize their media library or select what music to listen to depending on their mood; it can also be used by content providers to cluster together similar songs and thus be able to recommend music to their users based on their individual taste [2].

The most common way to automatically classify music seems to be based on audio signals with features extracted from the time- and/or frequency-domain representation [2, 3]. The approach I propose to take is a bit different: to classify the songs based on their lyrics. I think this is an interesting approach for several reasons:

1. It is simpler than the signal processing approach and requires less domain knowledge
2. It provides a different kind of insight into the music
3. It can generalize to other fields like classifying a book based on its content

Obvious limitations to this approach are the possible loss of accuracy that stems from using a simpler feature set and the fact that it can only classify vocal music (songs) as opposed to instrumental music. Such an approach has been attempted before [4] but the end goal was to classify the music into "happy" and "sad" categories, while the aim here is to classify the music into more specific categories like "folk rock", "pop", "speed metal", ...

2 Deliverables

By the end of this project, I aim to deliver an algorithm that, given a song's title, performing artist(s) and lyrics will output up to four sub-categories of music that the song most likely belongs to. As a by-product of its operation, the algorithm should provide insight into the different types of music by enumerating the different topics tackled by different musical genres.

3 Methodology

The following is a rough breakdown of the approach I intend to take:

3.1 Building the dataset

- The project was inspired by a dataset on Kaggle [5] containing the lyrics for over half a million songs along with the names of the artist(s) for each track.

- I will need to do a little scraping to label each song with some genres. For that purpose, I intend to use the musicbrainz database (https://musicbrainz.org) which exposes a useful searching API with a Python binding [6].
- The musicbrainz database provides "tags" for each song showing the categories it has been classified into.
- Any song which is not found on the musicbrainz database or has no tags on the database will be discarded.
- The tags will probably need some normalization (for example, the categories "alt-rock" and "alternative rock" are the same) so this step should follow.
- I will probably need to take a subset of the whole available data depending on memory availability. This step may need to be revisited later after the algorithm to use has been chosen and its complexity explored.
- The resulting list of chosen data points should be divided into training, validation and testing datasets.

3.2 Exploratory data analysis

I will need to perform some exploration to gain a feel for the dataset. Specifically, I will consider the following:

- Explore the top 10 words in each tag and the relative frequency of each of these words in songs tagged with that tag.
- Repeat the above exercise with 2-grams instead of words (1-grams).
- Explore if any topical patterns emerge (a certain topic or group of topics relate to a specific genre).

3.3 Feature extraction

First, the text data will need to be vectorized (represented as numerical vectors). For this step, depending on observations from the previous step, I will use either a bag-of-words (1-gram) approach or will attempt an n-gram approach. It may also be useful to apply term frequency-inverse document frequency (tfidf) to the data to emphasize only words (or n-grams) that are more representative of each genre.

Second, different numbers of features may be attempted to select a good compromise between complexity and accuracy. I will probably need to use some dimensionality reduction technique like PCA to remove irrelevant features.

3.4 Choice of algorithm

This step requires more reading of the available literature (textbooks and papers) on different classifier algorithms to choose one or two of the most suitable ones, use them to classify the songs and compare their respective performance. Some sort of Bayesian approach may need to be used as the end goal is a soft decision: a level of confidence for each tag being assigned to the given song. This approach enables the selection of multiple tags for the same song. One algorithm that seems to be promising is latent semantic analysis (LSA).

3.5 Validation and testing

The parameters of the chosen algorithm should be adjusted with the help of the validation dataset and its generalization error should be measured using the testing dataset.

4 Resources

Most of the resources I plan on using have been explained in the previous section. I summarize them again here:

4.1 Technical resources

These are resources needed for understanding the algorithm(s) to be used and any limitations and/or performance guarantees they might have.

- Course textbook(s)
- Recent papers
- Online tutorials

4.2 Computational resources

- `numpy` and `scipy` for scientific computation
- `scikit-learn` for algorithm implementation
- `pandas` for data manipulation
- `matplotlib` for data visualization
- `musicbrainzngs` for data scraping

5 Milestones

I am planning three milestones:

1. Dataset completed, started on exploratory data analysis (should be done in time for the midterm report)
2. Feature extraction and algorithm selection completed
3. Validation and testing done, conclusions to be drawn

References

[1] Music Genres List. The Most Definitive Music Genre List on the Web. `http://www.musicgenreslist.com/`. Accessed on 09-28-2017.

[2] Jonathan Baker. Methods of Music Classification and Transcription. Master's thesis, Brigham Young University, 2012.

[3] Changsheng Xu, N. C. Maddage, and Xi Shao. Automatic Music Classification and Summarization. *IEEE Transactions on Speech and Audio Processing*, 13(3):441–450, May 2005.

[4] Sebastian Raschka (rasbt). MusicMood. `https://github.com/rasbt/musicmood/blob/master/README.md`. Accessed on 09-28-2017.

[5] Soumitra Agarwal. Every song you have heard (almost)! `https://www.kaggle.com/artimous/every-song-you-have-heard-almost`. Accessed on 09-28-2017.

[6] Alastair Porter et al. musicbrainzngs 0.6 Documentation. `http://python-musicbrainzngs.readthedocs.io/en/v0.6/api/`. Accessed on 09-28-2017.