Can Song Lyrics Predict Genre?

Danny Diekroeger
Stanford University
danny1@stanford.edu

1. Motivation and Goal

Music has long been a way for people to express their emotions. And because we all have a wide variety of emotions, music comes in all types of styles. My personal iTunes library includes plenty of these different styles, ranging from Bob Marley's peaceful melodies to the smooth rap songs of Kanye West, to Beethoven's timeless Symphony Number 5. For my project, I was originally interested in developing a method to automatically classify music tracks into their different styles. This task, which has yet to be perfected, has direct real-world applicability. In a world where music is so readily accessible through the Internet, the ability to automatically classify music based solely on its content is becoming quite desirable. Automatic classification is particularly applicable to the task of suggesting songs for users, a task handled by programs such as Pandora Radio and Apple's iTunes Genius. But even both of these programs have not implemented a way to classify songs solely on content: Pandora determines music styles by having experts manually place "tags" on songs, while the iTunes Genius compiles data based on the contents of a user's entire library. Neither of these methods automatically classifies a song based on just the song itself. And that is exactly what I have tried to do.

In the attempt to classify a song's style simply by it's content, where does one start? The first step is to explore the smaller parts that make up the song. A song has two main components: instruments and vocals. And within each of these components are many different variations to consider: different instruments produce different sounds, and almost all instruments play many different combinations of notes and harmonies. Similarly vocals may vary in pitch, gender of singer, and lyrics. And on top of all this, a song may be fast or slow, and it may fall into certain patterns of chorus and verse. And it is all of these factors weaved together in an artistic manner that gives a song its own unique style.

In pursuit of my goal to classify music tracks, my initial plan was to represent as many of these factors as I could. I figured I would go straight to the contents of the audio file itself and conduct hardcore analysis using advanced mathematical methods such as

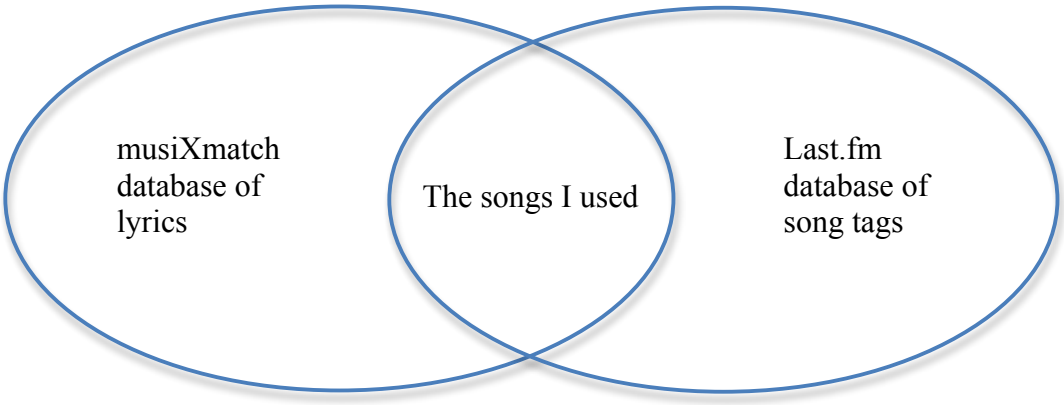
Fourier transforms to extract the information for my features. But after some effort, I found this process of audio analysis to be significantly beyond my current technical skill level, and I decided that pursuing this fully would simply take too much time for a single quarter-long project. Disappointed but not deterred, I instead decided to settle for something a little more manageable. I decided to see if a song's genre could be predicted solely by its lyrics.

2. Data

The first step in my project was to gather all the necessary data. This turned out be a challenging task, and a great learning process. The teaching staff thankfully pointed me towards the Million Song Dataset (MSD), a freely available collection of audio features and metadata for a million contemporary songs. In addition to the data they provide, MSD also links to other sibling datasets that contain more information. The first sibling dataset, from a site called musiXmatch, conveniently provides access to the lyrics for over 230,000 of the MSD tracks. The second dataset was from a site called Last.fm. It provides song "tags" for over 900,000 of the MSD tracks. Since my goal was to use lyrics to predict genre, I needed to combine the lyrical data from musiXmatch with the tags from Last.fm to compile my desired dataset.

One obstacle I faced was to extract genres from the Last.fm tags. Each track in this dataset is paired with several tags, which have been given over time by listeners on their site. The problem is that these tags were created by listeners, so they are not exactly equivalent to the genres that I wished to study. Some tags did refer to exact genres, such as "rock," "pop," and "alternative." But other tags were much more obscure, such as "subdued electronica" and "kinda sad." However upon further investigation, I found that the top five most popular tags were indeed relevant musical genres. Specifically they were rock, pop, alternative, indie, and electronic. So I decided to extract only the tracks that had been tagged with one of these five genres and ignore the rest. This left me with 320,452 tracks, each of which was labeled as one of the five genres.

At this point I had two datasets: one that paired a song with its lyrics, and one that paired a song with one of five genres. The next challenge was to combine these datasets, or rather, find and extract the tracks that appeared in both.



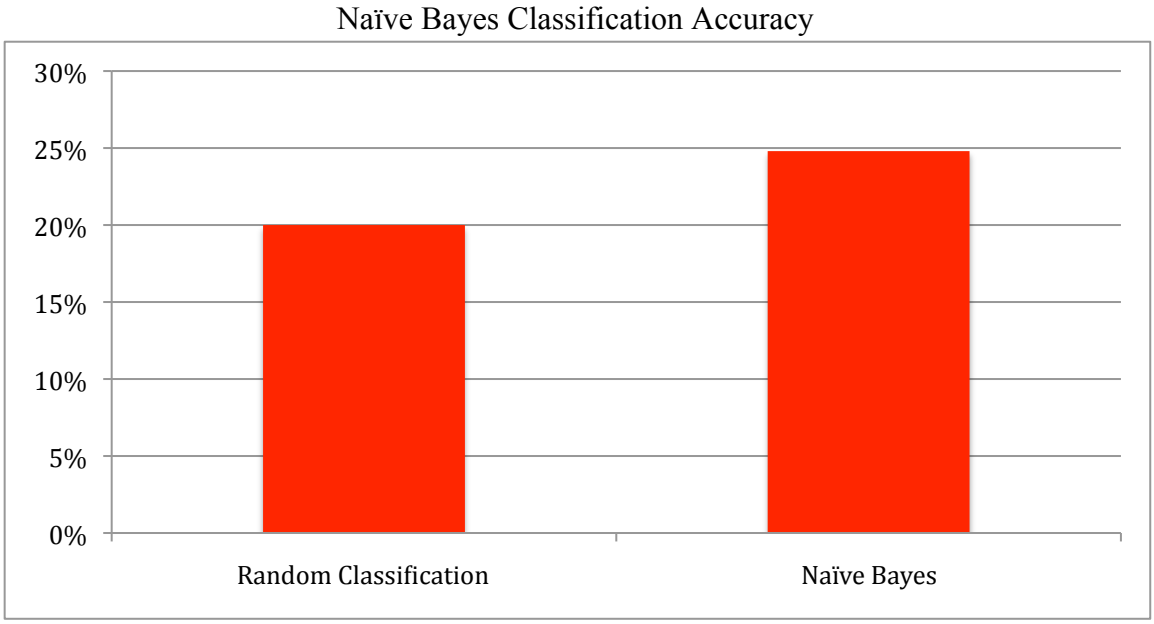
Conveniently, the "track IDs" for each of these datasets exactly corresponded to a track ID from the Million Song Dataset, so I was able find which tracks appeared in both datasets. Luckily I was enrolled in Introduction to Databases this quarter, so I had just recently learned the skills to perform the necessary data management through several SQLite queries. This was a challenge and a great learning process, and after some long hours, I successfully was able to compile the desired data of 186,380 tracks into text files that were ready for analysis.

3. Analysis and Results

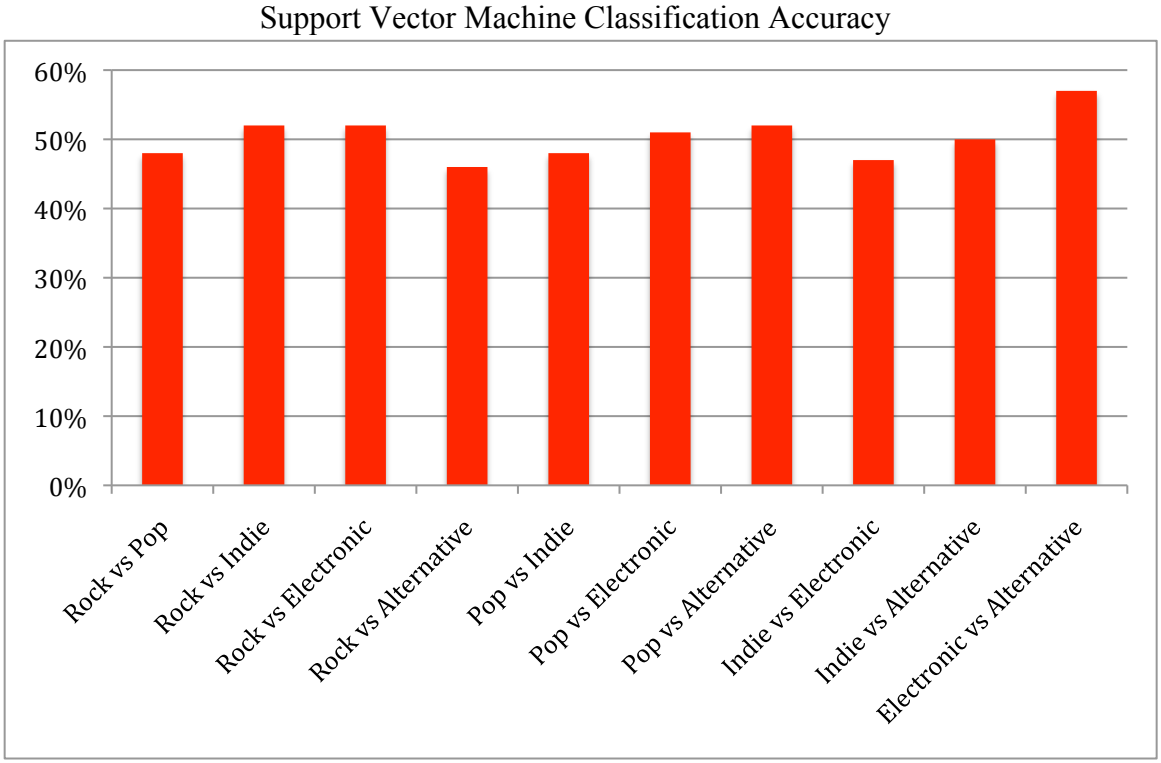
The analysis I conducted was very similar to the spam classifier that we implemented in the second problem set, because the data was so similar. For the spam classifier, our features consisted of emails paired with a dictionary containing the word counts for each word in the email. In my data, each song was paired with a dictionary of lyrics with the exact same format. Thus it seemed very natural, and of course convenient, to use the same techniques.

The first step was to run a Naïve Bayes classifier on a subset of the data. I randomly selected 3,206 of the tracks to be used as training data, and selected another 500 randomly for test data. The biggest change I made from the spam classifier was that I was classifying into one of five genres, rather than classifying into just two labels (spam vs. not spam). This required some extra coding.

The Naïve Bayes classifier did not perform as well as I had hoped. For the data specified, I achieved a success rate of 24.8%. This is only slightly better than random (for classifying into five genres, random classification should yield a success rate of ~20%).



Continuing to follow the structure of the spam classifier, the second step in my analysis was to try using Support Vector Machines on the same dataset. This time, I decided to simplify the problem and look at only two genres at a time. This was easier to actually implement because I was able to use the liblinear library. As a result, I could easily run multiple trials. I ran a total of ten trials, one for each pair of the five genres. My results, however, were again not very promising. All trials had accuracy rates close to 50%, and none were significantly better than random classification.



4. Conclusion

My results may have been disappointing, but that doesn't mean they aren't informative. While I had hoped my analysis would enable me to accurately predict genres, it instead shows that one cannot predict genre from lyrics alone. Let us examine what I have actually studied. My analysis, similar to spam classification, looked specifically at the word counts within a song's lyrics. I did not analyze any information regarding ordering or patterns in words, and I certainly did not analyze any other musical factors of the songs, such as instruments, pitch, etc. So what have I found? I have shown that word choice itself is *not* a good predictor of genre. In spam classification, there *are* keys words such as "save" or "Viagra" that predict spam emails well, which is why these algorithms work well. However, music genres don't seem to have those same indicator words. Instead, word choice is only one of many factors that determine a song's style, and to accurately predict genre we really need to look at the whole picture. Songs are

extremely complex, and thus predicting their style must take this complexity into account.

5. Further Research

Moving forward, I would stress the importance of examining all the other factors that contribute to a song's style when trying to accurately predict genres. Perhaps a combination of audio and lyrical analysis could produce better results. I am certainly interested in pursuing the goal of automatic music classification, but I would need to first acquire the necessary technical expertise to directly analyze audio files.

6. Acknowledgements

I would like to acknowledge Professor Ng and the entire teaching staff for their efforts in running the class this quarter. There were some daunting challenges along the way, but I really learned so much and am extremely happy that I decided to enroll, despite feeling somewhat under-qualified!

7. References

(1) Last.fm dataset, the official song tags and song similarity collection for the Million Song Dataset, available at: http://labrosa.ee.columbia.edu/millionsong/lastfm

(2) MusiXmatch dataset, the official lyrics collection for the Million Song Dataset, available at: http://labrosa.ee.columbia.edu/millionsong/musixmatch

(1) Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.