CMSC 360: Machine Learning
Project Proposal
Due November 13, 2019
Brian Becker and Hannah Beilinson

1. Dataset and Goal

We will use the 19000 Spotify songs dataset [**sp**]. Per the name, n for this dataset is 19,000 and p for the dataset is 18. The dataset has features that indicate the character of each song. This includes concrete measures such as tempo and key, as well as more subjective measures like "danceability." Our goal is to compare how different features affect the prediction of a song's popularity. Specifically, we hope to train an accurate classifier of song popularity and then use this to investigate the robustness of a song's popularity to various feature changes. We will also investigate if using different classifiers leads to a different level of robustness.

2. Software/Methods

We will use the sklearn python package to conduct our experiments. In particular, we will use CNNs and Random Forests as classifiers.

To examine the impact of features, we have two approaches. First, we will record which features provide the most information in each bootstrapped sample by looking at the chosen root for the stumps in the Random Forest. This will indicate which features are most likely to affect the data when perturbed, and will influence how we conduct our follow-up experiments.

Our second approach is to perturb one feature at a time in the test data and reclassify the perturbed examples using both Random Forests and CNNs. We will focus especially on features that showed up as frequent roots in the original Random Forest. The change in popularity after perturbations will show how important certain features are for song popularity.

3. Motivation and Scientific Question

Motivation: We are both interested in music and music theory. We hope to apply machine learning to understand which aspects of music make a song likable in a way not captured by traditional studies of music.

Scientific Question: Which features of a song have the most impact on its popularity? How robust is a song's popularity to small feature changes?

4. Results, Evaluation, and Interpretation

We plan to perturb features of the test data that we think might be meaningful (such as the probability that the song was performed live) and graph the predicted popularity versus values for a particular feature. This will show the effect that a feature has on predicted popularity. We will be able to see from the plots how far a feature can be changed before it has a significant impact, which would be an interesting result for applications such as writing song covers.

We can also evaluate results based off which features are frequently chosen by the Random Forest. This will tell us the importance of different features in determining the popularity of a song.

Finally, we will consider the overall accuracy of our classifier to confirm the validity of using this classifier as a measure of song popularity.

5. References

[sp] - https://www.kaggle.com/edalrami/19000-spotify-songs/data