Project 2

COSC 311

Kalyn Howes & Chloe VanCory

12.15.21

**1 Data Summary**

The Spotify Top 200 Charts (2020-2021) dataset from Kaggle, found here, [Spotify Top 200](https://www.kaggle.com/sashankpillai/spotify-top-200-charts-20202021), includes all of the songs that have been on the Top 200 Weekly (Global) charts of Spotify in 2020 & 2021. Each song has a number of features including: the highest charting position of the song, the number of times that the song has charted, the week when the song had the highest position, the name of the song, the song ID (provided by Spotify), the approximate number of streams the song has, the artist, the number of followers the artist has on Spotify, the genre, the release date, the weeks charted, popularity (between 0-100), danceability (how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity - a value of 0.0 is least danceable and 1.0 is most danceable), acousticness, energy, instrumentalness, liveness (audience during recording), loudness, speechiness (presence of spoken words), tempo, duration, valence (musical positiveness), and finally the main chord of the song. This dataset was mostly gathered from spotifycharts.com, with some features being calculated based on numerous other features.

**2. What “classes” are present within the data? Just by looking at some primitive plots or graphical breakdowns, are there features that “give away” which class a sample belongs to? For instance, in the adults dataset, the classes are whether the person makes more or less than $50k; a relevant question to ask is that, since this doesn’t correlate with any single attribute, perhaps there is a specific pairing of attributes for which a certain combination of values means that somebody is at a particular income level.**

The most obvious “classes'' in this data are the different genres of each song. However, the songs could also be classified by artist, tempo, energy, main chord, etc. Making the genres the classes would allow for easy classification in the learning algorithms. For example, songs with high levels of danceability, energy, tempo, and valence could be classified as the pop genre.

**3. What attributes look like they are parameters of an underlying population that could be learned by an unsupervised machine learning algorithm? For example, in the adults dataset, does it seem like any of the numerical parameters follow a specific distribution within different sub-populations? One example might be that if you know a person’s occupation, education, and income, can you effectively model their hours worked per week?**

The genre of a song could determine the number of streams it has and therefore how many times the song has charted. Maybe there is a correlation between the duration of a song and its highest charting position.

**4. Implement your own version (not using pre-built libraries) of two supervised or unsupervised machine learning algorithms, you may choose ones not discussed in class**

We will be implementing decision trees and the k nearest neighbors algorithm.

**5. Show a visualization of each feature in the dataset.**

**(a) For at least two of these, use a non-standard graphic (outside the standard bar, line, scatter plot) to represent and summarize it. Be creative! Think about visualizations that rely on maps or external structures to present spatial relationships, or perhaps an animated visual of some feature/structure changing over time (matplotlib.animation can help here).**

**(b) The visualization should emphasize both the values of the feature and their relationship to other aspects such as membership in a particular population within the data.**

**6. Show the behavior of your machine learning algorithms on the data. Specifically, consider the following:**

**(a) Show the “learning curve” of your algorithm: how the error changes with more training (either more training data or more epochs, depending on the algorithm). Include the performance on both testing and training datasets during this process, when applicable.**

**(b) Show the generalization error by using different sets of data for testing, training, and validation.**

**(c) How does the algorithm perform when you restrict the dimension of your data? This may be done by removing columns, or by using other dimension reduction algorithms (Principal Component Analysis, Multidimensional scaling, etc.). If you can get good behavior using only 2 or 3 features, try to visualize the algorithm's results on the full domain, to show the “decision surface”.**

**(d) Use more “classical” statistical techniques to motivate the success (or failure) of your machine learning algorithm: are the features uncorrelated with the label? Do hypothesis tests reject assumptions that would make prediction possible?**

**7. Finally, you will present this data and your results to your classmates on the final exam date, December 17. The presentation should be 10-15 minutes long, and will take the place of the final exam grade. Your presentation should focus on:**

**(a) Creative ways to visualize the data, and the results/reasoning behind your machine learning algorithms**

**(b) Clearly explaining why your results make sense in a boarder context, and whether they inform new information that is not directly captured by the dataset**

**(c) Providing “next steps” to somebody with more domain knowledge, that might lead to better or more meaningful outcomes.**