Outline

Becky Catlett, Harley Clifton, Natasha Gesker, Eliot Liucci

2023-11-28

Introduction

We will be using PCA and clustering analysis to find an model structure of the data that is more easily interpreted. The goal of this is to gain an understanding of music popularity in 2018 using Spotify data. We will be using PCA to reduce the number of dimensions since there are 16 different variables and only 100 observations. From there, we will explore k-means, hierarchical, and model-based clustering to analyze the data and see what relationships occur between variables that may not be obvious without the application of clustering methods.

Below is code for three different clustering techniques, from which we will chose one to use for our analysis of the Spotify data. Including all three methods ensures that the method we choose for our analysis is one that best minimizes the **sum of squares** of the data. Although all statistical results have uncertainty, the goal is to use the best statistical method to get rid of a lot of uncertainty in the results. Clustering/classification results are useful for artists, songwriters, producers, and many more individuals that are looking for qualities of a hit song.

Data Cleaning & PCA

The dataset being explored in this project is from Kaggle and is called "Top Spotify Tracks of 2018." This dataset includes numerous quantitative variables as well as some qualitative variables, such as the song title, artist, and the Spotify url ID. The quantitative variables are slightly less intuitive, such as danceability, energy, key, loudness (in dB), mode, speechiness, and acousticness. Most of these quantitative variables range from 0 to 1, where 1 is the highest measure and 0 indicates a low/null value. Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. Speechiness is how often words are present in the song, and mode indicates the modality, which is whether the song is written in major (1) or minor (0). Acousticness is how acoustic the song is, versus a non-acoustic song which would be composed of lots of electronics. Energy is essentially describing the tempo of the song, whether is song is slower and sleepier or more upbeat.

This data set contained no missing data. It did, however, have a lot of variables, meaning when trying to visualize the data using PCA made the most sense. This plot will be used when exploring cluster methods.

EXPLAIN OTHER VISUALS

Cluster Methods

Agglomerative Heirarchical Clustering

For the hierarchical clustering method clusters are based on the distance matrix. With the *single* method, it connects the points based on the minimum distance of the clusters. However, the strange shape of the dendrogram is due to the Single Linkage resulting in "chaining", meaning the clusters just grow, instead of new clusters being formed. This can be problematic because we only need a single pair of points to be

close to merge two clusters.Therefore, clusters can be too spread out and not compact enough. The *average* method, which connects the points based on average distances of the clusters, also has the same phenomenon occurring.

The *complete* method, which connects the points based on the maximum distance of the clusters does not have any "chaining" issues. When determining what height to cut off the groups, a height 7 was chosen as that is the first time we start to see a lot fo tiny groups forming into larger groups.

When visualizing these clusters over the PC Plot it seems that cluster 3 starts to take over most of the points. When comparing it to cluster 1, it is not entirely clear why a point should be in one group compared to another. There are also a lot of little groups that are in the outskirts of the plot. While you could argue that making less clusters by increasing the height, if the clustering logic already appears to be faulty at this point it may be detrimental to the process by making too few clusters.

K-Means Clustering

K-means clustering started with choosing a number of clusters before starting. To decide how many clusters would be idea a Within-Group Sum of Squares (WGSS) plot was created using the original data. From this plot one could argue many different number of clusters would be appropriate. However, for the purposes of this analysis 15 clusters will be created. This was chosen because having too many groups would get overly confusing, especially considering there are only 100 data points. While 15 is still quite large it does get a WGSS of under about 500, more than half of where the WGSS is with only 1 group.

The 'kmeans' function was used to create to assign each original data point to one of the 15 groups. When visualizing these clusters over the PC Plot, one can see that from a single glance these groups do not make a lot of sense for this data set. While groups tend to have clear centers and pretty consistent variability within groups, for interpreting what these results mean this method is not ideal for this data set.

Model-Based Clustering

For the model based clustering method assumes distributions for both the population and the sub population. While the pdfs are defined for these distributions for the purposes fo this data they will not be discussed in detail. The `Mclust` function created five separate clusters. Based on the classifications there is one point that stands out. Observation 44 is alone in its own cluster, this could either indicate that this observation is an outlier or that this clutering method is faulty.

When visualizing these clusters over the PC Plot, one can see that the the point in cluster 5 does appear to be out on its own compared to the other clusters. While there is some overlapping from cluster 1 across the other points, in this plot is easier to see the trends of the various clusters, especially cluster 2 and and 4.

Results

While one can look at the visualizations of these cluster methods and make gut dections about the quality of each method for this data set, we also want to look at the sums of squares of each.

It was found that the hierarchical method had a sum of squares that was almost twice the other two methods. While the model-based method does have a lower sum of squares than the k-means method, they do seem to be fairly close. Meaning, that if researches felt that the clusters formed by the k-means method were more useful for their analysis they would not be at a serious disadvantage for using k-means over model-based. Still, based these values and the vizulation of the different cluster method, we conclude that the model-based method is the best of this data set, provided that there was no model goal heading into forming these clusters.

Summary

Summary section that highlights the main points of your analysis

Citations