CPS 844 Assignment 2

04/12/2022
Section 1
Ryan Soliven
500840952

Background

Dataset Information

Two datasets were used in this assignment. The first dataset used was taken from the UCI Machine Learning Repository and is the Turkish Music Emotion Dataset[1]. Information taken from the website follows: The dataset is designed as a discrete model, and there are four classes in the dataset: happy, sad, angry, and relaxed. To prepare the dataset, verbal and non-verbal music are selected from different genres of Turkish music. A total of 100 music pieces are determined for each class in the database to have an equal number of samples in each class. There are 400 samples in the original dataset as 30 seconds from each sample. Number of Data in each class, Relax 100, Happy 100, Sad 100, Angry 100.

The second dataset used was taken from the UCI Machine Learning Repository and is the Tic-Tac-Toe Endgame Data Set[2]. Information taken from the website follows: This database encodes the complete set of possible board configurations at the end of tic-tac-toe games, where "x" is assumed to have played first. The target concept is "win for x" (i.e., true when "x" has one of 8 possible ways to create a "three-in-a-row").

Attribute Information

Attribute information for the first dataset found on the website follows: Features such as Mel Frequency Cepstral Coefficients (MFCCs), Tempo, Chromagram, Spectral

[1] https://archive.ics.uci.edu/ml/datasets/Turkish+Music+Emotion+Dataset#
[2] https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame

and Harmonic features have been extracted to analyze the emotional content in music signals. MIR toolbox is used for feature extraction.

For the second dataset, nine columns represent the position of the board and are filled with attributes “x”, “o”, or “b” for blank. The tenth column represents the class and is either “positive” or “negative”, with positive being the “x” player winning.

Methods

Three methods were used to analyze the first dataset and one method was used to analyze the second dataset.. Methods used to analyze the first dataset were K-means cluster analysis, Hierarchical Clustering methods, and Density-Based Clustering methods. Methods used to analyze the second dataset was Association Rules analysis.

Association Rules Analysis

Association rule mining was used to analyze the second dataset. Association rule mining is a technique to identify underlying relations between different items. We use the apyori library, where all of the code for generating the association rules has been implemented.

K-means Cluster Analysis

Cluster analysis is the task to partition a set of objects into groups of similar instances. We apply k-means clustering and use the scikit-learn library package to perform the analysis. The k-means clustering method consists of the iterative assignment and update steps: First we form k clusters by assigning each instance to its nearest centroid. Next, we recompute the centroid of each cluster.

Hierarchical Clustering

We apply the following hierarchical clustering methods. Single link (MIN). Complete link (MAX), Group average. We plot the associated dendrogram after each method.

Density-Based Clustering

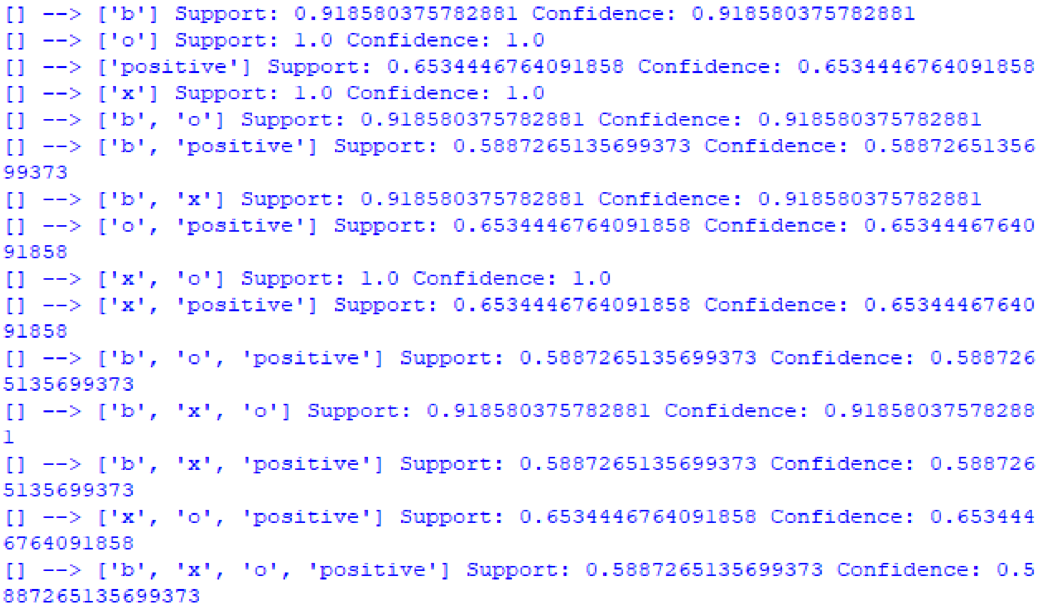
We perform Density-Based Clustering only on two columns of the first dataset. We analyze the "_RMSenergy_Mean" and "_Roughness_Mean" columns.

We identify high-density clusters separated by regions of low-density. In the popular DBScan method, data points are classified into 3 types: Core points, Border points, and Noise points.

Classification is applied as a function of two parameters: The radius of the neighborhood size (eps), the minimum number of points in the neighborhood (minpts).

Results and Conclusions

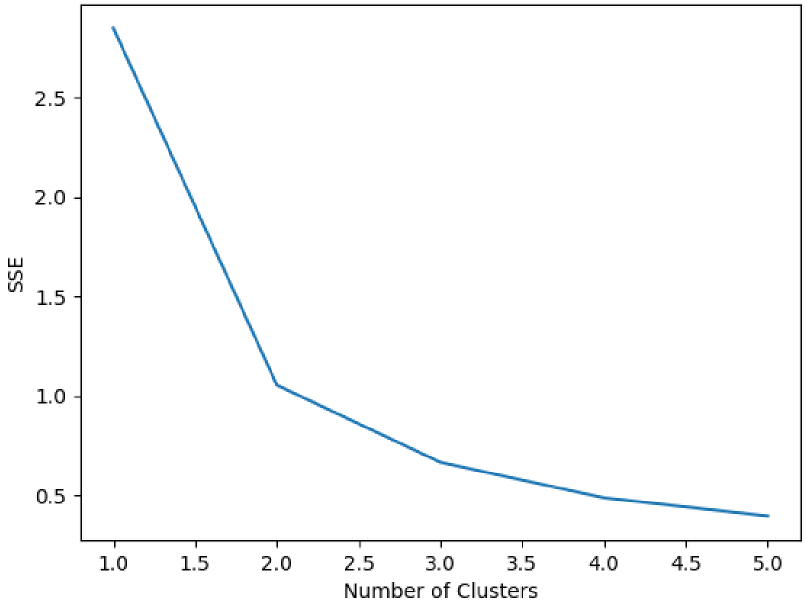
Association Rules Analysis



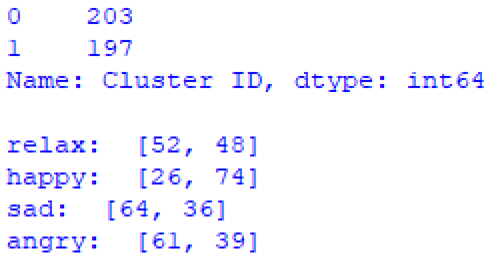
```
[] --> ['b'] Support: 0.918580375782881 Confidence: 0.918580375782881
[] --> ['o'] Support: 1.0 Confidence: 1.0
[] --> ['positive'] Support: 0.6534446764091858 Confidence: 0.6534446764091858
[] --> ['x'] Support: 1.0 Confidence: 1.0
[] --> ['b', 'o'] Support: 0.918580375782881 Confidence: 0.918580375782881
[] --> ['b', 'positive'] Support: 0.5887265135699373 Confidence: 0.58872651356
99373
[] --> ['b', 'x'] Support: 0.918580375782881 Confidence: 0.918580375782881
[] --> ['o', 'positive'] Support: 0.6534446764091858 Confidence: 0.65344467640
91858
[] --> ['x', 'o'] Support: 1.0 Confidence: 1.0
[] --> ['x', 'positive'] Support: 0.6534446764091858 Confidence: 0.65344467640
91858
[] --> ['b', 'o', 'positive'] Support: 0.5887265135699373 Confidence: 0.588726
5135699373
[] --> ['b', 'x', 'o'] Support: 0.918580375782881 Confidence: 0.91858037578288
1
[] --> ['b', 'x', 'positive'] Support: 0.5887265135699373 Confidence: 0.588726
5135699373
[] --> ['x', 'o', 'positive'] Support: 0.6534446764091858 Confidence: 0.653444
6764091858
[] --> ['b', 'x', 'o', 'positive'] Support: 0.5887265135699373 Confidence: 0.5
887265135699373
```

In the image above, we find each of the rules generated, along with their corresponding support and confidence values. We used a 40% support threshold and a 50% confidence threshold.

K-means Cluster Analysis



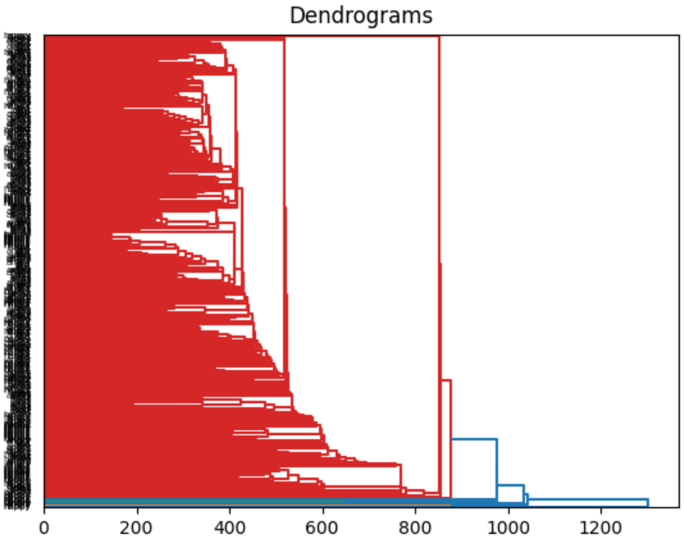
In the image above, we plot to find the SSE vs the Number of Clusters to visually find the "elbow" that estimates the number of clusters. We see that the elbow value is 2 and we can estimate that there will be two clusters.



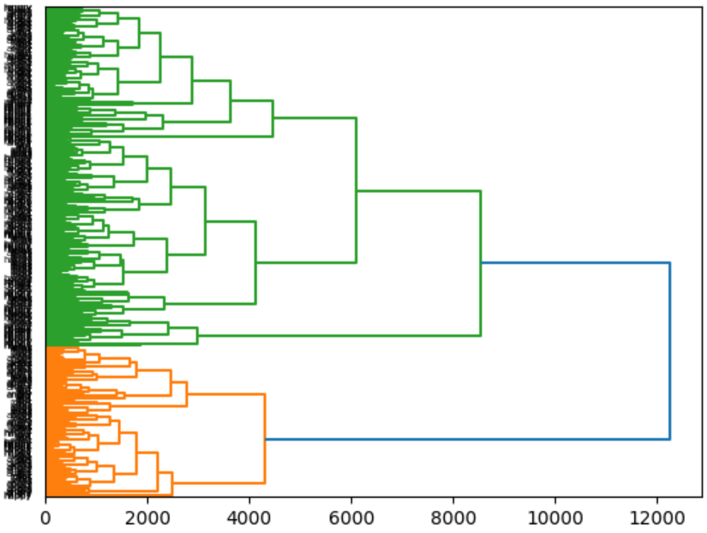
```
0     203
1     197
Name: Cluster ID, dtype: int64

relax:  [52, 48]
happy:  [26, 74]
sad:  [64, 36]
angry:  [61, 39]
```

Here we see that 203 of the rows belong to cluster 0 and 197 belong to cluster 1. 52 of the music classified as relax belong to cluster 0 and the other 48 belong to cluster 1. 26 of the music classified as happy belong to cluster 0 and the other 74 belong to cluster 1. 64 of the music classified as sad belong to cluster 0 and the other 36 belong to cluster 1. 61 of the music classified as angry belong to cluster 0 and the other 39 belong to cluster 1.

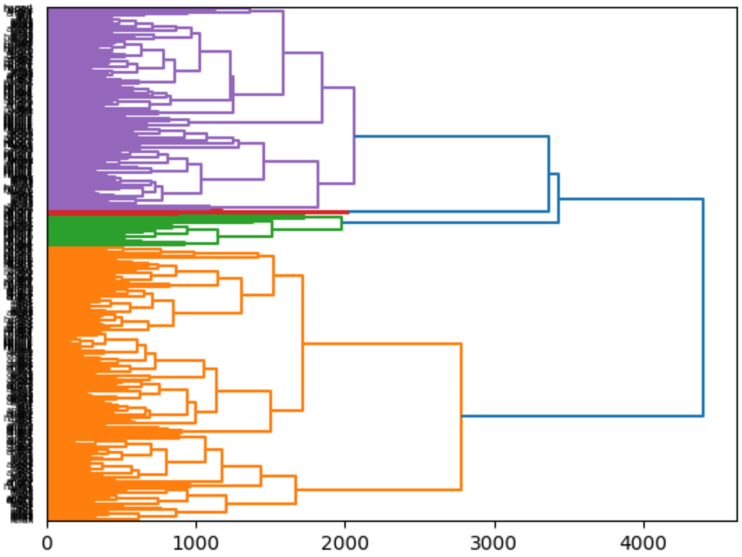
Hierarchical Clustering



In the image above is the resulting dendrogram after performing Single link (MIN) analysis.

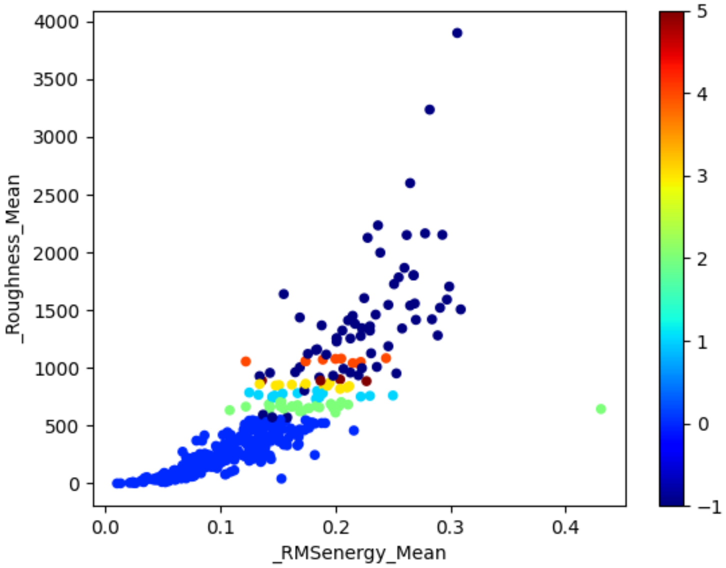


In this image, we see the plot associated dendrogram after completing Complete Link (MAX) analysis on the data.



In this final image, we see the plot associated dendrogram after completing Group Average analysis on the data.

Density-Based Clustering



In this image we see the scatter plot of the '_RMSenergy_Mean', and '_Roughness_Mean' columns from the first dataset. We see that there are four different clusters that were found, not including noise.

Sources

Dataset 1: https://archive.ics.uci.edu/ml/datasets/Turkish+Music+Emotion+Dataset#

Dataset 2: https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame