Classify Music into Genres

Team 7 (The Barbarians)

Kartik Gupta 201302008

Sagar Gaur 201364127

Vishal Thamizarasan 201302061

Introduction

The project aims to classify music into the following genres EDM, Classical, Rock and Jazz.

The task of genre classification in music is usually a tough task because of the ambiguity in the definitions of the genres themselves. For eg: Classical music of India is much different than the classical music of Europe.

Sometimes the definition of the genre is not based on the musical features but on the social features.
For e.g. Pop music stands for popular music, and there is no way to know from the musical features if a song can be classified as pop or not but over the time these definitions have faded and musical features got associated to these kind of genres as well but some ambiguities still exist.

The ground truth data is considered to be the one in Pandora. They have a team of musicologists who listen to each and every song and the give tags to each of them and then the final tags of a song are decided based on the majority.

Dataset

In our model each song is represented using 8 high level features namely, energy, liveness, speechiness, acousticness, danceability, instrumentalness, loudness and valence.

Along with that the training data is labeled either one of the genres 'rock', 'edm', 'jazz', 'classical'. A total of 4000 songs are present in the data, 1000 for each label.

70% of data was used for training and 30% for testing.

(The data is made available by The Echo Nest Project)

Low Level vs High Level Features

1. **Low level features:** This is the output of signal processing techniques when applied to the musical audio such as DFT. An example could be converting raw wave signals of songs into signal amplitude ordered by their frequencies. This may also include the pitch and timbre information at very small windows (0.2 seconds). Although this is the information using which all the other information about a musical piece is derived often this has very high dimensions and it can be too slow to use these as an input vector in the classification task. The original music files have 44100 samples per second.

2. **High Level Features:** These are the features derived from the low level features, and include various information about the track. These are more of subjective terms and more understandable to humans including musicologists (They would understand acousticness but not the DFT output). The have an advantage that they are very few in number and hence will be less time consuming to use them wherever possible.

Features

Acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.

Danceability: 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. A value of 0.0 is least danceable and 1.0 is most danceable.

Energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.

Instrumentleness: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the Instrumentleness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.

Liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.

Features (cont.)

Loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.

Speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.

Valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

Methodology

1. Simple KNN
2. KNN with Metric Learning
3. Logistic Regression
4. Growing Neural Gas with ANN

KNN

KNN was the most basic approach to set a lower limit ballpark that other algorithms should outperform

Neighbours: 5

Conclusion: Really bad. Needed to do better

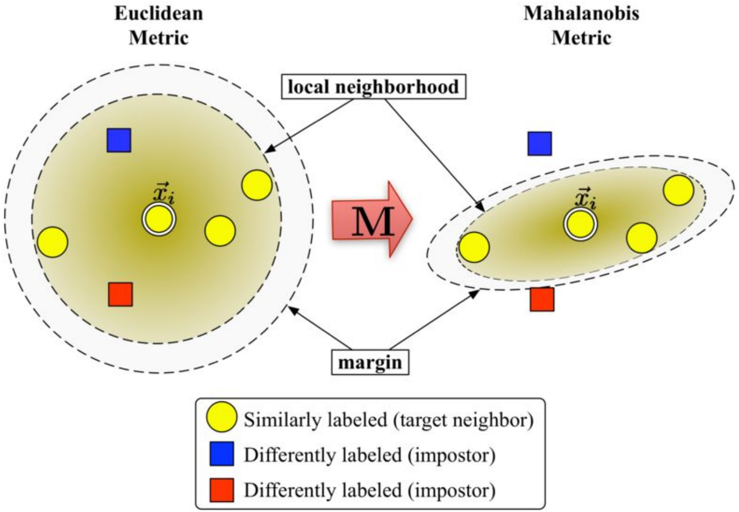
Classical	EDM	Jazz	Rock
547	75	125	86
34	509	51	172
64	22	490	40
47	108	26	404

Training set score: 0.696429

Classical	EDM	Jazz	Rock
217	20	74	42
23	186	19	97
43	14	196	29
25	66	19	130

Test set score: 0.607500

Metric Learning (Largest Margin NN)


Euclidean Metric
Mahalanobis Metric
local neighborhood
$\vec{x}_i$
M
$\vec{x}_i$
margin
● Similarly labeled (target neighbor)
🟦 Differently labeled (impostor)
🟥 Differently labeled (impostor)

Classical	EDM	Jazz	Rock
574	53	114	62
23	560	36	159
49	12	517	27
46	89	25	454

Training set score: 0.751786

Classical	EDM	Jazz	Rock
238	19	69	29
15	200	16	89
27	10	208	19
28	57	15	161

Test set score: 0.672500

Source: Wikipedia Commons

Logistic Regression

Conditional Probability

$$\mathcal{P}_{\boldsymbol{w}}(y=\pm 1|\boldsymbol{x}) \equiv \frac{1}{1+e^{-y\boldsymbol{w}^T\boldsymbol{x}}},$$

Minimizes the following regularized negative log-likelihood

$$P^{\mathsf{LR}}(\boldsymbol{w}) = C\sum_{i=1}^{l}\log\left(1+e^{-y_i\boldsymbol{w}^T\boldsymbol{x}_i}\right)+\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

Classical	EDM	Jazz	Rock
574	53	114	62
23	560	36	159
49	12	517	27
46	89	25	454

Training set score: 0.685357

Classical	EDM	Jazz	Rock
225	15	78	15
11	217	15	72
50	11	202	26
22	43	13	185

Test set score: 0.690833

Growing Neural Gas

1.Growing Neural Gas is a method to make a self transforming map of the data.

2.The idea is to make clusters of the data where the number of clusters is not known but there is a label (class) associated to the data points.

Algorithm

1.Start with two random vectors (of different classes). Consider them as the means of two different clusters.

2.While considering a new data point, try to fit it in the available clusters using a maximum threshold distance.

3.If the data point doesn’t fit in any of the available clusters then make a new cluster with that data point as the mean.

Healthy Clusters?

1.A cluster is said to be healthy if there is clear majority of a class in that cluster. The higher the majority the healthier the cluster is considered.

2.A genre can have multiple healthy clusters associated to them but each cluster should have only one major class in its data points.

3.The health of the clusters in our case depends on the distance ‘d’.

4.We have defined a unhealth function as

J(d) = Sigma(1-(no of points of the major class)/total points in that cluster)/Total number of clusters

Finding Best Value for ‘d’

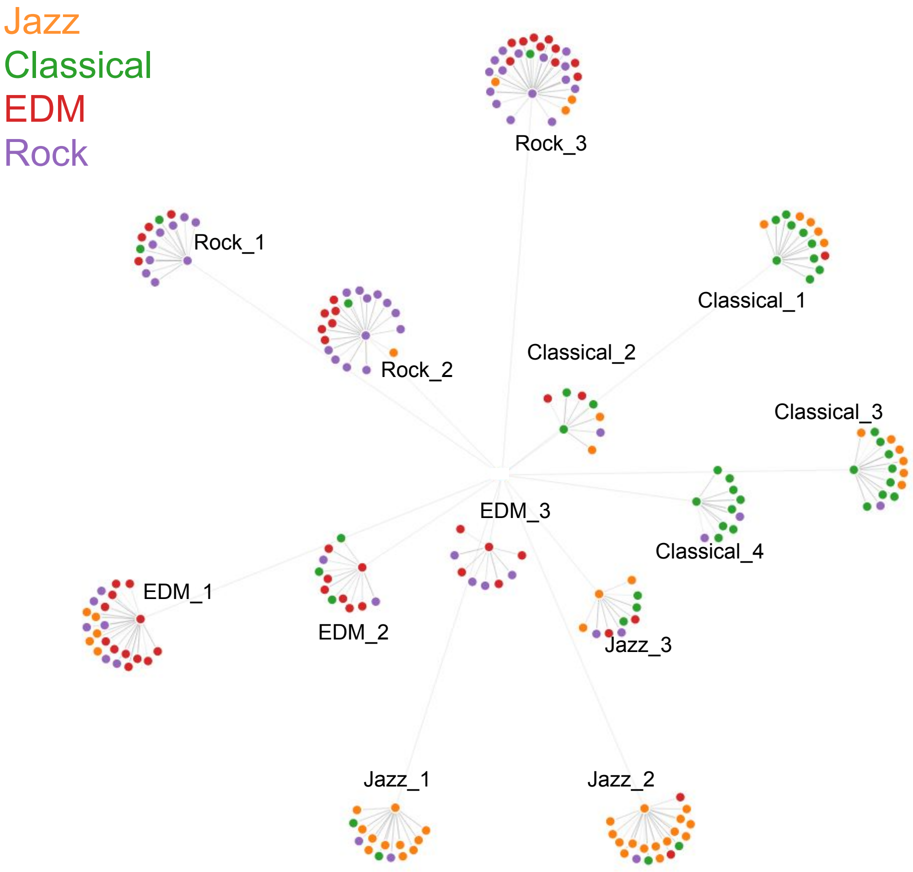
1.We used basic gradient descent function to minimise the unhealth function.

2.The best value found was around 1.1

3.The most of the clusters do have a clear and a high majority.

GNG Classes

Classical	EDM	Jazz	Rock	Top
9	1	5	0	**Classical**
9	0	0	2	**Classical**
3	2	2	1	**Classical**
9	0	6	1	**Classical**
0	12	5	6	**EDM**
0	5	0	4	**EDM**
3	7	0	2	**EDM**
2	2	15	1	**Jazz**
3	2	3	2	**Jazz**
2	0	11	2	**Jazz**
1	10	3	15	**Rock**
2	4	0	9	**Rock**
1	6	1	12	**Rock**



Artificial Neural Network

11 input nodes
13 hidden nodes
4 output nodes

Learning rate: 0.1
Convergence in about 850 epochs

Classical	EDM	Jazz	Rock
581	39	157	58
13	574	31	170
123	25	560	25
102	148	59	535

Training set score: 0.703125

Classical	EDM	Jazz	Rock
117	9	37	15
10	164	5	45
34	8	137	8
20	33	14	144

Test set score: 0.702500

Comparing Different Approaches

Algorithm	Accuracy (Train Set)	Accuracy (Test Set)
KNN	69%	60%
KNN with metric learning	75%	67%
Logistic Regression	68%	69%
ANN	70%	70%
ANN + GNN (S + 1MV)	74%	71%
ANN + GNN (S + 2MV)	76%	73%
ANN + GNN (S + 3MV)	77%	74%

Summary

1. We have compared 4 different approaches towards genre classification
2. The best was the GNG and ANN which gave an accuracy around 74%

References

Clark, Sam, Danny Park, and Adrien Guerard. "Music genre classification using machine learning techniques." (2012).

Haggblade, Michael, Yang Hong, and Kenny Kao. "Music genre classification." *Department of Computer Science, Stanford University* (2011).

Li, Tao, Mitsunori Ogihara, and Qi Li. "A comparative study on content-based music genre classification." *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2003