Predictive Modeling with The Million Song Dataset

Kevin Chou

Texas A&M University

Abstract

When we listen to music, we tend to only pay attention to the audio, and possibly the name of the artist. However, there is much more information in a digital song file than just the audio. A song's metadata describes every aspect and detail of the song except the actual audio. In this project, we focus in on those details and use statistical learning methods to establish relationships between songs. We extract all of the metadata from each song from a subset of the Million Song Dataset and perform various analyses to try and predict a information about a song based off the results obtained from the analyses.

Background

The Million Song Dataset is a collection of audio features and metadata for a million contemporary music tracks. The dataset was created by Thierry Bertin-Mahieux and Daniel P.W. Ellis from Columbia University and Brian Whitman and Paul Lamere from The Echo Nest under a grant from the National Science Foundation. The purposes of the Million Song Datset are

- To encourage research on algorithms that scale to commercial size
- To provide a reference dataset for evaluating research

album_id	album_name	artist_id
artist_latitude	artist_longitude	artist_location
arist_name	danceability	duration
key_signature	key_signature_confidence	num_of_songs
tempo	time_signature	time_signature_confidence
title	year	start_of_fade_out
end_of_fade_in	song_id	song_number

Table 1: List of the metadata fields for each song file used for testing in this project

Methods

Linear Regression for Year

The first approach to predicting the year of a song was attempting to fit a linear regression model of the Year variable against the other test variables. If a linear relationship is apparent and assumed from the fitted regression model, the relationship can be written mathematically as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p, \qquad (1)$$

where $\hat{y}$ is the predicted variable and $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2 ... \hat{\beta}_p$ are the parameters, and in our case, Year and variables tested respectively. After performing the linear regression tests for Year, we assess the accuracy of our models through leave-one-out cross-validation.

Classification of Regions

Using the Artist Latitude and Artist Longitude metadata, we created several regions—Asia, Europe, United States, United Kingdom, etc. Each song is then classified into a specific region based off its Artist Latitude and Longitude metadata.

A Bayesian approach is used in attempt to predict a song's region. The two methods primarily used in the analyses were LDA (linear discriminant analysis) and QDA (quadratic discriminant analysis).

LDA

The LDA method is simply using Bayes Rule to classify an object to a group, or a song to a region, but assuming that each group has a multivariate normal distribution and all groups have the same covariance matrix. LDA can be written mathematically as

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + log \pi_k \qquad (2)$$

Assign object x to group k.

QDA

The QDA method is similar to the LDA method in that it uses Bayes Rule and assumes multivariate normal distributions for each group, however, the groups are assumed to have different and unique covariance matrices.

Results

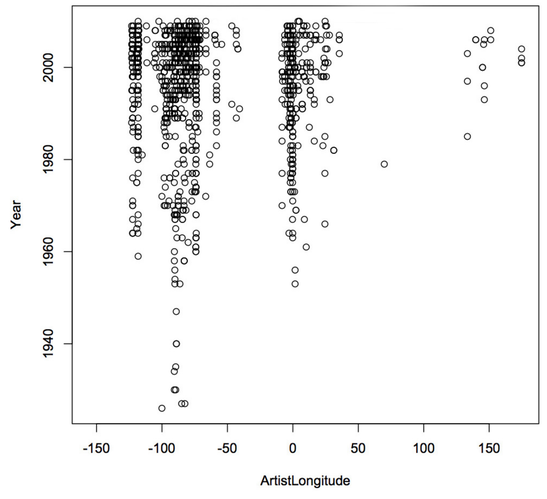

Figure 1: Linear regression of Year v. Artist Longitude

The scatter plot above suggests a non-linear relationship between the Year and Artist Longitude, and further testing on Year v. Other Variables produced similar, non-indicative results.

	Predicted Region							
Actual	Africa	Asia	Australia	Europe	Russia	SouthAmerica	UK	US
Africa	8	0	5	23	0	4	17	113
Asia	0	1	2	9	0	4	18	126
Australia	0	0	0	9	0	1	7	37
Europe	1	0	1	14	0	3	15	57
SouthAmerica	0	1	3	11	0	6	18	77
UnitedKingdom	2	0	1	9	0	2	32	61
UnitedStates	0	0	0	2	0	0	6	28

Table 2: LDA prediction for region confusion matrix

Linear discriminant analysis was used as an attempt to predict the region class of each song and the results are displayed in the table above. The QDA test produced similar results.

Discussion

The results obtained from both linear regression modeling and classification provided very little indication that there is a direct, significant relationship between the metadata fields of the song files. Observing the LDA confusion matrix (**Table 2**), it can be seen that the prediction were off and fairly inaccurate. The QDA confusion matrix (not shown) performed similarly with inaccurate predictions. Further work is being done to determine if a deeper relationship exists.

Additional Information

R code written for the various analyses and graphs and tables of the different results can be found here:

https://github.com/ardabney/millionsongs.git

More information on the Million Song Dataset can be found here:

http://labrosa.ee.columbia.edu/millionsong/

References

[1] Brian Whitman Paul Lamere Thierry Bertin-Mahieux, Daniel P.W. Ellis.
The million song dataset.

[2] Trevor Hastie Gareth James, Daniela Witten and Robert Tibshirani.
An Introduction to Statistical Learning.
Springer, New York, 2013.

Acknowledgements

This project is supported by Texas A&M University Department of Statistics and Dr. Alan Dabney.

TEXAS A&M UNIVERSITY