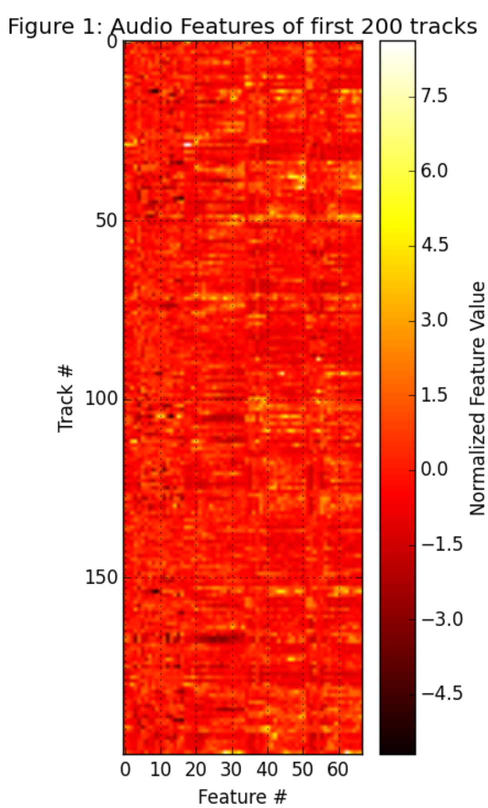
Application of Statistical Learning Models to Predict Geographical Origin of Music

Shifan Mao
CME 193 Final Project
05/18/15

Introduction

In the era of information technology, music becomes almost instantaneously accessible to more and more people, at a pace people would not have imagined in the past. The traditional way to appreciate music is almost singular - that is, to listen to the piece. As a result, the perception of music can be subjective, and can even vary with the listener's state of mind. However, nowadays music pieces can be analyzed digitally. Large sets of data can be compiled from analyzing large media libraries from music of different genres. The work by Zhou et al. [1] is based on such a dataset. Zhou et al. offers the data set about 1058 tracks from 33 countries and areas. Each track has its audio feature (such as spectrum centroid) and the latitude and longitude of its origin.

In this project, I propose to use different learning machines to predict the geographical origin of a music piece based on its audio features. Using the different learning models, a set of training data will be used to develop learning models to predict the geological origins of different music tracks. Such analysis will compliment to the work by Zhou et al. in finding the correlation between the audio features and geographical origins of music.



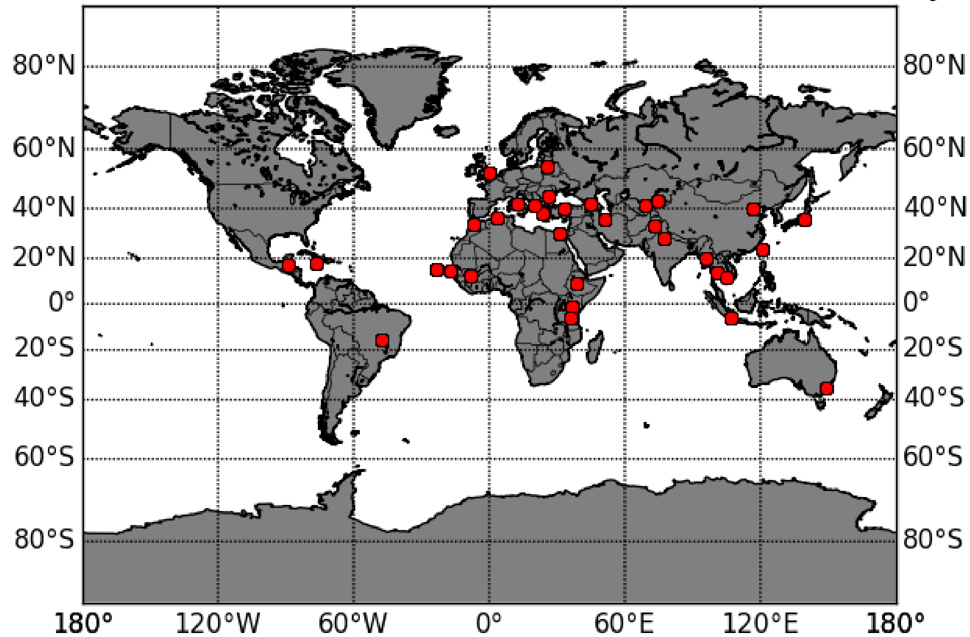
Dataset and Method

The audio features and geological location of a total of 1059 tracks is obtained from Zhou et al [1]. Each track contains 67 audio features extracted from software MARSYAS (http://marsyas.info) [2]. The audio tracks are from 33 distinct regions/countries. The audio features are normalized to have mean of 0 and standard deviation 1.

An illustration of the audio features of first 200 tracks is shown in the figure on the left. Some features are already noticeable by observation in Figure 1. For example, some tracks have highly negative values in some audio features, in contrast to other tracks without such signature. Such patterns can potentially be used to construct learning models for consistent predictions of geological locations of tracks based on audio features.

The geological locations of the 1059 tracks is shown in Figure 2. Note that 33 distinct features is widely distributed across the global. According to Zhao et al., tracks from Europe are statistically biasedly excluded due to the significant influence of western music across the world. Figure 2 is generated using Basemap toolkit in Python (http://matplotlib.org/basemap/).

Figure 2: Distribution of 1059 tracks (one location can have many tracks)



It needs to be pointed out that the problem is framed as a regression problem instead of a classification problem for the following reasons: (1) Number of distinct regions/counties is significantly smaller than the sample size/number of tracks (2) Difference between tracks in their geological locations can result into an error distance metric as the distance between locations. The underlying differences in the audio tracks should depend on their origin of region/country, rather than the true geological locations specified by longitudes and latitudes.

Different regression models are used to perform the predictions of geological locations of music tracks. In particular, ordinary linear regression (OLR), random forest (RF) and support vector machines (SVM) are used and their relative performances on this dataset are compared. All of the audio features are used in all regression models, without additional weightings. Scikit-learn package in Python is used to construct all the regression models (http://scikit-learn.org/stable).

Results

The datasets are split into training set and test set. Training set consists of ¾ of the total dataset. OLR, RF and SVM are trained to fit the training set. Test set is used to evaluate the performance of each regression model.

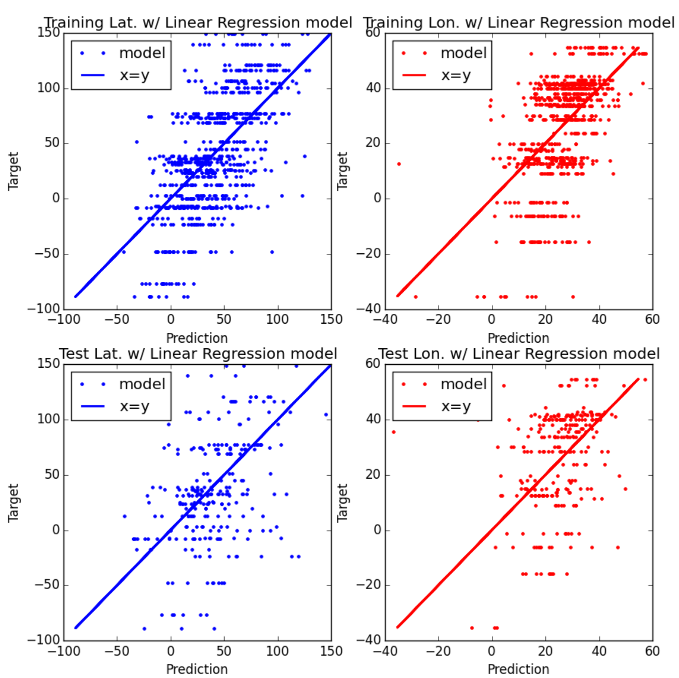
To simplify the problem, we treat longitude and latitude as evenly distributed real numbers, while in reality both longitude and latitude become more sparse closer to the equator.

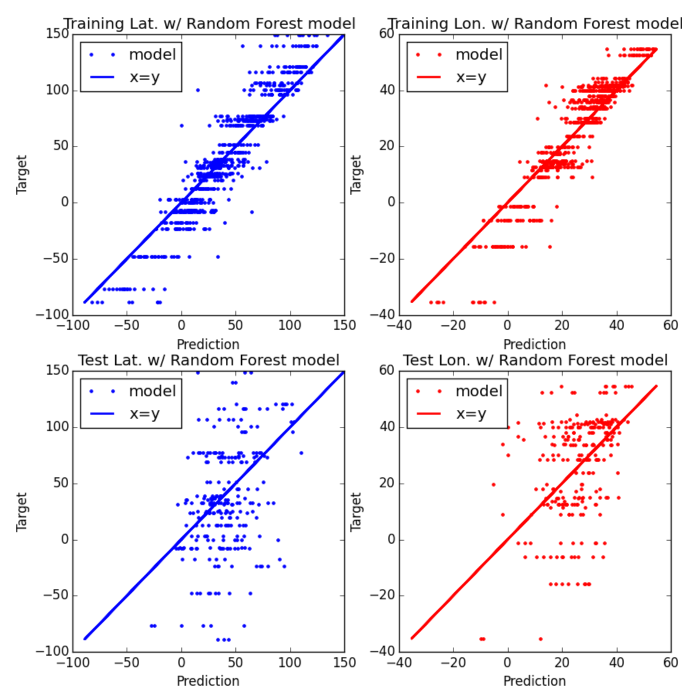
In order to visualize the performance of different learning models. I show the mean-squared errors of the predictions and observations from both training and test sets. In principle, OLR can give minimum training and test error if the data has underlying linear dependence of observations (longitude and latitude) on features (audio features). Similarly, SVM with particular choice of kernel can also show small training and test error for specific data sets with particular structures. On the other hand, RF can minimize training mean-squared error if the forest contains large size trees, but will result in large test mean-squared error with highly flexible learning models. Here I choose learning models OLR, SVM with linear kernal and RF with 10 trees in forest and default choice of tree parameters (http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html) in Python scikit-learn library.

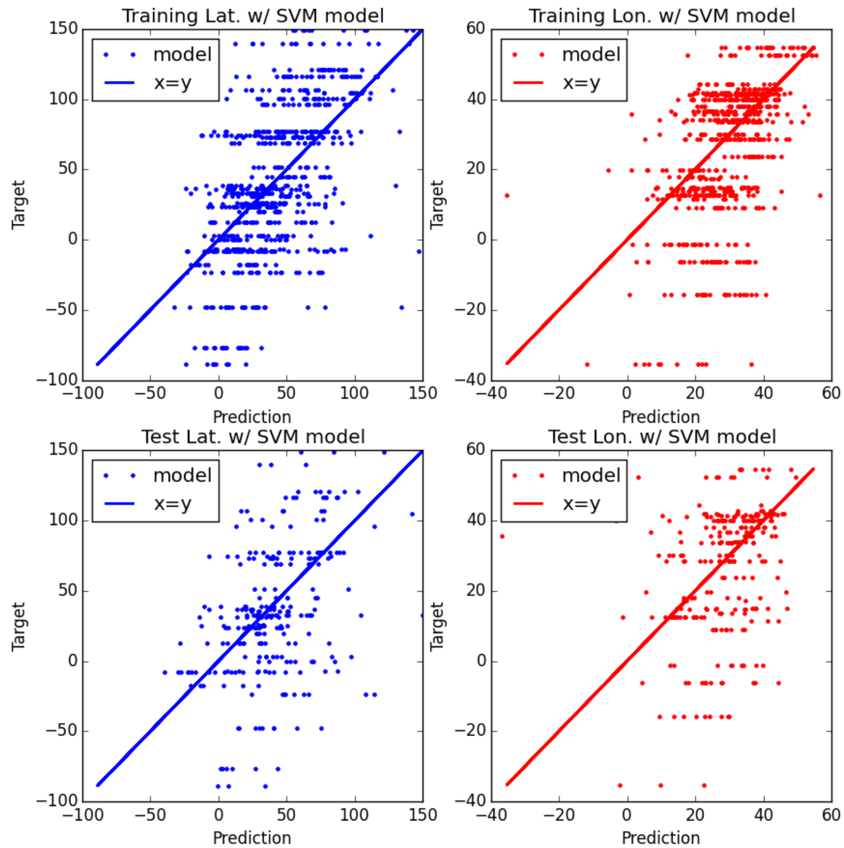
Model	OLR	RF	SVM
Longitude Training Error	1600.6	334.9	1754.5
Longitude Test Error	2253.4	2004.9	2231.6
Latitude Training Error	233.9	46.8	260.4
Latitude Test Error	309.4	274.6	327.7

Table1: Training and test error of longitude and latitude predictions of OLR, RF and SVM model with linear kernel. Values obtained from averaging 10 distinct trainings of each model.

The following figures show comparisons between predicted values from three models (dots) and actual values (lines) of longitude and latitude data in both training and test sets.







Comparing the performances of three models, training error is most minimized by RF model. This suggests that the outcome and predictor dependence is highly non-linear. Such complicated data structure is reflected in the worse performance of OLR and SVM with linear kernel. In comparison, OLR and SVM have comparable results with larger deviations from actual data in training set.

Observing test MSE with each learning model, OLR and SVM have similar performances. RF model, on the other hand, have slightly better performance with on average 11% lower longitude MSE and %11 lower latitude MSE than OLR and %10 lower longitude MSE and %16 lower latitude MSE. Such comparison suggests the outperformance of RF model in capturing the data structure and can be used to extend predictions of geological locations of music tracks based on their audio features.

Conclusions

In this project, I used Python scikit-learn library to perform statistical analysis of a dataset of geological locations and audio features of 1059 music tracks. Regression models based on ordinary linear regression, random forest and support vector machines are applied to predict geological locations of music tracks based on audio features. A comparison of the performances of three models shows that the dataset has highly nonlinear and such complexity is better captured in random forest model. In contrast, ordinary linear regression and support vector machines with linear kernel give ~10% larger MSE, i.e. worse predictions in both training and test sets.

References

[1] Zhou, Fang, Q. Claire, and Ross D. King. "Predicting the Geographical Origin of Music." Data Mining (ICDM), 2014 IEEE International Conference on. IEEE, 2014.
[2] https://archive.ics.uci.edu/ml/datasets/Geographical+Original+of+Music

Python Source Code https://github.com/shifanmao1989/Geological_Music.git