Yu Qin yq2186

Instruction of Project 4: Association mining of music and text

Feature Selection

I use only 6 features out of 16 for each songs to train the model.from the analysis of the length of the features, I give all the features the same length as 1375, which is the 95% quantile of the lengths. Then for these features, there are 2 matrix-value features, I treat each of these 2 matrix-value features as several individual features. So I have 28×1375 matrix for each song. With principle components analysis, I only choose the first principle component, since the first one take most amount of the total variance. So at the end, the final vision of the observation matrix is a 2350×1375 matrix, with each vector representing a song.

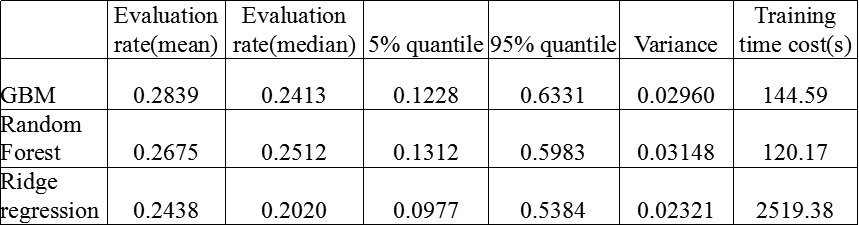
I also try 2 principle components for each song, it covers 90% variance. And the result shows a little improvement. Because the time cost is huge compared to the 1 PC case, I use 1 principle components for each songs.

Lyric Matrix

For lyric matrix, I use topic-model to divide the lyrics into 20 clusters. Instead of using the topic directly as my result, I think linear combination of the 20 topics may be the better approach. And I calculated 2 probability matrics: every word for each topic & every topic for each song.

Associated Patterns

Then I try 3 supervised methods to build the model. Since the number of parameters is unbalanced compared to the number of observations(songs), I tried Gradient Boosting Machine, Random Forest & Ridge Regression methods. I consider evaluation accuracy, stability and time cost as determined factors to judge if the method is good. And the results show below (with 1880 train observations and 470 test observations)



As a result, Ridge regression performs best out of these 3 methods.

Improvement

As far as I concerned, the above result may be improved from these aspect:

a) Information is lost too much in feature selection part. Maybe an extended observation matrix which contains more information is necessary for more accurate predict;

b) This Ridge Regression method may not be the “ideal’ modeling for this project. Deep learning and other supervised methods may have good performances.