M0444 Project One: Release Year Prediction for Songs

Marcos Teixeira
RA 209814
m209814@g.unicamp.br

Miguel Rodrguez
RA 192744
m.rodriguezs1990@gmail.com

I. Introduction

For our project one in M0444: Machine Learning and Pattern Recognition, we study on a subset of Million Song Dataset from UCI Machine Learning Repository [1] in order to obtain the best model for release year prediction. The features presented for this task are composed by 12 timbre average and 78 timbre covariance features. Initially, we did a exploratory analysis on the dataset in order to find some relationships between features. After this initial analysis, we evaluate Machine Learning models for year release prediction, including: Linear Regression with Gradient Descent, Linear Regression with Normal Equation and Polynomial Regression. Mean Absolute Error(MAE) was chosen to measure the model accuracy. We test many forms of model improvement, like: Regularization, Normalization, model parameters, dimensionality reduction, learning rates, etc. After all evaluations and tunings, the results indicated that Polynomial Regression with degree=2 is the best one with the MAE of 6.64008163 on test set.

II. Data

This data is a subset of the Million Song Dataset [1]: a collaboration between LabROSA (Columbia University) and The Echo Nest. Prepared by T. Bertin-Mahieux. The songs in dataset are mostly western, commercial tracks from 1922 to 2011. The distribution of dataset is 463,713 for training set and 36,285 for test set. The details are shown in Table I

Table I: Dataset distribution

Description	*inputs*
Entries in train set	*463,715*
Entries in test set	*36,285*
Total entries in dataset	*515,345*

Each line of the dataset are composed by 90 audio features, being 12 timbre average features and 78 timbre covariance features and the target (year released). This informations are summarized in Table. II

Table II: Feature distribution

Index	*Description*
0	*Release Year*
0 - 12	*Timbre Average*
13 - 90	*Timbre Covariance*

III. Data Analysis

In this section, we present some initial analysis on dataset. The purpose of this step is better understanding of data and features.

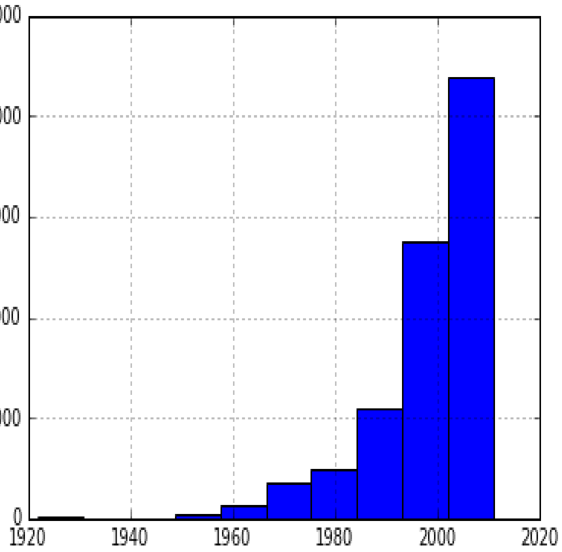


Figure 1: Histogram of music release by decades.

We start by plotting the release year histogram, presented on Fig. 1. Visually, we can see that 2000s years has more released musics than others periods. In contrast, data between 1920 to 1950 are basically irrelevant compared with other decades.

Another question that arose was about the range of the features in dataset. For better understand this, we plot in Fig. 2 the histograms of each one of the 12 timbre average features in dataset. As we can see, even 0-centered the features has different ranges. This suggests that a normalization may be necessary in future.

Another way that get information about relationships between features and our target variable is plotting the scatter matrix and that's what we did. The Figure 3 shows the scatter matrix of release year and 2 features from dataset. One of them is a timbre average and the other is a timbre covariance feature. This plot shows interesting and rich informations like the growth of the two features along the decades evolution.

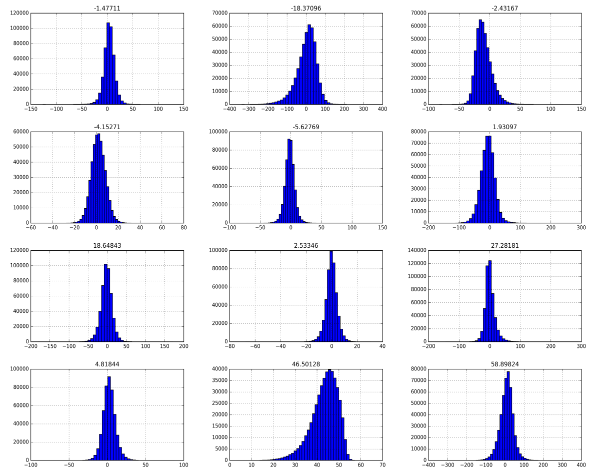


Figure 2: Histogram of 12 timbre average features

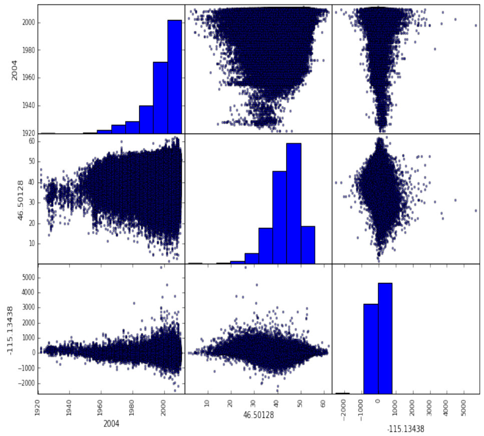


Figure 3: Scatter Matrix plot between release year, one timbre average feature and one timbre covariance feature

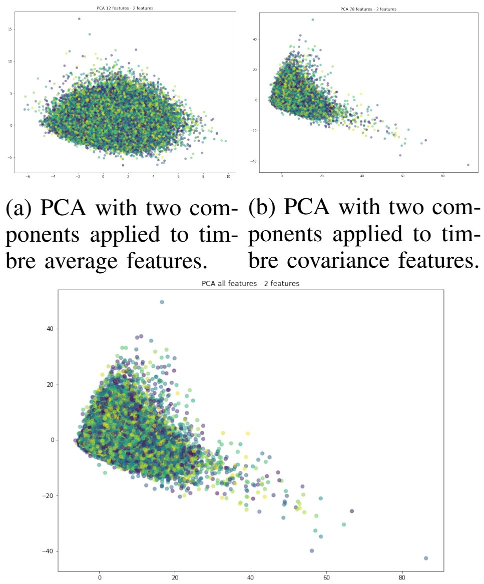
IV. Material and Methods

Since this first step of data analysis was done, we start to apply models that can be relevant for the desired task: release year prediction. In this work was employed five different models for this predictive task. Moreover, we also show an dimensionality reduction algorithm used to handle this large number of features.

A. *Data normalization*

The first procedure used is the data normalization, we decided preprocess the data using the Eq (1), which transforms the data to comparable scale. This process helps the models to better fit the dataset

$$Data = \frac{Data - \mu}{\sigma} \tag{1}$$



(a) PCA with two components applied to timbre average features.

(b) PCA with two components applied to timbre covariance features.

(c) PCA with two components applied to all features.

Figure 4: Application of PCA to different portions of the dataset.

B. *Principal Component Analysis (PCA)*

As we are dealing with a set of large features that describe our data, one way of working with a reduced set of these same features is applying PCA. This technique identifies new variables, the principal components, which are linear combinations of the original variables [2]. This technique can be used to explore data sets in which thousands of variables have been measured. Scikit-learn [3] also presents a probabilistic interpretation of PCA.

In order to visualize better the average features of the dataset, we performed three visualization experiments with PCA, the first was apply PCA with two components on 12 initial features (See Fig. 4a), the next one was apply PCA to 78 features of timbre covariance (See Fig. 4b) and the last one, we apply to all features in dataset (See Fig. 4c).

As can be seen in Fig 4, there is a big difference between the distribution of the data when analyzing the data of average and covariance separately. When visualizing the PCA performed on all the data, we can see that the adopted distribution is almost equal to the obtained when applying the covariance data, for which we believe that it is due to the amount of data is very unbalanced (12 features of average and 78 features of covariance). Our initial hypothesis is that we believe that the covariance data set contributes more noise than information because of the number of outliers.

C. *Baseline*

For the baseline model, we adopted the most frequent year in the training data as the predicted year in our test dataset. In other words:

$$\hat{y} = g(train[\text{`}year'']) \tag{2}$$

where $train[\text{`}year']$ is a vector that contains a histogram of years in train dataset, and g is a function that return the year with more probability in histogram vector. The most frequent year found was 2007.

This is the most naive approach, we use this as a baseline model to compare with another more complex models.

D. Linear Regression

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable y. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable [4] . The idea is formulate a hypothesis h that maps x's to y's. In order to evaluate the model, we want to minimize the following cost function:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i)^2 \qquad (3)$$

Two approaches for cost minimization was explored in this work. The first one was the Normal Equation method, where the problem is solved analytically. In other hand, Gradient Descent solves this optimization problem by choosing initial values for θ_0 and keep changing them to reduce the $J(\theta)$ until end up at a minimum. We choose the Stochastic Gradient Descent (SGD) for our Linear Regression gradient-based. Instead using all training set to compute the next updates, SGD uses a few training examples sampled of the total training set.

E. Polynomial Regression

Our hypothesis is evaluated in the same way as linear regression approach, but now we add some polynomial relationships on the equation. As an example, if in linear regression first model we had:

$$h_\theta = \theta_0 + \theta_1 x \qquad (4)$$

now, choosing degree = 3, we have:

$$h_\theta = \theta_0 + \theta_1 x + \theta_1 x^2 + \theta_1 x^3 \qquad (5)$$

V. Experiments and Discussion

In this subsection, we analyze the experimental evaluation results for the models presented previously. As a first step, we performed a simple normalization on dataset by reducing each value from features of mean and dividing by its standard deviation. This process was performed separately for timbre average and covariance. For all models, we start by splitting the data into random train (75%) and validation subsets (25%) from the original training set. The evaluation metric adopted for this predictive task was Mean Absolute Error(MAE) which is defined as:

$$MAE = \frac{1}{n} \sum_{t=1}^{n} |y_t - \hat{y}| \qquad (6)$$

We build and evaluate all methods cited previous using the scikit-learn library [3]. For Linear Regression (Normal Equation and Gradient-based) and Polynomial Regression approaches, we evaluate the performance of the models in 6 different ways:

I using all average features
II using 2-dimensional PCA in average features
III using 6-dimensional PCA in average features
IV using 10-dimensional PCA in covariance features
V using 20-dimensional PCA in covariance features
VI using the best of average and covariance approaches concatenated

For each analysis of Stochastic Gradient Descent we evaluate the performance of five different learning rates (0.0001,0.0003,0.0004,0.0005 and 0.0007).

For Linear Regression using SGD, our analysis presented best results approach (VII). Figure 5 shows the evaluation of the error along the learning rate changes on this approach. The results lie between 6.8819 to 6.8867. As the differences are so small, the y-axis presented small values. We also evaluate the error along the iterations of the model and this result is presented in Figure 6. The learning rate used for the SGD was the best obtained in previous analysis.

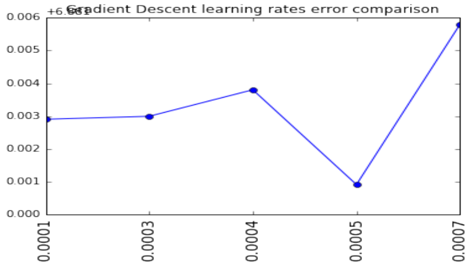


Figure 5: Learning rate vs error variations on Linear Regression with Gradient Descent

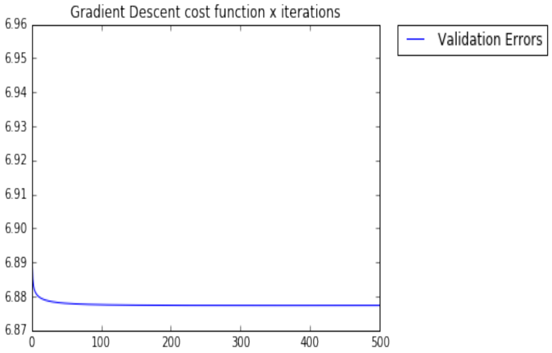


Figure 6: Cost functions x iterations using SGD

Next, we study some regularizations on Gradient-based Linear Regression for model improvement. Three types of regularization were analysed. L1 regularization adds an penalty equal to the absolute value of the magnitude of coefficient, L2 adds an penalty equal to the square of the magnitude of coefficients and Elasticnet linearly combines L1 and L2 regularizations. For these experiments, L2 regularization performed better than

Table III: Mean absolute errors for validation set

Models comparison		
Model	*approach*	*MAE*
Baseline	*(I-VI)*	*9.07861708459*
LR (SGD)	*(VI)*	*6.88191787269*
LR (SGD)+L2	*(VI)*	*6.88239829581*
LR (Normal Equation)	*(VI)*	*6.79413057717*
Polynomial Regression	*(VI)*	***6.5913741091***

Table IV: Mean absolute errors for test set

Models comparison		
Model	*approach*	*MAE*
Baseline	*(I-VI)*	*9.19424539742*
LR (SGD)	*(VI)*	*6.91772654527*
LR (SGD)+L1	*(VI)*	*6.91759536785*
LR (Normal Equation)	*(VI)*	*6.85143598996*
Polynomial Regression	*(VI)*	***6.64008162802***

other approaches but the differences and very small. The Figure 7 shows the evaluation of the regularizations along the learning rate changes on the approach (VI) that presented best results.

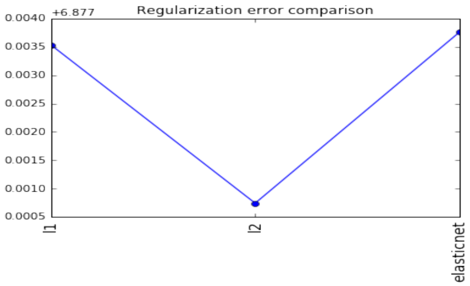


Figure 7: Regularization strategy vs error variations on Linear Regression with Gradient Descent

As mentioned before, all models was evaluated in I - VI approaches but we present in this report only the best result obtained. The Table III shows the comparison between all methods. Firstly, we evaluated our baseline, that predict always the most frequent year (2007) on train set as the output target, resulting in MAE=9.0786170. Consequently, if any posterior error is next of this error, the model can be rejected. An unexpected was the performance of Linear Regression with Normal Equation, presenting lower errors compared with GD-based approaches. Polynomial regression with degree=2 showed better results compared with all other methods and was obtained concatenating the features of all timbre averages with 20-Dimensional PCA in timbre covariances. This shows that our best model obtained this results representing the data in a quadratic way.

Finally, we predicted our best models presented before in the test set. Table IV show the results for the best models obtained in validation set. Polynomial Regression still getting better results than other models. It is important to note that the error for Polynomial Regression is smaller than for Linear Regression using Normal Equation in validation set, that was our second best model. Now the regularization L2 show a significant improvement compared to SGD approach without any normalization.

A. Experiments configurations

All the experiments performed with gradient decent used a maximum amount of iterations of 1000. The loss function used to solve the optimization problems was the squared loss function. To find the best learning rate we do a search of parameters, in this search we first test with classical values for learning rate (1, 0.1, 0.01, 0.001, 0.0001 y 0.00001). Then with the obtained results, we observed that the best results were on the scale of 0.0001, so we performed another more localized search using values 0.0001, 0.0002, 0.0003, 0.0004 y 0.0005. For the search of the polynomial models we used a search of models with degrees 1,2,3 and 4. This search was more limited because of the lack of computational processing of the machines used to run our experiments.

VI. Conclusion

The database used for this work has a very large number of features, as it was observed in the preliminary experiments with data, the dataset has many outliers that only provide noise. In general the models that obtained the smallest errors were created from the use of all features of average and the use of reduced version through PCA of the features of covariance, this because covariance features do not provide much information relevant to models.

In general the models will behave as expected, due to the great complexity of the data a model of linear adjustment should not have the best result at the time of predicting the year of the song. The polynomial models obtained the best results because the adjustment they perform is closer to the reality of the distribution of the data. Finally the solution of the problem through the normal equation had a great success when using few features.

When compare the errors obtained with the validation data and later with the test data, we can observe that the increase of the error in the test is not much, so we can say that our models succeed in generalizing the model without reaching have overfitting, which is very important, because these models could be used in real environments.

References

[1] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, “The million song dataset,” in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[2] M. Ringnér, “What is principal component analysis?” *Nature biotechnology*, vol. 26, no. 3, pp. 303–304, 2008.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[4] Y. Edu. (1998) Linear regression. [Online]. Available: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm