COMP90049 Project 2 Music Genre Classification

1 Introduction

To provide researchers with a reference data set for evaluating research value, the Columbia University Laboratory released an extensive music data set that includes the audio characteristics and metadata of one million songs (Bertin-Mahieux et al. 2011). Moreover, these data sets are almost entirely labelled, which makes it very suitable for unsupervised learning and the development and testing of classification methods. Then, to summaries the input characteristics on the time scale with musical meaning. Convolutional networks are proposed and used to classify the key, genre and author of songs in the million-song data set (A. Schindler and A. Rauber 2012). In the classification process, majority of data is used in the unsupervised learning phase. The experimental results can provide useful feedback on the realized benefits of the convolutional architecture and the additional benefits found in unsupervised pre-training.

Based on the above existing foundation, this paper is first to introduce support vector machines, logistic regression and Naive Bayes. Section 4 presents data preprocessing and feature selection. Section 5 shows the experimental results of different machine learning algorithms, and finally, use test validation and test data set to evaluate and draw conclusions.

2 Related literature review

The convolutional recurrent neural network shows strong performance in terms of the number of parameters and training time, which shows this hybrid structure has high effectiveness when it is necessary to extract and summarize music features (Fazekas et al. 2017).

The improvement of machine learning algorithms can improve the performance of machine learning to reduce the workload of researchers. Although significant progress has been made in this field, there are still apparent limitations in machine learning data sets. But in the continuous development process, these problems can be corrected by continually updating the data set (Dey et al. 2015).

3 Data set

Features and class labels are derived from the millions of songs data set of T. Bertin-Mahieux, D. P.W. Ellis, B. Whitman, and P. Lamere (2011) and the data set which was published by A. Schindler and A. Lauber in the 10th International Adaptive Multimedia Retrieval International Symposium (2012).

The data set contains the audio, metadata and text features of 8556 songs, as well as their genre tags. The songs are divided into a training set (7678 songs), development set (450 songs) and test set (428 songs). The detailed data is shown in table1.

Corpus	Features	
Training set	train_features	7678 instances
	train_labels	7678 instances
Development set	valid_ features	450 instances
	valid_ labels	450 instances
Test set	test_features	428 instances

Table 1- The feature of data set

4 Methodology

4.1 TF-IDF

Term frequency-inverse document frequency (TF-IDF) is a statistical method for evaluating the importance of words in a document set or corpus through a weighted approach. Among the data that needs to be processed in this assignment, tags and titles are both textual data. It is suitable for use as a characteristic word when identifying song types.

The expression process of the term frequency-inverse document frequency (TF-IDF) is shown below. The implementation of python code is edited based on this idea.

Term frequency (TF) indicates the frequency of the term keyword in the text.

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where

- $n_{i,j}$ is the number of times the term appears in the csvfile.
- $\sum_k n_{k,j}$ is the sum of the number of occurrences of all words in the csvfile.

Inverse Document Frequency (IDF) is calculated by dividing the total number of documents by the number of documents containing the term and then taking the logarithm of the obtained quotient.

$$idf_i = \log\frac{|D|}{|\{j: t_i \in d_j\}|}$$

where

$|D|$ is the total number of files in the corpus.

$|\{j: t_i \in d_j\}|$ is the number of files containing the term (t_i).

TF-IDF can be expressed as TF * IDF. Therefore, the word frequency and the file frequency in the entire file set jointly determine the weight of TF-IDF. In the data preprocessed by TF-IDF, common English words will be filtered, and keywords that can be classified will be retained.

4.2 Normalization

Normalization is to limit the required data to a specific range after processing. First of all, normalization can facilitate subsequent data processing, and secondly, it can ensure faster convergence when the program runs. The specific function is to summarize the statistical distribution of a unified sample (A. Singhal et al. 2017).

4.3 Zero-R baseline

Zero-R is a classifier that only performs classification based on historical data statistics. It will only select the category with the highest probability in the data set as the classification result. Therefore, for any data set, the accuracy of the classification prediction value is consistent. Consequently, it does not have any predictive power, but its baseline performance can be used as the standard of other classifiers. Only when the selected model outperforms a baseline, the model can be used for classification.

4.4 Naive Bayes

Under the assumption that the features are strongly independent, the Bayesian classifier will use Bayes' theorem to classify the data. The principle of classification is first to calculate the prior probability, and then use the Bayesian formula to calculate the posterior probability. The maximum posterior probability will be used as the classification result. The Bayesian formula is shown as follows:

$$P(C|X)=\frac{P(C)\,P(X|C)}{P(X)}$$

where

P(C) is the prior probability.

$P(C|X)$ is the class conditional probability of sample X relative to class label C.

P(x) is the evidence factor used for normalization.

4.5 Logistic regression

Logistic regression is a generalized linear regression analysis model. Its independent variables can be continuous or categorical. The hypothetical function of logistic regression is shown below:

$$\log\frac{P((Y|X_1X_2X_3))}{P((\bar{Y}|X_1X_2X_3))}=w_1x_1+w_2x_2+\ldots+b$$

where

w is the sum of feature weights.

x is the feature value.

5 Results and analysis

In this paper, zero-R is used as the baseline classifier. The statistics show that ‘classic pop and rock’ is the label with the most appearances, and it appeared in 1,629 songs in total. In the end, the accuracy of zero-R is 0.122. Meanwhile, to satisfy the normalized evidence factor used in the Bayesian classifier and facilitate data processing and observation. All data will be normalized, and the normalized range is set between 0 and 1.

5.1 The influence of TF-IDF parameters

Changes in the values of max_df and min_df in the TfidfVectorizer parameter will affect the accuracy of the classifier. max_df refers to ignoring the terms whose document frequency is higher than the given threshold. min_df refers to ignoring the terms whose document frequency is lower than the given threshold. Their value range is between 0.0 and 1.0, and the default value is 1.0. The floating-point value represents the scale of the document. We will use Multinomial Naive Bayes, logistic regression and the controlled variable method to calculate accuracy. The results are shown in table 2 and table 3 below.

max_df (default value)	Multinomial Naive Bayes (Accuracy)	Logistic regression (Accuracy)
min_df		
0.1	0.567	0.591
0.2	0.567	0.596
0.3	0.573	0.596
0.4	0.571	0.593
0.5	0.573	0.593
0.6	0.573	0.593
0.7	0.573	0.593

Table 2 The accuracy of logistic regression and Multinomial Naive Bayes under different min_df values.

After setting max_df as the default value, we tested the accuracy of different models with an interval of 0.1 from 0.1 to 0.7. The results show when the value of min_df increases from

0.1 to 0.3, the accuracy of both models is increasing. But when min_df exceeds 0.3, the accuracy gradually decreases, and after it exceeds 0.5, it remains unchanged.

The conclusion is when min_df is tiny, the words in the training set are still very mixed, and the interfering words are not removed, so the model cannot effectively learn information from the data. However, when min_df is too large, keywords with learning value have been removed, which leads to a decrease inaccuracy.

Therefore, we set 0.3 as the best min_df and set it as a fixed variable when measuring different max_df to find the best accuracy of the models.

min_df =0.3 max_df	Multinomial Naive Bayes (Accuracy)	Logistic regression (Accuracy)
0.1	0.467	0.631
0.09	0.464	0.6178
0.08	0.48	0.6333
0.07	0.53	0.62
0.06	0.54	0.64
0.05	0.56	0.6378
0.04	0.584	0.6333
0.03	0.5977	0.6267
0.02	0.5956	0.6289
0.01	0.6289	0.6311
0.005	0.6422	0.6311
0.001	0.6622	0.6111
0.0005	0.6444	0.6267
0.0001	0.5933	0.5956

Table 3 The accuracy of logistic regression and Multinomial Naive Bayes under different max_df values.

For Multinomial naive Bayes, when max_df is 0.001, its highest accuracy is 0.6622, and as max gradually decreases, the accuracy presents a trend of first increasing and then decreasing. But with the change of max, the accuracy of logistic regression did not show any regular trend.

By calculating the average of the accuracy of the two models, 0.005 is set as the best max value.

5.2 The influence of different Naive Bayes

The following table 4 shows the accuracy of the 3 different Naive Bayes models.

Naive Bayes models	Accuracy
Gaussian naïve Bayes	0.45
Multinomial naïve Bayes	0.64
Bernoulli naïve Bayes	0.56

Table 4 Accuracy of the different Naive Bayes models

Through the observation of the training set, we found that there are both continuous and discrete variables in the feature. However, continuous variables do not meet the ideal Gaussian distribution, and discrete variables do not meet the Bernoulli distribution. Therefore, compared with Multinomial naive Bayes, the accuracy of the other two models is lower.

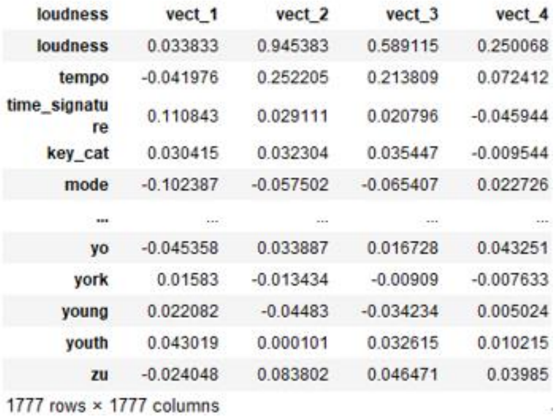
The data filtered by TF-IDF usually only has about 1 to 3 features. The remaining data features basically satisfy the Bernoulli distribution, which is why the accuracy of Bernoulli naive Bayes is higher than that of Gauss naive Bayes.

5.3 The influence of different models

Based on table 3, the accuracy of logistic regression and Naive Bayes is different under the same variable. The reason for the difference is that Naive Bayes requires conditional independence assumptions, which can directly use the logistic occurrence ratio of each feature as weights. In contrast, logistic regression does not need to meet the conditional independence assumptions, and the gradient descent method needs to be used to obtain coupling information between features，thereby getting the corresponding weight.

6 Error analysis

Figure 1 shows the correlation between some features.



loudness	vect_1	vect_2	vect_3	vect_4
loudness	0.033833	0.945383	0.589115	0.250068
tempo	-0.041976	0.252205	0.213809	0.072412
time_signature	0.110843	0.029111	0.020796	-0.045944
key_cat	0.030415	0.032304	0.035447	-0.009544
mode	-0.102387	-0.057502	-0.065407	0.022726
...	...	...	...	...
yo	-0.045358	0.033887	0.016728	0.043251
york	0.01583	-0.013434	-0.00909	-0.007633
young	0.022082	-0.04483	-0.034234	0.005024
youth	0.043019	0.000101	0.032615	0.010215
zu	-0.024048	0.083802	0.046471	0.03985

1777 rows × 1777 columns

Figure 1 Correlation between features

It can be seen from the figure that there are different degrees of correlation between various features. The individual correlation has reached a medium correlation or even a strong correlation. It seriously does not meet the conditional independence assumption of Naive Bayes. Therefore, it is not wise to use a Naive Bayes classifier to classify this data set.

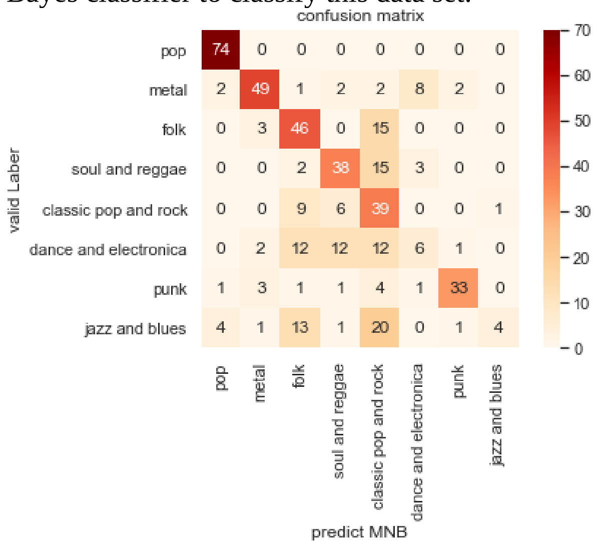


Figure 2 Confusion matrix of Multinomial naïve Bayes

The Multinomial Naive Bayes model has very terrible prediction results for 'dance and electronic music' and 'jazz and blues'. The 'Jazz and blues' is often mistakenly predicted as "classic pop and rock".

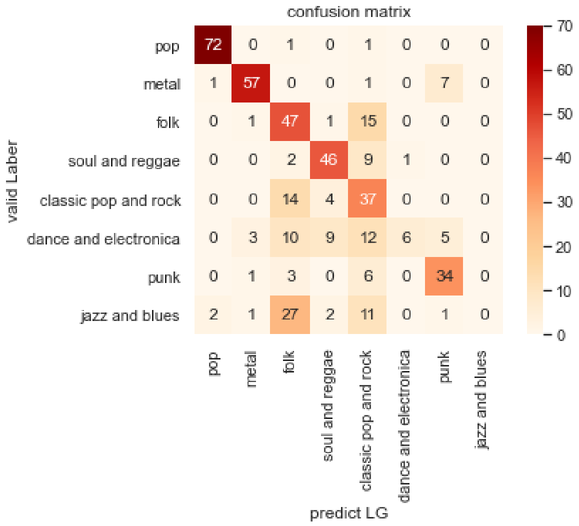


Figure 3 Confusion matrix of Logistic regression

The overall prediction result of Logistic regression model is slightly better than Multinomial Naive Bayes model. The Logistic regression model has incredibly poor prediction results for 'dance and electronica' and 'jazz and blues'. Especially for 'jazz and blues', the prediction results are all wrong, and "Jazz and blues" is often mistakenly predicted as "folk".

7 Conclusions

In this paper, we compared the performance of the naive Bayes model and the logistic regression model and discussed the impact of different TF-IDF parameters on accuracy. Finally, confusion matrix and correlations are used to perform error analysis on the model and data.

References

T. Bertin-Mahieux, D. P.W. Ellis, B. Whitman, and P. Lamere. The million-song dataset. In Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR), 2011.

A. Schindler and A. Rauber. Capturing the temporal domain in Echonest Features for improved classification effectiveness. In Proceedings of the 10th International Workshop on Adaptive Multimedia Retrieval (AMR), 2012.

K. Choi, G. Fazekas, M. Sandler, & K. Cho. Convolutional recurrent neural networks for music classification, 2017.

S. Das, A. Dey, A. Pal, and N. Roy. Applications of artificial intelligence in machine learning: review and prospect. International Journal of Computer Applications,115(9),2015.

A. Singhal, C. Buckley and M. Mitra. Pivoted document length normalization. In Acm sigir forum (Vol. 51, No. 2, pp. 176-184). New York, NY, USA: ACM, 2017