COMP90049 Project 2: Movie Genre Prediction

1 Introduction

In the past few years, the film industry has developed rapidly and produced many movies. So it has been much harder for people to choose movies that can fit their taste. As a result, movie genre prediction becomes more important for users to choose interesting movies. Many researchers have done some related work. Gabriel (2016) uses Neural Network to analyze movie videos for movie genre prediction. Deldjoo (2018) and Harper (2015) work on the movie feature datasets. And this report is based on these datasets.

This report will first discuss the predictability of subsets of given features. Then compare the performance of linear classifiers (Logistic Regression and Naive Bayes) and a non-linear classifier (Neural Network) by using evaluation metrics. And what learning rate and what hidden layer structure of Neural Network can achieve better prediction is also discussed in the report.

2 Feature Analysis

2.1 Feature Pre-processing

The original data set contains 130 features, which includes 3 metadata features, 107 visual features and 20 audio features. For visual and audio features, they are precomputed and not interpretable, so they do not need to be preprocessed. For tags feature in metadata features, it is a massive data feature and each tag feature contains one or multiple tags separated by commas. There are 200 unique kinds of tags in this column and it is meaningless if it is considered as a single feature. So It should be expanded to 200 features and if the movie contains a tag, the corresponding tag feature should be marked as 1 otherwise 0. For title feature, it can be converted from string into integer. And for year of release feature, there exists bad data and missing data, and the instance with bad data should be removed.

For now, the processed data set has 329 features, including 202 metadata features (title, year and 200 tag features), 107 visual features and 20 audio features.

2.2 Predictability of Feature Subsets

In the initial baseline experiment, Logistic Regression is implemented to train the model with all features. But the accuracy is not ideal. And this comes up with a question that the different subsets of features may have different predictability. And Table 1 shows the accuracy of different subsets of features with Logistic Regression Classifier.

Metadata Features	Visual Features	Audio Features	Accuracy
✓			38.13%
	✓		24.41%
		✓	19.40%
✓	✓		40.13%
	✓	✓	30.43%
✓		✓	38.80%
✓	✓	✓	39.46%

Table 1：Accuracy of different subsets of features with Logistic Regression Classifier

If only one of three different types of features can be chosen, metadata features perform best, visual features follow, audio features worst. Simply because the number of metadata features is much higher than the other two. And the audio features only have 20 features, which is not enough for training and leads to low accuracy.

As for the combination of two types of features,

the combination of metadata and visual features performs best with the highest accuracy over all kinds of combinations. This is mainly because these two features account for most of the number of feature values (309 out of 329) and contain enough information for prediction.

If all features are chosen (metadata, visual and audio), it performs worse than the combination of metadata and visual features. And the reason behind this is that audio is precomputed and unreadable by humans, which may entrain some noise that is not visible to humans, resulting in reduced accuracy.

In summary, the combination of metadata and visual features is most predictive.

3 Linear Model Comparison

Logistic Regression and Naive Bayes are both linear expressions of features. So which linear model can perform better will be discussed. And Table 2 shows the evaluation between Logistic Regression and Naive Bayes. Movie Genre Prediction is a multi-class problem, macro-averaging and weighted-averaging are used for evaluation.

	Logistic Regression	Naïve Bayes
Accuracy	38.8%	11.04%
Precision(Macro)	34.40%	27.79%
Recall(Macro)	28.28%	19.09%
F-Score(Macro)	31.00%	22.63%
Precision(Weighted)	42.38%	42.85%
Recall(Weighted)	38.80%	11.04%
F-Score(Weighted)	40.51%	17.56%

Table 2：Comparison between Logistic Regression Classifier and Naïve Bayes Classifier

The accuracy of Logistic Regression is much higher than that of Naive Bayes. There are mainly two reasons for that, one reason is because of conditional independence assumption for Naive Bayes. In such an assumption, features are all conditionally independent, but in the real world, it is almost impossible that we have a set of data in which features are completely independent, which causes low accuracy in prediction. Take this movie data set as an example. It cannot be said that visual and audio are completely independent because visual and audio are interrelated in a movie. Another reason is that the movie data set is massive and Logistic Regression can perform better, because Naive Bayes is suitable when the data set is relatively small, and Logistic Regression is suitable when the data set is relatively large.

The weighted-averaging precision of these two classifiers is higher than macro averaging precision. The main reason for this phenomenon is the movie data set is the validation data set is unevenly distributed. In the validation data set, “Romance” has 51 instances and “Adventure” has only 2 instances. The weighted-averaging takes the proportion of instances in every class into consideration and treat each class indiscriminately. And the difference between weighted-averaging precision and macro averaging precision also illustrates that the precision is very high in classes with a large number of validation instances.

For both Logistic Regression and Naive Bayes, the recall is always lower than precision. From the perspective of prediction results, Precision describes how many of the positive results predicted by the classifier are true positives, such as, how many of the action movies predicted by the classifier are real action movies; from the perspective of real results, Recall describes how many of the real positive examples are selected by the classifier, such as, how many of the real action movies are recalled by the classifier. Precision and Recall are a pair of conflicting performance metrics.

4 Non-linear Model Comparison

4.1 Performance of Neural Network

Logistic Regression and Naive Bayes both work in the linear field cannot learn the non-linear content in features. But Neural Network can learn the non-linear content and the main reason is that Neural Network imports non-linear activation function and gain higher learning ability. So whether a nonlinear

model with high learning ability can perform better will be discussed. The Neural Network used for comparison has a structure of two hidden layers, each with 50 neurons[1].

	Logistic Regression	Neural Network
Accuracy	38.8%	43.1%
Precision (Macro)	34.40%	31.41%
Recall (Macro)	28.28%	28.64%
F-Score (Macro)	31.00%	29.96%

Table 3：Comparison between Logistic Regression Classifier and Neural Network

From the above table, it can be seen that Neural Network has higher accuracy. And this is because Neural Network has the ability to learn and build models of non-linear complex relationships, such as in this movie data set, the relationship between input movie features and output movie genre is nonlinear and complex.

4.2 Effects of Learning Rate

The neural network model is trained by the gradient descent algorithm, and the learning rate determines how far the weight should move in the gradient direction in a mini-batch. To what extent the learning rate can affect neural network will be discussed.

Learning Rate	Loss	Accuracy
0.1	2.7	17.1%
0.01	2.14	28.1%
0.001	1.61	41.1%
0.0001	1.86	38.8%
0.00001	2.56	20.4%

Table 4：Effects of Learning Rate in Neural Network

Table 4 is the variation of the loss and accuracy with the learning rate under the premise of a fixed number of steps (2000 steps).

If the learning rate is high, for example, if the learning rate is set to 0.1, the training may not converge at all, or even diverge. The amount of weight change may be so large that the optimization crosses the minimum and makes the loss function worse.

When the learning rate is from 0.1 to 0.001, the loss value moves toward the lowest point and is more accurate. When the learning rate continues to decrease from 0.001, the training will become more reliable and accurate, but the optimization will take a longer time because each step towards the minimum of the loss function is small. When the number of steps is limited, the lowest point of loss cannot be reached.

In summary, The learning rate is one of the most important parameters for adjusting the neural network. Too large or too small will make the neural network perform poorly. Choosing an optimal learning rate can make the neural network perform well and fast.

4.3 Impact of Hidden Layer Structure

The structure of the hidden layer in the neural network determines the ability of the neural network to learn. The more hidden layers, the stronger learning ability of the neural network. So whether the more complex neural network can be more predictable for movie genres will be discussed.

Hidden Layer Structure	Loss	Training Accuracy	Validation Accuracy
10*10	2.02	37.5%	30.4%
20*20	1.81	45.1%	40.1%
40*40	1.44	60.1%	41.1%
60*60	1.44	60.1%	42.1%
100*100	1.32	67.9%	39.8%
80*80*80	1.24	69.5%	34.8%

Table 5：Effects of Hidden Layer Structure

[1] Implemented by TensorFlow Library (tensorflow.org)

As can be seen from Table 5 that a neural network with a complex hidden layer structure has a stronger learning ability and a lower loss value. However, a neural network with an excessive learning ability is not a good thing, which will make the neural network can learn more features in the training set to better fit the training set so that it performs well in the training set but performs poorly in the validation set. This phenomenon is called overfitting. The phenomenon of overfitting is due to the fact that during the training process because the training set includes sampling errors, the complex neural network model will also take sampling errors into account. This will lead to poor generalization ability of the model.

4.4 Error Analysis

There are many reasons that the accuracy is not high enough. One is that the learning ability of the neural network is too strong and then cause over-fitting, it will lead to a large gap between the accuracy of the training set and the accuracy of the validation set.

Another reason may be due to the given features. For visual and audio features, they are precomputed and not interpretable, so it is not clear whether these features contain noise. And for tags feature, some movies have a tag, and some movies have more than a dozen tags. This inequality of information will cause bias and lead to low accuracy.

5 Conclusions

In summary, this report evaluated different subsets of functions, and the combination of metadata and visual functions can achieve the best prediction. In linear classifiers, Logistic Regression outperforms than Naive Bayes for accuracy, recall and F-Score. And this is because conditional independence assumption may limit Naive Bayes and Logistic Regression is suitable for massive instances. And Neural Network, as a non-linear classifier, performs better than linear classifiers for movie genre prediction, for the reason that Neural Network can learn non-linear content existing in movie features and have stronger learning ability. Neural Network should take the appropriate learning rate and hidden layer structure in order to obtain a better prediction of the movie genres and avoid over-fitting.

References

Gabriel, S. Simões, J. Wehrmann, R. C. Barros and D. D. Ruiz. 2016. Movie genre classification with Convolutional Neural Networks. 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp.

Deldjoo, Yashar and Constantin, Mihai Gabriel and Schedl, Markus and Ionescu, Bogdan and Cremonesi, Paolo. MMTF-14K: A Multifaceted Movie Trailer Feature Dataset for Recommendation and Retrieval. Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018,Amsterdam, The Netherlands, June 12-15, 2018F.

Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015),19 pages.

Lea Frermann. 2020. COMP90049 Introduction to Machine Learning Lecture 5-9.

Hagan, Martin T., Howard B. Demuth and Mark Beale.1995. Neural Network Design.

Leslie N. Smith. 3 Jun 2015. Cyclical Learning Rates for Training Neural Network.