

Self-reflection of Movie Genre Classifier Project

Name: Zongcheng DU

Student ID: 1096319

After reading the reports of the two classmates and thinking about the reviews of the two peers, I realize that my models and reports may be improved in many ways.

In the model, I directly use the combination of textual features, visual features and audio features as input. In subsequent experiments, I will try to separate the data set to only textual data, visual data, audio data or a combination of both. By comparing the prediction results of different training sets, I can choose the best training set, which can improve the accuracy of my model. Besides, during data preprocessing, it is dangerous to delete some columns just because I think they have nothing to do with the result. I can use principal component analysis to generate combined features. In this way, I can not only reduce the dimensions of the training set but also reduce the possibility of overfitting. Finally, I should try to analyze the training difference of the same model under different parameters. By adjusting the training function or kernel, my model can perform better when processing the dataset containing multiple data types.

In the report, I should clearly state what features I used as input to avoid misunderstandings when reading. Besides, I should explain the difference between various parameters. It might be better to use a mathematical model to describe the logic. Numerical results such as accuracy are better demonstrated in tables, which enables the reader to have a distinct view. Any graphs or charts also make the report more vivid, understandable and readable. I might try to apply the learning curve which assists the parameter tuning process a lot. When representing parameters and formulas, a specific format should be used. For example, the parameters are in italics and the formulas are in templates. These are very small details but can make it easier for readers to understand the report.

Finally, I must reflect on the time allocation in this assignment. I have a month to conduct the experiment and complete the report. However, due to unreasonable time allocation, I only conducted a detailed and in-depth study on the two models of Naive Bayes and SVM. For KNN and random forest, although I have implemented the code, my understanding of the model parameters is not deep enough.