Machine Learning Assignment 3 Reflection on the individual classification task

Date: May 2021, Semester 1

Author: Aoi Fujii (1086220)

Word count: total 574 words

1. The process of completing this project. (281 words)

- Firstly, I have read relevant past research to get some idea on conducting a research project with an effective analytical report.

- I also went through lectures, tutorials and online to improve the conceptual understanding of different classifiers.

- Then, I looked into the dataset to check its distribution and further researched the transformation methods for text features, including CountVectorizer, TfidfTransformer and doc2vec. The imbalanced distribution became a factor to use f1-score as the main performance measure.

- The steps to complete this project was planned out with sufficient analysis of the performance of the respective classifiers (including how to split the data, type of feature selections and text feature conversion method to use).

- Feature selection was conducted, and the performance was compared for each method. The reasons for the resulting performance were considered in depth.

- Next, I experimented with parameter tuning and visualised f1-score with different feature size. The best feature size and parameters were utilised for the final model.

- The learning curve of each method on train and test set with different train size was visualised, and the error analysis was conducted for each model with a confusion matrix.

- I also experimented with the stratified sampling method to reduce bias and variance, but this was not useful as the dataset chosen by random holdout was already stratified.

- Lastly, the process in the report was summarised with graphs. I presented to show the result of each step separately so that the comparison would be clear. For example, in the evaluation section, f1-score for each feature selection methods were compared in Table 4.1.1.

- The grammar, spelling, formatting and references were checked in the end.

2. things that are satisfied with and those can be improved in your Stage I deliverables (293 words)

I am satisfied with the clear and concise delivery of the report with effective graphs and proper headings. The analysis of each step of this project was in-depth considering the given time and word limit. The model evaluation was conducted from various aspects including error analysis and visualising learning curves. The suggestion to improve the performance on the imbalanced data (the Synthetic Minority Oversampling Technique) was indicated which was a good extension for future research. The theoretical properties of each model were explained with its performance concisely in the final performance section.

Regarding the feature selection, the sequential feature selection including backward and forward selections could have been conducted. However, this was skipped due to the limited processing power of my computer. In addition, it could have been interesting to explore the correlation of each feature and removed the highly correlated feature for further comparison. These ideas would help investigate more variety of feature selection methods to test the effect on the model performance if time allowed. Furthermore, the analysis of each feature selection methodology and the transformation methods for text features could be discussed in more depth (how each methodology affected the result). However, this was not included as the focus in this report should be more on the model analysis. The distribution of the numerical features could be analysed with numbers to show that it is a normal curve. For a better performance measure, cross-validation could have been used instead of a random holdout as it gives a lower variance. However, considering the significantly long computation time for the SVM classifier on my laptop, the random holdout was utilised for the evaluation. Lastly, the related past research could be analysed and discussed more in-depth to support the argument of this project.