Machine Learning 1-2 Assignment 2021-2022

By Yeram Isaac Cho

**Introduction**

This paper discusses the process and result involved in the Machine Learning 1-2 Assignment. The research is based on the Income Dataset provided by Mustaffa Fatakdawal in Kaggle.com. It consists of predictive features such as an individual’s education, employment status, marital status, etc. and was originally meant to be used to predict whether one’s salary is greater than $50,000. In contrast, the assignment attempts to see whether it is possible to predict one’s gender from this dataset.

**Pre-processing**

The first task was to drop certain features that may interfere with the prediction. Initially, some features were redundant. The relationship of the individual, for instance, consisted of gender-related information such as whether the person is a husband or wife. The next factor was the lack of presence. For example, the income data was dropped because it was not given in the test dataset that is provided along with the train dataset. The feature ‘fnlwgt’ was dropped because there was insufficient information on what is exactly was referring to. Missing values were dropped because there were not that much in terms of ratio.

Considering that the task would be about binary classification, the gender values were converted into 0 and 1. The rest of the categorical values were also converted into numerical binary values with the dummies function of Pandas. This increased the number of columns, which also made the train data have one feature that the test data did not have: ‘workclass-Without-pay’. As such, this was dropped as well.

**Classical Method 1: Logistic Regression**

Initially, a pipeline that consists of a standard scaler and logistic regression feature was created. As this was the first method being tested, a baseline comparison was also done and revealed that the logistic regression model is at least showing a positive result. The pipeline was then applied in a cross-validated grid search with an f1 macro scoring, considering the dataset’s characteristic of having multiple binary labels. This resulted into an accuracy of 0.79 with a relatively better result with predicting males (0.84) over females (0.68).

**Classical Method 2: K-NN Classification**

Like the first classical method, the k-nearest neighbours algorithm was applied with the training data that is scaled with standard scaler. This resulted to a relatively lower accuracy of 0.76 and a similar male to female unbalance (0.82 to 0.62). Therefore, compared with logistic regression, k-nearest neighbours performed less accurately in general. Nevertheless, its results are still notable.

**Classical Method 3: Support Vector Machine**

Like the previous two, standard scaler was applied to the training data before the method. The results of support vector machine were identical with logistic regression. It had the accuracy of 0.78 with an overlapping male to female discrepancy of 0.84 to 0.68.

**Neural Method 1: Shallow Model**

For the neural methods, the train data was initially spitted into train and validation data. The validation data was assigned with the same size as the test data. A shallow model with one dense layer was designed with its activation set as sigmoid, considering that the assignment involves binary classification. Accordingly, a model with stochastic gradient descent and binary cross entropy was set and was fitted with the train and validation data. The validation data showed a stark contrast against the train data with a notable fluctuation in both its accuracy and loss values throughout the epochs. This was a concerning indication in contrast with the model’s resulting accuracy of 0.71.

**Neural Method 2: Deep Model**

For the deep model, a hidden layer was added with the relu activation and 500 neurons. The rest of the settings were identical to the previous shallow model. Besides that, the settings were identical to the previous method with stochastic gradient descent and binary cross entropy. While this lead to an accuracy of 0.73, drastic fluctuations of the validation data were present. Moreover, the precision of female prediction was significantly low with 0.29, making its overall f1-score 0.42.

**Neural Method 3: 1D Convolutions Network**

Compared to the previous two methods, 2 blocks of 1d convolution layer and a hidden layer was added to the model. While the rest was similar to the other methods, Adam optimizer was added instead of stochastic gradient descent for a quicker result. While the validation data still showed fluctuations throughout the epochs, the accuracy of the result was much better than the previous two with 0.79.

**Conclusion**

Considering that the dummy classifier with the strategy of frequency scored 0.67, the shallow and deep neural models do not seem to be performing very well. Logistic regression and support vector machine showed an identical performance that were even higher than K-NN classification. While 1d convolutions network falls short slightly in the female f1 score, its overall result is close. The overall results, however, indicate that the dataset may not be sufficient enough to accurately predict one’s gender.