Logistic regression

Machine Learning for Data Science Homework #2

Vid Stropnik; 63200434

ABSTRACT

In the second homework in the Machine Learning for Data Science class, I worked with logistic regression. More specifically, I was to implement models for Ordinal and Multinomial logistic regression. To analyze the models' accuracy and interpretability, I fitted them on a data set of questionnaire results. In this report, I highlight the lessons learned through solving the proposed exercises.

1 DATASET & FEATURE PREPARATION

Most results presented in this report are derived from experiments, conducted on the provided `datset.csv` file. From the dataset, the features `sex` and `response` were transformed to numerical space, with the target variable `response` taking values in ascending order $\{\text{'very poor', 'poor', 'average', 'good', 'very good'}\} \xrightarrow[map]{} \{0, 1, 2, 3, 4\}$ and `sex` equalling *1* when the respondent answered with *M* and *0* otherwise. From there, the data was pre-processed to translate its mean to 0 and so that it's variance equalled *1*; for easier interpretability of the parameters achieved with optimization.

2 MODEL COMPARISON

As was asked of us in the description, a multinomial and an ordinal logistic regression model was fitted on the described dataset, both including the intercept to correct for the inherent bias of our dataset. Figure 1 compares the average classification accuracies and mean log-losses of the `5-fold` cross-validation, performed on the data.

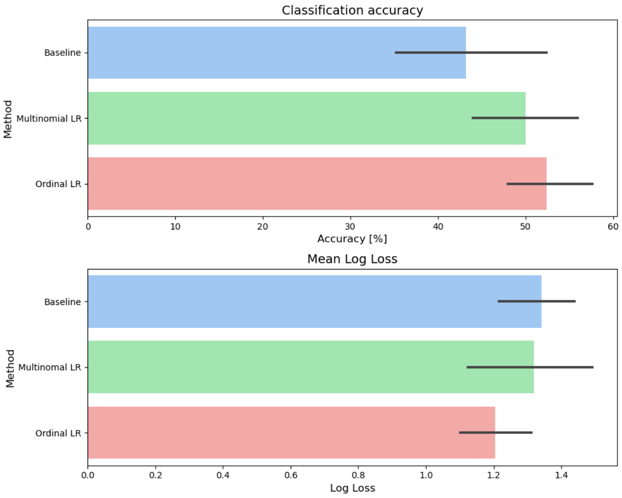

Figure 1: Comparative evaluation of CA & Log Loss on both models, with a provided naive baseline. We observe that the Ordinal LR model is the most suitable to use on the provided dataset, due to the inherent ordering of its classes.

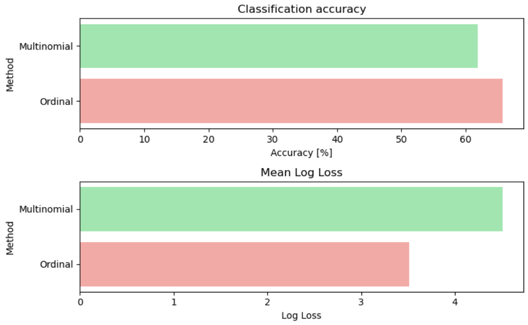

Figure 2: Comparative evaluation of both model on the Intermezzo dataset.

3 INTERMEZZO: CUSTOM DATASET

To find a Dataset that performs better on ordinal data, I randomly sampled 250 values from a normal distribution for each of the four classes. Each class had a mean, higher than the first, while I made sure that the distances between the means were not proportional to one another. For more details, see implementation in Listing 1.

When training the model on just the mean for each class (and scaling the data in the same way as described in Section 1), we achieve results, presented in Figure 2. We can observe that the Ordinal model was able to infer the natural ordering of the input categories. For the intermezzo set, the multinomial log loss is improved upon by *22.3 %* and the accuracy of prediction is *4 %* better.

```r
> c0 <- rnorm(250, mean=4)
> c1 <- rnorm(250, mean=5)
> c2 <- rnorm(250, mean=7)
> c3 <- rnorm(250, mean=8.5)
> df = data.frame(c0, c1, c2, c3)
```

Listing 1: Testing data generation process in R. The dataset was later melted to long format using Python.

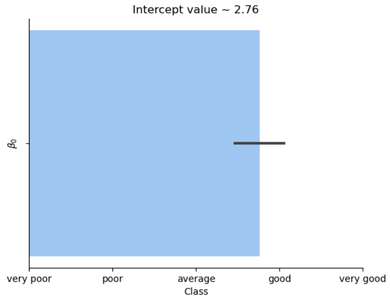

Figure 3: The value of the intercept parameter is useful to quantify the expected questionnaire response.

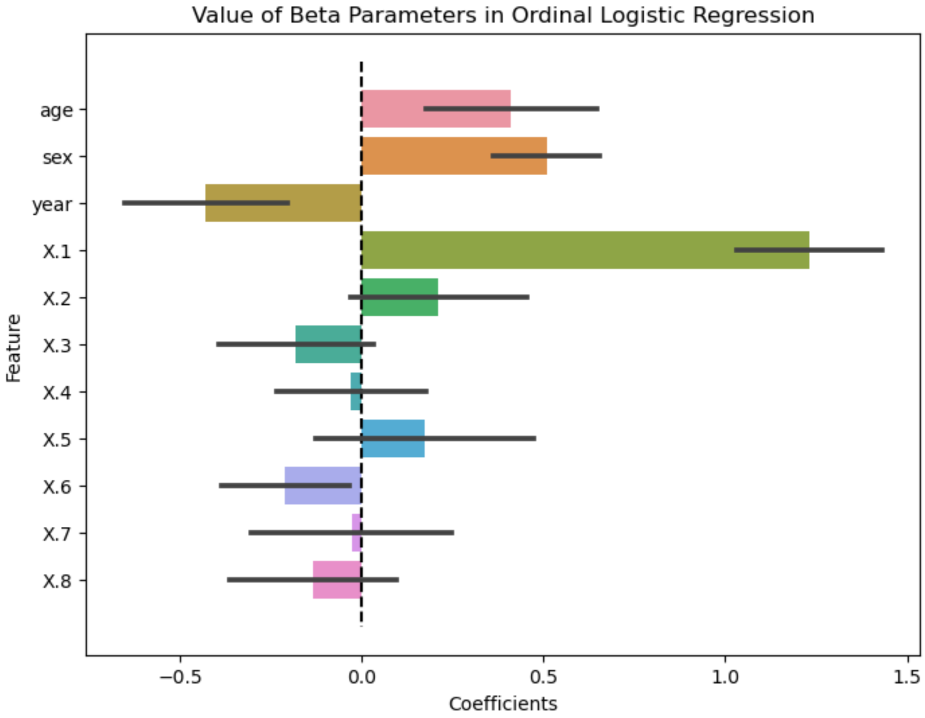

Figure 4: Evaluation of optimized parameters for the Ordinal Regression model. The size of bar denotes feature importance, while the polarity represents positive/negative (decrease/decrease) effect on final score (and classification).

4 INTERPRETATION OF OLR PARAMETERS

Figure 4 shows the values of β parameters in the Ordinal regression model. The uncertainty is quantified by Bootstraping the results from 100 iterations of ordinal regression. The confidence intervals shown are the standard deviations of the bootstrapped parameters, corresponding to each feature. From there, we can observe several interesting factors, the most notable of which I list here:

- A student's success in the `X.1` course highly correlates with their liking of the evaluated one and is by far the most influential feature.
- Older, male students evaluated the course better than their female and/or younger colleagues.
- Students in year 2 liked the course less than those in year 1.
- The uncertainty in all other Features is too high to infer any sort of influence on the respondent's answer.

The intercept parameter is excluded from the visualization due to size. We can use it to analyse the expected value (wrt. to our class) of a respondent's answer to this questionnaire. Figure 3 shows that this is somewhere between `'Average'` and `'Good'`.

5 CONCLUSION

One of the key takeaways from this homework in the realization that one must consider the nature of the analyzed data before selecting the model with which they might analyze it. Here, we showed significant performance improvements after our realization that the provided dataset was, in fact, ordinal.

The results achieved in the last section make sense, seeing as a student's liking of a given course might be highly correlated with their preexisting knowledge in a specific field, while not presenting the classes as complex equations of different fields.

Using bootstrap on the Beta parameters of Ordinal logistic regression turned out to be a great method for not only showing uncertainty, but also interpreting the data.

Having completed this homework, I now feel my understanding of Logistic regression has improved significantly.