Logistic regression

Žiga Trojer - 63200440

1. Models

We implemented multinomial logistic regression and ordinal logistic regression as described in the lecture notes. We used log-likelihood $\text{LL} = \sum_i log(Cat(y_i|\theta))$, where $Cat(y_i|\theta)$ is the categorical distribution. For the optimization of the likelihood, we used fmin_l_bfgs_b method from the scipy.optimize library. Numerical gradients were used.

2. Intermezzo

We created a data set, where ordinal logistic regression performs better than multinomial logistic regression. We sampled from two distributions - normal and exponential. We took 4 values for the response *variance*: {large, medium, low and very low}. For our features, we took:

- feature *normal*: 1025 data points from the normal distribution with the means $\mu \in \{37,42,42,42\}$ and the variance with values $\sigma \in \{7,5,1,0.2\}$, which represents the classes {large, medium, low, very low},
- feature *other_normal*: 1025 data points from the normal distribution with the means $\mu \in \{2,3,4,3\}$ and variance $\sigma \in \{7,5,1,0.2\}$,
- feature *exponential*: 1025 data points from the exponential distribution with the $\frac{1}{\lambda} \in \{\sqrt{7},\sqrt{5},\sqrt{1},\sqrt{0.2}\}$,
- feature *rand*: 1025 data points randomly chosen from range(10).

Basically, we would like to predict how big the class of variance. We provided example from our data set in Table 1.

variance	*normal*	*other_normal*	*exponential*	*rand*
large	39.33	-9.28	3.21	6
mid	46.47	4.44	2.82	4
low	41.95	3.24	2.90	0
very_low	42.37	2.77	0.22	5

Table 1. Sample from our data set.

File multinomial_bad_ordinal_good_train.csv contains data set for training. It contains only 25 rows. File multinomial_bad_ordinal_good_test.csv contains data set for testing (of size 1000). We fitted parameters for both algorithms on train data and predicted the variance class for test data.

On our data set, it makes sense to use the ordinal logistic regression, since there is an order of the response. It is important that this order has a meaning - in our case, the *very_low* class has the lowest variance of data and *large* has the highest variance. Those are two extremes for the variance, therefore, the distance between them is the largest. Class *low* is closer to *very_low* than to *large* class, etc. The data set must have these properties so that ordinal logistic regression can work better than multinomial, or that it makes sense to use this model at all. The next 'condition' for ordinal logistic regression to perform better is to have a small train data set. If we have a large amount of data available, usually multinomial logistic regression performs better, because the model is more complex (we have more parameters that needs to be fitted).

Then why use ordinal logistic regression at all when we have a lot of data? The reason is that the interpretation is very simple. Interpretation of the model is often very important, so we use ordinal logistic regression when possible.

Table 2 shows the performance results for both models. Both in terms of accuracy and log-loss, ordinal logistic regression performed better (we had a very small data set). We tested both algorithms on the similar data set, where we took a larger train set (of size 500) and the multinomial model outperformed the ordinal one.

model	accuracy	log-loss
multinomial	38.6%	1.9712
ordinal	44.5%	1.7971

Table 2. Performance of multinomial and ordinal logistic regression on our data set.

3. Application

For this part, we used the dataset.csv data set. We first preprocessed the data: we transformed feature sex into is_female feature - 0 if *M* and 1 if *F*. Other features were numerical, so there was no need for transforming those. Then we standardized values by removing the mean and scaling to unit variance.

3.1 log-loss

We estimated log-loss of both models using 10-fold cross validation. Indexes for folds are chosen randomly (of course, without repetition). Table 3 shows 95% confidence interval of log-loss for 3 models: multinomial, ordinal and naive. Naive model always predicts $p_i = (0.15,0.1,0.05,0.4,0.3)$. Best performing model is ordinal logistic regression, which has the lowers bounds for confidence interval. Multinomial logistic regression is also better than the naive model, which is a good

sign that the algorithm works as it should. This is a great example where ordinal logistic regression performs better than multinomial logistic regression.

model	log-loss CI
multinomial	1.3003 ± 0.0916
ordinal	1.2131 ± 0.0771
naive	1.3418 ± 0.0621

Table 3. Log-loss 95% confidence interval for 3 models.

3.2 interpretation

We build a model on a whole data set and got the following coefficients for β and average values - see Table 4 and thresholds - Table 5. We start with interpreting the intersection. Because we removed the mean and scaled to unit variance, our job is easier. Coefficient for $\beta_0 = 1.312$, which falls into response *good* - whoever got similar values of parameters like in Table 4, his response was *good*. This means that on average, the response is *good*.

We will try to find out by how much the value of a feature must increase/decrease so that only with this change, the response would be different. With the average values, our model predicts the response *good*. By how much does the *age* have to increase from the average for the model to predict the response *very good* if the other values are not changed? We think that it must increase for 2.5 years (could be wrong, as 2.5 could have some other unit). Such analysis is hard to do, so we will explain it in some other way.

We observe from Table 4 that the highest value for β is for the interception. This indicates that it is very important to include it! Satisfaction is greater for older students on average, but students in second year are less satisfied on average, which is interesting. Maybe there are some students in the first year of study that applied into program later - maybe they took additional year. We don't know the reason behind it. Also, men on average are more satisfied with the course on average (we transformed *sex* into *is_female*). Second largest absolute value of β belongs to *X1*. This feature represents the final grade of the course that students got. It makes sense that this one is very important, since students with better grade are more satisfied with the course than students with lower grade. Grades from other courses are not so important for predictions, as their β's are small in absolute value. We also looked at correlation between grades of this course and other courses, and there is a slightly positive correlation between them. Whoever got good grade at other courses means that he probably got good grade for this course. But those grades are not a good predictor for the satisfaction.

feature	average	β_i
interception		1.312
age	23.54	0.193
sex	M	-0.241
year	1.54 (2nd year)	-0.189
X1	71.34	0.564
X2	77.50	0.105
X3	80.94	-0.098
X4	81.64	-0.012
X5	82.56	0.100
X6	85.37	-0.093
X7	86.75	-0.016
X8	88.79	-0.054

Table 4. β coefficient transformed into odds ratio for the ordinal logistic regression.

response	interval
very poor	(-inf, 0]
poor	(0, 0.274]
average	(0.274, 0.584]
good	(0.584, 1.794]
very good	(1.794, inf]

Table 5. Decision intervals.