MLDS: Homework 3

Andraž De Luisa

March 18, 2020

Given a data set with numeric input (independent) variables x_i in and ordinal target (dependent) y_i with k levels, we decide to use ordinal regression: $(y_i|x_i, \beta, t) \sim \text{Categorical}(p_i)$, where β is the vector of coefficients, t is a vector of thresholds and p_i is the probability vector defined for each class j as:

$$p_i(j) = \phi(t_j - \beta^T x_i) - \phi(t_{j-1} - \beta^T x_i),$$

where ϕ is the inverse logit function.

Part 1 Derive the log likelihood:

$$L(\beta, t, y, x) = \prod_{i=1}^{n} \prod_{j=1}^{k} p_i(j)^{[y_i == j]}$$

$$l(\beta, t, y, x) = \sum_{i=1}^{n} \sum_{j=1}^{k} [y_i == j] \log(p_i(j))$$
$$= \sum_{i=1}^{n} \log(\phi(t_{y_i} - \beta^T x_i) - \phi(t_{y_i - 1} - \beta^T x_i))$$

Part 2 We try to predict the students rating of a completed Master's course from the available information. For comparison, beside the ordinal logistic regression model we use also a simple baseline model (which always predicts $p_i = (.15, .1, .05, .4, .3)$) and a random forest. For evaluation and comparison we use the misclassification rates and (mean of) log-losses. The train-test split is set at 80%.

	Miscl. rate	**Log-loss**
Ordinal regression	0.38	1.039
Baseline	0.42	1.115
Random forest	0.38	1.018

Table 1: Evaluation and comparison of fitted models

In table 1 the performances of the models are shown. One single evaluation is not enough to make some precise considerations, but it seems that both the ordinal regression and random forest perform slightly better than the baseline. Some more precise estimations are shown in table 2, which presents the results of a 10-fold cross-validation. It becomes clearer that the ordinal regression is actually performing better than the baseline, while the random forests seem worse (and less stable, note the bigger standard error).

	Log-loss	**Std. error**
Ordinal regression	1.183	0.100
Baseline	1.342	0.078
Random forest	1.593	0.678

Table 2: Log-loss estimation obtained from 10-fold cross validation

In table 1 the coefficients of the ordinal regression are presented. As could be expected, the grade obtained in the observed course has the most influence to the students' rating. The grades from the other courses are not so important. Note also the positive influence of the sex feature (male students rate the course higher than female) and negative of the year of studying.

Feature	**Coefficient**
Intercept	2.136
Age	0.271
Sex	1.096
Year	-0.615
X.1	1.136
X.2	0.170
X.3	-0.063
X.4	-0.127
X.5	0.201
X.6	-0.072
X.7	-0.051
X.8	-0.027

Table 3: Ordinal logistic regression model coefficients