Ordinal Regression

Course of *Statistical Machine Learning*

Sandro Barissi Marin
Francesco Bollero
Anna Peruso

21st January 2022

1 Introduction

Machine learning scientific community has paid great attention to methods for classification of labelled patterns. However, less consideration has been given to those problems where labels are naturally ordered. For instance, satisfaction surveys could request to rate a film, a restaurant, a teaching assistant, *etc.* on an ordinal scale such as $\{bad, average, good, excellent\}$. Apart from social sciences, naturally ordered classification problems are found in medical research, text classification, facial beauty assessment and many other fields. Precisely because this is a widespread problem, it is important to understand both how to use order information in models and how to appropriately evaluate their performances. In particular, in the following report, logistic regression model and related methods will be investigated and compared through datasets created on purpose; plus, a second part will deal with the application of those models to a real dataset (*white wine quality* dataset [1]).

2 Problem description and Methods

An ordinal regression problem consists on predicting the labels $y \in \mathcal{Y} = \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_Q\}$ of an input vector $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^n$. In addition, it exists an order relation ($\prec$) among the labels, *i.e.* $\mathcal{C}_1 \prec \mathcal{C}_2 \prec \cdots \prec \mathcal{C}_Q$. Under this assumption, each pair of labels can be compared through the order relation ($\prec$): $\forall i, j = 1, ..., Q, i \neq j$ either $\mathcal{C}_i \prec \mathcal{C}_j$ or $\mathcal{C}_j \prec \mathcal{C}_i$. If compared to regression ($y \in \mathbb{R}$), real values can be ordered by the standard operator ($<$). However, in ordinal problems, labels do not carry metric information, *e.g.* you may not be able to measure the distance between *very unlikely* and *somewhat likely* or between *somewhat likely* and *likely*. In the following sections we will present models that exploit the order information to generate more or less complex algorithms.

2.1 Naïve methods

Ordinal regression can be easily reduced to standard problems by making some assumptions. Even if the assumptions may seem strict, these approaches can be very competitive since they inherit the performance of already consolidated models. In particular, one could use regression or nominal classification. In the former case, labels $\{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_Q\}$ are cast into real values $\{r_1, r_2..., r_Q\}$ and then the standard regression is applied. As previously stressed, r_i is arbitrarily selected, thus the performance of the method can be affected by the choice of these values. In the latter case, labels are predicted as if there were not an order. Since the ordering is ignored, more data could be necessary to train the model. In subsubsection 2.2.1, classification through logistic models are further analyzed.

2.2 Cumulative Link Models

The idea behind cumulative link models is to estimate the probabilities $P(y \prec C_q|\mathbf{x})$ in such a way that $g^{-1}(P(y \preceq \mathcal{C}_q|\mathbf{x})) = b_q - \mathbf{w}^T\mathbf{x}$, where $q = 1, ..., Q-1$ and $g^{-1} : [0, 1] \to (-\infty, +\infty)$ is a monotonic function and b_q is the threshold defined for class C_q. In our project, we aim to investigate the logistic model (later referred to as *Ordinal Logistic Regression*, or *Ordinal LM*):

$$\ln\left(\frac{P(y \preceq \mathcal{C}_q|\mathbf{x})}{P(y \succ \mathcal{C}_q|\mathbf{x})}\right) = b_q - \mathbf{w}^T\mathbf{x}, \ q = 1, ..., Q-1. \tag{1}$$

The assumption behind this model is that the classes are ordered along a direction ($\mathbf{w}$) and thus the input space can be split by parallel hyperplanes. Moreover, given two patterns $\mathbf{x}_0$ and $\mathbf{x}_1$, it results:

$$\frac{odds(y \preceq \mathcal{C}_q|\mathbf{x}_0)}{odds(y \preceq \mathcal{C}_q|\mathbf{x}_1)} = \exp\left(-\mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_0)\right) \tag{2}$$

for all q. This is called the **proportional odds assumption** and it implies that input variables have the same effect on the odds regardless of the threshold. There are many ways to test this hypothesis and see whether the model could have a good performance in fitting data points. Statistical hypothesis tests have been developed, but it has been observed that they tend to reject the parallel assumption even when it holds.

2.2.1 Logistic models comparison

In this section, three logistic models will be compared, from which only the last will be an ordinal regression (*Ordinal LM*). The other two are *One vs. Rest Logistic Model* (*OVR LM*) and *Multinomial Logistic Model* (*Multinomial LM*).

(i) *Multinomial LM.* This is the natural extension from binary logistic classification. One class is fixed, which we can assume that it is class Q without loss of generality, and it is used as pivot. Then, for all the remaining class $q = 1, \dots, Q-1$ we perform a binary classification as follows:

$$\ln\left(\frac{P(y=q|\mathbf{x})}{P(y=Q|\mathbf{x})}\right) = b_q + \mathbf{w}_q^T\mathbf{x}. \tag{3}$$

The pivot can be also estimated as a hyperparameter by repeating Q times this procedure using each class as a pivot and picking the one that yields a better result for its regression.

(ii) *OVR LM.* Similarly to the previous case, the problem is simplified to a binary classification one. For each $q = 1, \dots, Q$ class, given a class q, the remaining $Q-1$ are merged into another class and the logistic model is applied on these two classes:

$$\ln\left(\frac{P(y=q|\mathbf{x})}{P(y\neq q|\mathbf{x})}\right) = b_q + \mathbf{w}_q^T\mathbf{x},\ q = 1, ..., Q. \tag{4}$$

In order to compare these three models we considered four bivariate Gaussian distributions, with $\sigma = \mathbb{I}_2$ and centered in $(0, 0)$, $(0, 0.5)$, $(0, 2)$ and $(0.5, 2.5)$; class $0, 1, 2, 3$ were respectively assigned. Train and test dataset (Figure 1) were respectively made of 4000 and 1000 point. As reported in Figure 2, the predicted output differs in all the three cases. As expected, the *Ordinal LM* draws three parallel lines, whereas there is no parallel constraint for the other two models. Since the centers of the distribution are not aligned, there is no straight direction along which classes grow, and non parallel hyperplanes seem to better separate classes. However, despite an overall equal or lower accuracy, the *Ordinal LM* has an important benefit; each class only borders on its own upper and lower class. This guarantees a lower misclassification rate among distant classes, as explained in the next section.

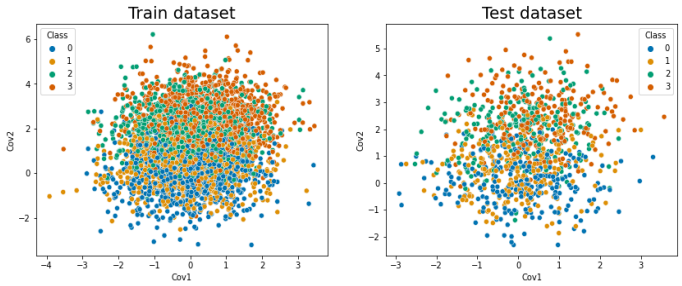


Figure 1: *Logistic models comparison.* Train and test dataset.

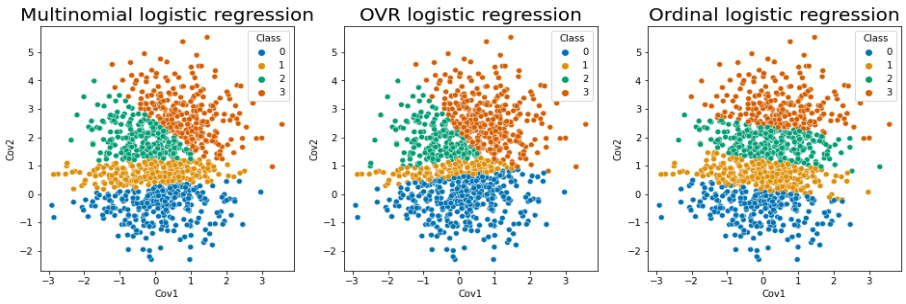


Figure 2: *Logistic models comparison.* Models output.

2.2.2 Performance Evaluation

In order to properly compare these three methods, we need a measure that takes into account how far did the prediction fall from the real value in terms of the **order**. We introduce two measures which are the most common for ordinal regression problems [2]. The Mean Zero-One Error (MZE) (Equation 5) considers a zero-one loss for misclassification, not penalizing the points predicted *far away* from their true class (*e.g.* assigning class 0 to class 2 instead of assigning it to class 1). On the other hand, the Mean Absolute Error (MAE) (Equation 6)

measures the average deviation in absolute value from the predicted rank, so it partially fix this issue.

$$MZE = \frac{1}{N}\sum_{i=1}^{N}\{y_i^* \neq y_i\} \tag{5}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\mathcal{O}(y_i) - \mathcal{O}(y_i^*)| \tag{6}$$

In both of the previous equations, y_i represents the prediction for item i and y_i^* its true class. Using the example introduced before in subsubsection 2.2.1, we can also observe the difference between these two errors. As we can see from Table 1, *Ordinal LM* has the largest MZE. This model is the worst when predicting the actual class of a given point, but this does not mean that it is a bad model. In fact, the MAE is essentially the same (and the lowest) for the *Multinomial* and *Ordinal LM*, so it is also more sensitive to the order than the *OVR LM* model for instance. Moreover, the *Ordinal* model swapped class 0 and class 3 only 1 time, when the multinomial also made this error 7 times. However, *Multinomial LM* has the lowest errors for both the MZE and the MAE. This can be due to the particularities of the data chosen; indeed, as already mentioned, the centers of the distributions are not aligned. Advantages of *Ordinal LM* will be highlighted in subsection 2.4.

	Multinomial LM	*OVR LM*	*Ordinal LM*
MZE	0.510	0.515	0.525
MAE	0.599	0.646	0.600

Table 1: MZE and MAE for the three models.

2.3 Related methods

What if the parallel assumption introduced in subsection 2.2 does not hold? There are many options to solve this problem, the first one could be to find a function that projects the point in a space where the point are separable by parallel hyperplanes. Another option could be to create a model that separates the different classes with non-parallel hyperplanes. The idea of this model is to perform Q-1 independent logistic binary classifications (Q being the number of classes). This new model is specified by:

$$\ln\left(\frac{P(y > q|\mathbf{x})}{P(y <= q|\mathbf{x})}\right) = b_q + \mathbf{w}_q^T\mathbf{x},\ \ q = 1, ..., Q-1. \tag{7}$$

In particular, for each class k, the dataset is divided into two: those which belong to a class bigger than k and those to a lower or equal class. For each division a logistic regression is performed. The different logistic regressions give the margins between the different classes. The logistic regression for data points belonging to class bigger than k versus the others, for instance, will give the margin between class k and class $k+1$. Once obtained all the margins each class will have its assigned part of the hyperspace and consequently it will be sufficient to find the location of a data point to assign him a predicted class. On one hand, this model (from now on *Non-Parallel Ordinal LM* or *NP Ordinal LM*) should perform better than the non-ordered models because it takes into account the ordered information and could perform better than the parallel one because it relies only on classes linear separability and not on parallel separability. However, this model has also its drawbacks. The main issue is that it yields to meaningless cases which cannot be classified. For instance, many points could be classified in the same time as bigger of a certain class k but smaller than the class $k-1$, and thus they would not belong to any actual ordered class. In Figure 3 there is one particular example of this situation.
To solve this issue one could perform a monotone (or *isotonic*) regression to fit the predicted probabilities of being bigger than the different classes with a decreasing function, since it is not possible to have $P(y > q) > P(y > q-1)$. More precisely, one should first find the points that present this issue; for those points take the probabilities obtained by the different logistic regressions and then apply an isotonic decreasing regression on these probabilities. Doing so the predicted probabilities $P(y > q)$ will decrease as q increases as it should be. Figure 4 displays how the previous example looks like after the correction. This procedure can be done both before and after having the predictions of the single logistic regressions. In the former case, isotonic regression will be applied to all those points that present a non monotone succession of predicted probabilities. In the latter case, the isotonic regression will be applied only on those point that are in the region where the result cannot be interpreted.

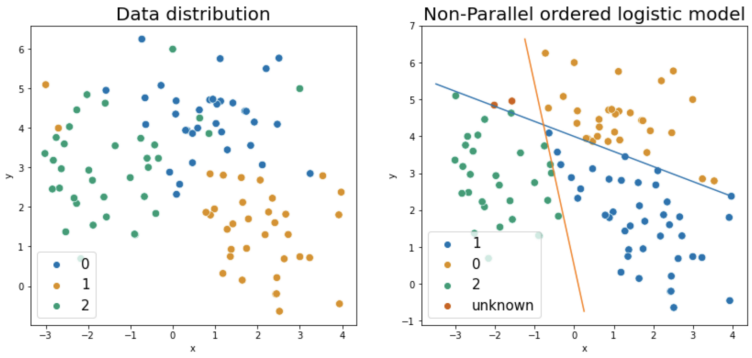


Figure 3: *NP Ordinal LM.* Predicted classes.

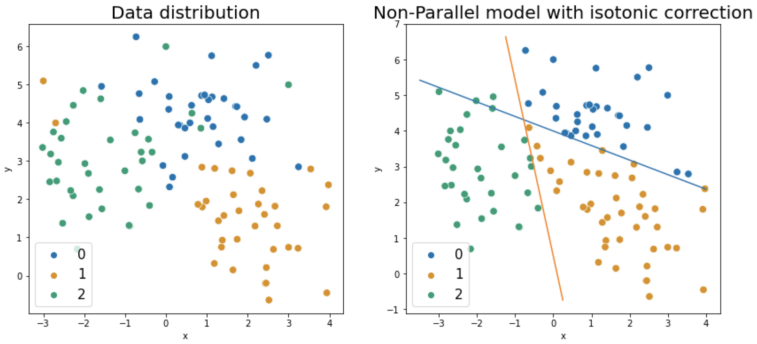


Figure 4: *NP Ordinal LM.* Predicted classes with isotonic correction.

2.4 Overfitting Issues

The *Ordinal LM* has another advantage when the sample size is small and the classes number large. In fact, the parameters to be estimated are significantly less than both in the non-ordinal and the ordinal but non-parallel models. For the latter models, this can lead to overfitting as shown in the following example.

Given a sample size of 100 and four different classes $(0, 1, 2, 3)$, all the aforementioned models were performed. Afterwards, two points in the dataset were swapped (between classes 0 and 1) and each model was run again. The two data distributions are shown in Figure 5, while Figure 6 plots the different regions predicted by each model before and after the modification. *OVR LM* is subject to significant changes, but even the *Multinomial LM* and *NP Ordinal LM* are influenced by this slight alteration. Finally, *Ordinal LM* has not undergone any perceptible change.

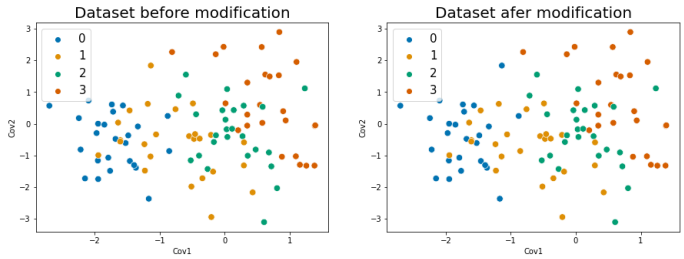
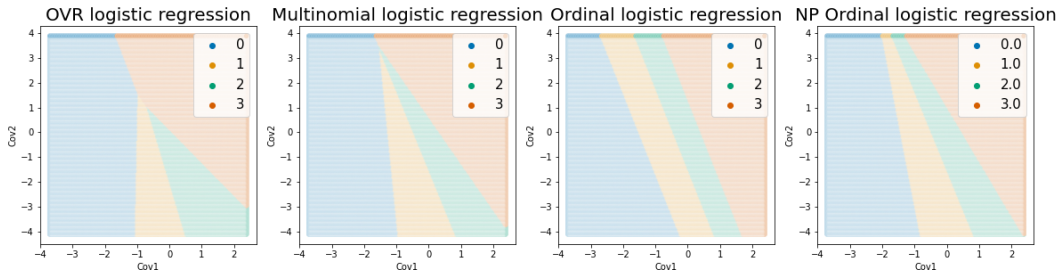


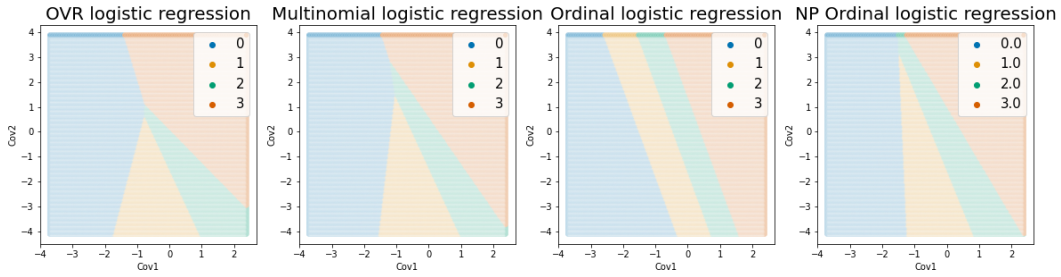
Figure 5: *Overfitting.* Dataset.

3 Experiments

Let us now focus on a particular example to illustrate the previous models. We will consider data to determine *quality* of white wines, for which we will consider 11 attributes as described in [1] (including alcohol, acidity and sugars). Our objective would be to classify the quality of the wine, and thus this is a well posed problem for our ordinal regression models.



(a)



(b)

Figure 6: *Overfitting.* Figure 6a and Figure 6b show the models results before and after data modification.

3.1 Data Processing

The *white wines* dataset we considered is made of 4898 observations and the output consists of 10 classes. Each predictor is continuous and approximately Gaussian around its mean, as it could be seen from histogram plots. The only exception is *residual sugar*, which has lots of observations near the value 0. Afterwards, we analyzed the correlations between the covariates through a scatterplot. A few points are very distant from the rest, hence they were removed in order to better analyze if linear dependence could be found between covariates. In total, 6 points were deleted. Since it is a really small fraction of the total sample, we believed that the removal could not modify the results. The new scatterplot highlighted linear dependencies between *density* and *residual sugar* and *density* was removed. Finally, we noticed that very few observations were related to the highest and lowest class, while the majority of data was classified as 5, 6 or 7. Hence, we opted for grouping the lower classes together as well as the higher, in order not to have unbalanced output.

3.2 Performances of the models

Once we have our cleaned data, we are going to compare the performances of three models: *Multinomial LM*, *Ordinal LM* and the *NP Ordinal LM*. *OVR LM* was avoided since it has previously given the worst performance. To do so we made a simple split of the total amount of data to perform a simple validation for the different models. Simple validation consists on splitting data and construct a model with only a part of the total amount of them. The other part of data is then used to test whether the model we are constructing fits our data-set well. We decided to use simple validation, because the number of data points in the data-set is very high and in such a situation it is sufficient to show the performances of the different models. The size of the training data was set to 70% of the total amount of data. Additionally we set a seed on the random split to be sure that the train-test split was always the same for the three models. We then applied the models to our set and we noticed that the performances of all of them were not high. To improve them we decided to standardize data because from the processing we noticed that almost all the predictors have almost a normal distribution. This change didn't bring any improvements. Consequently we decided to report the original results. Performance results on the test set are shown in Figure 7 and Table 2.

	Multinomial	*Ordinal*	*NP Ordinal*
Accuracy	0.5735	0.544	0.5715
MAE	0.44	0.46	0.43

Table 2: Accuracy and MAE of the three models.

In terms of accuracy the *Multinomial LM* is slightly the best model and, as discussed above, such a result is because the aim of this model is only to predict the right class, without considering the order. Consequently

	0	1	2
0	275	175	7
1	140	489	51
2	19	234	78

Multinomial

	0	1	2
0	260	189	8
1	176	431	73
2	39	184	108

Ordinal

	0	1	2
0	236	218	3
1	117	517	46
2	14	231	86

NP Ordinal

Figure 7: Confusion matrices for the three models with predicted values on the columns and true values on the rows. Class 0 indicates the bad quality, 1 indicates medium quality and 2 good quality.

for that model there is no difference in predicting class 0 with class 2 or with class 1. However, together with a similar accuracy, the *NP Ordinal LM* is the one that yields the best prediction with respect to the **ordered information**; in fact, it is the model with the smallest number of class 2 predicted as class 0 and vice versa. Computing the MAEs here confirms what we observed with the confusion matrices. In fact the model with the highest MAE is the *Ordinal LM*, while the *Multinomial* and the *NP Ordinal* have a MAE of 0.44 and 0.43 respectively. Regarding the *Ordinal LM*, it has the worst performance, for the accuracy is the lowest and MAE the largest. Therefore, the parallel assumption revealed to be too strict again. Additionally, we can observe that classes were neither well separable by linear but non-parallel hyperplanes, reason why all the proposed models were not able to predict the different classes. Further tests should be carried removing any linear assumption.

4 Conclusions

In this report, we have described the main hypothesis and assumptions for the ordinal regression models, including the *proportional odds assumption* and its geometric relation with the direction of the hyperplanes.
We have also compared three logistic models, for which we have shown some of the pathological issues regarding their performance using the concepts of MZE and MAE, together with the notion of *order* in ordinal regression. An extension of these models has been provided when the parallel assumption is relaxed, introducing the non-parallel models. The problem of *meaningless data* has been discussed and we presented a possible way to fix it using *isotonic regression*. For all models we have considered overfitting as a possible cause of error.
Finally, in the last section we have tested ordinal regression on real data from wine quality. After processing the data, we have found that the most accurate model was the *Multinomial Logistic* one, but nonetheless the Non-Parallel Logistic provided the best prediction when taking into account the order of the data. However, none of the models provided a satisfying division of the data given that this was not separable.
Further research could focus on other datasets and testing the performance on models which are separable. Despite its interpretability, logistic regression revealed to be not always inefficient and thus not suitable when dealing with complex datasets. Another possible direction would be to explore the proportional odd assumption by building a hyperspace in which the datapoints are separable by linear hyperplanes and thus check whether the performance improves or not.

References

[1] *White wine dataset*. https://archive.ics.uci.edu/ml/datasets/wine+quality.

[2] Pedro Antonio Gutiérrez et al. 'Ordinal Regression Methods: Survey and Experimental Study'. In: *IEEE Transactions on Knowledge and Data Engineering* 28.1 (2016), pp. 127–146. DOI: 10.1109/TKDE.2015.2457911.

[3] R. Williams. 'Generalized ordered logit/partial proportional odds models for ordinal dependent variables'. In: *Stata J.* 6.1 (2006), pp. 205–217.

[4] Hall M. Frank E. 'A Simple Approach to Ordinal Classification'. In: *ECML 2001. ECML 2001. Lecture Notes in Computer Science* 2167.1 (2001), pp. 127–146. DOI: 10.1007/3-540-44795-4_13.

[5] *Ordinal logistic regression*. https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/.

[6] *Ordinal regression*. https://en.wikipedia.org/wiki/Ordinal_regression.

[7] *Isotonic regression*. https://en.wikipedia.org/wiki/Isotonic_regression.