ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

Feature Relevance Bounds for Ordinal Regression

Lukas Pfannschmidt[1], Jonathan Jakob[1], Michael Biehl[2],
Peter Tino[3], Barbara Hammer[1]

1 - Machine learning group, Bielefeld University, DE
2 - Intelligent Systems Group, University of Groningen, NL
3 - Computer Science, University of Birmingham, UK

Abstract. The increasing occurrence of ordinal data, mainly sociodemographic, led to a renewed research interest in ordinal regression, i.e. the prediction of ordered classes. Besides model accuracy, the interpretation of these models itself is of high relevance, and existing approaches therefore enforce e.g. model sparsity. For high dimensional or highly correlated data, however, this might be misleading due to strong variable dependencies. In this contribution, we aim for an identification of feature relevance bounds which – besides identifying all relevant features – explicitly differentiates between strongly and weakly relevant features.[1]

1 Introduction

Ordinal data often occur in sociodemographic, financial or medical contexts where it is hard to give absolute qualitative measurements, but it is easily possible to compare samples. The *ordinal regression problem* (ORP) is the task to embed given data in the real numbers such that they are ordered according to their label i.e. the target variable. Although the problem can be attempted with a regular regression or classification method, dedicated techniques are to be preferred, which can account for the fact that the distance between ordinal classes in the data is unknown and not necessarily evenly distributed. Examples of ordinal regression include treatments as multiclass classification problem [1], and extensions of standard models such as the support vector machine (SVM) or learning vector quantization (LVQ) to ordinal regression tasks [2, 3, 4, 5]

Besides a mere classification prescription, practitioners are often interested in the relevance of input features i.e. the relevance of ordinal explanatory variables for the given task. This is particularly relevant when the objective exceeds mere diagnostics, such as safety-critical decision making, or the design of repair strategies. There do exist a few approaches which address such feature selection for ordinal regression: The approach [6] uses a minimal redundancy formulation based on a feature importance score to find the subset of relevant features. The work in [7] focuses on multiple filter methods which are adapted to ranking data. These models deliver sparse ordinal regression models which enable some insight into the underlying classification prescription. Yet, their result is arbitrary in the case of correlated and redundant features, where there does not exist a unique minimum relevant feature set.

[1]Funding by the DFG in the frame of the graduate school DiDy (1906/3) and by the BMBF (grant number 01S18041A) is gratefully acknowledged.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

The so-called *all* relevant feature selection problem deals with the challenge to determine all features, which are potentially relevant for a given task – a problem which is particularly relevant for diagnostics purposes if it is not priorly clear which one of a set of relevant, but redundant features to choose. Finding this subset is generally computationally intractable. For standard classification and regression schemes, a few efficient heuristics have recently been proposed: A recent approach focuses on the case of linear mappings and phrases the problem to determine the interval of possible variable relevances as a linear optimization problem [8].

In this paper we introduce an extension of the feature-relevance-interval-computation scheme as proposed in [8] to the context of ordinal regression data. For this purpose, we recapture a large margin ordinal regression formalization in section 2. This is extended to an optimization scheme to determine feature relevance bounds in section 3, which can be transferred to a linear programming problem. In section 4 we compare our method to classical approaches for artificial data with known ground truth and real life benchmarks.

2 Large Margin Ordinal Regression

Given ordered class labels $L = \{1, 2, \ldots, l\}$. Given training data $X = \{\mathbf{x}_i^j \in \mathbb{R}^n \mid i = 1, \ldots, m_i,\ j \in L\}$ where data point x_i^j is assigned the class label $j \in L$. The full data set has size $m := m_1 + \ldots + m_l$. Here the index j refers to the ordinal target variable (represented by b_j) the data point x_i^j belongs to. The ORP can be phrased as the search for a mapping $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(\mathbf{x}_{i_1}^{j_1}) < f(\mathbf{x}_{i_2}^{j_2})$ for all i_1, i_2 and all class labels $j_1 < j_2$.

We will restrict to the case of a linear function, i.e. $f(\mathbf{x}) = \mathbf{w}^t\mathbf{x}$ with parameter $\mathbf{w} \in \mathbb{R}^n$. A popular formulation which is inspired by support vector machines imposes a margin for the embedding [3]:

$$\min_{\mathbf{w}, b_j, \chi_i^j, \xi_i^j} \quad 0.5 \cdot \|\mathbf{w}\|_1 + C \cdot \sum_{i,j} \left(\chi_i^j + \xi_i^j\right) \tag{1}$$

$$\text{s.t. for all } i, j \quad \begin{aligned} &\mathbf{w}^t\mathbf{x}_i^j \leq b_j - 1 + \chi_i^j \\ &\mathbf{w}^t\mathbf{x}_i^{j+1} \geq b_j + 1 - \xi_i^{j+1} \\ &b_j \leq b_{j+1} \text{ and } \chi_i^j \geq 0, \xi_i^j \geq 0 \end{aligned} \tag{2}$$

where χ_i^j and ξ_i^j are slack variables, and the thresholds b_j for $j = 1, \ldots, l-1$ determine the boundaries which separate l classes. The hyper-parameter $C > 0$ controls the trade off of the margin and number of errors and can be chosen through cross validation. Unlike [3], which uses L_2 regularisation, we will use L_1 regularisation in (Eq. 1), aiming for sparse solutions.

3 Feature Relevance Bounds for Ordinal Regression

Assume a training set X. Denote an optimum solution of problem (1) as $(\tilde{\mathbf{w}}, \tilde{b}_j, \tilde{\xi}_i^j, \tilde{\chi}_i^j)$. This solution induces the value $\mu_X := \|\tilde{\mathbf{w}}\|_1 + C \cdot \sum_{i,j} \left(\tilde{\chi}_i^j + \tilde{\xi}_i^j\right)$, which is uniquely determined by X. We are interested in the class of equivalent

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

good hypotheses, i.e. all weight vectors $\mathbf{w}$ which yield (almost) the same quality as regards the regression error and generalisation ability as the function induced by $\tilde{\mathbf{w}}$. For the sake of feasibility, we use the following proxy induced by μ_X

$$\begin{aligned} F_\delta(X) \quad := \quad & \{\mathbf{w} \in \mathbb{R}^n \mid \exists \xi_i^j, \chi_i^j, b_j \text{ such that constraints (2) hold,} \\ & \|\mathbf{w}\|_1 + C \cdot \sum_{i,j} \left(\xi_i^j + \chi_i^j\right) \leq (1+\delta) \cdot \mu_X\} \end{aligned} \tag{3}$$

These constraints ensure: 1. The empirical error of equivalent functions in $F_\delta(X)$ is minimum, as measured by the slack variables. 2. The loss of the generalisation ability is limited, as guaranteed by a small L_1-norm of the weight vector and learning theoretical guarantees as provided e.g. by Theorem 7 in [9] and Corollary 5 in [10]. The parameter $\delta \geq 0$ quantifies the tolerated deviation to accept a function as yet good enough, C is chosen according to the solution of Problem (1).

Solutions $\mathbf{w}$ in $F_\delta(X)$ are sparse in the sense that irrelevant features are uniformly weighted as 0 for all solutions in $F_\delta(X)$. Relevant but potentially redundant features can be weighted arbitrarily, disregarding sparsity, similar in spirit to the elastic net, which weights mutually redundant features equally [11]. In this contribution we are interested in the relevance of features for forming good hypotheses, where we are interested in the following more specific characteristics:

- **Strong relevance** of feature I for $F_\delta(X)$: Is feature I relevant for all hypotheses in $F_\delta(X)$, i.e. all weight vectors $\mathbf{w} \in F_\delta(X)$ yield $w_I \neq 0$?
- **Weak relevance** of feature I for $F_\delta(X)$: Is feature I relevant for at least one hypothesis in $F_\delta(X)$ in the sense that one weight vector $\mathbf{w} \in F_\delta(X)$ exists with $w_I \neq 0$, but this does not hold for all weight vectors in $F_\delta(X)$?
- **Irrelevance** of feature I for $F_\delta(X)$: Is feature I irrelevant for every hypothesis in $F_\delta(X)$, i.e. all weight vectors $\mathbf{w} \in F_\delta(X)$ yield $w_I = 0$?

A feature is irrelevant for $F_\delta(X)$ if it is neither strongly nor weakly relevant. The questions of strong and weak relevance can be answered via the following optimisation problems:

Problem minrel(I):

$$\begin{aligned} \min_{\mathbf{w}, b_j, \chi_i^j, \xi_i^j} \quad & |w_I| && (4) \\ \text{s.t. for all } i,j \quad & \text{conditions (2) hold and} \\ & \|\mathbf{w}\|_1 + C \cdot \sum_{k,l} \left(\chi_k^l + \xi_k^l\right) \leq (1+\delta) \cdot \mu_X && (5) \end{aligned}$$

Feature I is strongly relevant for $F_\delta(X)$ iff minrel(I) yields an optimum larger than 0.

Problem maxrel(I):

$$\begin{aligned} \max_{\mathbf{w}, b_j, \chi_i^j, \xi_i^j} \quad & |w_I| && (6) \\ \text{s.t. for all } i,j \quad & \text{conditions (2) and (5) hold and} \end{aligned}$$

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Feature I is weakly relevant for $F_\delta(X)$ iff minrel(I) yields an optimum 0 and maxrel(I) yields an optimum larger than 0

These two optimisation problems span a real-valued interval for every feature I with the result of minrel(I) as lower and maxrel(I) as upper bound. This interval characterises the range of weights for I occupied by good solutions in $F_\delta(X)$. Hence, besides information about a features relevance, some indication about the degree up to which a feature is relevant or can be substituted by others, is given. Note, however, that the solutions are in general not consistent estimators of an underlying 'true' weight vector as regards its exact value, as has been discussed e.g. for Lasso [12]. Similar to [8], these optimization problems can be rephrased as equivalent linear optimization problems which can be solved efficiently in polynomial time. Here, we omit this formulation and proof due to page limitations.

Threshold Selection: To estimate a threshold for a feature to be considered as weakly relevant, we generate features with the same statistical properties but no relevance by design, sometimes referred to as probe or shadow features [13]: We permute feature I and compute the relevance interval bounds using I_P instead of I. We repeat this d times. For a cutoff, we accept a certain rate $r_{FP} := 0.01$ of false positives. From the descending list of all upper probe interval bounds we pick the threshold with index $\lfloor r_{FP} \times d \rfloor$. To determine, if a feature is strongly relevant, it is sufficient to check whether the problem for the lower bound and feature I_P is deemed infeasible by the solver, since the accuracy of the model without feature I degrades significantly.

4 Experiments

Artificial Data We adapt the generation method presented in [8] for ordinal regression. By using equal frequency binning we converted the continuous regression variable into an ordered discrete target variable. Gaussian noise as well as additional noisy features are added. Several data sets with different characteristics as regards the amount of strongly, weakly and irrelevant variables are created this way, see. Table 1.

For evaluation, we consider cross-validated feature elimination. We compare our model (dubbed feature relevance interval - FRI)[2] to ordinal regression with the Lasso penalty [14] and the ElasticNet [15] penalty with ratio of $0.5L_1 + 0.5L_2$. Hyperparameters are selected according to 5-fold cross validation. We evaluate whether the method is able to identify all relevant features, whereby we evaluate the correspondence of the sets by the true and detected set of relevant features by the F-measure, see Table 1. In all cases where weakly relevant features are involved, FRI provides significantly better accuracy than elastic net and Lasso to detect all relevant features.

Benchmark Data We test the proposed method on benchmark data as described in [16]. These data sets are imbalanced. The proposed model provides a mapping $\mathbf{x} \to f(\mathbf{x}) = y \in \{1, \ldots, l\}$ where $f(\mathbf{x}) = \text{argmin}_i\{\mathbf{w}^t\mathbf{x} \leq b_i\}$ where $b_l := \infty$. The

[2]Implementation in Python: https://github.com/lpfann/fri

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Table 1: Artificially created data sets with known ground truth and evaluation of the identified relevant features by the methods as compared to all relevant features. The score is averaged over 30 independent runs.

Data Set Characteristics				Results (F-measure)		
Points	*Strong*	*Weak*	*Irrelevant*	*ElasticNet*	*FRI*	*Lasso*
150	6	0	6	**0.91**	0.87	0.88
150	0	6	6	0.67	**0.93**	0.64
256	6	6	6	0.88	**0.97**	0.87
512	1	2	11	0.81	**0.87**	0.76
200	1	20	0	0.36	**1.00**	0.26
200	1	20	20	0.36	**0.90**	0.41

Table 2: Left: MMAE of FRI, elastic net (EN) and Lasso along with standard deviations ($\pm$) across 30 folds of data sets from [15, 16]. Right: Mean feature set size of the methods. FRI allows extra discrimination between strong (FRI_s) relevant and weak (FRI_w) relevance.

		Average Feature Set Size			
	MMAE	FRI_s	FRI_w	EN	Lasso
Automobile	0.661 $\pm$ 0.129	4.5	12.6	4.0	4.8
Bondrate	1.36 $\pm$ 0.122	0.0	5.4	2.0	2.0
Contact-lenses	0.914 $\pm$ 0.206	0.9	1.1	2.0	2.0
Eucalyptus	0.406 $\pm$ 0.027	2.1	33.2	15.6	15.7
Newthyroid	0.667 $\pm$ 0.0	0.0	4.7	2.0	2.0
Pasture	0.367 $\pm$ 0.121	0.0	15.5	6.0	5.1
Squash-stored	0.39 $\pm$ 0.164	2.4	7.9	11.1	6.6
Squash-unstored	0.317 $\pm$ 0.168	1.8	3.3	8.0	7.3
TAE	0.621 $\pm$ 0.153	1.9	5.4	16.8	13.7
Winequality-red	1.081 $\pm$ 0.037	0.0	7.6	5.4	5.3

output is evaluated by the macro-averaged absolute error: $\text{MMAE} = \sum_j^l \frac{\sum_i |j - f(\mathbf{x}_i^j)|}{m_j} / l$ where m_j is the number of examples in class j. We replicated the experiments which have been presented in [4, 5], whereby results are averaged over 30 folds. Results are reported in Tab. 2. Since all methods rely on linear models, their MMAE is comparable. For these data, no ground truth as regards the relevant features is available, so we can only compare the amount of information provided by the methods. We report the average number of features identified as relevant by the techniques. For three data sets (Squash-stored, Squash-unstored, TAE), FRI identifies a smaller number of relevant features than the alternatives, yielding the same accuracy. For three further data sets (Automobile, Eucalyptus, Pasture), FRI identifies more (weakly relevant) features. In all cases, FRI offers more information than direct Lasso or ElasticNet by identifying weakly relevant features.

5 Conclusions

In this paper we presented the adaption of the feature relevance bounds approach to ordinal regression data. Based on the experiments we showed that the method

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

can provide a good all-relevant feature set approximation in this new setting. These feature sets represent additional information useful in analytic use cases for model and experiment design, subject for further evaluation.

References

[1] Eibe Frank and Mark Hall. A Simple Approach to Ordinal Classification. In Luc De Raedt and Peter Flach, editors, *Machine Learning: ECML 2001*, pages 145–156, 2001.

[2] Amnon Shashua and Anat Levin. Ranking with large margin principle: Two approaches. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, pages 961–968, Cambridge, MA, USA, 2002. MIT Press.

[3] Wei Chu and S. Sathiya Keerthi. Support Vector Ordinal Regression. *Neural Comput.*, 19(3):792–815, March 2007.

[4] Shereen Fouad and Peter Tiňo. Adaptive Metric Learning Vector Quantization for Ordinal Classification. *Neural computation*, 24 11:2825–51, 2012.

[5] Fengzhen Tang and Peter Tiňo. Ordinal regression based on learning vector quantization. *Neural Networks*, 93:76–88, 2017.

[6] Xiubo Geng, Tie-Yan Liu, Tao Qin, and Hang Li. Feature Selection for Ranking. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 407–414, New York, NY, USA, 2007. ACM.

[7] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Feature Selection for Ordinal Regression. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, pages 1748–1754, New York, NY, USA, 2010. ACM.

[8] Christina Göpfert, Lukas Pfannschmidt, Jan Philip Göpfert, and Barbara Hammer. Interpretation of linear classifiers by means of feature relevance bounds. *Neurocomputing*, 298:69–79, 2018.

[9] Shivani Agarwal. Generalization Bounds for Some Ordinal Regression Algorithms. In Yoav Freund, László Györfi, György Turán, and Thomas Zeugmann, editors, *Algorithmic Learning Theory, 19th International Conference, ALT 2008, Budapest, Hungary, October 13-16, 2008. Proceedings*, pages 7–21, 2008.

[10] Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

[11] Hui Zou and Trevor Hastie. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

[12] Peng Zhao and Bin Yu. On Model Selection Consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, December 2006.

[13] Hervé Stoppiglia, Gérard Dreyfus, Rémi Dubois, and Yacine Oussar. Ranking a random feature for variable and feature selection. *Journal of machine learning research*, 3(Mar):1399–1414, 2003.

[14] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[15] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[16] Javier Sánchez-Monedero, Gutiérrez, Pedro Antonio, Peter Tino, and César Hervás-Martínez. Exploitation of Pairwise Class Distances for Ordinal Classification. *Neural Computation*, 25(9), 2013.