A Comparison of Different ROC Measures for Ordinal Regression

Willem Waegeman WILLEM.WAEGEMAN@UGENT.BE
Department of Electrical Energy, Systems and Automation, Ghent University, Technologiepark 913, B-9052 Ghent, Belgium

Bernard De Baets BERNARD.DEBAETS@UGENT.BE
Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, B-9000 Ghent, Belgium

Luc Boullart LUC.BOULLART@UGENT.BE
Department of Electrical Energy, Systems and Automation, Ghent University, Technologiepark 913, B-9052 Ghent, Belgium

Abstract

Ordinal regression learning has characteristics of both multi-class classification and metric regression because labels take ordered, discrete values. In applications of ordinal regression, the misclassification cost among the classes often differs and with different misclassification costs the common performance measures are not appropriate. Therefore we extend ROC analysis principles to ordinal regression. We derive an exact expression for the volume under the ROC surface (VUS) spanned by the true positive rates for each class and show its interpretation as the probability that a randomly drawn sequence with one object of each class is correctly ranked. Because the computation of VUS has a huge time complexity, we also propose three approximations to this measure. Furthermore, the properties of VUS and its relationship with the approximations are analyzed by simulation. The results demonstrate that optimizing various measures will lead to different models.

1. Introduction

In multi-class classification labels are picked from a set of unordered categories. In metric regression labels might take continuous values. Ordinal regression can be located in between these learning problems because here labels are chosen from a set of ordered categories. Applications of ordinal regression frequently arise in domains where humans are part of the data generation process. When humans assess objects for their beauty, quality, suitability or any other characteristic, they really prefer to qualify them with ordinal labels instead of continuous scores. This kind of datasets is obtained in information retrieval and quality control, where the user or the human expert frequently evaluates objects with linguistic terms, varying from "very bad" to "very good" for example. Also in medicine and social sciences, where many datasets originate by interaction with humans, ordinal regression models can be used.

In these applications of ordinal regression one is often the most interested in a subset of the classes. In many cases these classes of interest are the "extreme" categories, such as the documents with the highest relevance to the query or the products with the lowest quality. Moreover, there is often an unequal number of training objects for the different categories in real-world ordinal regression problems. The overall classification rate or mean absolute error are in these cases not the most pertinent performance measures. Criteria such as the area under the *receiver operating characteristics* (ROC) curve — which is related to defining an optimal ranking of the objects — are more appropriate. This article aims to discuss possible extensions of ROC analysis for ordinal regression.

Nowadays the area under the ROC curve is used as a standard performance measure in many fields where a binary classification system is needed. A ROC curve is created by plotting the *true positive rate* (TPR) versus

Appearing in *Proceedings of the ICML 2006 workshop on ROC Analysis in Machine Learning*, Pittsburgh, USA, 2006. Copyright 2006 by the author(s)/owner(s).

	$\widehat{y}=-1$	$\widehat{y}=1$	
$y=-1$	TN	FP	n_-
$y=1$	FN	TP	n_+
	NP	PP	n

Table 1: Confusion matrix for a two class classification problem of size n

the *false positive rate* (FPR). The TPR (or *sensitivity*) and the FPR (also known as 1 - *specificity*) are computed from the confusion matrix or contingency table (shown in Table 1). Sensitivity is defined as the number of positive predicted examples from the positive class TP divided by the number of positive examples n_+ and specificity is defined as the number of negative predicted examples TN from the negative class divided by the number of negative examples n_-:

$$Sens = TPR = \frac{TP}{TP+FN} \tag{1}$$

$$Spec = TNR = 1 - FPR = \frac{TN}{TN+FP} \tag{2}$$

With a classifier that estimates a continuous function f, the class prediction $\widehat{y}$ for an object x is obtained by the following rule:

$$\widehat{y} = \mathrm{sgn}(f(x)+b) \tag{3}$$

The points defining the ROC curve can then be computed by varying the threshold b from the most negative to the most positive function value and the *area under the ROC curve* (AUC) gives an impression of quality of the classifier. It has been shown [Cortes & Mohri, 2003,Yan et al., 2003] that the AUC is equivalent to the *Wilcoxon-Mann-Whitney* statistic:

$$WMW = AUC(f) = \frac{1}{n_- n_+}\sum_{i=1}^{n_-}\sum_{j=1}^{n_+} I_{f(x_i)<f(x_j)} \tag{4}$$

The value of the indicator function I will be one when its argument is true and zero otherwise. The measure $AUC(f)$ can be seen as a nonparametric estimate for the probability that the function value of an object randomly drawn from the negative class is strictly smaller than the function value of an object randomly drawn from the positive class:

$$AUC(f) = P(f(x_i) < f(x_j) \mid y_i = -1 \wedge y_j = 1)) \tag{5}$$

2. ROC Measures for Ordinal Regression

Recently, different approaches have been proposed to extend ROC analysis for multi-class classification. In the most general case, the *volume under the ROC surface* (VUS) has to be minimized in multi-class classification. The ROC surface can be seen as a Pareto front, where each objective corresponds to one dimension. In case there are more then two classes (let's say r), then the number of objectives depends on the multi-class method that is used:

- For a *one-versus-all* method, r functions f_k are estimated that try to separate objects of class k from the other classes. As a consequence misclassification costs for each class are fixed and the corresponding ROC surface will have r dimensions representing the true positive rates TPR_k for each class [Flach, 2004]. ROC points are here obtained by varying the thresholds b_k in the prediction rule $\widehat{y} = \mathrm{argmax}_k f_k(x) + b_k$.

- For a *one-versus-one* method, a function f_{kl} is estimated for each pair of classes, which allows to specify the cost for a misclassification of an object of class k predicted as class l. The corresponding ROC space is in this case spanned by $\frac{r(r-1)}{2}$ objectives [Ferri et al., 2003]. A prediction for new instances is done by majority voting over all $\frac{r(r-1)}{2}$ classifiers based on the outcomes $\mathrm{sgn}(f_{kl} + b_{kl})$.

In ordinal regression the picture is slightly different. The vast majority of existing methods for ordinal regression — including traditional statistical methods like cumulative logit models and their variants [Agresti, 2002], kernel methods [Chu & Keerthi, 2005,Shashua & Levin, 2003] and bayesian approaches [Chu & Gharhamani, 2005] — fit in general one function f to the data together with $r-1$ thresholds b_k for r ordered classes. New observations can then be classified by predicting them into the class k for which it holds that

$$b_{k-1} < f(x) \leq b_k \text{ with } b_0 = -\infty \text{ and } b_r = +\infty. \tag{6}$$

The simplicity of this kind of models has as disadvantage that one can not control the cost of misclassifying an object of a given class into another specified class. In other words, like in *one-versus-all* multi-class classification only r objectives can be simultaneously minimized. Therefore one could wonder whether a *one-versus-one* approach could be useful for ordinal regression. However, the answer is negative because it would lead to a more complex model with more variables to be estimated. Fortunately, Misclassification costs are always proportional to the absolute difference between the real and the predicted class, so defining a loss function with this property will solve the problem [Rennie & Srebro, 2005].

We will further assume that the misclassification costs are fixed for each class (they are always to proportional to the absolute difference between the real and the predicted label). Like in binary classification, we want a model f that imposes an optimal ranking of the data objects. There are several ways to define an optimal ranking. By analogy with (5) an optimal ranking could here be defined as a ranking that maximizes the joint probability that an r-tuple $(x_1, ..., x_r)$ is correctly ordered where each element x_k is randomly drawn from class k. This probability is given by

$$P(\bigwedge_{k=1}^{r-1} (f(x_k) < f(x_{k+1}) \mid y_k = k) \tag{7}$$

and it can be estimated for a given model by counting the number of ordered r-tuples occurring in the training dataset, i.e.

$$OrdTuples(f) = \frac{1}{\prod_{k=1}^{r} n_k} \sum_{y_{j_1}<...<y_{j_r}} I_{f(x_{j_1})<...<f(x_{j_r})} \tag{8}$$

Here n_k stands for the number of objects with label k. It is straightforward to see that $OrdTuples(f)$ reduces to (4) in case of two classes. Furthermore, we can show the following.

Theorem 2.1 *Given a continuous function f that imposes a ranking over a dataset with r ordered classes, OrdTuples(f) is the volume under the ROC surface ($VUS_{ord}(f)$) spanned by the true positive rates for each class.*

In statistics there has some related work on this topic. [Dreisetl et al., 2000] derive formulas for the variance of VUS_{ord} and the covariance between two volumes in the three class case. This work is extended to the general r-class case in [Nakas & Yiannoutsos, 2004]. They conclude that bootstrapping is preferred over U-statistics for large values of n and r. In this article we focus more on the use of $VUS_{ord}(f)$ as performance measure for ordinal regression problems.

For three ordered classes the ROC surface can be visualized. We have constructed this ROC surface for a synthetic dataset. We sampled $3 * 100$ instances from 3 bivariate Gaussian clusters with consecutive ranks. The mean of the clusters was set to (10,10), (20,10) and (20,20) respectively, σ_1 and σ_2 were set to 5 for the first two clusters and were set to 7 for the last cluster. ρ was fixed to 0. This dataset is visualized in Figure 1. We used the support vector ordinal regression algorithm of [Chu & Keerthi, 2005] to estimate the function f, without looking at the thresholds. The ROC surface is shown in Figure 2.

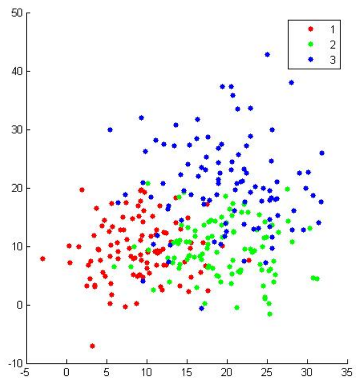


Figure 1: Synthetic dataset

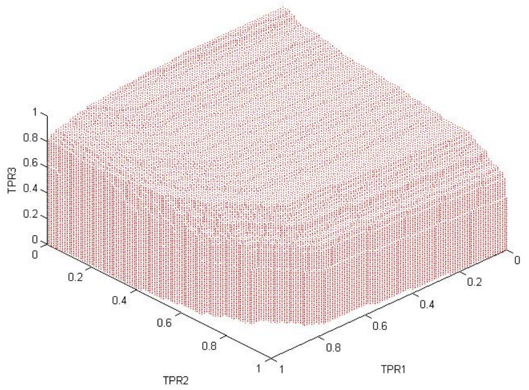
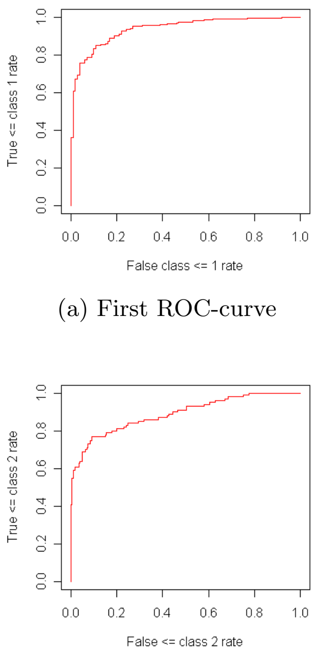


Figure 2: 3D ROC surface for the synthetic dataset

Optimizing the AUC instead of accuracy has been suggested for binary classification, for example with gradient descent or a quadratic solver. However, the computation of $VUS_{ord}(f)$ has a large time complexity. The function I is evaluated $\prod_{k=1}^{r} n_k$ times, which is exponential in the number of classes r. As a consequence, minimizing $VUS_{ord}(f)$ will lead to a hard optimization problem.

We will look at approximations of $VUS_{ord}(f)$ which can be more easily transformed into a suitable loss function. The biggest problem is that all $r - tuples$ need to be verified. Much would be gained if only pairs of function values have to be correctly ranked in each evaluation of I. This is another way of evaluating the imposed ranking. We discuss here three approximations of VUS_{ord} that all reduce to I-evaluations



(a) First ROC-curve

(b) Second ROC-curve

Figure 3: The three dimensional ROC surface approximated by a set of two ROC-curves for the synthetic dataset.

with only one condition.

The first approximation $Cons(f)$ is directly deduced from the way the majority of existing ordinal regression models are constructed. With a function f and $r-1$ thresholds one could look at threshold b_k as providing the separation between the consecutive ranks k and $k+1$. Varying this threshold will change the proportion between objects predicted lower than or equal to class k and objects predicted higher than class k. This corresponds to measuring the non-weighted sum of $r-1$ two-dimensional ROC curves representing the trade-off between consecutive classes:

$$Cons(f) = \frac{1}{r-1}\sum_{l=1}^{r-1} AUC_l(f) \tag{9}$$

$$AUC_l(f) = \frac{1}{\sum_{i=1}^{l} n_i \sum_{j=l+1}^{n} n_j} \sum_{i:y_i \leq l} \sum_{j:y_j > l} I_{f(x_i)<f(x_j)}$$

The two ROC curves belonging to the synthetic dataset are shown in figure 3.

For a second approximation of $VUS_{ord}(f)$ we looked at the statistical literature. In nonparametric statistics the *Jonckheere-Terpstra* test is known as a more powerful alternative for a *Kruskal-Wallis* test for testing

$$H_0 : \quad \mu_1 \leq \mu_2 \leq ... \leq \mu_r \tag{10}$$

versus the one side alternative

$$H_a : \quad \mu_1 \geq \mu_2 \geq ... \geq \mu_r \tag{11}$$

if there is a cdf F for which $F_k(x) = F(x - \mu_k))$. It is composed of a set of one sided WMW-tests:

$$JT = \sum_{i<j} WMW_{ij} \tag{12}$$

JT computes the WMW statistic for all possible pairs of classes, which is the same as computing the AUC for each pair of classes. This has been done for *one-versus-one* multi-class classification [Hand & Till, 2001], which gives rise to the following approximation:

$$Ovo(f) = \frac{2}{r(r-1)} \sum_{l<k} AUC_{lk}(f) \tag{13}$$

$$AUC_{lk}(f) = \frac{1}{n_l n_k} \sum_{i:y_i=l} \sum_{j:y_j=k} I_{f(x_i)<f(x_j)}$$

A third measure could exist of counting the number pairs that are correctly ranked among all possible pairs of data objects:

$$Pairs(f) = \frac{1}{\sum_{k<l} n_k n_l} \sum_{i=1}^{n} \sum_{j=1;y_i<y_j}^{n} I_{f(x_i)<f(x_j)} \tag{14}$$

A loss function based on (14) is used in the ordinal regression method of [Herbrich et al., 2000]. The difference with $Ovo(f)$ is that here a weighted average of the ROC areas for each of pair of classes is taken. The weights are the prior π_k probabilities of observing an object of class k, i.e.

$$Pairs(f) = \frac{2}{r(r-1)} \sum_{l<k} \pi_k \pi_l AUC_{lk}(f) \tag{15}$$

3. Simulation experiments

To see the characteristics of the different measures, we conducted some simulation experiments. In the first experiment we wanted to find out which values are obtained for different levels of separability and for an increasing number of classes. Therefore we assume that the function values of the model f can be represented by a distribution with cdf $F(x)$, in which the function values for the objects of class k are distributed with cdf $F_k(x) = F(x - kd)$. Furthermore we chose to sample

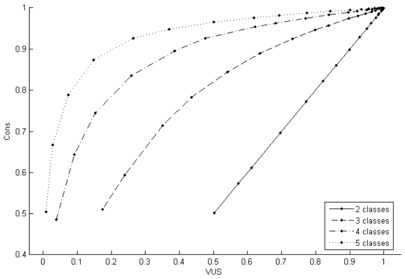


Figure 4: Relation between $VUS_{ord}(f)$ and $Cons(f)$ for $r = 1, ..., 5$ and $d = 0, ..., 5$ with step size 0.25. The values are averaged over 20 runs.

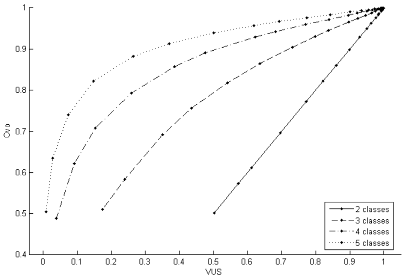


Figure 5: Relation between $VUS_{ord}(f)$ and $Ovo(f)$ for $r = 1, ..., 5$ and $d = 0, ..., 5$ with step size 0.25. The values are averaged over 20 runs.

from a Gaussian distribution with standard deviation $\sigma = 1$. So the function values conditioned on the labels are normally distributed with equidistant ordered means. Repeatedly 100 data points were sampled from each class while we increased the distance d between the means of consecutive clusters. We started at $d = 0$ (random classifier) and stopped at $d = 5$ (as good as perfect separation) with step size 0.25.

The results obtained for $VUS_{ord}(f)$, $Cons(f)$ and $Ovo(f)$ are graphically compared. In this simulation all classes have the same prior of occurring, so $Ovo(f)$ and $Pairs(f)$ will always have the same value. Consequently the results for $Pairs(f)$ are omitted. The relationship between $VUS_{ord}(f)$ and $Cons(f)$ on the one side and between $VUS_{ord}(f)$ and $Ovo(f)$ on the other side are shown in Figures 4 and 5. One can see that, as expected, the relation between $VUS_{ord}(f)$ and the other two measures is without doubt nonlinear. The expected value for $VUS_{ord}(f)$ heavily depends on the number of classes, while this is not the case for the approximations. The approximations all take an average over a set of two dimensional ROC-curves, so their expected value is never lower than a half, irrespective

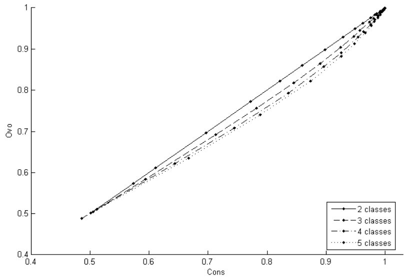


Figure 6: Relation between $Cons(f)$ and $Pairs(f)$ for $r = 1, ..., 5$ and $d = 0, ..., 5$ with step size 0.25. The values are averaged over 20 runs.

of the number of classes. Nevertheless, one can also see that $VUS_{ord}(f)$ converges rapidly to one when the distance between the subsequent means increases. In addition, $Cons(f)$ and $Ovo(f)$ behave quite similar in this simulation. This is also shown in Figure 6. Their observed values become more dissimilar when the number of classes increases.

In a second experiment we wanted to investigate whether optimizing the various performance measures would lead to the same model. For two measures M_1 and M_2 this implies that

$$\forall f, f^* \in \mathcal{H} : M_1(f) < M_1(f^*) \Leftrightarrow M_2(f) < M_2(f^*) \quad (16)$$

$$\forall f, f^* \in \mathcal{H} : M_1(f) = M_1(f^*) \Leftrightarrow M_2(f) = M_2(f^*) \quad (17)$$

The following experiment was set up to test whether this property holds for the four measures. All measures only quantify the quality of the ordering of a dataset for a function f. For a dataset of size n there are $n!$ possible rankings of the objects, so evaluating them all is computationally intractable. Therefore we sampled randomly 1000 rankings from all possible orderings of the dataset. We assumed we had 50 samples per class with four ordered classes, resulting in a sample size of 200 objects and 200! possible rankings. The results are given in Figure 7, which shows the distributions of all measures together with pairwise scatter plots. All classes again have the same prior of occurring, so $Ovo(f)$ and $Pairs(f)$ have a perfect correlation. This is however not true for the other measures. One can clearly see that for no pair of measures conditions (16) or (17) hold. In general, $VUS_{ord}(f), Cons(f)$ and $Ovo(f)$ will have different maxima over a hypothesis space $\mathcal{H}$ and a given dataset. So, optimizing one of the proposed approximations of $VUS_{ord}(f)$ will give rise to different classifiers.

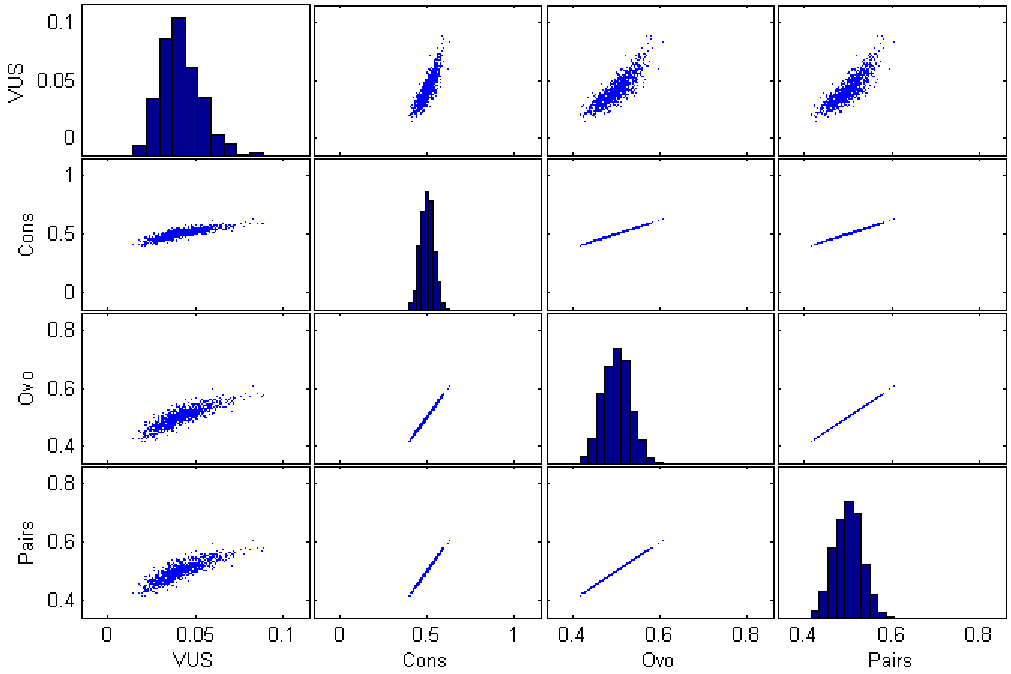


Figure 7: Histograms and pairwise scatter plots for all the measures.

4. Discussion and further research

In this article we argued that accuracy or mean absolute error are not the most powerful performance measures to evaluate ordinal regression models when misclassification costs are not equal for each class or when the data is unbalanced. Therefore we proposed some new measures, which extend binary and multiclass ROC analysis to ordinal regression. They all measure the quality of the ranking imposed by an ordinal regression model. First of all we showed that counting the number of ordered r-tuples in the ranking is equivalent to the area under the r-dimensional ROC curve spanned by the true positive rates of all classes. However, $VUS_{ord}(f)$ can't be transformed easily into a suitable loss function for learning algorithms, so three approximations were also analyzed. By simulation we showed that these four measures in general have a different distribution and that none of them is a monotone function of another. Further research will be devoted to converting measures like the area under the ROC curve into a loss function for a learning algorithm and to further analyse the characteristics of the presented measures.

Acknowledgments

The authors would like to thank Bruno Bauwens for useful discussions. Willem Waegeman is supported by a grant of the "Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen)".

References

Agresti, A. (2002). *Categorical Data Analysis, 2nd version*. John Wiley and Suns Publications.

Chu, W., & Gharhamani, Z. (2005). Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, *6*, 1019–1041.

Chu, W., & Keerthi, S. (2005). New approaches to support vector ordinal regression. *Proceedings of the International Conference on Machine Learning, Bonn, Germany* (pp. 321–328).

Cortes, C., & Mohri, M. (2003). AUC optimization versus error rate minimization. *In Advances in Neural Information Processing Systems, Vancouver, Canada*. The MIT Press.

Dreisetl, S., Ohno-Machado, L., & Binder, M. (2000).

Comparing three-class diagnostic tests by three-way roc analysis. *Medical Decision Making*, *20*, 323–331.

Ferri, C., Hernandez-Orallo, J., & Salido, M. (2003). Volume under ROC surface for multi-class problems. *In Proceedings of the European Conference on Machine Learning, Dubrovnik, Croatia* (pp. 108–120).

Flach, P. (2004). The many faces of ROC analysis in machine learning. Tutorial presented at the European Conference on Machine Learning, Valencia, Spain.

Hand, D. J., & Till, R. J. (2001). A simple generalization of the area under the ROC curve for multiple class problems. *Machine Learning*, *45*, 171–186.

Herbrich, R., Graepel, T., & Obermayer, K. (2000). Large margin rank boundaries for ordinal regression. *Advances in Large Margin Classifiers* (pp. 115–132). The MIT Press.

Nakas, C. T., & Yiannoutsos, C. T. (2004). Ordered multiple-class roc analysis with continuous measurements. *Statistics in Medicine*, *22*, 3437–3449.

Rennie, J. D. M., & Srebro, N. (2005). Loss functions for preference levels: Regression with discrete, ordered labels. *Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling, Edinburgh, Scotland* (pp. 180–186).

Shashua, A., & Levin, A. (2003). Ranking with large margin principle: Two approaches. *Proceedings of the International Conference on Neural Information Processing Systems, Vancouver, Canada* (pp. 937–944). Cambridge MA: MIT Press.

Yan, L., Dodier, R., Mozer, M. C., & Wolniewiecz, R. (2003). Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. *Proceedings of the International Conference on Machine Learning, Washington D. C., USA* (pp. 848–855).