Binary Classification Approach to Ordinal Regression

Jianhao Peng
jianhao2@illinois.edu

Abstract

In the lecture we mainly focus on questions about binary classification or regression. However in many cases we may want the results to show preference in terms of several discrete, but ordered, values, like ratings of a restaurant, disease condition and students' letter grades. Such problems are usually known as ordinal regression or ordinal classification. Standard approaches to ordinal regression in statistics require a assumption on the distribution of a latent variable while many other learning methods treat this as a multiclass problem, which may lose the order information in the data. In this paper I first present the probability model of ordinal regression in traditional statistics, then briefly go through different types of loss function defined from two perspectives in order to reduce the problem to simple binary case. From the loss function we can also get performance bounds similar to those we learnt in binary classification.

1 Introduction

Compared to the standard machine learning tasks of classification or regression, there exist many other systems in which users specify preferences by giving each sample one of several discrete values, e.g. one through five stars in movie or restaurant rating and A through F in letter grades. Those problems are often considered as multiclass classification, which treated the labels as a finite unordered nominal set. However in that case the natural order structure of the data is ignored and may not be as accurate. Therefore one nature question arises: can we make use of the lost ordinal information to potentially improve the predictive performance of the classifier or simplify the learning algorithm?

Such question has been studied in statistics for decades. Notice that the ordinality also bring us the capability of grouping the outcome into binary classification problem, i.e. $(\mathbf{Y} \le j$ versus $\mathbf{Y} > j)$. Many approaches apply regression method in these questions, P. McCullagh first proposed a ordered logistic regression model in 1980 [4], which is also known as the proportional odds model or cumulative logit model. One interpretation of ordered logistic regression is by using a latent variable $\mathbf{Z}$. Let θ_j be $K-1$ different threshold then we have the relationship $P\{\mathbf{Y} \le j\} = P\{\mathbf{Z} \le \theta_j\}$. A major drawback of this is that it relies on probability model of a latent variable which means it is distribution dependent.

In order to use the analytic tools in learning theory, we need to carefully design the loss function in ordinal regression. In Alexander's work [5] their exploit the ordinal nature by introducing a 'preference' function which compare true ranks of two example with their predicted ranks to act like a 0-1 loss function. They also provided a convergence bound on expected loss and presented a large margin algorithm. But in their implement they expand the size of example to $\mathcal{O}(N^2)$ and their bound was restricted to hard-margin cases, i.e., for all example(x, y) there exist a f such that $yf(x) \geq \Delta > 0$. Another approach is provided by Ling and Hsuan-Tien in [3], where they use a 'cost matrix' to define the expected loss. Instead of directly comparing each pair of examples, they extended the origin example first then reduced the problem back to a binary classification. By a pre-defined ranking rule their framework combined results from the binary classifiers and predict a rank. It turns out many existing ordinal regression algorithms can be unified in their framework [1]. Since the reduced framework is just binary classification, we can apply the results in the course and I derive a confidence bound for it, which is consistent with their result in [3].

The paper is organized as follows. I briefly go through the traditional statistics model of ordinal regression first in section 2, which also includes some stochastic ordering assumption. In section 3 I give a review on Alexander's results based on his paper [5]. A confidence bound based on the result of RKHS with hinge loss and the framework of [3] is provided in section 4, and section 5 is the conclusion.

2 Statistics Model for Ordinal Regression

Let $\mathcal{X}$ and $\mathcal{Y}$ be the feature space and label space. Assume the label $\mathcal{Y} = \{y_1, y_2, \dots, y_K\}$ with ordered ranks $y_k \succ y_{q-1} \succ \cdots \succ y_1$ where $\succ$ stands for the users' preference. Since $\mathcal{Y}$ is a finite set, $P\{Y = y_i|\mathbf{x}\} = \pi_i(\mathbf{x})$ is a multinomial distribution. In statistics we made the assumption that the cumulative probability $P\{Y \leq y_i|\mathbf{x}\} = \sum_{j=1}^{i} \pi_j(\mathbf{x})$ is a logistic function with a linear model.

$$P\{Y \leq y_i|\mathbf{x}\} = \phi(\theta_i - \mathbf{w}^T\mathbf{x}) = \frac{1}{1 + exp(\mathbf{w}^T\mathbf{x} - \theta_i)} \tag{1}$$

where $\mathbf{w}, \theta_i$ are unknown parameters to be estimated (with $\theta_0 = -\infty, \theta_k = \infty$ by definition) and ϕ is the logistic function $\phi(z) = \frac{1}{1+e^{-z}}$. Unlike the general polytomous regression, where each category y_i has different $\mathbf{w}_i$ and θ_i, in ordinal regression there is only one weight vector $\mathbf{w}$ for every category the only difference is the threshold θ_i as shown in figure 1. Which means, the hyperplanes that separate different labels are parallel for all classifiers. That gives us one effective way to construct the ordinal regression model with a latent variable $Z = \mathbf{w}^T\mathbf{x} + \epsilon$ where ϵ is a mean zero random variable. So we have the following monotone relationship:

$$Y = y_j \iff Z \in [\theta_{j-1}, \theta_j] \tag{2}$$

It follows from (2):

$$\begin{aligned} P\{Y \leq y_i|\mathbf{x}\} &= \sum_{j=1}^{i} \pi_j(\mathbf{x}) = P\{Z \leq \theta_i\} \\ &= P\{\mathbf{w}^T\mathbf{x} + \epsilon \leq \theta_i\} = P\{\epsilon \leq \theta_i - \mathbf{w}^T\mathbf{x}\} \\ &= P_\epsilon\{\theta_i - \mathbf{w}^T\mathbf{x}\} \end{aligned}$$

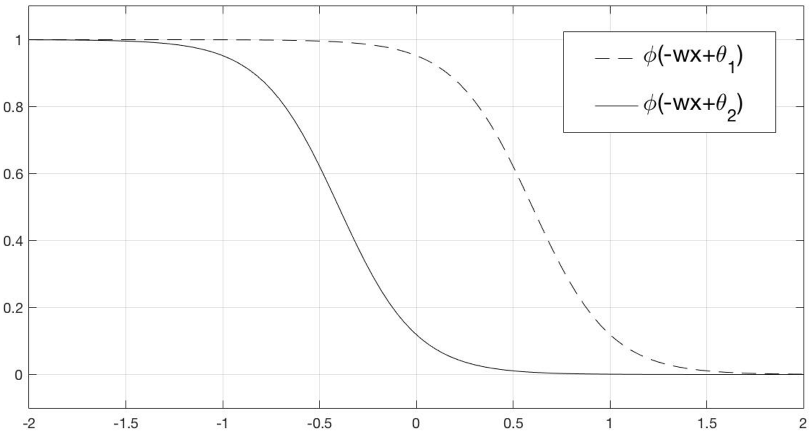

Figure 1: logit function with 2 different thresholds.

When P_ϵ is a logistics distribution, we get model in equation (1). Other distributional assumptions like normal distribution, yield different generalized linear model with the corresponding link function. After choosing the distribution, we can use maximum likelihood estimation to get $\hat{\mathbf{w}}$ and $\hat{\theta}_i$.

So far we have made two major assumptions in the traditional model: (i) all the weight vectors $\mathbf{w}$ for every category are the same and (ii) the distribution of a latent variable Z. In Eibe Frank and Mark Hall's work [2] they used the cumulative property of this model that $\pi_i(\mathbf{x}) = P\{Y \leq y_i|\mathbf{x}\} - P\{Y \leq y_{i-1}|\mathbf{x}\}$ and trained $K-1$ classifiers each handle the binary task: is the predicted value of example $\mathbf{x}$ larger than y_i or not? Then combined the results from $k-1$ classifiers and gave a final decision. It's a easy approach for the problem but each classifier is independent of the others hence it's difficult to analyze the performance. In the next section, I will show the work from [5] where they provided a distribution-free confidence bound by using the 'preference' function.

3 Confidence Bound with Preference Function

This section in a summary of results in [5]. One major issue we need to overcome with ordinal regression is the definition of loss function and the corresponding empirical loss. Unlike the binary case, where we can use 0-1 loss to determine whether the result is good or bad, in ordinal regression the simple indicator function $\mathbb{1}_{\{f(x)\neq y\}}$ does not show the difference of closeness, e.g. for a test example $(x, 4)$, we may consider the prediction $(x, 3)$ is closer than $(x, 1)$. Let $\mathcal{G}$ be all the mappings from $\mathcal{X}$ to $\mathcal{Y}$. From the ordinality of outcome($y_i \succ y_j$) they induced a ordering $\succ_{\mathcal{X}}$ on the feature space:

$$\mathbf{x}_i \succ_{\mathcal{X}} \mathbf{x}_j \iff g(x_i) \succ g(x_j) \tag{3}$$

Consider a mapping $g \in \mathcal{G}$ and two examples $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$. They first defined the rank difference $\ominus : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{Z}$ by: $y_i \ominus y_j \triangleq i - j$. Then from (3) we can compare the rank difference $y_1 \ominus y_2$ and $g(\mathbf{x}_1) \ominus g(\mathbf{x}_2)$ and determine whether the mapping is consistent with the ordering or not. We said that g violates the ordering if $\{y_1 \ominus y_2\} \times \{g(\mathbf{x}_1) \ominus g(\mathbf{x}_2)\} \leq 0$. Thus we have a preference function as below:

$$c_{pref}(\mathbf{x}_1, \mathbf{x}_2, y_1, y_2, g(\mathbf{x}_1), g(\mathbf{x}_2)) = \begin{cases} 1 & \text{if } \{y_1 \ominus y_2\} \times \{g(\mathbf{x}_1) \ominus g(\mathbf{x}_2)\} \leq 0; \\ 0 & \text{else.} \end{cases} \tag{4}$$

Notice that by this definition they expanded the sample space to $\mathcal{O}(N^2)$, and those examples are not iid samples. Furthermore they introduced the 'loss function' c_g by:

$$c_g(\mathbf{x}, y, g(\mathbf{x})) = \mathbb{E}_{X_1,Y_1}[c_{pref}(\mathbf{x}, X_1, y, Y_1, g(\mathbf{x}), g(X_1))] \tag{5}$$

Recall that in the binary case, our goal is to minimize the expected loss of a given function:

$$L_P(f) = P\{f(X) \neq Y\} = \mathbb{E}_{X,Y}[\mathbb{1}_{\{f(X)\neq Y\}}] \leq \mathbb{E}_{X,Y}[l(X, Y, f(X))] \tag{6}$$

where l is 0-1 loss or some surrogate loss function. Similarly, they defined the risk function to be minimized as:

$$R_{pref} = \mathbb{E}_{X_1,Y_1,X,Y}[c_{pref}(X_1, X, Y_1, Y, g(X_1), g(X))] = \mathbb{E}_{X_1,Y_1}[c_g(X_1, Y_1, g(X_1))] \tag{7}$$

Although equation (7) looks very like the right hand side of expected loss we use in (6) and each c_g function has already included the ordering information in it, the fact that it is actually a expectation over $\mathcal{X} \times \mathcal{Y} \times \mathcal{X} \times \mathcal{Y}$ and those examples violate the iid assumption in learning theory, makes it more complicate to analyze and minimize than the standard 0-1 loss. Hence they also provided a slightly redefined empirical loss in order to relate the R_{pref} to the standard classification task. For simplicity, define the new training set derived from pairs of $\mathbf{x}$ and y with different ranks by:

$$(X', Y') = \{((\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}), \Omega(y_i^{(1)}, y_i^{(2)}))\}_{i=1}^{m'}, \text{ for } \forall|y_i^{(1)} - y_i^{(2)}| > 0 \tag{8}$$

where $\Omega(y_i, y_j) = \text{sgn}(y_i \ominus y_j)$ and $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ denoted the first and second object of a pair, m' is the cardinality of new set (X', Y'). Let $\mathbf{z} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ and $t = \Omega(y^{(1)}, y^{(2)})$, then the risk function (7) can be written as:

$$R_{pref}(g) = \mathbb{E}_{Y_1,Y_2}[|\Omega(Y_1, Y_2)|]\, \mathbb{E}_{\mathbf{Z},t}[c_g(Z, t, \Omega(g(X^{(1)}), g(X^{(1)}))] \tag{9}$$

with sample size $N = m$ the empirical loss can be defined as:

$$R_{emp}(g) = \frac{m'}{m^2} \sum_{i=1}^{m'} c_g((\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}), \Omega(y_i^{(1)}, y_i^{(2)}), \Omega(g(\mathbf{x}_i^{(1)}), g(\mathbf{x}_i^{(2)}))) \tag{10}$$

The rest of their paper they provided a performance bound based on the zero training error case. i.e there exists a $g \in \mathcal{G}$ such that the R_{emp} can be 0. Then with probability $1 - \delta$:

$$R_{pref}(g) \leq \frac{2}{m-1}(k \log_2(\frac{8e(m-1)}{k}) \log_2(32(m-1)) + \log_2(\frac{8(m-1)}{\delta})) \tag{11}$$

where $k \leq e(m-1)$. The major drawback of this bound is that it was restricted to the separable case, which means minimal empirical error must be zero. When there doesn't exists such $g \in \mathcal{G}$, we cannot guarantee the performance of the classifier even if it minimized the empirical loss. Not to mention that it exploded the feature space so was not very efficient when size N is very large. In the next session I will talk about another approach to solve the problem which turns out to be quite universal.

4 Analysis with Cost Martix

Let P be the hidden distribution under $\mathcal{X} \times \mathcal{Y}$. Recall the expected loss in binary classification, $L_P(f) = \mathbb{E}_P[\mathbb{1}_{\{f(X) \neq Y\}}]$. One can think of the this as the expectation of elements in a 2×2 matrix where the rows and columns represent the possible output and label respectively. Similarly we can introduce the $K \times K$ cost matrix $\mathcal{C}$ with $\mathcal{C}_{y,k}$ being the cost of predicting an example $(\mathbf{x}, y)$ as rank k. Then we can define the generalization error of a ranking rule $r : \mathcal{X} \mapsto \mathcal{Y}$ as:

$$C(r, P) = \mathbb{E}_P[C_{Y,r(X)}] \tag{12}$$

where we assume $\mathcal{C}_{y,y} = 0$ and $\mathcal{C}_{y,k} > 0$ for $k \neq y$. Usually a cost matrix with V-shaped rows are preferred in order to interpret the ordinal information. That means $\mathcal{C}_{y,k-1} > \mathcal{C}_{y,k}$ for $k \leq y$ and $\mathcal{C}_{y,k} < \mathcal{C}_{y,k+1}$ for $k \geq y$. The absolute cost matrix $\mathcal{C}_{y,k} = |y - k|$ is a popular choice. Given a data set $S = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ the goal is to find a ranking rule r that has a small generalization error $C_{r,P}$. According to the ordinal property of the data, one way to construct a ranking rule is by adding the result from $K - 1$ binary classifiers $\mathbb{1}_{\{f(\mathbf{x},k)>0\}}$, each of them determines whether the rank of $\mathbf{x}$ is greater than k or not, i.e.

$$r(\mathbf{x}) = 1 + \sum_{k=1}^{K} \mathbb{1}_{\{f(\mathbf{x},k)>0\}} \tag{13}$$

we also want f to be rank-monotonic, i.e. $f(\mathbf{x}, 1) \geq f(\mathbf{x}, 2) \geq \ldots \geq f(\mathbf{x}, K-1)$ for every example. Then we can rewrite the cost $C_{y,r(\mathbf{x})}$ as:

$$\mathcal{C}_{y,r(\mathbf{x})} = \sum_{k=r(\mathbf{x})}^{K-1} (\mathcal{C}_{y,k} - \mathcal{C}_{y,k+1}) + \mathcal{C}_{y,K} = \sum_{k=1}^{K-1} (\mathcal{C}_{y,k} - \mathcal{C}_{y,k+1}) \mathbb{1}_{\{f(\mathbf{x},k) \leq 0\}} + \mathcal{C}_{y,K} \tag{14}$$

For notational simplicity, let $\mathbf{x}^{(k)} = (\mathbf{x}, k)$ and $y^{(k)} = 2 * \mathbb{1}_{\{k<y\}} - 1$ so we can define the extended example $(\mathbf{x}^{(k)}, y^{(k)})$ with a weights $w_{y,k} = |C_{y,k} - C_{y,k+1}|$. And (14) can be bounded by:

$$\begin{aligned} \mathcal{C}_{y,r(\mathbf{x})} &= \sum_{k=1}^{y-1} w_{y,k} y^{(k)} \mathbb{1}_{\{f(\mathbf{x}^{(k)}) \leq 0\}} + \sum_{k=y}^{K-1} w_{y,k} y^{(k)} (1 - \mathbb{1}_{\{f(\mathbf{x}^{(k)})>0\}}) + \mathcal{C}_{y,K} \qquad (15) \\ &\leq \sum_{k=1}^{K-1} w_{y,k} \mathbb{1}_{\{y^{(k)} f(\mathbf{x}^{(k)}) \leq 0\}}. \qquad (16) \end{aligned}$$

Equation (15) shows that the cost of $r(\mathbf{x})$ is bounded by the weighted sum of 0-1 loss of the binary classifier $f(\mathbf{x}, k) = f(x^{(k)})$! That gives us the ability to bound the expected loss. One choice of the function f is to use a threshold model: $f(\mathbf{x}, k) = g(\mathbf{x}) - \theta_k$ and in [3] they shows that when minimizing the surrogate loss of such f with regularization, there exists a optimal solution $(g^*, \boldsymbol{\theta}^*)$ such that $\boldsymbol{\theta}^*$ is ordered.

4.1 Generalization Bound

From the definition of $w_{y,k}$ and V-shaped property, we know that $\sum_k w_{y,k} = \mathcal{C}_{y,1} + \mathcal{C}_{y,K} = c_y$. Then $P_{k|y}\{k|y\} = \frac{w_{y,k}}{c_y}$ is a probability mass function. From (16) we have:

$$\mathcal{C}_{y,r(\mathbf{x})} \leq c_y \sum_{k=1}^{K-1} \frac{w_{y,k}}{c_y} \mathbb{1}_{\{y^{(k)} f(\mathbf{x}^{(k)}) \leq 0\}} = c_y \, \mathbb{E}_{P_k}[y^{(k)} f(\mathbf{x}^{(k)}) \leq 0] \tag{17}$$

Therefore we can construct a probability distribution $\hat{P}$ on $\{(\mathbf{x}^{(k)}, y^{(k)})\}$ that generate a new example $(\mathbf{x}^{(k)}, y^{(k)})$ by choosing the example $(\mathbf{x}, y)$ first from P and then choosing k from $P_{k|y}$. By which we have the following theorem:

Theorem 1. *if f is rank monotonic, let $c = \max_y c_y$. Then there exists a distribution $\hat{P}$ on $(\mathbf{X}^{(k)}, Y^{(k)})$ such that*

$$\mathbb{E}_P[\mathcal{C}_{y,r(\mathbf{x})}] \leq c \, \mathbb{E}_{\hat{P}}[Y^{(k)} f(\mathbf{X}^{(k)}) \leq 0]$$

Proof. with the $\hat{P}$ constructed as above, taking expectation on both side of (17) leads to:

$$\mathbb{E}_P[\mathcal{C}_{y,r(\mathbf{x})}] \leq \mathbb{E}_P \, c_y \cdot \mathbb{E}_{P_k}[y^{(k)} f(\mathbf{x}^{(k)}) \leq 0] \leq c \, \mathbb{E}_{\hat{P}}[Y^{(k)} f(\mathbf{X}^{(k)}) \leq 0]$$

□

And $\{(\mathbf{x}^{(k)}, y^{(k)})\}$ from $\hat{P}$ are iid. Notice that the RHS of the inequality is just the expected 0-1 loss of a binary classifier f times a constant c. Hence we can use the bound in learning theory to guarantee the performance of the extended binary classifier.

Theorem 2. *(Bounds for extended binary classification with surrogate loss) Suppose $\mathcal{F}$ and the penalty function φ are chosen so that the following conditions are satisfied:*

- *$\varphi(-yf(x)) \leq B$ for some constant B fo all $(x, y) \in (\mathbf{X}^{(k)}, Y^{(k)})$ and $f \in \mathcal{F}$*
- *φ is M_φ Lipschitz continuous.*

then for any n and $\delta \in (0, 1)$, and any learning algorithm, the following bounds holds with probability $1 - \delta$

$$\mathbb{E}_P[\mathcal{C}_{y,\hat{r}_n(\mathbf{x})}] \leq c \left\{ A_{\varphi,n}(\hat{f}_n) + 4M_\varphi \, \mathbb{E}[R_n(\mathcal{F}(X^n))] + B\sqrt{\frac{log(1/\delta)}{2n}} \right\} \tag{18}$$

4.2 Application of Theorem 2 in RKHS by using Support Vector Machine

Consider the kernel $K(x, x) = \langle x, x \rangle + 1$ and the RKHS associated with K is $\mathcal{H}_K$. If we constrained the function f to a closed ball of radius 1 and a closed $\sqrt{K(x, x)} \leq C_K$, i.e. :

$$f(\mathbf{x}, k) \in \{f : (\mathbf{x}, k) \mapsto \langle \mathbf{u}, \mathbf{x} \rangle - \theta_k, \|f\|^2 = \|\mathbf{u}\|^2 + \|\boldsymbol{\theta}\|^2 \leq 1, \|\mathbf{x}\|^2 + 1 \leq C_k^2\}$$

with the ramp loss $\varphi(x) = \min\{1, (1+x)_+\}$So the expected loss of ranking rule $\hat{r}_n(\mathbf{x})$ generating by $\hat{f}_n$ from (13) is bounded by:

$$\mathbb{E}_P[\mathcal{C}_{y,\hat{r}_n(\mathbf{x})}] \leq c \left\{ \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{\{y_n^{(kn)} \hat{f}(x_n^{(kn)}) \leq 1\}} + 4\frac{C_K}{\sqrt{N}} + \sqrt{\frac{log(1/\delta)}{2N}} \right\} \tag{19}$$

where $\mathbb{1}_{\{y_n^{(kn)} f(x_n^{(kn)}) \leq 1\}}$ is a random variable introduced by $P_{k|y}$. Its mean $\frac{1}{c_{y_n}} \sum_{k=1}^{K-1} w_n^{(k)} \mathbb{1}_{\{y_n^{(k)} f(x_n^{(k)}) \leq 1\}}$ By Huffding's inequality, we have:

$$\frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{\{y_n^{(kn)} \hat{f}(x_n^{(kn)}) \leq 1\}} \leq \frac{1}{N} \sum_{n=1}^{N} \frac{1}{c_{y_n}} \sum_{k=1}^{K-1} w_n^{(k)} \mathbb{1}_{\{y_n^{(k)} \hat{f}(x_n^{(k)}) \leq 1\}} + \sqrt{\frac{\log 1/\delta}{N}}, \text{ w.p. at least 1 - } \delta \tag{20}$$

Combine (19) and (20), we have the same bound w.p. at least $1 - \delta$ in the form:

$$\mathbb{E}_P[\mathcal{C}_{y,\hat{r}_n(\mathbf{x})}] \leq \frac{\beta}{N} \sum_{n=1}^{N} \sum_{k=1}^{K-1} w_{y_n}^{(k)} \mathbb{1}_{\{y_n^{(k)} \hat{f}(\mathbf{x}_n^{(k)}) \leq 1\}} + \mathcal{O}(\frac{C_k}{\sqrt{N}}, \sqrt{\frac{log\frac{1}{\delta}}{N}}) \text{ ,where } \beta = \frac{\max_y c_y}{\min_y c_y}$$

5 Conclusion

In the paper, I first go through the traditional statistics setup of ordinal regression problem which makes several distribution assumptions. Although the model is fairly simple, it does not always make sense to have those assumptions. Then I review two major distribution independent methods towards the problem. In section 3 where they tried to interpret the ordering information by a preference loss function. The idea behind this is straightforward but it's too complicate to analyze the performance and even their results are restricted to the hard margin case only. In last section I summarize some key definitions and results from [3], then verify the result using a SVM example.

References

[1] Wei Chu and S Sathiya Keerthi. New approaches to support vector ordinal regression. In *Proceedings of the 22nd international conference on Machine learning*, pages 145–152. ACM, 2005.

[2] Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *European Conference on Machine Learning*, pages 145–156. Springer, 2001.

[3] Ling Li and Hsuan-Tien Lin. Ordinal regression by extended binary classification. *Advances in neural information processing systems*, 19:865, 2007.

[4] Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):109–142, 1980.

[5] Alexander J Smola. Large margin rank boundaries for ordinal regression. In *Advances in large margin classifiers*, pages 115–132. MIT press, 2000.