Regression Models for Ordinal Data: A Machine Learning Approach

Ralf Herbrich, Thore Graepel, Klaus Obermayer

Technische Universität Berlin
Fachbereich Informatik
Franklinstraße 28/29
10587 Berlin

Bericht 1999-3

Abstract

In contrast to the standard machine learning tasks of classification and metric regression we investigate the problem of predicting variables of ordinal scale, a setting referred to as *ordinal regression*. The task of ordinal regression arises frequently in the social sciences and in information retrieval where human preferences play a major role. Also many multi–class problems are really problems of ordinal regression due to an ordering of the classes. Although the problem is rather novel to the Machine Learning Community it has been widely considered in Statistics before. All the statistical methods rely on a probability model of a latent (unobserved) variable and on the condition of *stochastic ordering*. In this paper we develop a distribution independent formulation of the problem and give uniform bounds for our risk functional. The main difference to classification is the restriction that the mapping of objects to ranks must be transitive and asymmetric. Combining our theoretical framework with results from measurement theory we present an approach that is based on a mapping from objects to scalar utility values and thus guarantees transitivity and asymmetry. Applying the principle of Structural Risk Minimization as employed in Support Vector Machines we derive a new learning algorithm based on large margin rank boundaries for the task of ordinal regression. Our method is easily extended to nonlinear utility functions. We give experimental results for an Information Retrieval task of learning the order of documents with respect to an initial query. Moreover, we show that our algorithm outperforms more naive approaches to ordinal regression such as Support Vector Classification and Support Vector Regression in the case of more than two ranks[1].

[1] This paper is a preliminary version of (Herbrich et al. 1999)

1 Introduction

Given a training sample of objects and their associated target, machine learning — as well as statistics — focuses on finding dependencies between objects and target values. This process is called *induction* and requires to make assertions given only a finite set of observations. It is widely accepted that any kind of induction requires some restriction on the assumed dependencies. These restrictions reflect the prior knowledge about the problem at hand and are chosen beforehand. In contrast to statistics where prior knowledge is expressed in the form of probability distributions over the observations and over the assumed dependencies (Bayesian school), machine learning techniques make their prior knowledge explicit by restricting the space of assumed dependencies without making any distributional assumptions (Vapnik 1982; Vapnik 1998). In the past, machine learning mainly focused on the problems of classification and regression estimation. As will be seen later, the problem of ordinal regression shares characteristics of both these tasks.

Let us consider the basic assumptions made in (supervised) machine learning (Vapnik 1998): Given an i.i.d. sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell} \sim P_{XY}^{\ell}$ where $P_{XY}^{\ell} = \prod_{i=1}^{\ell} P_{XY}$, and a set $\mathcal{H}$ of mappings h from X to Y, a learning procedure selects one mapping h^{ℓ} such that — using a predefined loss $l : Y \times Y \mapsto \mathcal{R}$ — the risk functional $R(h^{\ell})$ is minimized. In Statistical Learning Theory (SLT) the risk functional $R(h)$ under consideration is the expectation value of the loss $l(y, h(\mathbf{x}))$, i.e., the loss at each point $(\mathbf{x}, y)$ weighted by its (unknown) probability $P_{XY}(\mathbf{x}, y)$. Using the principle of Empirical Risk Minimization (ERM), one chooses that function h^{ℓ} which minimizes the mean of the loss $R_{\text{emp}}(h^{\ell})$ given the sample S. Introducing a quantity which characterizes the "capacity" of $\mathcal{H}$, bounds for the deviation $|R(h^{\ell}) - \inf_{h \in \mathcal{H}} R(h)|$ can be derived (Vapnik and Chervonenkis 1971; Pollard 1984; Shawe-Taylor et al. 1996). Two main scenarios were considered in the past: (i) If Y is a finite unordered set (nominal scale), the task is referred to as *classification learning*. Since Y is unordered, the $0-1$ loss, i.e., $l_{0-1}(y, \hat{y}) = 0$ iff $y = \hat{y}$, and $l_{0-1}(y, \hat{y}) = 1$ iff $y \neq \hat{y}$, is adequate to capture the loss at each point $(\mathbf{x}, y)$. (ii) If Y is a metric space, e.g., the set of real numbers, the task is referred to as *regression estimation*. In this case the loss function can take into account the full metric structure. Different metric loss functions have been proposed which are optimal under given probability models $P_{Y|X=\mathbf{x}}(y|\mathbf{x})$ (Huber 1981). Usually, optimality is measured in term of the mean squared error (MSE) of h^{ℓ}. For the class of unbiased estimators the Gauss–Markov theorem states that using the loss function $L_2(y, \hat{y}) = (y - \hat{y})^2$ results in h^{ℓ} with the smallest MSE (see, e.g., (McCullagh and Nelder 1983)).

In this paper, we consider a problem which shares properties of both cases (i) and (ii). Like in (i) Y is a finite set and like in (ii) there exists an ordering among the elements of Y. In contrast to regression estimation we have to deal with the fact that Y is a non–metric space. A variable of the above type exhibits an *ordinal scale* and can be considered as the result of a coarsely measured continuous variable (Anderson and Philips 1981). The ordinal scale leads to problems in defining an appropriate loss function for our task (see McCullagh 1980 and Anderson 1984): On the one hand, there exists no metric in the space Y, i.e., the distance $(y - \hat{y})$ of two elements is not defined. On the other hand, the simple $0 - -1$ loss does not reflect the ordering in Y. Since no loss function $l(y, \hat{y})$ can be found that acts on true ranks y and predicted ranks $\hat{y}$, we suggest to exploit the ordinal nature of the elements of Y by considering the order on the space X induced by each mapping $h : X \mapsto Y$. Thus our loss function $l_{\text{pref}}(\hat{y_1}, \hat{y_2}, y_1, y_2)$ acts on pairs of true ranks (y_1, y_2) and predicted ranks $(\hat{y_1}, \hat{y_2})$. Such an approach makes it possible to formulate a distribution independent theory of ordinal regression and to give uniform bounds for the risk functional. Roughly speaking, the proposed risk functional measures the probability of misclassification of a randomly drawn pair $(\mathbf{x}_1, \mathbf{x}_2)$ of observations, where the two classes are $\mathbf{x}_1 \succ_{\mathrm{X}} \mathbf{x}_2$ and $\mathbf{x}_2 \succ_{\mathrm{X}} \mathbf{x}_1$ (see Section 3). Problems of ordinal regression arise in many fields, e.g., in information retrieval (Wong et al. 1988; Herbrich et al. 1998), in econometric models (Tangian and Gruber 1995; Herbrich et al. 1998), and in classical statistics (McCullagh 1980; Fahrmeir and Tutz 1994; Anderson 1984; de Moraes and Dunsmore 1995; Keener and Waldman 1985).

As an application of the above–mentioned theory, we suggest to model ranks by intervals on the real line. Then the task is to find a latent utility function that maps objects to scalar values. Due to the ordering of ranks, the function is restricted to be transitive and asymmetric, because these are the defining properties of a *preference relation*. The resulting learning task is also referred to as *learning of preference relations* (see Herbrich et al. 1998; Wong et al. 1988). One might think that learning of preference relations reduces to a standard classification problem if pairs of objects are considered. This, however, is not true in general because the properties of transitivity and asymmetry may be violated by traditional Bayesian approaches which may violate stochastic transitivity (Suppes et al. 1989). Considering pairs of objects, the task of learning reduces to finding a utility function that best reflects the preferences induced by the unknown distribution P_{XY}. Our learning procedure on pairs of objects is an application of the large margin idea known from data–dependent Structural Risk Minimization (Shawe-Taylor et al. 1996; Cristianini et al. 1998). The resulting algorithm is similar to Support Vector Machines (SVM) (Cortes and Vapnik 1995; Vapnik 1995), which became popular in recent years due to their good generalization properties (Cortes and Vapnik 1995; Smola 1996; Schölkopf 1997; Girosi 1997; Wahba 1997; Weston et al. 1997; Pontil and Verri 1998; Weston and Watkins 1998). Since during learning and application of SVM's only inner products of object representations $\mathbf{x}_i$ and $\mathbf{x}_j$ have to be computed, the method of potential functions (the so called "kernel trick") can be applied (Aizerman et al. 1964). This makes it possible to model non–linear utility functions and thus to incorporate correlations between features of X at minimal computational costs for predicting the ranks of objects.

In Section 2 we introduce the setting of ordinal regression and present well known results and algorithms from the field of Statistics. In Section 3 we present a distribution independent model for ordinal regression. We give explicit bounds for the proposed loss function and show the relation of ordinal regression problems to preference learning problems. To justify the choice of loss, we prove some basic properties of the loss in Section 4. Here we argue that neither classification nor metric regression can solve the ordinal regression problem in a satisfactory way. We give the training error minimization method (ERM) another interpretation by deriving the distance measure captured by our loss on a finite sample. In Section 5 we derive an algorithm for ordinal regression for a particular modeling of a rank and application of large margin techniques which is based on a mapping of objects to an underlying utility. We extend our approach to non–linear utility functions by applying the "kernel trick". In Section 6 we present learning curves of our approach in a controlled experiment and in a real–world experiment on data from Information Retrieval.

Notation Throughout the paper vectors are denoted by lower case bold letter, e.g. $\mathbf{x}$, whereas the i-th element of a vector is denoted by subscript, e.g. x_i. For the sake of notational simplicity, the symbol P simultaneously denotes a probability distribution and probability density, whereas $P(x)$ is an abbreviation of $P(X = x)$. The symbol $\sim$ means "is distributed according to". The expectation value of f w.r.t. the distribution P is denoted by $\mathbf{E}_P\left[f\right] = \int P(x)f(x)\ dx$. The superscript $^\ell$ often denotes an estimator based on a randomly drawn sample of size ℓ, whereas the superscript * is used to refer to the optimal function or value. The set of real and natural numbers are denoted by $\mathcal{R}$ and $\mathcal{N}$, respectively.

2 A Distribution Dependent Model of Ordinal Regression

In this section let us recall the well–known cumulative or threshold model for ordinal regression (McCullagh 1980). Consider an input space $X \subset \mathcal{R}^n$ with objects being represented by feature vectors $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathcal{R}^n$, where n denotes the number of features. Furthermore, let us assume that there is an outcome space $Y = \{r_1, \ldots, r_q\}$ with ordered ranks $r_q \succ_{\mathrm{Y}} r_{q-1} \succ_{\mathrm{Y}} \cdots \succ_{\mathrm{Y}} r_1$. The symbol $\succ_{\mathrm{Y}}$ denotes the ordering between different ranks and can be interpreted as "is prefered to". Suppose that an i.i.d. sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell} \subset X \times Y$ is given[2], where $S \sim P_{XY}^{\ell} = \prod_{i=1}^{\ell} P_{XY}$. Because Y contains only a finite number of ranks, $P(y = r_i|\mathbf{x})$ is a multinomial distribution.

Stochastic ordering and the cumulative model Since Y is an ordered space, we make the assumption of stochastic ordering of the related space X, i.e., for all pairwise different $\mathbf{x}_1$ and $\mathbf{x}_2$ either

$$P(y \leq r_i|\mathbf{x}_1) \quad \geq \quad P(y \leq r_i|\mathbf{x}_2) \qquad \text{for all } r_i \in Y\,, \tag{1}$$

or

$$P(y \leq r_i|\mathbf{x}_1) \quad \leq \quad P(y \leq r_i|\mathbf{x}_2) \qquad \text{for all } r_i \in Y\,. \tag{2}$$

Stochastic ordering is satisfied by a model of the form

$$g^{-1}(P(y \leq r_i|\mathbf{x})) \quad = \quad \theta(r_i) - \mathbf{w}^T\mathbf{x}\,, \tag{3}$$

[2] For better readability we will drop the subscripts on P if the follow from the context

where $g^{-1} : [0,1] \mapsto (-\infty, +\infty)$ is a monotonic function often referred to as the inverse link function. The stochastic ordering follows from the fact that

$$\begin{aligned} P(y \le r_i|\mathbf{x}_1) \ge P(y \le r_i|\mathbf{x}_2) &\Leftrightarrow P(y \le r_i|\mathbf{x}_1) - P(y \le r_i|\mathbf{x}_2) \ge 0 \\ &\Leftrightarrow g^{-1}(P(y \le r_i|\mathbf{x}_1)) - g^{-1}(P(y \le r_i|\mathbf{x}_2)) \ge 0 \\ &\Leftrightarrow \mathbf{w}^T(\mathbf{x}_2 - \mathbf{x}_1) \ge 0\,, \end{aligned}$$

which no longer depends on r_i (the same applies to $P(y \le r_i|\mathbf{x}_1) \le P(y \le r_i|\mathbf{x}_2)$). Such a model is called cumulative or threshold model and can be motivated by the following argument: Let us assume that the ordinal response is a coarsely measured *latent* continuous variable $U(\mathbf{x})$. Thus, we observe rank r_i in the training set iff

$$y = r_i \quad \Leftrightarrow \quad U(\mathbf{x}) \in [\theta(r_{i-1}), \theta(r_i)]\,, \tag{4}$$

where the function U (latent utility) and $\boldsymbol{\theta} = (\theta(r_0), \ldots, \theta(r_q))^T$ are to be determined from the data. By definition $\theta(r_0) = -\infty$ and $\theta(r_q) = +\infty$. We see that for the real line $U(\mathbf{x}), \mathbf{x} \in X$ is divided into q consecutive intervals, where each interval corresponds to a rank r_i. Now let us make a linear model of the latent variable $U(\mathbf{x})$

$$U(\mathbf{x}) \quad = \quad \mathbf{w}^T\mathbf{x} + \epsilon\,, \tag{5}$$

where ϵ is the random component of zero expectation, $\mathbf{E}[\epsilon] = 0$, and distributed according to F_ϵ. It follows from Equation (4) that

$$\begin{aligned} P(y \le r_i|\mathbf{x}) &= \sum_{j=1}^{i} P(y = r_j|\mathbf{x}) = \sum_{j=1}^{i} P(U(\mathbf{x}) \in [\theta(r_{j-1}), \theta(r_j)]) \\ &= P(U(\mathbf{x}) \in [-\infty, \theta(r_i)]) \\ &= P(\mathbf{w}^T\mathbf{x} + \epsilon \le \theta(r_i)) \\ &= P(\epsilon \le \underbrace{\theta(r_i) - \mathbf{w}^T\mathbf{x}}_{\eta}) = F_\epsilon(\theta(r_i) - \mathbf{w}^T\mathbf{x})\,. \end{aligned}$$

If we now make a distributional assumption F_ϵ for ϵ we obtain the cumulative model by choosing as the inverse link function g^{-1} the inverse distribution function F_ϵ^{-1} (quantile function). Note that each quantile function $F_\epsilon^{-1} : [0,1] \mapsto (-\infty, +\infty)$ is a monotonic function. Different distributional assumptions for ϵ yield the logit, probit, or complementary log–log model (see Table 1).

Estimation in the cumulative model In order to estimate $\mathbf{w}$ and $\boldsymbol{\theta}$ we conclude from model (3) for the observations $(\mathbf{x}_i, y) \in S$ with equal $\mathbf{x}_i$

$$\underbrace{\begin{pmatrix} o_1(\mathbf{x}_i) \\ o_2(\mathbf{x}_i) \\ \vdots \\ o_{q-2}(\mathbf{x}_i) \\ o_{q-1}(\mathbf{x}_i) \end{pmatrix}}_{\mathbf{o}(\mathbf{x}_i)} = \underbrace{\begin{pmatrix} -\mathbf{x}_i & 1 & 0 & \cdots & 0 & 0 \\ -\mathbf{x}_i & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ -\mathbf{x}_i & 0 & 0 & \cdots & 1 & 0 \\ -\mathbf{x}_i & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}}_{\mathbf{Z}(\mathbf{x}_i)} \underbrace{\begin{pmatrix} \mathbf{w} \\ \theta(r_1) \\ \theta(r_2) \\ \vdots \\ \theta(r_{q-2}) \\ \theta(r_{q-1}) \end{pmatrix}}_{\tilde{\mathbf{w}}}\,,$$

model	inverse link function $F_\epsilon^{-1}(\Delta)$	distribution $P_\epsilon(\epsilon = \eta)$
logit	$\ln \frac{\Delta}{1-\Delta}$	$\frac{\exp(\eta)}{(1+\exp(\eta))^2}$
probit	$N^{-1}(\Delta)$	$\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\eta^2}{2}\right\}$
complementary log–log	$\ln(-\ln(1-\Delta))$	$\exp\{\eta - \exp(\eta)\}$

Table 1: Inverse link functions for different models for ordinal regression (taken from McCullagh and Nelder, 1983). Here, N^{-1} denotes the inverse normal function.

where $o_j(\mathbf{x}_i) = F_\epsilon^{-1}(P(y \le r_j|\mathbf{x}_i))$ is the transformed probability of ranks less than or equal r_j given $\mathbf{x}_i$, which will be estimated from the sample by the transformed frequencies of that event. Note that the complexity of the model is determined by the linearity assumption (5) and F_ϵ^{-1} which can be thought of as a regularizer in the resulting likelihood equation. For the complete training set we obtain

$$\underbrace{\begin{pmatrix} \mathbf{o}(\mathbf{x}_1) \\ \vdots \\ \mathbf{o}(\mathbf{x}_\ell) \end{pmatrix}}_{g^{-1}(\mathbf{y}) \text{ (observable random variables)}} = \underbrace{\begin{pmatrix} \mathbf{Z}(\mathbf{x}_1) & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{Z}(\mathbf{x}_\ell) \end{pmatrix}}_{\mathbf{Z} \text{ (observable)}} \underbrace{\begin{pmatrix} \tilde{\mathbf{w}} \\ \vdots \\ \tilde{\mathbf{w}} \end{pmatrix}}_{\tilde{\mathbf{w}} \text{ (parameters)}} . \tag{6}$$

The last equation is called the design matrix of a multivariate generalized linear model (GLM). A generalized linear model $\mathbf{y} = g(\mathbf{Z}\tilde{\mathbf{w}})$ is mainly determined by the design matrix **Z** (see Equation (6)) and the link function $g(\cdot) = F_\epsilon(\cdot)$. Then given a sample S and a link function — which coincides with a distributional assumption about the data — methods for calculating the maximum likelihood estimate $\tilde{\mathbf{w}}^\ell$ exist (see McCullagh and Nelder 1983 or Fahrmeir and Tutz 1994 for a detailed discussion). The main difficulty in maximizing the likelihood is introduced by the nonlinear link function.

To conclude this review of classical statistical methods we want to highlight the two main assumptions made for ordinal regression: (i) a distributional assumption on the unobservable latent variable (ii) and the assumption of stochastic ordering of the space X.

3 A Distribution Independent Model of Ordinal Regression

In this section we present a distribution independent model for ordinal regression. Anderson and Philips 1981 doubted the feasibility of formulating a distribution independent model for ordinal regression. Also McCullagh 1980 in his seminal work claimed that "...an appealing requirement for ordinal data is that the model should not be invariant under arbitrary permutations." This statement is the starting point for our analysis. Instead of the distributional assumptions made in the last section, we now consider a parameterized model space $\mathcal{H} = \{h(\cdot;\mathbf{w}) : X \mapsto Y|\mathbf{w} \in \Lambda\}$ of mappings from objects to ranks. Each such function h induces an ordering $\succ_X$ on the elements of the input space by the following rule

$$\mathbf{x}_i \succ_{\mathrm{X}} \mathbf{x}_j \quad \Leftrightarrow \quad h(\mathbf{x}_i;\mathbf{w}) \succ_{\mathrm{Y}} h(\mathbf{x}_j;\mathbf{w})\,. \tag{7}$$

If we neglect the ordering of the space Y, the Bayes–optimal function h^*_{class} can be written as

$$h^*_{\text{class}}(\mathbf{x}) \quad = \quad \arg\max_{r_i \in Y} P(r_i|\mathbf{x})\,. \tag{8}$$

This function h^*_{class} has the following well known property.

Theorem 1. *Given an unknown probability measure P_{XY} the function h^*_{class} minimizes the risk functional R_{0-1} defined by*

$$R_{0-1}(h) \quad = \quad \mathbf{E}_{P_{XY}}\left[l_{0-1}(h(\mathbf{x}),y)\right]\,. \tag{9}$$

where l_{0-1} is given by

$$l_{0-1}(\hat{y},y) \quad = \quad \begin{cases} 1 & \hat{y} \neq y \\ 0 & \hat{y} = y \end{cases}\,. \tag{10}$$

Proof. If we use the relationship $P(\mathbf{x},r_i) = P(r_i|\mathbf{x})P(\mathbf{x})$, Equation (9) becomes

$$\begin{aligned} R_{0-1}(h) &= \int \left[\sum_{i=1}^{q} l_{0-1}(r_i, h(\mathbf{x}))P(r_i|\mathbf{x})\right] P(\mathbf{x})\, d\mathbf{x} && (11)\\ &= \int \mathcal{Q}_{0-1}(\mathbf{x},h)P(\mathbf{x})\, d\mathbf{x}\,. && (12)\end{aligned}$$

Then use

$$\begin{aligned} \mathcal{Q}_{0-1}(\mathbf{x},h) &= \sum_{i=1}^{q} P(r_i|\mathbf{x}) - P(h(\mathbf{x})|\mathbf{x}) && (13)\\ &= 1 - P(h(\mathbf{x})|\mathbf{x})\,, && (14)\end{aligned}$$

which is minimized for every $\mathbf{x}$ iff $h(\mathbf{x}) = \arg\max_{r_i \in Y} P(r_i|\mathbf{x})$. Inserting (14) into (12) gives

$$R_{0-1}(h) \quad = \quad 1 - \int P(h(\mathbf{x})|\mathbf{x})P(\mathbf{x})\, d\mathbf{x}\,. \tag{15}$$

□

A closer look at Equation (15) shows that a sufficient condition for two mappings h_1 and h_2 to incur equal risks $R_{0-1}(h_1)$ and $R_{0-1}(h_2)$ is given by $P(h_1(\mathbf{x})|\mathbf{x}) = P(h_2(\mathbf{x})|\mathbf{x})$ for every $\mathbf{x}$. Assuming that $P(r_i|\mathbf{x})$ is one for every $\mathbf{x}$ at a certain rank r_k the risks are equal — independent of how "far away" (in terms of rank difference) the mapping $h_1(\mathbf{x})$ and $h_2(\mathbf{x})$ are from the optimal rank $\arg\max_{r_i \in Y} P(r_i|\mathbf{x})$. This property does not take into account that Y is an ordered space. Thus, R_{0-1} is inappropriate for the case where a natural ordering is defined on the elements of Y.

The distribution independent risk functional Taking into account the abovementioned requirement we argue that a distribution independent model of ordinal regression has to single out that function h^*_{pref} which induces the ordering of the space X that incurs the smallest number of inversions on pairs $(\mathbf{x}_1, \mathbf{x}_2)$ of objects (for a similar reasoning see Sobel 1990). To model this property we note that due to the ordering of the space Y, each mapping h induces an ordering on the space X by Equation (7). Let use define the rank difference $\ominus : Y \times Y \mapsto \mathcal{N}$ by

$$r_i \ominus r_j \quad \equiv \quad i - j \,. \tag{16}$$

Now given a pair $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$ of objects we distinguish between two different events: $y_1 \ominus y_2 > 0$ and $y_1 \ominus y_2 < 0$. According to Equation (7) a function h violates the ordering if $y_1 \ominus y_2 > 0$ and $h(\mathbf{x}_1) \ominus h(\mathbf{x}_2) \leq 0$, or $y_1 \ominus y_2 < 0$ and $h(\mathbf{x}_1) \ominus h(\mathbf{x}_2) \geq 0$. Additionally taking into account that each weak order $\succ_{\text{Y}}$ induces an equivalence $\sim_{\text{Y}}$ (see Appendix A) the case $y_1 \ominus y_2 = 0$ is automatically taken care of. Thus, a risk functional that takes into account the ordering of the space Y is given by

$$R^s_{\text{pref}}(h) \quad = \quad \mathbf{E}_{P_{XXYY}} \left[l^s_{\text{pref}}(h(\mathbf{x}_1), h(\mathbf{x}_2), y_1, y_2) \right] \,. \tag{17}$$

with

$$l^s_{\text{pref}}(\hat{y_1}, \hat{y_2}, y_1, y_2) \quad = \quad \begin{cases} 1 & \text{if} \quad 0 < y_1 \ominus y_2 \leq s \quad \text{and} \quad \hat{y_1} \ominus \hat{y_2} \geq 0 \\ 1 & \text{if} \quad 0 < y_2 \ominus y_1 \leq s \quad \text{and} \quad \hat{y_2} \ominus \hat{y_1} \geq 0 \\ 0 & \text{else} \end{cases} \tag{18}$$

Note that by definition

$$P_{XXYY}(\mathbf{x}_1, \mathbf{x}_2, y_1, y_2) \quad = \quad P_{XY}(\mathbf{x}_1, y_1) P_{XY}(\mathbf{x}_2, y_2) \,, \tag{19}$$

and thus given ℓ i.i.d. samples drawn according to P_{XY} we obtain ℓ^2 i.i.d. samples drawn according to P_{XXYY}, a fact to be exploited later.

The role of the parameter s and pointwise losses The role of the parameter s is to allow for an incorporation of a priori knowledge about $P(r_i|\mathbf{x})$. To highlight this characteristic rewrite Equation (17) by using Equation (19), expanding the expectation value, and making use of $P_{XY}(y, \mathbf{x}) = P_{Y|X=\mathbf{x}}(y|\mathbf{x}) P_X(\mathbf{x})$ to

$$R^s_{\text{pref}}(h) \quad = \quad \int \sum_{i=1}^{q} \left[\int \sum_{j=1}^{q} l^s_{\text{pref}}(h(\mathbf{x}_1), h(\mathbf{x}_2), r_i, r_j) P_{Y|X=\mathbf{x}_2}(r_j|\mathbf{x}_2) P_X(\mathbf{x}_2) \, d\mathbf{x}_2 \right] \times$$
$$P_{Y|X=\mathbf{x}_1}(r_i|\mathbf{x}_1) P_X(\mathbf{x}_1) \, d\mathbf{x}_1 \,. \tag{20}$$

Now assume that for each $\mathbf{x}$, $P_{Y|X=\mathbf{x}}(r_i|\mathbf{x})$ is peaked at a certain rank r_k and zero everywhere else. Thus, the variation of $P_{Y|X=\mathbf{x}}(r_i|\mathbf{x})$ is as small as possible. Then, independent of l^s_{pref} the term in the brackets is unequal to zero only for rank r_k. If we now exchange the role of $P_{Y|X=\mathbf{x}}(r_i|\mathbf{x})$ and l^s_{pref} in the argument, we see how s controls the amount of assumed variation of $P_{Y|X=\mathbf{x}}(r_i|\mathbf{x})$. This can be treated as an assumption on the concentration the probability $P_{Y|X=\mathbf{x}}(r_i|\mathbf{x})$ around a “true” rank. We will later see that s effectively controls the complexity of the learning procedure.

The risk functional (20) allows for another interpretation if one takes into account that the term in the brackets only depends on r_i, $\mathbf{x}_1$, and h. Thus, if we define

$$l_h^s(\hat{y},\tilde{y}) \quad = \quad \mathbf{E}_{P_{XY}}\left[l_{\text{pref}}^s(h(\mathbf{x}),\hat{y},y,\tilde{y})\right] \,, \tag{21}$$

the risk functional (17) can be rewritten as the expectation value of l_h^s

$$R_{\text{pref}}^s(h) \quad = \quad \mathbf{E}_{P_{XY}}\left[l_h^s(h(\mathbf{x}),y)\right] \,. \tag{22}$$

Although Equation (22) shows great similarity with the classification learning risk functional (9) we see that due to the loss function l_{pref}^s, which exploits the ordinal nature of Y, we have for each h a different pointwise loss function. Thus we have found a risk functional which can be used for ordinal regression and takes into account the ordering as proposed by McCullagh 1980. We will present more appealing properties of the risk functional (17) in Section 4.

Application of the ERM principle The ERM principle recommends to take that mapping h^ℓ which minimizes the empirical risk $R_{\text{emp}}^s(h;S)$. While $R_{\text{pref}}^s(h)$ is the expectation value of the function $l_{\text{pref}}{}^s$, $R_{\text{emp}}^s(h;S)$ is defined by

$$R_{\text{emp}}^s(h;S) \quad = \quad \frac{1}{\ell^2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell} l_{\text{pref}}^s(y_i,y_j,h(\mathbf{x}_i;\mathbf{w}),h(\mathbf{x}_j;\mathbf{w}))\,. \tag{23}$$

As one can see, the size of the new training set derived from the ℓ–sized training set S is ℓ^2 — which is unfavorable for practical applications. In order to reduce the complexity during learning we apply the following equivalence which will later be used to derive uniform convergence bounds on $R_{\text{pref}}^s(h)$. For notational simplification let us define the space $\mathcal{E}$ of events of pairs $\mathbf{x}$’s an y’s with unequal ranks (up to a rank difference of s) by

$$\mathcal{E} = \{(\mathbf{x}_i,\mathbf{x}_j,y_k,y_l)|\mathbf{x}_i \in X, \mathbf{x}_j \in X, y_l \in Y, y_l \in Y, 0 < |y_k \ominus y_l| \le s\}\,, \tag{24}$$

and define the risk of misclassifying a randomly drawn pair from $\mathcal{E}$ by h by

$$R_{\text{pref}}^{0-1}(h) \quad = \quad \mathbf{E}_{P_{\mathcal{E}}}\left[l_{0-1}(\Omega(h(\mathbf{x}_1),h(\mathbf{x}_2)),\Omega(y_1,y_2))\right] \,. \tag{25}$$

Furthermore, using the shorthand notation $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ to denote the first and second object of a pair a new training set $S' : X \times X \times \{-1,+1\}$ can be derived from S if we use all 2–sets $\left\{(\mathbf{x}_i^{(1)},y_i^{(1)}),(\mathbf{x}_i^{(2)},y_i^{(2)})\right\}$ from S

$$\forall\ 0 < |y_i^{(1)} - y_i^{(2)}| \le s \qquad S' \quad = \quad \left\{\left((\mathbf{x}_i^{(1)},\mathbf{x}_i^{(2)}),\Omega\left(y_i^{(1)},y_i^{(2)}\right)\right)\right\}_{i=1}^{t} \tag{26}$$

$$\Omega(y_1,y_2) \quad = \quad \text{sign}(y_1 \ominus y_2)\,, \tag{27}$$

where Ω is an indicator function for rank differences and t is the cardinality of S'.

Theorem 2. *Assume an unknown probability measure P_{XY} on XY is given. Furthermore assume that $P_{Y|X=\mathbf{x}}(y|\mathbf{x})$ is unequal to zero only for a sequence of less or equal to s ranks. Then for each $h : X \mapsto Y$ the following equality holds true*

$$R^s_{\mathit{pref}}(h) = (1 - P(y_1 \sim_{\scriptscriptstyle Y} y_2))R^{0-1}_{\mathit{pref}}(h)\,, \tag{28}$$

where

$$P(y_1 \sim_{\scriptscriptstyle Y} y_2) = \mathbf{E}_{P_{XXYY}}\left[1 - |\Omega(y_1, y_2)|\right] . \tag{29}$$

Corollary 1. *Assume a training set S of size ℓ drawn i.i.d. according to an unknown probability measure P_{XY} on $X \times Y$ is given. Then for each $h : X \mapsto Y$ the following equality holds true*

$$R^s_{\mathit{emp}}(h; S) = \frac{t}{\ell^2} R^{0-1}_{\mathit{emp}}(h; S') \tag{30}$$

where

$$R^{0-1}_{\mathit{emp}}(h; S') = \frac{1}{t}\sum_{i=1}^{t} l_{0-1}\left(\Omega\left(h(\mathbf{x}_i^{(1)}), h(\mathbf{x}_i^{(2)})\right), \Omega\left(y_i^{(1)}, y_i^{(2)}\right)\right) . \tag{31}$$

Proof. The theorem is proved by explicitly deriving $P_{\mathcal{E}}$ for a given P_{XY}. If we expand the probability of occurrence of any event from $XXYY$ we obtain

$$\begin{aligned} &P_{XXYY}(\mathbf{x}_1 \in X, \mathbf{x}_2 \in X, y_1 \in Y, y_2 \in Y) = \\ &\qquad \underbrace{\int\int \sum_{i=1}^{q} \sum_{\substack{j=1 \\ i \neq j}}^{q} P_{XXYY}(\mathbf{x}_1, \mathbf{x}_2, r_i, r_j)\; d\mathbf{x}_1\; d\mathbf{x}_2}_{1-\delta} + \\ &\qquad \underbrace{\int\int \sum_{i=1}^{q} P_{XXYY}(\mathbf{x}_1, \mathbf{x}_2, r_i, r_i)\; d\mathbf{x}_1\; d\mathbf{x}_2}_{\delta} . \end{aligned} \tag{32}$$

If $P_{Y|X=\mathbf{x}}(y|\mathbf{x})$ is unequal to zero only for a sequence of less or equal to s ranks then by definition events that account for the second term are excluded from $\mathcal{E}$ and thus the probability $P_{\mathcal{E}}$ derived from P_{XY} is given by

$$P_{\mathcal{E}}(\mathbf{x}_1, \mathbf{x}_2, y_1, y_2) = \begin{cases} 0 & y_1 = y_2 \\ P_{XXYY}(\mathbf{x}_1, \mathbf{x}_2, y_1, y_2)/(1-\delta) & y_1 \neq y_2 \end{cases} . \tag{33}$$

Using the fact that $R^s_{\text{pref}}(h) \leq (1-\delta)$ (since every time $y_i = y_j$, $l^s_{\text{pref}}(\cdot, \cdot, y_i, y_j) = 0$ (see Equation (18))) and $P_{\mathcal{E}}$ is equal to $P_{XXYY}/(1-\delta)$ we know that $R^s_{\text{pref}}(h) = (1-\delta)R^{0-1}_{\text{pref}}$. Finally we see that δ is given by $P(y_1 \sim_{\scriptscriptstyle Y} y_2)$ (see Equation (29)) which proves the second statement. The corollary follows directly from the theorem if we use $P_{XY} = P_S$. Then, the term $P(y_1 \sim_{\scriptscriptstyle Y} y_2)$ becomes $1 - t/\ell^2$ which proves the corollary. □

The importance of the parameter s becomes apparent as the size t of the new training set S' mainly depends on s due to the restriction to pairs with rank difference less than or equal to s, i.e. the higher s the higher the size t of the new training set S'. Taking into account that each function $h \in \mathcal{H}$ defines a function $p : X \times X \mapsto \{-1, 0, +1\}$ by

$$p(\mathbf{x}_1, \mathbf{x}_2) \quad = \quad \Omega(h(\mathbf{x}_1), h(\mathbf{x}_2)) \,, \tag{34}$$

Corollary 1 states that the empirical risk of a certain mapping h on a sample S is equivalent to the l_{0-1} loss of the related mapping p on the sample S' up to a constant factor t/ℓ^2 which depends neither on h nor on p. Thus, the problem of distribution independent ordinal regression can be reduced to a classification problem on pairs of objects. It is important to emphasize the chain of argumentation that lead to this equivalence. The original problem was to find a function h^ℓ that maps objects to ranks given a sample S. By taking the ordinal nature of ranks into account this results in the equivalent formulation of finding a function p^ℓ that maps pairs of objects to the three classes $\succ_{\mathrm{Y}}$, $\prec_{\mathrm{Y}}$, and $\sim_{\mathrm{Y}}$. Reverting the chain of argumentation may lead to difficulties by observing that only those p are admissible — in the sense that there is a function h that fulfills Equation (34) — which define an asymmetric, transitive relation on X (see Appendix A). Therefore we call this problem also the problem of *preference learning*. It was shown that the Bayes optimal decision function on pairs of objects can result in a function p^ℓ which is no longer transitive on X (Herbrich et al. 1998). This is also known as the problem of stochastic transitivity (Suppes et al. 1989). Note also that the demand of transitivity and asymmetry effectively reduces the space of admissible classification functions p acting on pairs of objects.

Distribution independent bounds on R^s_{pref} The main advantage of the distribution independent model for ordinal regression is given by the following theorem which is an application of uniform convergence bounds (Vapnik and Chervonenkis 1971; Pollard 1984; Vapnik 1982; Vapnik 1998)

Theorem 3. *Let P be a probability measure on XY, let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$ be an i.i.d. sample from P and let d be the VC dimension of the set of functions $\{l^s_{\mathit{pref}}(h(\cdot), h(\cdot), \cdot, \cdot) | h \in \mathcal{H}\}$. Then*

$$P\left\{\sup_{h\in\mathcal{H}} \left|R^s_{\mathit{pref}}(h) - R^s_{\mathit{emp}}(h;S)\right| > \varepsilon\right\} \quad < \quad 4\exp\left\{\left(\frac{d(1+\ln(2\ell^2/d))}{\ell^2} - \varepsilon^2\right)\ell^2\right\} ,$$

where the probability is taken over the random choice of S.

Corollary 2. *With probability $1-\delta$ the following inequality holds simultaneously for all $h \in \mathcal{H}$*

$$R^s_{\mathit{pref}}(h) \quad \leq \quad R^s_{\mathit{emp}}(h;S) + \sqrt{\frac{d(\ln 2\ell^2/d + 1) - \ln\delta/4}{\ell^2}} \,. \tag{35}$$

With probability $1-2\delta$ the following inequality holds simultaneously for $h^\ell = \arg\min_{h\in\mathcal{H}} R^s_{\mathit{emp}}(h;S)$ and $h^* = \arg\min_{h\in\mathcal{H}} R^s_{\mathit{pref}}(h)$

$$R^s_{\mathit{pref}}(h^\ell) \quad \leq \quad R^s_{\mathit{pref}}(h^*;S) + \sqrt{\frac{d(\ln 2\ell^2/d + 1) - \ln\delta/4}{\ell^2}} + \sqrt{-\frac{\ln\delta}{2\ell^2}} \,. \tag{36}$$

Proof. By virtue of (19) we obtained ℓ^2 i.i.d. samples out of ℓ i.i.d. samples drawn according to P_{XY}. Thus the proof of the theorem is a mere application of the uniform convergence bounds for the set of indicator functions (see Theorem 3.1 in Vapnik 1995). The first statement of Corollary 2 is obtained if we set the right–hand side of Theorem 3 to δ and solve for ε. The second statement is valid is we use

$$R^s_{\text{emp}}(h^\ell; S) \quad \leq \quad R^s_{\text{emp}}(h^*; S)\,, \tag{37}$$

and by virtue of Hoeffdings inequality (Hoeffding 1963) with probability $1-\delta$

$$R^s_{\text{emp}}(h^*; S) \quad \leq \quad R^s_{\text{pref}}(h^*) + \sqrt{-\frac{\ln\delta}{2\ell^2}}\,. \tag{38}$$

If we combine the first statement of Corollary 2 for h^ℓ with these two inequalities we obtain the second statement of Corollary 2. □

Note, that in the above theorem the VC dimension of the induced loss function class $\{l^s_{\text{pref}}(h(\cdot), h(\cdot), \cdot, \cdot)|h \in \mathcal{H}\}$ is involved which can be seen as the maximum number of objects that can be arranged in any order by the set of function $\mathcal{H}$ using Equation (7). The major merit of Theorem 3 is given by the fact that the VC dimension need not coincide with the number of free parameters and thus it allows to derive algorithms which overcome the "curse of dimensionality" for the problem of ordinal regression (see Section 5). Moreover, Theorem 3 bounds the risk of a maximum likelihood solution $\tilde{\mathbf{w}}$ (see Section 2) — even if the distributional assumption does not hold true. A drawback of the given theorem is that it is a "worst case" bound and thus it is very loose, i.e., to bound the true risk $R^s_{\text{pref}}(h^\ell)$ it assumes the most pathological distribution P_{XY} underlying the data.

A much tighter bound can be found if we make the following assumption: In addition to the set $\mathcal{H}$ of functions that map objects to ranks we are given a set $\mathcal{F} = \{f : X \mapsto \mathcal{R}\}$ of functions that map objects to the real line. Now thresholding the value $f(\mathbf{x})$ at certain points $\theta(r_1), \ldots, \theta(r_{q-1})$ gives

$$h(\mathbf{x}) = r_i \quad \Leftrightarrow \quad f(\mathbf{x}) \in [\theta(r_{i-1}), \theta(r_i)]\,. \tag{39}$$

Note the close relationship to the cumulative model given in Section 2. Using $\mathcal{F}$ and $\boldsymbol{\theta}$ instead of $\mathcal{H}$ we can give a much tighter bound which also justifies the algorithm presented in Section 5. Conceptually, the following theorem is an application of the data–dependent structural risk minimization bounds (see, e.g., (Shawe-Taylor et al. 1996) and (Alon et al. 1997)).

Theorem 4. *Assume that for a given set $\mathcal{H}$ of mappings from objects to ranks there exists a set $\mathcal{F}$ of mappings from objects to $\mathcal{R}$ such that for each function $h \in \mathcal{H}$ there is a function $f \in \mathcal{F}$ (and vice versa) with*

$$h(\mathbf{x}) = r_i \quad \Leftrightarrow \quad f(\mathbf{x}) \in [\theta(r_{i-1}), \theta(r_i)]\,. \tag{40}$$

Let P be a probability measure on XY, let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^\ell$ be an i.i.d. sample from P, S' be derived from S by Equation (26) *and the fat–shattering dimension of the set of* *functions $\mathcal{F}$ be bounded above by the function afat* $: \mathcal{R} \mapsto \mathcal{N}$*. Then for each function* h^ℓ *with* $R^{0-1}_{\text{emp}}(h^\ell; S') = 0$ *and* $\gamma = \min_{S'} |f^\ell(\mathbf{x}^{(1)}) - f^\ell(\mathbf{x}^{(2)})|$ *with probability* $1 - \delta$

$$R^s_{\text{pref}}(h^\ell; S) \leq \frac{2}{t}\left(k \log_2\left(\frac{8et}{k}\right)\log_2(32t) + \log_2\left(\frac{8t}{\delta}\right)\right)(1 - P(y_1 \sim_Y y_2)) \,, \quad (41)$$

where $k =$ *afat*$(\gamma/8) \leq et$ *and* $t = |S'|$*.*

Proof. Let us recall a theorem given in Shawe-Taylor et al. 1996.

Theorem 5. Consider a real valued function class $\mathcal{F}$ having fat shattering function bounded above by the function afat $: \mathcal{R} \mapsto \mathcal{N}$ which is continuous from the right. Fix $\theta \in \mathcal{R}$. The with probability $1 - \delta$ a learner who correctly classifies ℓ i.i.d. generated examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)$ with $h^\ell = T_\theta(f) \in T_\theta(\mathcal{F})$ such that $h^\ell(\mathbf{x}_i) = y_i, i = 1, \ldots, \ell$ and $\gamma = \min |f(\mathbf{x}_i) - \theta|$ will have error of h^ℓ bounded from above by

$$\frac{2}{\ell}\left(k \log_2\left(\frac{8e\ell}{k}\right)\log_2(32\ell) + \log_2\left(\frac{8\ell}{\delta}\right)\right) \,, \quad (42)$$

where $k = \text{afat}(\gamma/8) \leq \ell$.

Now taking into account that from the ℓ^2 i.i.d. samples from P_{XXYY} we obtain t i.i.d. samples drawn from $P_{\mathcal{E}}$ and that a classification of a pair is carried out by decision based on the difference $f(\mathbf{x}_i^{(1)}) - f(\mathbf{x}_i^{(2)})$ we can bound $R^{0-1}_{\text{pref}}(h^\ell)$ above by substituting each ℓ with t and using $\theta = 0$. Utilizing the result of Theorem 2 proves the theorem. □

The afat(γ)–shattering dimension of $\mathcal{F}$ can be thought of as the maximum number of objects that can be arranged in any order using functions from $\mathcal{F}$ and a minimum margin γ of $|f(\mathbf{x}_1) - f(\mathbf{x}_2)|$ (utilizing Equation (7) together with (40)). It is worthwhile to see that maximizing the margin $\min_{S'} |f^\ell(\mathbf{x}^{(1)}) - f^\ell(\mathbf{x}^{(2)})|$ decreases the bound on the true risk while keeping $R^s_{\text{emp}}(h^\ell; S') = 0$ constant for some functions h.

4 Properties of the Proposed Risk

In this section we give some properties of the proposed risk functional (17). First of all we give necessary and sufficient conditions of the function h^*_{pref} to minimizes the risk functional. Then, we show that $R^s_{\text{pref}}(h)$ measures the rank coefficient between the ranking induced by h^*_{pref} and h on the space X (see Equation (7)). Finally, we argue why more naive approaches to ordinal regression problems can only insufficiently capture the existing dependency within the data.

Theorem 6. *The function* $h^*_{\text{pref}} : X \mapsto Y$ *obeys*

$$h^*_{\text{pref}}(\mathbf{x}_1) \succ_Y h^*_{\text{pref}}(\mathbf{x}_2) \quad \Leftrightarrow \quad P(y_1 \succ_Y y_2 | \mathbf{x}_1, \mathbf{x}_2) > P(y_2 \succ_Y y_1 | \mathbf{x}_1, \mathbf{x}_2) \quad (43)$$

$$h^*_{\text{pref}}(\mathbf{x}_2) \succ_Y h^*_{\text{pref}}(\mathbf{x}_1) \quad \Leftrightarrow \quad P(y_2 \succ_Y y_1 | \mathbf{x}_1, \mathbf{x}_2) > P(y_1 \succ_Y y_2 | \mathbf{x}_1, \mathbf{x}_2) \quad (44)$$

$$h^*_{\text{pref}}(\mathbf{x}_1) \sim_Y h^*_{\text{pref}}(\mathbf{x}_2) \quad \Leftrightarrow \quad P(y_2 \succ_Y y_1 | \mathbf{x}_1, \mathbf{x}_2) = P(y_1 \succ_Y y_2 | \mathbf{x}_1, \mathbf{x}_2) \,, \quad (45)$$

where

$$
\begin{aligned}
P(y_1 \succ_Y y_2 | \mathbf{x}_1, \mathbf{x}_2) &= \sum_{i=1}^{q} P(r_i|\mathbf{x}_1) \sum_{j=i+1}^{q} P(r_j|\mathbf{x}_2) && (46)\\
P(y_2 \succ_Y y_1 | \mathbf{x}_1, \mathbf{x}_2) &= \sum_{i=1}^{q} P(r_i|\mathbf{x}_1) \sum_{j=1}^{i-1} P(r_j|\mathbf{x}_2)\,. && (47)
\end{aligned}
$$

minimizes the risk functional (17) *over all functions* $h : X \mapsto Y$.

Proof. If we rewrite the risk functional (17) by

$$
\begin{aligned}
R^s_{\text{pref}}(h) &= \int\int \mathcal{Q}(h, \mathbf{x}_1, \mathbf{x}_2) P(\mathbf{x}_1) P(\mathbf{x}_2)\; d\mathbf{x}_1\; d\mathbf{x}_2 && (48)\\
\mathcal{Q}(h, \mathbf{x}_1, \mathbf{x}_2) &= \sum_{i=1}^{q} P(r_i|\mathbf{x}_1) \sum_{j=1}^{q} P(r_j|\mathbf{x}_2)\; l^s_{\text{pref}}(h(\mathbf{x}_1), h(\mathbf{x}_2), r_i, r_j)\,, && (49)
\end{aligned}
$$

we can decompose $\mathcal{Q}(h, \mathbf{x}_1, \mathbf{x}_2) = \mathcal{Q}_1(h, \mathbf{x}_1, \mathbf{x}_2) + \mathcal{Q}_2(h, \mathbf{x}_1, \mathbf{x}_2) + \mathcal{Q}_3(h, \mathbf{x}_1, \mathbf{x}_2)$

$$
\begin{aligned}
Q_1(h, \mathbf{x}_1, \mathbf{x}_2) &= \sum_{i=1}^{q} P(r_i|\mathbf{x}_1) \sum_{j=i+1}^{q} P(r_j|\mathbf{x}_2)\; l^s_{\text{pref}}(h(\mathbf{x}_1), h(\mathbf{x}_2), r_i, r_j) && (50)\\
Q_2(h, \mathbf{x}_1, \mathbf{x}_2) &= \sum_{i=1}^{q} P(r_i|\mathbf{x}_1) \sum_{j=1}^{i-1} P(r_j|\mathbf{x}_2)\; l^s_{\text{pref}}(h(\mathbf{x}_1), h(\mathbf{x}_2), r_i, r_j) && (51)\\
Q_3(h, \mathbf{x}_1, \mathbf{x}_2) &= \sum_{i=1}^{q} P(r_i|\mathbf{x}_1)^2\; l^s_{\text{pref}}(h(\mathbf{x}_1), h(\mathbf{x}_2), r_i, r_j)\,. && (52)
\end{aligned}
$$

By definition h^*_{pref} minimizes $\mathcal{Q}(h^*_{\text{pref}}, \mathbf{x}_1, \mathbf{x}_2)$ for each $\mathbf{x}_1$ and $\mathbf{x}_2$ and thus the whole integral is minimized. □

Using this result we conclude that minimizing $R^s_{\text{pref}}(h)$ for a restricted hypothesis space $\mathcal{H}$ gives that function h^* which best matches the ordering induced by the unknown P_{XY}. In order to measure the amount of discrepancy between h^*_{pref} and any h we proposed the risk functional (17). To substantiate this choice, we give the following result.

Theorem 7. *Let A and B be a set of ranks. The rank coefficient* Kendall's tau $\tau(A, B)$ *is defined by (see Kendall 1970)*

$$
\tau(A, B) = \frac{P - Q}{P + Q}\,, \tag{53}
$$

where $P = |\{a \succ_Y b | a \in A, b \in B\}|$ is the number of pairs (a, b) with higher rank of the first component of the pair and $P = |\{b \succ_Y a | a \in A, b \in B\}|$ is the number of pairs (a, b) with higher rank of the second component of the pair. Then, for every finite sample S and the sample S' derived from it (see Equation (26)*)*

$$
R^{0-1}_{\mathit{emp}}(h; S') = \frac{1 - \tau(A, B)}{2} \left(1 - R^{0-1}_{\mathit{emp}}(h^{\ell}_{\mathit{pref}}; S')\right) + R^{0-1}_{\mathit{emp}}(h^{\ell}_{\mathit{pref}}; S')\,, \tag{54}
$$

where $h^{\ell}_{\mathit{pref}} = \arg\min_{h\in\mathcal{H}} R^{0-1}_{\mathit{emp}}(h;S')$ is the finite–sample counterpart of h^{*}_{pref} (see Theorem 6) and $A = \{h(\mathbf{x})|(\mathbf{x},y)\in S\}$ and $B = \{h^{\ell}_{\mathit{pref}}(\mathbf{x})|(\mathbf{x},y)\in S\}$ are the sets of predicted ranks for the sample S by h and h^{ℓ}_{pref}, respectively.

Proof. By definition (31) of $R^{0-1}_{\text{emp}}(h;S')$ we see that

$$0 \quad \le \quad \frac{R^{0-1}_{\text{emp}}(h;S') - R^{0-1}_{\text{emp}}(h^{\ell}_{\text{pref}};S')}{1 - R^{0-1}_{\text{emp}}(h^{\ell}_{\text{pref}};S')} \le 1\,. \tag{55}$$

Whilst in the definition of Kendall's tau every inversion of a rank counts as minus one, the l_{0-1} loss gives every time an error of one. Thus, the quantity has to be rescaled to the interval starting from plus one to minus one. This is obtained by

$$\tau(A,B) \quad = \quad -2\frac{R^{0-1}_{\text{emp}}(h;S') - R^{0-1}_{\text{emp}}(h^{\ell}_{\text{pref}};S')}{1 - R^{0-1}_{\text{emp}}(h^{\ell}_{\text{pref}};S')} + 1\,. \tag{56}$$

Rearranging all terms proves the theorem. □

From that point of view we see that minimization of $R^{s}_{\text{emp}}(h;S) \propto R^{0-1}_{\text{emp}}(h;S')$ is equivalent to minimizing Kendall's tau between h and h^{ℓ}_{pref} since all term on the r.h.s. of (54) do not depend on h. One can think of Kendall's tau $\tau(A,B)$ as a linear function of the number of steps bubble sort would need to resort A into B. In the line of Kendall (Kendall 1970) we argue that our loss constitutes a natural measures for comparing mapping from objects to ranks.

It is often argued that simpler approaches like (i) treating the ordinal regression problem as a multi–class classification problem or (ii) treating the rank indices as instantiations of a real valued function (Caruana et al. 1995) can be used to solve the ordinal regression problem. It should be clear, however, that option (i) neglects information contained in the training set due to the ordering of the classes or ranks, consequently considers an unnecessarily large hypothesis space and thus cannot achieve satisfactory generalization. The consequence of this deficiency will be demonstrated in the experimental section 6. On the other hand, option (ii) assumes a metric structure in the space Y of ranks and is thus not invariant under the choice of representation of ranks by real numbers. An example will be given in Section 6. It can be concluded that only the above defined loss which captures the number of incurred inversion and thereby uses only order information (see the previous theorem) strikes a balance between the insufficient assumption on which the classification approach is based and the too strong assumption of the metric regression approach.

5 An Algorithm for the Latent Utility Model

In this section we apply the theory from Section 3 to derive an algorithm for ordinal regression. By Equation (7) and the strict ordering of Y we know that $\succ_{\mathrm{X}}$ induced by each h is a weak order. Using the representation theorem for weak orders (see Appendix A) there must be a function $\mu^* : X \mapsto \mathcal{L}$ such that

$$\mathbf{x}_i \succ_{\mathrm{X}} \mathbf{x}_j \quad \Leftrightarrow \quad \mu^*(\mathbf{x}_i) \succ_{\mathcal{L}} \mu^*(\mathbf{x}_j)\,, \tag{57}$$

where $\succ_{\mathcal{L}}$ is weak order on the set $\mathcal{L}$. In measurement theory the function μ^* is called a *scale*. The relation $\succ_{\mathrm{X}}$ is a weak order with a finite number of ranks. Now similar to the cumulative model we suggest to model ranks as disjunctive intervals on the real line. We denote such a set of disjunctive intervals by $\mathcal{I}$. The relation $\succ_{\mathcal{I}}$ and the induced equivalence $\sim_{\mathcal{I}}$ are defined as follows

$$[a,b] \succ_{\mathcal{I}} [c,d] \quad \Leftrightarrow \quad a \geq d \tag{58}$$

$$[a,b] \sim_{\mathcal{I}} [c,d] \quad \Leftrightarrow \quad a = c \wedge b = d\,. \tag{59}$$

If follows from these definitions that

$$\forall x \in [a,b], y \in [c,d] \qquad [a,b] \succ_{\mathcal{I}} [c,d] \quad \Rightarrow \quad x > y \tag{60}$$

The Linear Utility Model Let us introduce a linear function $U : X \mapsto \mathcal{R}$

$$U(\mathbf{x};\mathbf{w}) \quad = \quad \mathbf{w}^T\mathbf{x} \tag{61}$$

which relates to the above mentioned scale $\mu : X \mapsto \mathcal{L}$ by

$$\mu(\mathbf{x}) \quad = \quad [\theta(r_{i-1}), \theta(r_i)] \Leftrightarrow U(\mathbf{x};\mathbf{w}) \in [\theta(r_{i-1}), \theta(r_i)]\,, \tag{62}$$

where we assume that $\theta(r_0) = -\infty$ and $\theta(r_q) = +\infty$. Such a function is called a *utility function*. To relate this measurement theoretic approach to our problem note that each scale induces a mapping h of objects to ranks by the following rule

$$h(\mathbf{x};\mathbf{w}) = r_i \quad \Leftrightarrow \quad \mu(\mathbf{x}) = [\theta(r_{i-1}), \theta(r_i)] \Leftrightarrow U(\mathbf{x};\mathbf{w}) \in [\theta(r_{i-1}), \theta(r_i)]\,. \tag{63}$$

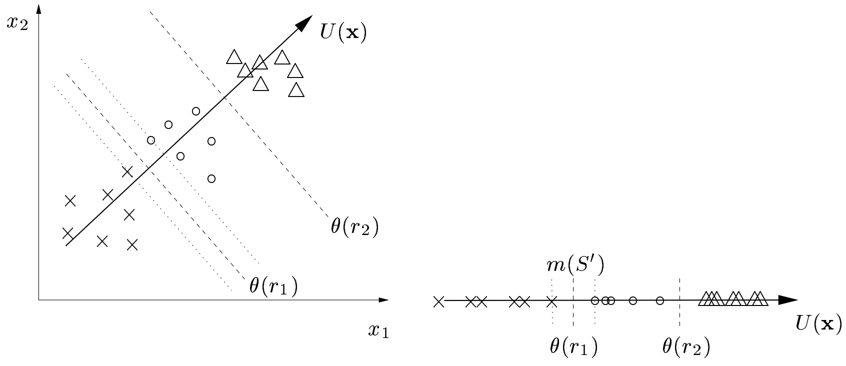
From Equations (60) and (62) we know that $U(\mathbf{x};\mathbf{w})$ incurs no error for the i-th example in the training set S' (see Equation (26)) iff

$$z_i = +1 \quad \Rightarrow \quad \mathbf{w}^T\mathbf{x}_i^{(1)} > \mathbf{w}^T\mathbf{x}_i^{(2)} \Leftrightarrow \mathbf{w}^T\left(\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)}\right) > 0 \tag{64}$$

$$z_i = -1 \quad \Rightarrow \quad \mathbf{w}^T\mathbf{x}_i^{(2)} > \mathbf{w}^T\mathbf{x}_i^{(1)} \Leftrightarrow \mathbf{w}^T\left(\mathbf{x}_i^{(2)} - \mathbf{x}_i^{(1)}\right) > 0\,, \tag{65}$$

where $z_i = \Omega(y_i^{(1)}, y_i^{(2)})$ was used. Note that the preference relation is expressed in terms of the difference between feature vectors $\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)}$, which can be thought of as the combined feature vector of the pair of objects. By assuming a finite margin between the n-dimensional feature vectors $\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)}$ of classes $z_i = +1$ and $z_i = -1$, we make the constraints (64) and (65) stronger and obtain:

$$z_i \left[\mathbf{w}^T\left(\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)}\right)\right] \quad \geq \quad 1 - \xi_i \qquad i = 1, \ldots, t\,, \tag{66}$$



(a) (b)

Figure 1: (**a**) Mapping of objects from rank r_1 ($\times$), rank r_2 ($\circ$), and rank r_3 ($\triangle$) to the axis $U(\mathbf{x})$, where $\mathbf{x} = (x_1, x_2)^T$. Note that by $\theta(r_1)$ and $\theta(r_2)$ two coupled hyperplanes are defined. (**b**) The margin of the coupled hyperplanes $m(S') = \min_{S'} |U(\mathbf{x}_i^{(1)}) - U(\mathbf{x}_i^{(2)})|$ is this time defined at the rank boundaries $\theta(r_i)$.

where the positive ξ_i-s measure the degree of violation of the i-th constraint. Assume that there are vectors $\mathbf{w}$ which fulfill the t constraints given by Equation (66). The weight vector $\mathbf{w}^\ell$ which maximizes the margin — this time at the rank boundaries $\theta(r_i)$ (see Equation (62) and Figure 1) — can now be determined by minimizing the squared norm $\|\mathbf{w}\|^2 + C \sum_{i=1}^t \xi_i$ under the constraints (66). This approach is closely related to the idea of canonical hyperplanes used in Support Vector Classification (Vapnik 1995). A theoretical justification for the applicability of SRM is given by Theorem 3, whereas Theorem 4 bounds $R^s_{\text{pref}}(\mathbf{w})$ above for the particular case of $C = \infty$. Introducing Lagrangian multipliers the primal optimization problem becomes

$$\begin{aligned}(\mathbf{w}^\ell, \boldsymbol{\xi}^\ell, \boldsymbol{\alpha}^\ell, \boldsymbol{\beta}^\ell) &= \min_{\mathbf{w}, \boldsymbol{\xi} \geq \mathbf{0}} \max_{\boldsymbol{\alpha} \geq \mathbf{0}, \boldsymbol{\beta} \geq \mathbf{0}} \left[\frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^t \xi_i - \right. \\ &\left. \sum_{i=1}^t \alpha_i \left[z_i \mathbf{w}^T \left(\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)} \right) - 1 + \xi_i \right] - \sum_{i=1}^t \beta_i \xi_i \right] . \quad (67)\end{aligned}$$

Performing unconstrained optimization with respect to $\mathbf{w}$ leads to the dual problem

$$\boldsymbol{\alpha}^\ell = \max_{\substack{C \geq \boldsymbol{\alpha} \geq \mathbf{0} \\ \boldsymbol{\alpha}^T \mathbf{z} = 0}} \left[\sum_{i=1}^t \alpha_i - \frac{1}{2} \sum_{i,j=1}^t \alpha_i \alpha_j z_i z_j \left(\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)} \right)^T \left(\mathbf{x}_j^{(1)} - \mathbf{x}_j^{(2)} \right) \right] \quad (68)$$

with $\mathbf{z} = (z_1, \ldots, z_t)^T$. This problem is a standard QP–problem and can efficiently be solved using techniques from mathematical programming (Hadley 1964; Mangasarian 1969; Vanderbei 1997; Joachims 1998). Note, however, that due to the expansion of the last term in (68)

$$\left(\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)} \right)^T \left(\mathbf{x}_j^{(1)} - \mathbf{x}_j^{(2)} \right) = \mathbf{x}_i^{(1)T} \mathbf{x}_j^{(1)} - \mathbf{x}_i^{(1)T} \mathbf{x}_j^{(2)} - \mathbf{x}_i^{(2)T} \mathbf{x}_j^{(1)} + \mathbf{x}_i^{(2)T} \mathbf{x}_j^{(2)} ,$$

the solution $\boldsymbol{\alpha}^\ell$ to this problem can be calculated solely in terms of the inner products between the feature vectors without reference to the feature vectors themselves. This fact will be exploited in the following section for the generalization of the method to nonlinear utility functions.

Reconstruction of the Utility function Given the optimal vector $\boldsymbol{\alpha}^\ell$ as solution to the problem (68), the optimal weight vector $\mathbf{w}^\ell$ can be written as a linear combination of differences of feature vectors from the training set (Kuhn–Tucker conditions):

$$\mathbf{w}^\ell \quad = \quad \sum_{i=1}^{t} \alpha_i^\ell z_i \left(\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)}\right) . \tag{69}$$

All pairs of objects with $\alpha_i^t \neq 0$ support the construction of the optimal scale $U(\mathbf{x}; \mathbf{w}^\ell)$ in the space of pairs of objects, and are therefore referred to as support vectors (Cortes 1995). After learning, the utility function is represented by the vector $\boldsymbol{\alpha}^\ell$ together with the training set S'. Using (69) it is also possible to reconstruct the utility function up to an additive constant b as

$$U(\mathbf{x}; \mathbf{w}^\ell) \quad = \quad \mathbf{x}^T \mathbf{w}^t = \sum_{i=1}^{t} \alpha_i^\ell z_i \left(\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)}\right)^T \mathbf{x} \, . \tag{70}$$

This calculation benefits from the sparseness of the expansion (69), which significantly reduces its computational costs (Girosi 1997).

Estimation of the rank boundaries To estimate the rank boundaries we note that due to Equations (66) the difference in utility is greater or equal to one for all training examples with $\xi_i = 0$. Thus if $\Theta(k) \subset S'$ is the fraction of objects from the training set with $\xi_i = 0$ and rank difference $\ominus$ exactly one starting from rank r_k, i.e.

$$\Theta(k) \quad = \quad \left\{ \left(x_i^{(1)}, x_i^{(2)}\right) \middle| y_i^{(1)} = r_k \wedge y_i^{(2)} = r_{k+1} \wedge \xi_i = 0 \right\} \tag{71}$$

then the estimation of $\theta(r_k)$ is given by

$$\theta(r_k) \quad = \quad \frac{U(\mathbf{x}_1; \mathbf{w}^\ell) + U(\mathbf{x}_2; \mathbf{w}^\ell)}{2} \, , \tag{72}$$

where

$$(\mathbf{x}_1; \mathbf{x}_2) \quad = \quad \arg \min_{(\mathbf{x}_i, \mathbf{x}_j) \in \Theta(k)} \left[U(\mathbf{x}_i; \mathbf{w}^\ell) - U(\mathbf{x}_j; \mathbf{x}^\ell) \right] . \tag{73}$$

In other words, the optimal threshold $\theta(r_k)$ for rank r_k lies in the middle of the utilities of the closest (in the sense of their utility) objects of rank r_k and r_{k+1}. To detect those training examples with $\xi > 0$, we recall that in the unconstrained optimization of the primal problem given by Equation (67) $\alpha_i + \beta_i = C$. Therefore $\alpha_i = C$ is a necessary condition for $\xi_i > 0$. Thus, the condition $\xi_i = 0$ can be replaced by $\alpha_i < C$ which can be evaluated after solving the QP problem. After the estimation of the rank boundaries $\theta(r_k)$ a new object is classified according to Equation (63).

Extension to Nonlinear Utility Functions Now let us make a non–linear model of $U(\mathbf{x})$ by introducing a mapping $\Phi : X \mapsto \mathcal{X}$ of X into the so called "feature space" $\mathcal{X}$

$$U(\mathbf{x};\mathbf{w}) \quad = \quad \mathbf{w}^T\Phi(\mathbf{x})\,. \tag{74}$$

The computation of $\boldsymbol{\alpha}^\ell$ involves computation of the inner product of two difference vectors (see Equation (68)) and thus using the non–linear model (74) we have to replace each occurrence of $\mathbf{x}$ by $\Phi(\mathbf{x})$. This gives

$$\boldsymbol{\alpha}^\ell \quad = \quad \max_{\substack{C\geq\boldsymbol{\alpha}\geq\mathbf{0}\\ \boldsymbol{\alpha}^T\mathbf{z}=0}} \left[\sum_{i=1}^t \alpha_i - \frac{1}{2}\sum_{i,j=1}^t \alpha_i\alpha_j z_i z_j \mathcal{K}\left(\mathbf{x}_i^{(1)},\mathbf{x}_i^{(2)},\mathbf{x}_j^{(1)},\mathbf{x}_j^{(2)}\right)\right]\,, \tag{75}$$

where $\mathcal{K}$ is defined by

$$\mathcal{K}(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4) \quad = \quad (\Phi(\mathbf{x}_1)-\Phi(\mathbf{x}_2))^T(\Phi(\mathbf{x}_3)-\Phi(\mathbf{x}_4))\,. \tag{76}$$

Let us assume that the feature space $\mathcal{X}$ is a reproducing kernel hilbert space (rkhs) (Whaba 1990). Each rkhs is uniquely determined by the so called *kernel* $K : X\times X \mapsto \mathcal{R}$ which has the following property

$$K(\mathbf{x}_i,\mathbf{x}_j) \quad = \quad \Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j)\,, \tag{77}$$

where the inner product is taken in the space $\mathcal{X}$. Then the function $\mathcal{K}$ simplifies to

$$\mathcal{K}(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4) \quad = \quad K(\mathbf{x}_1,\mathbf{x}_3)-K(\mathbf{x}_1,\mathbf{x}_4)-K(\mathbf{x}_2,\mathbf{x}_3)+K(\mathbf{x}_2,\mathbf{x}_4) \tag{78}$$

which avoids explicit computation of the mapping $\mathbf{x} \mapsto \Phi(\mathbf{x})$. Finally, in order to estimate $\theta(r_k)$ (see Equation (72)) we have to compute $U(\mathbf{x};\mathbf{w}^\ell)$ which becomes

$$U(\mathbf{x};\mathbf{w}^\ell) = U(\mathbf{x};\boldsymbol{\alpha}^\ell) \quad = \quad \sum_{i=1}^t \alpha_i^\ell z_i\left(K\left(\mathbf{x}_i^{(1)},\mathbf{x}\right)-K\left(\mathbf{x}_i^{(2)},\mathbf{x}\right)\right)\,. \tag{79}$$

Thus by using the "kernel trick" we are able to deal with nonlinear utility functions $U(\mathbf{x})$ without running into computational difficulties. Moreover, as stated in Theorem 4 the bound on the risk $R^s_{\text{pref}}(\mathbf{w}^\ell)$ does not depend on the dimension of $\mathcal{X}$ but on the margin quantity $\min_{S'}|U(\mathbf{x}_i^{(1)})-U(\mathbf{x}_i^{(2)})|$ and thus our algorithm can even generalize in infinite–dimensional feature spaces $\mathcal{X}$.

6 Experimental Results

In this section we present some experimental results for the algorithm presented in Section 5. We start by giving results for artificial data which allows us to analyse our algorithm in a controlled setting. Then we give learning curves for an example from the field of Information Retrieval. In our retrieval model we assume that after an initial query the user is asked to rank a small number of documents returned by the IR system. From this ranking we want to find the ranking of the remaining documents which is in best accordance with the ranking given by the user.

6.1 Learning Curves for Ordinal Regression

In this experiment we want to compare the generalization behavior of our algorithm with the multi-class SVM and Support Vector regression (SVR) — the methods of choice, if one does not pay attention to the ordinal nature of Y and instead treats ranks as classes (classification) or continuous response values (regression estimation). Another reason to choose those algorithms for comparison is due to their similar regularizer $\|\mathbf{w}\|^2$ and hypothesis space $\mathcal{H}$ which makes them as comparable as possible. We generated 1000 observations $\mathbf{x} = (x_1, x_2)^T$ in the unit square $[0,1] \times [0,1] \subset \mathcal{R}^2$ according to a uniform distribution. We assigned to each observation $\mathbf{x}$ a value y according to

$$y = i \quad \Leftrightarrow \quad \underbrace{10((x_1 - 0.5) \cdot (x_2 - 0.5))}_{U(\mathbf{x})} + \epsilon \in [\theta(r_{i-1}), \theta(r_i)]\,, \qquad (80)$$

where ϵ was normally distributed, i.e. $\epsilon \sim N(0, 0.125)$, and $\boldsymbol{\theta} = (-\infty, -1, -0.1, 0.25, 1, +\infty)^T$ is the vector of predefined thresholds. In Figure 2 (a) the points $\mathbf{x}_i$ which are assigned to a different rank after the addition of the normally distributed quantity ϵ_i are shown. If we treat the whole task as a classification problem, we would call them incorrectly classified training examples. The solid lines in Figure 2 (a) indicate the "true" rank boundaries $\boldsymbol{\theta}$ on $U(\mathbf{x})$.

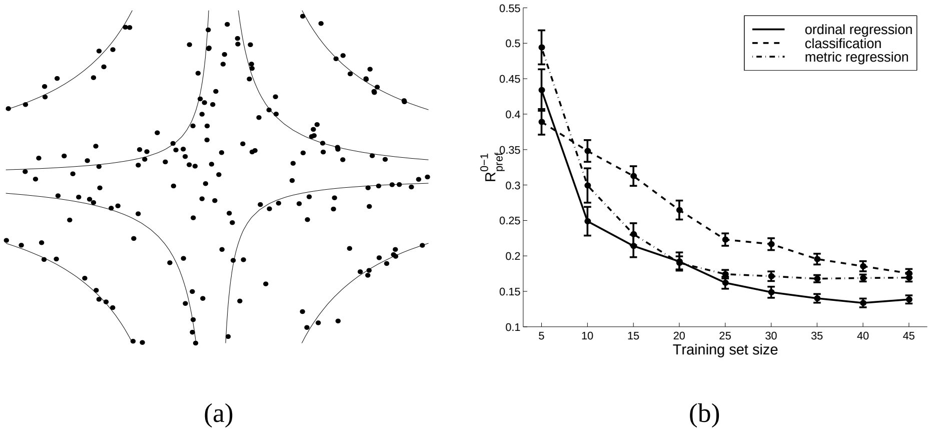


Figure 2: (**a**) Scatter plot of data points $\mathbf{x}$ which $U(\mathbf{x})$ maps to a different interval than $U(\mathbf{x}) + \epsilon$ (see Equation (80)). (**b**) Learning curves for multi-class SVM (dashed lines), SV regression (dashed–dotted line) and the algorithm for ordinal regression (solid line) if we measure R^{0-1}_{pref}. The error bars indicate the 95% confidence intervals of the estimated risk R^{0-1}_{pref}.

In order to compare the three different algorithms we randomly drew 100 training samples $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$ of training set sizes ℓ ranging from 5 to 45 and thereby making sure that at least one representative of each rank is within the drawn training set. Classification with multi-class SVM's was carried out by computing the pairwise $5 \cdot 4/2 = 10$ hyperplanes using the algorithm presented in Weston and Watkins 1998. For all algorithms, i.e. multi-class SVM's, SVR, and the algorithm presented in Section 5, we chose the kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + 1)^2$ and a trade-off parameter $C = 1000000$. In the particular case of Support Vector regression we used an $\varepsilon = 0.5$ for the ε–intensive loss function (see (Vapnik 1995) for the definition of this loss function) and thresholds $\boldsymbol{\theta} = (0.5, 1.5, 2.5, 3.5, 4.5)^T$ to transform real valued predictions into "ranks".

From the remaining 995 to 955 data points we estimated the risk R_{pref}^{0-1} and averaged over all 100 results for a given training set size. Thus we obtained the three learning curves shown in Figure 2 (b). We used R_{pref}^{0-1} instead of R_{pref}^{s} — which is larger by a constant factor of $(1 - P(y_1 \sim_{\text{Y}} y_2))$ (see Theorem 2). It can be seen that the algorithm proposed for ordinal regression generalizes much faster by exploiting the ordinal nature underlying Y compared to classification. This can be explained by the fact that due to the model of a latent utility all "hyperplanes" $U(\mathbf{x}) = \theta(r_k)$ are coupled (see Figure 1) which does not hold true for the case of multi-class SVM's. Furthermore, the learning curves for SVR and the proposed ordinal regression algorithm are very close which can be explained by the fact that the predefined thresholds $\theta(r_k)$ are defined in such a way that their pairwise difference is about 0.5 — the size of the ε–tube chosen beforehand. Thus the utility and the continuous ranks estimated by the regression algorithm are of the same magnitude which results in the same generalization behavior.

In Figure 3 we plotted the assignments of the unit square to ranks r_1 (black areas) to ranks r_5 (white areas) for the functions $h^\ell(\mathbf{x}; S')$ learned from randomly drawn training set ranging from size $\ell = 5$ (top row) to $\ell = 25$ (bottom row). We used the same parameters as for the computation of the learning curves. In the rightmost column (e) the true assignment, i.e. $y = r_i \Leftrightarrow U(\mathbf{x}) \in [\theta(r_{i-1}), \theta(r_i)]$ is shown. In the first column (a) we can see how the algorithm presented in Section 5 performs for varying training set sizes. As expected, for the training set size $\ell = 25$, the method found an utility function together with a set of thresholds which represent the true ranking very well. The second column (b) shows the results of the abovementioned multi-class SVM on the task. Here the pairwise hyperplanes are not coupled since the ordinal nature of Y is not taken into account. This results in a worse generalization, especially in regions, where no training points were given. The thirt column (c) gives the assignments made by the SVR algorithm if we represent each rank r_i by i. Similar to the good results seen in the learning curve, the generalization behavior is comparable to the ordinal regression method (first column). The deficiency of SVR for this task becomes apparent when we change the representation of ranks. In the fourth column (d) we applied the same SVR algorithm, this time on the representation $\exp(i)$ for rank r_i. As can be seen, this dramatically changes the generalization behavior of the SVR method. We conclude that the crucial task for application of metric regression estimation methods to the task of ordinal regression is the definition of the representation of ranks (see Section 4). This is automatically — but more time–consuming — solved by the proposed algorithm.

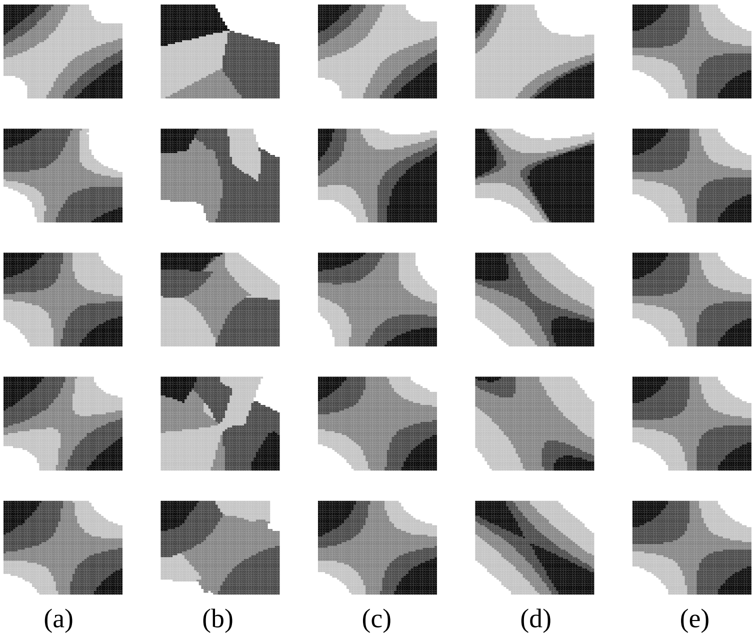
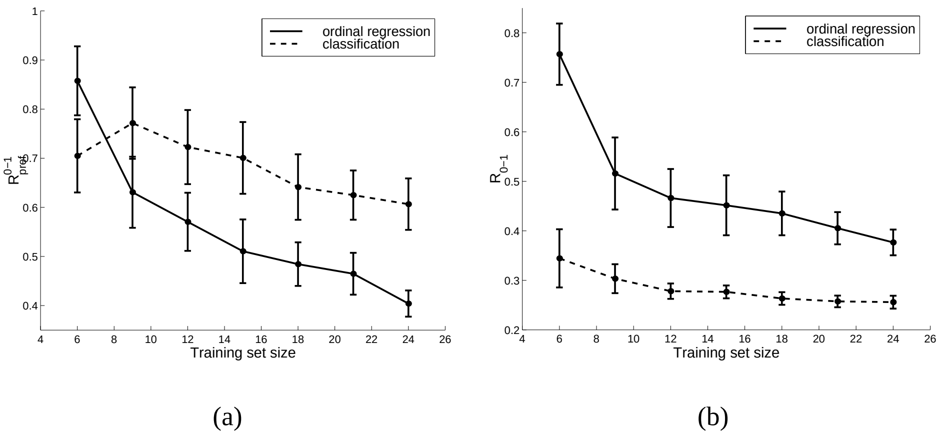


Figure 3: Assignments of points to ranks r_1 (black area) to r_5 (white area) by the learned function $h(\mathbf{x})$ based on randomly drawn training samples of size $5, 10, 15, 20,$ and 25 (top row to bottom row). (**a**) Results of the algorithm presented in Section 5. (**b**) Results of multi-class SVM if we treat each rank as a class. (**c**) Results of SVR if we assign rank r_i to number i. (**d**) Results of SVR if we assign rank r_i to real number $\exp(i)$. (**e**) Underlying assignment uncorrupted by noise.

6.2 An Application to Information Retrieval

In this experiment we make the following assumption: After an initial (textual) query a user makes to an IR system, the system returns a bundle of documents to the user. Now the user assigns ranks to a small fraction of the returned documents and the task for the learning algorithm is to assign ranks to the remaining unranked documents in order to rank the remaining documents. We assume that the quantity of interest is the number of inversions incurred by the ranking induced by the learning algorithm. This quantity is captured by R^{0-1}_{emp} (see Equation (31)) and thus after using $\ell = 6$ up to $\ell = 24$ documents and their respective ranking we measure R^{0-1}_{emp} on the remaining documents. For this experiment we used the same parameters as in the last experiment. The only publicly available dataset we found was the OHSUMED dataset collected by William Hersh, which consists of $348\ 566$ documents and 106 queries with their respective ranked results. There are three ranks: “document is relevant”, “document is partially relevant”, and “irrelevant document” w.r.t. the given textual query. For our experiments we used the results of query 1 (“Are there adverse effects on lipids when progesterone is given with estrogen replacement therapy?") which consists of 107 documents taken from the whole database. In order to apply our algorithm we used the "bag–of–words" representation (Salton 1968), i.e., we computed for every document the vector of "term–frequencies–inverse–document–frequencies" (TFIDF) components . We restricted ourselves to each term that appears at least three times in the whole database. This results in $\approx$ 1700 terms which leads for a certain document to a very high–dimensional but sparse vector. We normalized the length of each document vector to unity (see Joachims 1997).



(a)

(b)

Figure 4: Learning curves for multi-class SVM (dashed lines) and the algorithm for ordinal regression (solid line) for the OHSUMED dataset query 1 if we measure (**a**) R_{pref}^{0-1} and (**b**) R_{0-1}. Error bars indicate the 95% confidence intervals.

Figure 4 (a) shows the learning curves for multi-class SVM's and our algorithm for ordinal regression measured in term of the number of incurred inversions. As can be seen from the plot, the proposed algorithm shows very good generalization behavior compared to the algorithm which treats each rank as a separate class. Note that we plotted R_{pref}^{0-1} which is the proportion of misclassified pairs if we restrict ourself to pairs with different ranks. This quantity is much larger than the estimated probability of an incurred inversion on a *randomly drawn* pair because we exclude the documents with equivalent rank from the evaluation set (see Theorem 2). Figure 4 (b) shows the learning curves for both algorithms if we measure the number of misclassifications — treating the ranks as classes. As expected, the multi-class SVM's perform much better than our algorithm. It is important to note again, that minimizing the zero–one loss R_{0-1} does not automatically leads to a minimal number of inversions and thus and optimal ordering.

Figure 5 (a) shows the learning curves for SVR and our algorithm for ordinal regression. if we measure the number of incurred inversions. While the former performs quite well on the artificial dataset, in the real world dataset the SVR algorithm fails to find a ranking which minimizes the number of inversions. This can be explained

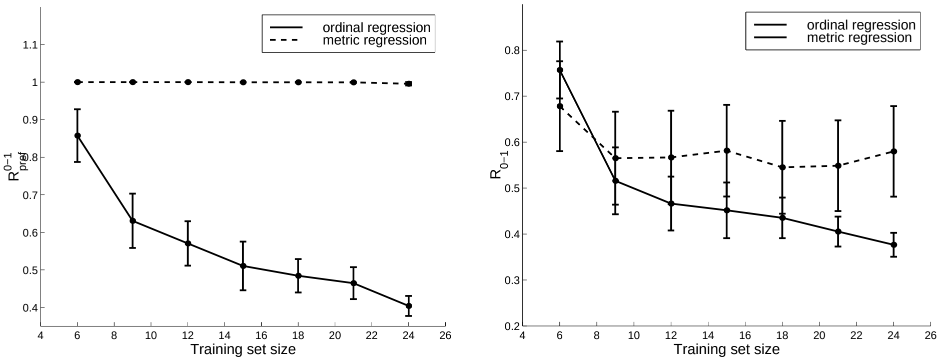


Figure 5: Learning curves for SVR (dashed lines) and the algorithm for ordinal regression (solid line) for the OHSUMED dataset query 1 if we measure (**a**) R^{0-1}_{pref} and (**b**) R_{0-1}. Error bars indicate the 95% confidence intervals.

by fact that for the real–world example the equidistance in the assumed utility may no longer hold — especially taking into account that the data space is very sparse for this type of problem. Similarly, Figure 5 (b) shows the learning curves for both algorithms if we measure the number of misclassifications. As expected from the right curves the SVR algorithm is worse even on that measure. Note that the SVR algorithm minimizes neither R_{pref} nor R_{0-1} which may explain its worse generalization behavior. Also note that we made no adaptation of the parameter ε — the size of the "tube". The reason is that in this particular task there would not be enough training examples available to set aside a reasonable portion of them for validation purposes.

7 Discussion and Conclusion

In this paper we introduced a new learning task to the ML Community — the task of ordinal regression — which is mainly characterized by the ordinal nature of the outcome space Y. Such a task is of particular interest in, e.g., Information Retrieval, economics, and social sciences. Although this problem is very novel to the Machine Learning community, there exist several approaches to this problem in the field of Statistics (see Section 2). All known approaches to this problem make distributional assumptions on an underlying continuous random variable. In contrast, we proposed a loss function which allows for application of distribution independent methods to solve ordinal regression problems. By exploiting the fact that the induced loss function class is a set of indicator functions we could give distribution independent bounds on our proposed risk. Moreover, we could show that to each ordinal regression problem there exists a unique preference learning problem on pairs of objects. This result built the link between ordinal regression and classification methods — this time on pairs of objects. For a particular representation of ranks by intervals on the real line, we could give tighter bounds on our proposed risk. This bound involves a quantity which is known as the margin — this time at the rank boundaries. Based on this result we presented an algorithm which is very similar to the well known Support Vector algorithm. Nevertheless, we would like to stress that this algorithm is only a particular instantiation of the presented theory. Retaining the proposed model of a rank we could also apply Gaussian Processes (MacKay 1997; Zhu et al. 1997; Gibbs and MacKay 1997) or other type of linear classifiers in order to derive the utility $U(\mathbf{x})$. Moreover, other modeling approaches for the ranks would result in completely different algorithms.

Noting that our presented loss involves pairs of objects we see that the problem of multi-class classification can also be reformulated on pairs of objects which leads to the problem of learning an *equivalence relation*. Usually, in order to extend a binary classification method to multiple classes, “one–against–one” or “one–against–all” techniques are devised (Hastie and Tibshirani 1997; Weston and Watkins 1998). Such techniques increase the size of the hypothesis space quadratically or linearly in the number of classes, respectively. Recent work (Phillips 1998) has shown that learning equivalence relation can increase the generalization behavior of binary–class methods when extended to multiple classes.

Further investigations include the following question[3]: Does the application of the GLM methods presented in Section 2 lead automatically to large margins (see Theorem 4)? The answer to such a question would finally close the gap between methods extensively used in the past to theories developed currently in the field of Machine Learning.

Acknowledgments

First of all we are indebted to our collaborator Peter Bollmann-Sdorra who first stimulated research on this topic. We also thank U. Kockelkorn, N. Cristianini, and G. Tutz for fruitful discussions, and Craig Saunders, Mark O. Stitson, Jason Weston, Leo Bottou, Bernhard Schölkopf and Alex Smola for providing their Support Vector implementation (Saunders et al. 1998). This project was funded by the Technical University of Berlin via the Forschungsinitiativprojekt FIP 13/41.

A The Relationship between Measurement Theory and Preference Learning

We give some results from measurement theory which are useful for modeling preference structures. Most of this material is taken from Fishburn 1985. After recalling some basic notation on relational systems we identify preference relations and use

[3] In 1996, P. Bartlett could prove the good generalization behavior of the backpropagation method applied to classification problems — 10 years after their extensive and successive application. This was made possible by the development of a theory of data–dependent capacity concepts known as the fat–shattering dimension (see Bartlett 1998).

the representation theorem of weakly ordered sets to motivate the utility function approach. Suppose we are given a set A together with a relation $R \subseteq A \times A$. We will call the tuple (A, R) a *relational system* and write $(a, b) \in R$ as aRb. Let us recall some general properties of relations.

Definition 1 (Properties of relations). *A binary relation R on a set A is*

$$\textit{reflexive} \qquad \textit{if} \quad aRa \qquad \forall a \in A \tag{81}$$

$$\textit{irreflexive} \qquad \textit{if} \quad \neg aRa \qquad \forall a \in A \tag{82}$$

$$\textit{symmetric} \qquad \textit{if} \quad aRb \Rightarrow bRa \qquad \forall a, b \in A \tag{83}$$

$$\textit{asymmetric} \qquad \textit{if} \quad aRb \Rightarrow \neg bRa \qquad \forall a, b \in A \tag{84}$$

$$\textit{transitive} \qquad \textit{if} \quad aRb \wedge bRc \Rightarrow aRc \qquad \forall a, b, c \in A \tag{85}$$

$$\textit{negatively transitive} \qquad \textit{if} \quad aRb \Rightarrow aRc \vee cRb \qquad \forall a, b, c \in A \tag{86}$$

$$\textit{complete} \qquad \textit{if} \quad aRb \vee bRa \qquad \forall a, b \in Y \tag{87}$$

$$\textit{weakly connected} \qquad \textit{if} \quad a \neq b \Rightarrow aRb \vee bRa \qquad \forall a, b \in A \tag{88}$$

Definition 2 (Ordered set). *A binary relation R on a set A is*

$$\textit{a}\ \text{weak order} \quad \Leftrightarrow \quad R \textit{ on } A \textit{ is asymmetric and negatively transitive} \tag{89}$$

$$\textit{a}\ \text{strict order} \quad \Leftrightarrow \quad R \textit{ on } A \textit{ is a weakly connected weak order} \tag{90}$$

$$\textit{an}\ \text{equivalence} \quad \Leftrightarrow \quad R \textit{ on } A \textit{ is reflexive, symmetric, and transitive}\,. \tag{91}$$

For weakly ordered sets A a relation $\sim_R$ defined by

$$x \sim_R y \quad \Leftrightarrow \quad \neg xRy \wedge \neg yRx \quad \forall x, y \in A \tag{92}$$

is an equivalence relation. Now let us recall the so called representation theorem for weakly ordered sets.

Theorem 8 (Representation of weakly ordered sets (Fishburn 1985)). *If $\succ$ on the set A is a weak order such that $A/\sim$ is countable, then there exists $U : A \mapsto \mathcal{R}$ such that for all $x, y \in A$*

$$x \succ y \quad \Leftrightarrow \quad U(x) > U(y)\,. \tag{93}$$

From the definition of $\succ_{\mathrm{Y}}$ we know that $\succ_{\mathrm{Y}}$ on $Y = \{r_1, \ldots, r_q\}$ is an asymmetric and transitive relation. Thus it can easily be shown that $(Y, \succ_{\mathrm{Y}})$ is an empirical system with a weak order. Using the property that for each pair of ranks r_i and r_j where $i \neq j$, either $r_i \succ_{\mathrm{Y}} r_j$ or $r_j \succ_{\mathrm{Y}} r_i$, $(Y, \succ_{\mathrm{Y}})$ is also a strict order. Given a mapping $h : X \mapsto Y$, we can look at the relational system $(X, \succ_{\mathrm{X}})$ where $\succ_{\mathrm{X}}$ is defined by Equation (7). It is obvious that $(X, \succ_{\mathrm{X}})$ for a given h is at least a weak order.

References

Aizerman, M., E. Braverman, and L. Rozonoer (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control 25*, 821–837.

Alon, N., S. Ben-David, N. Cesa-Bianchi, and D. Haussler (1997). Scale sensitive dimensions, uniform convergence and learnability. *Journal of the ACM 44*(4), 615–631.

Anderson, J. (1984). Regression and ordered categorical variables (with discussion). *Journal of the Royal Statistical Society – Series B 46*, 1–30.

Anderson, J. and P. Philips (1981). Regression, discriminantion and measurement models for ordered categorial variables. *Applied Statistics 30*, 22–31.

Bartlett, P. L. (1998). The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory 44*(2), 525–536.

Caruana, R., S. Baluja, and T. Mitchell (1995). Using the future to sort out the present: Rankprob and multitask learning for medical risk evaluation. In *Advances in Neural Information Processing Systems*, Volume 8, Denver, pp. 959–965.

Cortes, C. (1995). *Prediction of Generalization Ability in Learning Machines*. Ph. D. thesis, University of Rochester, Rochester, USA.

Cortes, C. and V. Vapnik (1995). Support Vector Networks. *Machine Learning 20*, 273–297.

Cristianini, N., J. Shawe-Taylor, and P. Sykacek (1998). Bayesian classifiers are large margin hyperplanes in a hilbert space. Technical report, Royal Holloway, University of London. NC2–TR–1998–008.

de Moraes, A. and I. R. Dunsmore (1995). Predictive comparisons in ordinal models. *Communications in Statistics – Theory and Methods 24*(8), 2145–2164.

Fahrmeir, L. and G. Tutz (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer–Verlag.

Fishburn, P. C. (1985). *Interval Orders and Interval Graphs*. Jon Wiley and Sons.

Gibbs, M. and D. J. MacKay (1997). Efficient implementation of gaussian processes. available at http://wol.ra.phy.cam.ac.uk/mng10/GP/.

Girosi, F. (1997). An equivalence between sparse approximation and Support Vector Machines. Technical report, Massachusetts Institute of Technology, Artificial Intelligence Laboratory. AI Memo No. 147.

Hadley, G. (1964). *Nonlinear and Dynamic Programming*. London: Addison–Wesley.

Hastie, T. and R. Tibshirani (1997). Classification by pairwise coupling. In *Advances in Neural Information Processing Systems*, San Mateo, CA. Morgan Kaufmann.

Herbrich, R., T. Graepel, P. Bollmann-Sdorra, and K. Obermayer (1998). Learning a preference relation for information retrieval. In *Proceedings of the AAAI Workshop Text Categorization and Machine Learning*, Madison, USA.

Herbrich, R., T. Graepel, and K. Obermayer (1999). Regression models for ordinal data: A machine learning approach. submitted to *Machine Learning*.

Herbrich, R., M. Keilbach, T. Graepel, P. Bollmann-Sdorra, and K. Obermayer (1998). *Computational Techniques for Modelling Learning in Economics*, Chapter 3, Neural Networks in Economics: Background, Applications and New Developments, pp. 42. Kluwer Academics. accepted for publication.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*, 13–30.

Huber, P. J. (1981). *Robust Statistics*. John Wiley and Sons.

Joachims, T. (1997). Text categorization with Support Vector Machines: Learning with many relevant features. Technical report, University Dortmund, Department of Artifical Intelligence. LS–8 Report 23.

Joachims, T. (1998). *Advances in Kernel Methods — Support Vector Learning*, Chapter 11, Making Large–Scale SVM Learning Practical. MIT Press.

Keener, R. W. and D. M. Waldman (1985). Maximum liklihood regression of rank-censored data. *Journal of the American Statistical Association 80*, 385–392.

Kendall, M. G. (1970). *Rank Correlation Methods*. Griffin–London.

MacKay, D. J. (1997). Gaussian Processes. Tutorial for the NIPS–97.

Mangasarian, O. L. (1969). *Nonlinear Programming*. New York: McGraw–Hill.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society – Series B 42*, 109–142.

McCullagh, P. and J. A. Nelder (1983). *Generalized Linear Models*. London: Chapman & Hall.

Phillips, P. J. (1998). Support Vector Machines applied to face recognition. In *Proceedings of the Neural Information Processing Conference*, Denver, USA. in press.

Pollard, D. (1984). *Convergence of Stochastic Processess*. New York: Springer–Verlag.

Pontil, M. and A. Verri (1998). Properties of Support Vector Machines. *Neural Computation 10*, 955–974.

Salton, G. (1968). *Automatic Information Organization and Retrieval*. New York: McGraw–Hill.

Saunders, C., M. O. Stitson, J. Weston, L. Bottou, B. Schölkopf, and A. Smola (1998). Support Vector Machine reference manual. Technical report, Royal Holloway, University of London. CSD–TR–98–03.

Schölkopf, B. (1997). *Support Vector Learning*. Ph. D. thesis, Technische Universitä Berlin, Berlin, Germany.

Shawe-Taylor, J., P. L. Bartlett, R. C. Williamson, and M. Anthony (1996). Structural risk minimization over data–dependent hierarchies. Technical report, Royal Holloway, University of London. NC–TR–1996–053.

Smola, A. (1996). Regression estimation with Support Vector Learning Machines. Master's thesis, Technical University Munich.

Sobel, M. (1990). Complete ranking procedures with appropriate loss functions. *Communications in Statistics – Theory and Methods 19*(12), 4525–4544.

Suppes, P., D. H. Krantz, R. D. Luce, and A. Tversky (1989). *Foundations of Measurement Vol. II*. San Diego: Academic Press Inc.

Tangian, A. and J. Gruber (1995). Constructing quadratic and polynomial objective functions. In *Proceedings of the 3rd International Conference on Econometric Decision Models*, Schwerte, Germany, pp. 166–194. Springer.

Vanderbei, R. J. (1997). *Linear Programming: Foundations and Extensions*. Hingham: Kluwer Academics.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. New York: Springer–Verlag.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer–Verlag.

Vapnik, V. (1998). *Statistical Learning Theory*. New York: John Wiley and Sons.

Vapnik, V. and A. Chervonenkis (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Application 16*(2), 264–281.

Wahba, G. (1997). Support Vector Machines, Reproducing Kernel Hilbert Spaces and the randomized GACV. Technical report, Department of Statistics, University of Wisconsin, Madison. TR–NO–984.

Weston, J., A. Gammerman, M. Stitson, V. Vapnik, V. Vovk, and C. Watkins (1997). Density estimation using Support Vector Machines. Technical report, Royal Holloway, University of London. CSD–TR–97–23.

Weston, J. and C. Watkins (1998). Multi-class Support Vector Machines. Technical report, Royal Holloway, University of London. CSD–TR–98–04.

Whaba, G. (1990). *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.

Wong, S. K. M., Y. Y. Yao, and P. Bollmann (1988). Linear structure in information retrieval. In *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 219–232.

Zhu, H., C. K. Williams, R. Rohwer, and M. Morciniec (1997). Gaussian regression and optimal finite dimensional linear models. Technical report, Neural Computing Research Group, Aston University. NCRG/97/011.