Support Vector Learning for Ordinal Regression

Ralf Herbrich, Thore Graepel, Klaus Obermayer
Technical University of Berlin
Department of Computer Science
Franklinstr. 28/29
10587 Berlin
ralfh|graepel2|oby@cs.tu-berlin.de

Abstract

We investigate the problem of predicting variables of ordinal scale. This task is referred to as *ordinal regression* and is complementary to the standard machine learning tasks of classification and metric regression. In contrast to statistical models we present a distribution independent formulation of the problem together with uniform bounds of the risk functional. The approach presented is based on a mapping from objects to scalar utility values. Similar to Support Vector methods we derive a new learning algorithm for the task of ordinal regression based on large margin rank boundaries. We give experimental results for an information retrieval task: learning the order of documents w.r.t. an initial query. Experimental results indicate that the presented 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 Introduction

Problems of ordinal regression arise in many fields, e.g., in information retrieval (Herbrich et al. 1998), in econometric models (Tangian and Gruber 1995), and in classical statistics (McCullagh 1980; Anderson 1984). They can be related to the standard machine learning paradigm as follows:

Given an i.i.d. sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell} \sim P_{XY}^{\ell}$ 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 $l : Y \times Y \mapsto \mathbb{R}$ — the risk functional $R(h^{\ell})$ is minimized. Typically, in machine learning 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. 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*. 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 (see Smola (1998) for a detailed discussion of loss functions for regression).

In ordinal regression, we consider a problem which shares properties of both classification (i) and metric regression (ii). Like in (i) Y is a finite set and like in (ii) there exists an ordering among the elements of Y. A variable of the above type exhibits an *ordinal scale* and can be thought of as the result of coarse measurement of a continuous variable (Anderson 1984). The ordinal scale leads to problems in defining an appropriate loss function for our task (see McCullagh 1980). In Section 2 we present a distribution independent model for ordinal regression, which is based on a loss function that acts on pairs of ranks. We give explicit uniform convergence bounds for the pro-

posed loss function and show the relation between ordinal regression and preference learning.

As an application of the theory we derive an algorithm for ordinal regression in Section 3 by modeling ranks as intervals on the real line. 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}. The resulting algorithm is similar to Support Vector Machines (SVM) (Vapnik 1998) and enforces large margin rank boundaries. It is easily extended to non–linear utility functions using the "kernel trick" (Smola 1998).

Finally, in Section 4 we present learning curves of our approach in a controlled experiment and in a real–world experiment on data from information retrieval.

2 A Risk Formulation for Ordinal Regression

Consider an input space $X \subset \mathbb{R}^n$ with objects being represented by feature vectors $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$, 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 $r_q \succ_{\text{Y}} r_{q-1} \succ_{\text{Y}} \cdots \succ_{\text{Y}} r_1$. The symbol $\succ_{\text{Y}}$ 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. Let us consider a model space $\mathcal{H} = \{h(\cdot) : X \mapsto Y\}$ of mappings from objects to ranks. Moreover, each such function h induces an ordering $\succ_{\text{X}}$ on the elements of the input space by the following rule

$$\mathbf{x}_i \succ_{\text{X}} \mathbf{x}_j \quad \Leftrightarrow \quad h(\mathbf{x}_i) \succ_{\text{Y}} h(\mathbf{x}_j)\,. \tag{1}$$

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; McCullagh 1980). Given a pair $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$ of objects we have to distinguish between two different outcomes: $y_1 \succ_{\text{Y}} y_2$ and $y_2 \succ_{\text{Y}} y_1$. Thus, the probability of incurred inversion is given by the following risk functional

$$R_{\text{pref}}(h) \quad = \quad \mathbf{E}\left[l_{\text{pref}}(h(\mathbf{x}_1), h(\mathbf{x}_2), y_1, y_2)\right]\,, \tag{2}$$

with

$$l_{\text{pref}}(\hat{y_1}, \hat{y_2}, y_1, y_2) \quad = \quad \begin{cases} 1 & \text{if } y_1 \succ_{\text{Y}} y_2 \\ & \text{and } \neg(\hat{y_1} \succ_{\text{Y}} \hat{y_2}) \\ 1 & \text{if } y_2 \succ_{\text{Y}} y_1 \\ & \text{and } \neg(\hat{y_2} \succ_{\text{Y}} \hat{y_1}) \\ 0 & \text{else} \end{cases} \tag{3}$$

The ERM principle recommends to take that mapping h^ℓ which minimizes the empirical risk $R_{\text{emp}}(h; S)$,

$$R_{\text{emp}}(h; S) \quad = \quad \frac{1}{\ell^2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} l^s_{\text{pref}}(h(\mathbf{x}_i), h(\mathbf{x}_j), y_i, y_j)\,,$$

which is effectively based on a new training set whose elements are pairs of objects. Using the shorthand notation $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ to denote the first and second object of a pair, the 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 where either $y_i^{(1)} \succ_{\text{Y}} y_i^{(2)}$ or $y_i^{(2)} \succ_{\text{Y}} y_i^{(1)}$, i.e.

$$S' \quad = \quad \left\{\left((\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}), \Omega\left(y_i^{(1)}, y_i^{(2)}\right)\right)\right\}_{i=1}^{t} \tag{4}$$

$$\Omega(y_1, y_2) \quad = \quad \text{sign}(y_1 \ominus y_2)\,, \tag{5}$$

where $\ominus$ is the rank difference and t is the cardinality of S'.

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

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

$$\frac{1}{t} \sum_{i=1}^{t} l_{0-1}(\Omega(h(\mathbf{x}_i^{(1)}), h(\mathbf{x}_i^{(2)})), \Omega(y_i^{(1)}, y_i^{(2)}))$$

Taking into account that each function $h \in \mathcal{H}$ defines a function $p : X \times X \mapsto \{-1, 0, +1\}$ by

$$p(\mathbf{x}_1, \mathbf{x}_2) \quad = \quad \Omega(h(\mathbf{x}_1), h(\mathbf{x}_2))\,, \tag{6}$$

Theorem 1 states that the empirical risk of a certain mapping h on a sample S is equivalent to the empirical risk based on 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 ordinal regression can be reduced to a classification problem on pairs of objects. 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. 1999). Note also that the requirements of transitivity and asymmetry effectively reduce the space of admissible classification functions p acting on pairs of objects.

The following bound on $R_{\text{pref}}(h^\ell)$ gives a justification for the large-margin algorithm to be presented in Section 3. A proof based on a result of (Shawe-Taylor et al. 1998) can be found in (Herbrich et al. 1999).

Theorem 2. *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 $\mathbb{R}$ such that for each function $h \in \mathcal{H}$ there exists a function $U \in \mathcal{F}$ (and vice versa) with*

$$h(\mathbf{x}) = r_i \quad \Leftrightarrow \quad U(\mathbf{x}) \in [\theta(r_{i-1}), \theta(r_i)]\,. \tag{7}$$

Let P_{XY} be a probability measure on XY, let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$ be an i.i.d. sample from P_{XY}, S' be derived from S by Equation (4) *and the fat–shattering dimension of the set of functions $\mathcal{F}$ be bounded above by the function afat : $\mathbb{R} \mapsto \mathbb{N}$. Then for each function h^ℓ with $R^{0-1}_{emp}(h^\ell; S') = 0$ and $\gamma = \min_{S'} |U^\ell(\mathbf{x}^{(1)}) - U^\ell(\mathbf{x}^{(2)})|$ with probability $1 - \delta$*

$$R_{pref}(h^\ell) \leq \frac{2}{t}\left(k \log_2\left(\frac{8et}{k}\right)\log_2(32t) + \log_2\left(\frac{8t}{\delta}\right)\right),$$

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

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

3 Support Vector Machines for Ordinal Regression

In this section we apply the theory from Section 2 to derive an algorithm for ordinal regression. Let us consider a linear function $U : X \mapsto \mathbb{R}$

$$U(\mathbf{x}) \quad = \quad \mathbf{w}^T\mathbf{x} \tag{8}$$

which is related to a mapping h from objects to ranks by (7). We assume that $\theta(r_0) = -\infty$ and $\theta(r_q) = +\infty$. Such a function is commonly called a *utility function*. We know that $U(\mathbf{x})$ incurs no error for the i-th example in the training set S' (see Equation (4)) iff

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

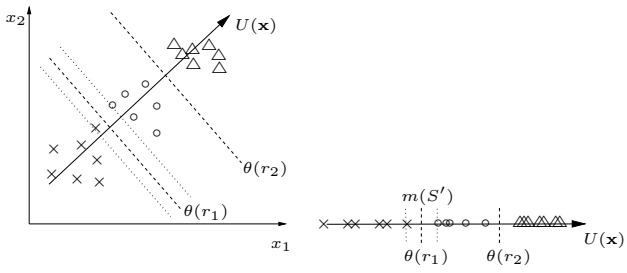
where $z_i = \Omega(y_i^{(1)}, y_i^{(2)})$ was used. Note, that the preference relation is expressed in terms of the difference $\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)}$ of feature vectors, which can be thought of as the combined feature vector of the pair of objects. By assuming a finite margin between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

$$z_i[\mathbf{w}^T(\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)})] \quad \geq \quad 1 - \xi_i\,, \quad i = 1, \ldots, t\,, \tag{9}$$

where the non-negative ξ_i measure the degree of violation of the i-th constraint. The weight vector $\mathbf{w}^\ell$ which maximizes the margin — this time at the rank boundaries $\theta(r_i)$ (see Equation (7) 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 (9). This approach is closely related to the idea of canonical hyperplanes used in Support Vector classification (Vapnik 1998). A theoretical justification for the applicability of SRM is given by Theorem 2. Introducing Lagrangian multipliers and performing unconstrained optimization with respect to $\mathbf{w}$ leads to the dual problem of finding $\boldsymbol{\alpha}^\ell$ such that

$$\boldsymbol{\alpha}^\ell \quad = \quad \max_{\substack{\mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{1} \\ \boldsymbol{\alpha}^T\mathbf{z} = 0}} \left[\mathbf{1}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Z}^T\mathbf{Q}\mathbf{Z}\boldsymbol{\alpha}\right], \tag{10}$$

with $\mathbf{z} = (z_1, \ldots, z_t)^T$, $\mathbf{Z} = \text{diag}(\mathbf{z})$, and $\mathbf{Q}_{ij} = (\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)})^T(\mathbf{x}_j^{(1)} - \mathbf{x}_j^{(2)})$. This is a standard QP–problem and



(a) (b)

Figure 1: (**a**) Mapping of objects from ranks r_1 ($\times$), r_2 ($\circ$), and r_3 ($\triangle$) to the axis $U(\mathbf{x})$. $\theta(r_1)$ and $\theta(r_2)$ define two coupled hyperplanes. (**b**) The margin $m(S') = \min_{S'} |U(\mathbf{x}_i^{(1)}) - U(\mathbf{x}_i^{(2)})|$ of the hyperplanes is defined at the rank boundaries $\theta(r_i)$.

can efficiently be solved using techniques from mathematical programming (Mangasarian 1969). Given the optimal vector $\boldsymbol{\alpha}^\ell$ as solution to (10), 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):

$$\mathbf{w}^\ell \quad = \quad \sum_{i=1}^{t} \alpha_i^\ell z_i \left(\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)} \right) . \qquad (11)$$

To estimate the rank boundaries we note that due to Equations (9) the difference in utility is greater or equal to one for all training examples with $\xi_i = 0$ (or equivalently $\alpha_i < C$). Thus if $\Theta(k) \subset S'$ is the fraction of objects from the training set with $\xi_i = 0$ and rank difference of exactly one starting from rank r_k, then the estimation of $\theta(r_k)$ is given by

$$\theta(r_k) \quad = \quad \frac{U(\mathbf{x}_1; \mathbf{w}^\ell) + U(\mathbf{x}_2; \mathbf{w}^\ell)}{2} , \qquad (12)$$

where

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

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}. After the estimation of the rank boundaries $\theta(r_k)$ a new object is classified according to Equation (7).

The extension to non–linear utility functions follows the same reasoning as with non–linear SVM (Vapnik 1998). Note, that $\mathbf{Q}_{ij}$ can be expanded into four inner products between objects $\mathbf{x}$. Thus, given a function $K : X \times X \mapsto \mathbb{R}$ which is symmetric and positive definite, each calculation of an inner product in X is replaced by $K(\cdot, \cdot)$. This corresponds to a non–linear utility $U(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x})$ with $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$ and using (11) for the resulting $\mathbf{w}^\ell$. Here $\Phi : X \to \mathcal{X}$ is a mapping from input space X to a reproducing kernel Hilbert space $\mathcal{X}$ often refered to as "feature space". Details can be found in (Smola 1998).

4 Experimental Results

4.1 Learning Curves for Ordinal Regression

In this experiment we compare the generalization behavior of the presented algorithm with the multi-class SVM and Support Vector regression (SVR). Those algorithms were chosen for comparison due to their similar regularizer $\|\mathbf{w}\|^2$ and hypothesis space $\mathcal{H}$. We generated 1000 observations $\mathbf{x} = (x_1, x_2)^T$ in the unit square $[0, 1] \times [0, 1] \subset \mathbb{R}^2$ according to a uniform distribution. We assigned to each observation $\mathbf{x}$ a value y according to

$$y = i \quad \Leftrightarrow \quad U(\mathbf{x}) + \epsilon \in [\theta(r_{i-1}), \theta(r_i)] , \qquad (13)$$

$$U(\mathbf{x}) \quad = \quad 10((x_1 - 0.5) \cdot (x_2 - 0.5)) , \qquad (14)$$

where $\epsilon \sim N(0, 0.125)$, and $\boldsymbol{\theta} = (-\infty, -1, -0.1, 0.25, 1, +\infty)^T$ is the vector of predefined thresholds.

We randomly drew 100 training samples of sizes ranging from 5 to 45 making sure that at least one representative of each rank was in the 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, 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 $C = 1000000$. For Support Vector regression we used $\varepsilon = 0.5$ for the ε–insensitive loss function (see (Vapnik 1998) for the definition of this loss function) and thresholds $\boldsymbol{\theta} = (0.5, 1.5, 2.5, 3.5, 4.5)^T$.

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

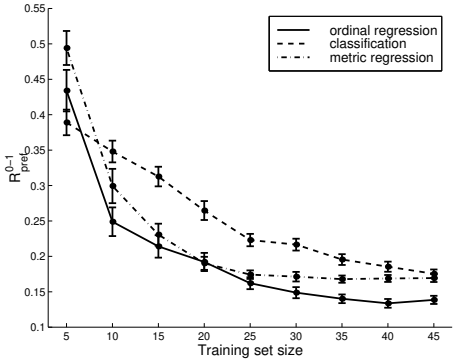


Figure 2: Learning curves for multi-class SVM (dashed lines), SV regression (dashed–dotted line) and the algorithm for ordinal regression (solid line) if we measure R^{0-1}_{pref}. The error bars indicate the 95% confidence intervals of the estimated risk R^{0-1}_{pref}.

training set size (see Figure 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. 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. The learning curves for SVR and the proposed ordinal regression algorithm are very close, because the predefined thresholds $\theta(r_k)$ are defined at a distance of 0.5 — the size of the ε–tube chosen beforehand.

4.2 An Application to Information Retrieval

In this experiment we made the following assumption: After an initial (textual) query to an IR system, the system returns a bundle of documents, of which the user ranks a small fraction. The task for the learning algorithm is to assign ranks to the remaining unranked documents. After using $\ell = 6$ up to $\ell = 24$ documents and their respective ranking for training we measure R^{0-1}_{emp} on the remaining documents. For this experiment the same parameter values were used as in the last experiment. The simulations were carried out on the OHSUMED dataset, which consists of $348\,566$ documents and 106 queries with their respective ranked results ("document is relevant", "document is partially relevant", "irrelevant document"). 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 consisted of 107 documents taken from the whole database. The documents were represented as "bag–of–words" (Salton 1968), with the resulting document vectors normalized to unit length.

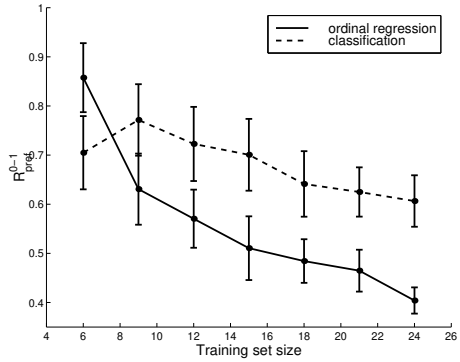


Figure 3: Learning curves for multi-class SVM (dashed lines) and the algorithm for ordinal regression (solid line) on the OHSUMED dataset query 1 as measured by R^{0-1}_{pref}. Error bars indicate 95% confidence intervals.

As can be seen from the results (see Figure 3), the proposed algorithm shows very good generalization behavior compared to the multi-class SVM, which treats each rank as a separate class. Note that the plotted R^{0-1}_{pref} is the proportion of misclassified pairs if we restrict ourselves to pairs with different ranks. This quantity is much larger than the estimated probability of an incurred inversion on a *randomly drawn* pair because documents of equivalent ranks were excluded from the evaluation set.

5 Discussion and Conclusion

We introduced a new learning task to the ML community: ordinal regression. The task is complementary to classification and metric regression due to its discrete and ordered outcome space Y. We showed that every ordinal

regression problem corresponds to a unique preference learning problem on pairs of objects. This result builds the link between ordinal regression and classification methods on pairs of objects and allows for a theoretical treatment in the framework of classification. We would like to stress that the presented algorithm is only a particular instantiation of the theory. Retaining the proposed model of a rank we could also apply Gaussian Processes (MacKay 1997; Zhu et al. 1997) or other types of classifiers baed on thresholded real valued functions.

Noting that our presented loss involves pairs of objects it is interesting to note that the problem of multi-class classification can also be reformulated on pairs of objects. This leads to the problem of learning an *equivalence relation* between the objects. Recent work (Phillips 1999) shows that learning an equivalence relation can increase the generalization behavior of binary–class methods when extended to multiple classes.

Acknowledgments

We are indebted to our collaborator Peter Bollmann-Sdorra who first stimulated research on this topic. We would also like to thank Matthias Burger, Vladimir Vapnik, Nello Cristianini, Ulrich Kockelkorn, and Gerhard Tutz for helpful comments. This project was funded by the Technical University of Berlin via the Forschungsinitiativprojekt FIP 13/41.

References

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

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. Technical report, TU Berlin. TR-99/03.

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.

Phillips, P. J. (1999). Support Vector Machines applied to face recognition. In *Proceedings of the Neural Information Processing Conference*, Denver, USA, pp. 803–809.

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

Shawe-Taylor, J., P. L. Bartlett, R. C. Williamson, and M. Anthony (1998). Structural risk minimization over data–dependent hierarchies. *IEEE Transactions on Information Theory 44*(5), 1926–1940.

Smola, A. J. (1998). *Learning with Kernels*. Ph. D. thesis, Technische Universität Berlin.

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

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.

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

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

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.