ORDINAL LEAST SQUARES SUPPORT VECTOR MACHINES - A DISCRIMINANT ANALYSIS APPROACH

Kristiaan Pelckmans, Peter Karsmakers, Johan A.K. Suykens, Bart De Moor

K.U. Leuven, ESAT/SCD-SISTA
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

ABSTRACT

This paper explores the extension of the classical ideas behind linear discriminant analysis to the problem of ordinal regression. It is shown how this reasoning fits in a framework of Least Squares Support Vector Machines (LS-SVMs) and kernel machines, hereby allowing for a nonlinear extension. The resulting method is conceived as a practical alternative to recently proposed, computationally demanding formulations based on maximal margin.

1. INTRODUCTION

Discriminant analysis is involved with the task of learning optimal binary classification rules from a given set of class distributions (parametric approach) or given finite set of observations (empirical approach). Linear Discriminant Analysis (LDA) and its extensions constitute the basic methodology of the statistician [12] and the machine learner [8] in the first setting, while Support Vector Machines [2, 18] gain increasing popularity in the empirical setting due to its intuitive appeal, the proper extension to the nonlinear case using kernels, its theoretical foundations and convex formulation.

While the binary case can be extended straightforwardly towards the multi-class case using coding schemes (see e.g. [5, 10]), the special case of multi-class classification where the different class-labels possess an intrinsic order is recognized as a learning task on its own, denominated as the task of ordinal regression [11]. The crucial issue discriminating this task from classical function approximation tasks is that in case of ordinal data, the response labels possesses no (given) metric (norm). (e.g. the difference between the 1st and the 2nd class is not necessarily the same as the difference between the 10th and the 11th class). The statistical approach as described in [11] proceeds using a generalized linear model where the link function provides the parametric cumulative distribution of the class distributions. Recently, maximal margin methods where described for the similar task of ranking [9] or [13, 4] with successful applications to information retrieval tasks and collaborative filtering amongst others. The Gaussian process framework and Bayesian inference techniques where studied in this context in [3].

Least Squares Support Vector Machines (LS-SVMs) [16, 15] constitute a method for building a binary classification rule from observed labeled samples. The naming convention is motivated by the similarity of the derivation with $L2$-SVMs, where additionaly inequalities are substituted by equalities. Other cornerstones are the use of the kernel trick and the fact that the final solution follows from a well-conditioned linear system. The formulation of ordinal LS-SVMs (or oLS-SVMs) retains those advantages, and additionally shows that the derivation underlying LS-SVMs will not necessarily lead to a simple *regression on the labels* in all cases. On a theoretical account, SVMs however do differentiate from LS-SVMs in the sense of the distribution-free learning theory supporting the former. The dependence of the least squares approach on the Gaussian assumption, however, can be relaxed theoretically [14] and practically (using robust versions as described in [15]). The view essentially differs from the approach taken by the Gaussian process approach [3] as it does not concern modelling of the data-generation process.

This paper is organized as follows. Section 2 discusses the principles behind the discriminant analysis for ordinal data. Section 3 implements those ideas in the form of ordinal LS-SVMs, and topics as model selection are discussed briefly.

2. AN ORDINAL CLASSIFIER

2.1. ORDINAL LINEAR DISCRIMINANT ANALYSIS

Consider a set of samples $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \{1, \ldots, L\}$ sampled i.i.d. from a joint distribution $F_X \times F_Y$ with $L \in \mathbb{N}_0$. The labels y_i reflect an ordering of the samples - but not a metric - such that if $y_i > y_j$ and $y_j > y_l$, then $y_i > y_l$ for all $i, j, l = 1, \ldots, N$. Let $\mathcal{S}_l$ denote the indices of the data-points with label l such that $y_i = l$ for all $i \in \mathcal{S}_l$. Assume that the samples of each class $\{x_i\}_{i \in \mathcal{S}_l}$ are distributed as a Gaussian distribution with mean $m_l \in \mathbb{R}^d$ and variance $\Sigma_l = \Sigma \in \mathbb{R}^{d \times d}$, formally

$$\begin{aligned} &x_i \sim N(x_i; m_l, \Sigma_l) \\ &= \frac{1}{(2\pi)^{p/2}|\Sigma_l|^{1/2}} \exp\left(-\frac{1}{2}(x_i - m_l)^T \Sigma_l^{-1}(x_i - m_l)\right). \end{aligned} \tag{1}$$

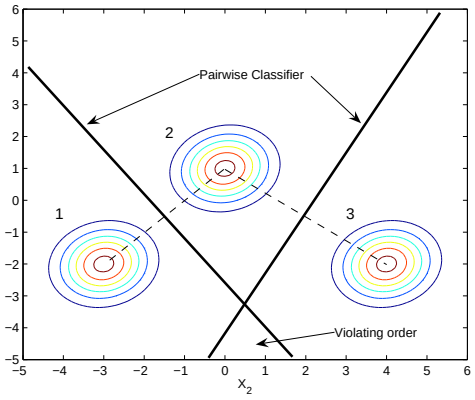

Fig. 1. *A multi-class approach via the combination of atomic pairwise classification rules (solid lines) of the classes (contour lines) in general does not lead to estimates which preserve in any case the same order (the decision lines do often cross each other).*

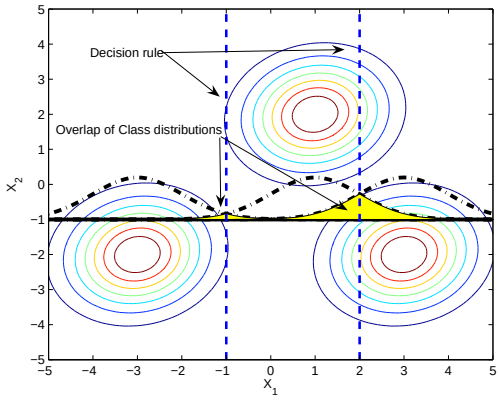

Fig. 2. *A schematical view of the principles behind oLDA. Assuming that the class distributions are approximatively Gaussian, the marginal class distributions (dashed-dotted line) obtained by projecting the class distributions (contour lines) on the utility function (solid line) can be computed explicitly. The risk will then coincide with the overlap between the different distributions (shaded areas). The ordinal rule with minimal risk corresponds with a set of threshold rules on the utility function (dashed lines).*

Let the vector $M = (m_1^T, \ldots, m_L^T)^T \in \mathbb{R}^{L\times d}$ contain the mean vectors. Instead of studying the optimal classifiers between each pair of distributions (LDA, [8]), we are concerned with the problem of deriving the best predictor reflecting the ordering of the labels. Therefor, the linear utility function underlying statistical ordinal regression [11] is employed: the data are projected as follows

$$U(x; w) = w^T x \triangleq u, \tag{2}$$

since the resulting sample $\{u_i\}_{i=1}^N$ is univariate it possesses a well-defined ordinal scale ($u_i < u_j$ or $u_i = u_j$ or $u_i > u_j$ for all $i, j = 1, \ldots, N$). The use of a global linear utility function results in parallel decision planes, which is a necessary condition for obeying the ordering in all cases (see Fig. 1).

Now we investigate the marginal distribution of $\{u_i\}_{i\in\mathcal{S}_l}$ by projecting the corresponding distribution on w:

$$\begin{aligned} u_i &\sim N(u; w^T m_l, w^T \Sigma w) \\ &= \frac{1}{\sqrt{2\pi(w^T\Sigma_l w)}} \exp\left(-\frac{(u - w^T m_l)^2}{2(w^T\Sigma_l w)}\right). \end{aligned} \tag{3}$$

Let $\mathcal{I}$ be the indicator function such that $\mathcal{I}(z) = 1$ if z is true and zero otherwise. Given those univariate distributions, the rules deciding whether a point belongs to the lth class become

$$c_l(x; w, M) = \mathcal{I}\left(v_{l-1,l} < U(x; w) \leq v_{l,l+1}\right), \tag{4}$$

where $v_{l,l+1}$ is defined as the mean $\frac{1}{2}(w^T m_l + w^T m_{l+1})$ for all $l = 1, \ldots, L-1$, and where we define $v_{0,1} = -\infty$ and $v_{L,L+1} = +\infty$. This rule is consistent with the given ordering whenever $v_{1,2} \leq \cdots \leq v_{L-1,L}$. Now we turn to the task of estimating the parameter vector w such that the above rules will perform optimally. Herefor, we have to quantify the probability mass which violates the rules, i.e. the risk $\mathcal{R}(w; M, \Sigma)$ defined as (see e.g. [12])

$$\begin{aligned} &\frac{1}{L+1}\sum_{l=1}^{L-1}\int_{u\geq v_{l,l+1}} N(u; w^T m_l, w^T\Sigma w)\, du \\ &\qquad + \int_{u\leq v_{l,l+1}} N(u; w^T m_{l+1}, w^T\Sigma w)\, du, \end{aligned} \tag{5}$$

reflecting only the overlap of two consecutive distributions. Using the definition of the cumulative distribution function of the univariate normal distribution in terms of the monotonous function $\mathrm{erf} : \mathbb{R} \to]0, 1[$, one gets for each $l = 1 \ldots, L-1$

$$\begin{aligned} &\int_{u>v_{l,l+1}} N(u, w^T m_l, w^T\Sigma w)\, du \\ &= \int_{u<w^T m_l - (v_{l,l+1} - w^T m_l)} N(u, w^T m_l, w^T\Sigma w)\, du \\ &= \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{v_{l,l+1} - w^T m_l}{\sqrt{2w^T\Sigma w}}\right)\right), \end{aligned} \tag{6}$$

due to symmetry of the normal distribution around the mean $w^T m_l$. Similarly $\int_{u<v_{l,l+1}} N(u, w^T m_{l+1}, w^T\Sigma w)du$ $= \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{v_{l,l+1} - w^T m_{l+1}}{\sqrt{2w^T\Sigma w}}\right)\right)$. This expression can be simplified by using the fact that erf is symmetric, and by combining both equalities and substitution of $v_{l,l+1} = \frac{1}{2}(w^T m_l + w^T m_{l+1})$ such that:

$$\mathcal{R}(w; M, \Sigma) \quad \propto \quad \sum_{l=1}^{L-1} -\mathrm{erf}\left(\frac{w^T(m_{l+1} - m_l)}{\sqrt{2\, w^T\Sigma w}}\right), \tag{7}$$

assuming that the ordering inequalities $w^T m_1 \leq \cdots \leq w^T m_L$ hold. It is interesting to note at this point that the above risk is also useful when the Gaussian assumption does not hold. As long as the second moment of the class distribution $w^T \Sigma w$ exists, Chebychev's inequality can be used to bound the probability mass beyond a threshold given the variance, see e.g. [6]. Furthermore, note that the specific form of the true distribution translates into a corresponding function similar to erf which is monotonously increasing by construction.

Remark further that this risk function is invariant w.r.t. the norm of the parameter vector $\|w\|$. An optimal predictor is obtained by minimizing the theoretical risk (or the measure of overlap between the distributions of the different classes) w.r.t. the parameters w of the utility function. A learning machine implementing minimization of $\mathcal{R}(w; M, \Sigma)$ for given values of M and Σ may be referred to as to *ordinal Linear Discriminant Analysis (oLDA)*. Note that in general, the risk minimization does not follow straightforwardly from an analytical expression or a standard convex optimization problem due to the use of the erf function.

2.2. THE BINARY CASE

The case of $L = 2$ is considered here in some detail. Since erf is a monotonically increasing function, the risk minimization may be implemented as follows

$$\min_w \mathcal{R}_2(w; m_1, m_2, \Sigma) \Leftrightarrow \max_w \frac{w^T(m_2 - m_1)}{\sqrt{2\, w^T \Sigma w}}. \quad (8)$$

Since the risk as defined here is not dependent on the norm $\|w\|$, we fix $\|w\|$ such that the values $w^T m_1 + w_0 = -1$ and $w^T m_2 + w_0 = 1$, for a corresponding $w_0 \in \mathbb{R}$, resulting in

$$\min_{w, w_0} \sqrt{2 w^T \Sigma w} \quad \text{s.t.} \quad w^T m_1 + w_0 = -1,\ w^T m_2 + w_0 = 1. \quad (9)$$

At this point, the following insight avoids the need for having an accurate description of the multivariate means and the covariance matrix. As one is only interested in the marginal distribution of the projection, knowledge of the variance of the univariate sets $\{w^T x_i\}_{i \in \mathcal{S}_l}$ is sufficient for minimizing the risk as previously. Furtheron, we substitute the sample mean and variances of those sets in the estimator: (i) the shifted marginal mean $w^T m_l + w_0$ is substituted by the sample mean $\frac{1}{|\mathcal{S}_l|} \sum_{i \in \mathcal{S}_l} w^T x_i + w_0$, and (ii) the variance $w^T \Sigma w$ is replaced by the sample estimate $\frac{1}{N} \sum_{i \in \mathcal{S}_1} (w^T x_i + w_0 + 1)^2 + \sum_{i \in \mathcal{S}_2} (w^T x_i + w_0 - 1)^2$. To counteract the finite sample effect of the minimum sample variance to deviate too much from the true marginal variance (or *overfitting*), regularization is applied. Including a regularized version of the sample variance defined as $w^T(\hat{\Sigma} + 1/\gamma I_d) w$ for a fixed $\lambda > 0$ (see Regularized Discriminant Analysis [7]) yields following modified optimization problem

$$(w^*, w_0^*) = \arg\min_{w, w_0} \mathcal{J}_2(w, w_0) =$$
$$\sum_{i \in \mathcal{S}_1} (w^T x_i + w_0 + 1)^2 + \sum_{i \in \mathcal{S}_2} (w^T x_i + w_0 - 1)^2 + \frac{1}{\gamma} w^T w$$
$$\text{s.t.} \quad \begin{cases} \frac{1}{|\mathcal{S}_1|} \sum_{i \in \mathcal{S}_1} w^T x_i + w_0 = -1 & (a) \\ \frac{1}{|\mathcal{S}_2|} \sum_{i \in \mathcal{S}_2} w^T x_i + w_0 = +1, & (b) \end{cases} \quad (10)$$

which establishes a connection between LDA, RDA and LS-SVMs [15] where the constraints (a) and (b) are omitted. For additional links of LS-SVMs and FDA, see e.g. [17].

2.3. ORDINAL LEAST SQUARES CLASSIFIER

Given only a set of samples $\mathcal{D}$, the accurate estimation of the class means $\{m_l\}_{l=1}^L$ and the covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ often prohibits the construction of an effective rule, especially when d is high. However, this can be avoided using the fact that in the end one is only interested in the means $\{b_l \triangleq w^T m_l\}_{l=1}^L$ and the variance $w^T \Sigma w$ of the marginal distributions. Hereto, the formulation of oLDA is modified in different ways as in the previous subsection using the sample class mean ($b_l = \frac{1}{|\mathcal{S}_l|} \sum_{i \in \mathcal{S}_l} w^T x_i + w_0$) and the sample variance ($frac1\gamma w^T w + \frac{1}{N} \sum_{i=1}^N (w^T x_i + w_0 - b_{l|i})^2$) where $b_{l|i}$ denotes the class l containing the ith sample. Risk minimization is obtained as

$$\min_{w,b} \frac{1}{2} \hat{\mathcal{R}}_\gamma(w, b; \hat{\Sigma}) \propto \sum_{l=1}^{L-1} \mathrm{erf}\left(\frac{-(b_{l+1} - b_l)}{\sqrt{2\, w^T (\hat{\Sigma} + 1/\gamma I) w}} \right), \quad (11)$$

such that $b_1 \leq \cdots \leq b_L$ holds, and where $b = (b_1, \ldots, b_L)^T \in \mathbb{R}^L$. Instead of solving this problem, the following approximation is used

$$\min_{w,b} \mathcal{J}_{\gamma,\lambda}(\hat{\Sigma}, w, b) = w^T(\gamma \hat{\Sigma} + I_d) w + \frac{\lambda}{2} \sum_{l=1}^{L-1} (b_{l+1} - b_l)^2, \quad (12)$$

with $\lambda > 0$ an appropriate constant (see e.g. Fig. 3). Although this approximation is only based on an intuitive ground, the resulting rule turns out to be a good approximation if the interval $b_1, \ldots, b_L$ is hinged on -1 and 1 similar as in the binary case. In this case the different values of b are forced towards an equidistantly spaced configuration preserving the specified ordering. Combining those ideas motivates the following ordinal classifier, only making use of least squares criteria and equality constraints.

Definition 1 (Ordinal Least Squares Classifier) *Let $\gamma > 0$ and $\lambda > 0$ be given hyper-parameters. Given the set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{1, \ldots, L\}$. Then the above principles (minimal overlap) are implemented as fol-*

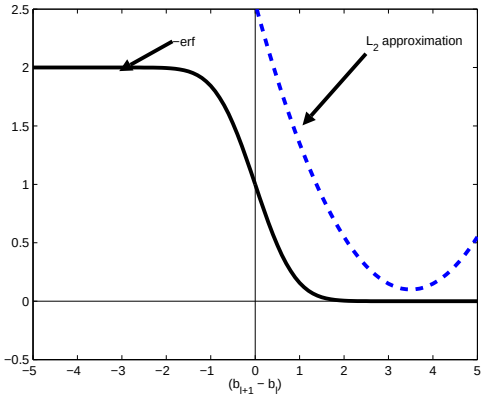

Fig. 3. *The* erf*-function (solid line) is convex in the case the domain is restricted to the positive half plane. The ordinal least squares classifier approximates this effect in terms of a L_2 function (dashed line).*

lows

$$(w^*, w_0^*, b^*) = \arg\min_{w_0,w,b} \mathcal{J}_{\gamma,\lambda}(w_0, w, b) =$$

$$\frac{\lambda}{2}\sum_{l=1}^{L-1} \|b_{l+1} - b_l\|^2 + \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{i=1}^{N} \|w^T x_i + w_0 - b_{l|i}\|^2$$

$$\text{s.t.} \quad \begin{cases} b_1 = -1, \;\; b_L = 1 \\ \frac{1}{|\mathcal{S}_l|}\sum_{i\in\mathcal{S}_l} w^T x_i + w_0 = b_l \quad \forall l = 1, \ldots, L \end{cases} \tag{13}$$

where $b_{l|i}$ denotes the mean $b_l \sim w^T m_l$ corresponding with the lth class containing the ith sample. Let $C_{\mathcal{S}} \in \mathbb{R}^{N\times L}$ be defined as the class indicator matrix such that $C_{\mathcal{S},il} = 1$ iff $i \in \mathcal{S}_l$ and zero otherwise. Then we can also note $b_{l|i}$ as $b_{l|i} = C_{\mathcal{S},i}b$.

Most importantly, this learning machine does not implement a simple regression on the labels, but allows instead a different metric of the responses (as parameterized by the vector $b = (b_1, \ldots, b_L)$). When the parameter $\lambda \to \infty$, it can be shown that the vector b is spread equidistantly over the interval $[-1, 1]$ and effectively equals the approach of regression on the labels. When $\lambda \to 0$, the ordering constraints are in general not satisfied. A proper value of λ results then in a vector b obeying the ordering while the values of b are varying reasonable flexible on the interval $[-1, 1]$.

3. A KERNEL-BASED APPROACH TO ORDINAL REGRESSION TASKS

3.1. ORDINAL LEAST SQUARES SUPPORT VECTOR MACHINES

A primal-dual derivation of the ordinal least squares classifier presented in the previous subsection is given, hereby allowing for the kernel trick as standard in LS-SVMs and kernel machines [18, 15]. Herefor, we reformulate the ordinal least squares classifier as a constrained optimization problem as follows:

$$\min_{w,w_0,b,e} \mathcal{J}'_{\gamma,\lambda}(w, b, e) = \frac{\lambda}{2}\sum_{l=1}^{L-1} \|b_{l+1} - b_l\|^2 + \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{i=1}^{N} e_i^2$$

$$\text{s.t.} \quad \begin{cases} w^T\varphi(x_i) + w_0 + e_i = b_{l|i} & \forall i = 1, \ldots, N \\ b_1 = -1 & \\ b_L = 1 & \\ \sum_{i\in\mathcal{S}_l}(w^T\varphi(x_i) + w_0 - b_l) = 0 & \forall l = 1, \ldots, L, \end{cases} \tag{14}$$

where $\varphi : \mathbb{R}^d \to \mathbb{R}^{D_\varphi}$ denotes the feature map induced by a positive definite kernel. The solution of this optimization problem can be found by solving a linear system.

Proposition 1 (Ordinal LS-SVM) *Define* [1]

- $C'_{\mathcal{S}} = C_{\mathcal{S}}(:, 2 : L-1) \in \mathbb{R}^{N\times L-2}$ *let* $C''_{\mathcal{S}} = C_{\mathcal{S}}(:, L) - C_{\mathcal{S}}(:, 1) \in \mathbb{R}^N$.
- *Let* $D \in \{-1, 0, 1\}^{L-1\times L}$ *be the first order difference matrix with* $D_{l,l} = 1$, $D_{l,l+1} = -1$ *for all* $l = 1, \ldots, L$ *and zero elsewhere.*
- *Let* $\Delta = D^T D \in \mathbb{R}^{L\times L}$, *then* $\Delta' = \Delta(:, 2 : L-1) \in \mathbb{R}^{L,L-2}$, *and let* $\Delta'' = \Delta(2 : L-1, 2 : L-1) \in \mathbb{R}^{L-2,L-2}$.
- $I_L \in \mathbb{R}^{L\times L-2}$ *equals the identity matrix of size* L, *then* $I'_L = I_L(:, 2 : L-1)$.
- $b'' \in \mathbb{R}^L$ *be defined as the vector* $b = (-1, 0, \ldots, 0, 1)^T$.
- *The vector* $\delta(\lambda) \in \mathbb{R}^{L-2}$ *is defined as* $\delta(\lambda) = \lambda(\Delta_L(2 : L-1, L) - \Delta_L(2 : L-1, L))$.

The optimum to (14) is characterized by the following dual set of linear equations

$$A \begin{bmatrix} \alpha \\ w_0 \\ \hline b'' \\ \xi \end{bmatrix} = \begin{bmatrix} I''_{\mathcal{S}} \\ 0 \\ \hline \delta(\lambda) \\ b \end{bmatrix}. \tag{15}$$

Let $1_N = (1, \ldots, 1)^T \in \mathbb{R}^N$ *and* $I_N \in \mathbb{R}^{N\times N}$ *be the identity matrix. The symmetric positive semidefinite matrix* $A \in \mathbb{R}^{(N+2*L-1)\times(N+2*L-1)}$ *is defined as*

$$A = \left[\begin{array}{cc|cc} \Omega + \frac{1}{\gamma}I_N & 1_N & -C'_{\mathcal{S}} & \bar{\Omega}_N \\ 1_N^T & 0 & 0 & 1_L^T \\ \hline -C'^T_{\mathcal{S}} & 0 & -\lambda\Delta'' & -I'^T_L \\ \bar{\Omega}_N^T & 1_L & -I'_L & \bar{\Omega} \end{array}\right], \tag{16}$$

[1] The matlab notation $A(:, i : l)$ denotes columns i to l of the matrix A, analoguously $A(i : l, :)$ denotes rows i to l of the matrix A.

where Ω is the kernel matrix such that $\Omega_{ij} = \varphi(x_i)^T\varphi(x_j)$ for all $i,j = 1,\ldots,N$. Let $\bar{\Omega}_L \in \mathbb{R}^{N\times L}$ be defined as $\bar{\Omega}_{L,il} = \frac{1}{|\mathcal{S}_i|}\sum_{j\in\mathcal{S}_l}\Omega_{ij}$ and $\bar{\Omega} \in \mathbb{R}^{L\times L}$ be defined as $\bar{\Omega} = \bar{\Omega}_L^T\bar{\Omega}_L$.

Proof: The Lagrangian of the constrained problem (14) becomes

$$\mathcal{L}(w,w_0,b,e;\alpha,\beta) = \frac{\lambda}{2}b^T\Delta b + \frac{1}{2}w^Tw + \frac{\gamma}{2}\sum_{i=1}^N e_i^2$$
$$-\sum_{i=1}^N \alpha_i(w^T\varphi(x_i)+w_0+e_i-b_{l|i}) - \beta_1(b_1+1) - \beta_L(b_L-1)$$
$$-\sum_{l=1}^L \xi_l\frac{1}{|\mathcal{S}_l|}\left(\sum_{i\in\mathcal{S}_l}(w^T\varphi(x_i)+w_0-b_l)\right), \quad (17)$$

with multipliers $\alpha = (\alpha_1,\ldots,\alpha_N)^T \in \mathbb{R}^N, \xi = (\xi_1,\ldots,\xi_L)^T \in \mathbb{R}^L$ and $\beta_1,\beta_L \in \mathbb{R}$. The first order conditions for optimality become

$$\begin{cases} \frac{\partial\mathcal{L}}{\partial w} = 0 \rightarrow & w = \sum_{i=1}^N \alpha_i\varphi(x_i) + \sum_{l=1}^L \xi_l\bar{x}_l \\ \frac{\partial\mathcal{L}}{\partial w_0} = 0 \rightarrow & \sum_{i=1}^N \alpha_i + \sum_{l=1}^L \xi_l = 0 \\ \frac{\partial\mathcal{L}}{\partial b_1} = 0 \rightarrow & \lambda\Delta_1 b = -\sum_{i\in\mathcal{S}_1}\alpha_i + \beta_1 - \xi_1 \\ \frac{\partial\mathcal{L}}{\partial b_L} = 0 \rightarrow & \lambda\Delta_L b = -\sum_{i\in\mathcal{S}_L}\alpha_i + \beta_L - \xi_L \\ \frac{\partial\mathcal{L}}{\partial b_l} = 0 \rightarrow & \lambda\Delta_l b = -\sum_{i\in\mathcal{S}_l}\alpha_i - \xi_l \quad (1<l<L) \\ \frac{\partial\mathcal{L}}{\partial e_i} = 0 \rightarrow & \gamma e_i = \alpha_i \qquad\qquad \forall = 1,\ldots,N \\ \frac{\partial\mathcal{L}}{\partial \alpha_i} = 0 \rightarrow & w^T\varphi(x_i) + e_i = b_{l|i} \quad \forall = 1,\ldots,N \\ \frac{\partial\mathcal{L}}{\partial \beta_1} = 0 \rightarrow & b_1 = -1 \\ \frac{\partial\mathcal{L}}{\partial \beta_L} = 0 \rightarrow & b_L = 1 \\ \frac{\partial\mathcal{L}}{\partial \xi_l} = 0 \rightarrow & w^T\bar{\varphi}_l + w_0 = b_l, \quad \forall l = 1,\ldots,L \end{cases} \quad (18)$$

where $\bar{\varphi}_l \in \mathbb{R}^d$ is defined as $\bar{\varphi}_l = \frac{1}{|\mathcal{S}_l|}\sum_{i\in\mathcal{S}_l}\varphi(x_i)$ for all $l = 1,\ldots,L$. Eliminating the primal variables w, b_1, b_L and e_i, and the dual variables β_1 and β_2 yields the dual system (15) where the symmetric positive semi-definite matrix $\Omega \in \mathbb{R}^{N\times N}$ contains the inner products between every pair of data-points $\Omega_{ij} = \Omega_{ji} = \varphi(x_i)^T\varphi(x_j)$. This can be written entirely in terms of the inner product hereby allowing for the application of the kernel trick $K(x_i,x_j) \triangleq \varphi(x_i)^T\varphi(x_j)$ where $K : \mathbb{R}^d\times\mathbb{R}^d \rightarrow \mathbb{R}$ is a proper inner product function referred to as the kernel. Note that one can also eliminate the primal variables b from the dual set of equations, resulting in smaller dual systems. ■

Using the condition $w = \sum_{i=1}^N \alpha_i\varphi(x_i) + \sum_{l=1}^L \xi_l\bar{\varphi}_l$, and by switching sums, the estimated utility function can be evaluated as follows

$$\hat{U}(x_*) = \sum_{i=1}^N\left(\alpha_i + \frac{\xi_{l|i}}{|\mathcal{S}_l|}\right)K(x_i,x_*), \quad (19)$$

where α and ξ are the solutions to (15) and $\xi_{l|i}$ denotes the multiplier belonging to the lth class containing the ith sample. The resulting classifier rules becomes

$$\left\{\hat{c}_l(x) = I(\hat{v}_{l-1,l} < \hat{U}(x) + w_0 \leq \hat{v}_{l,l+1})\right\}_{l=1}^L, \quad (20)$$

where again $\hat{v}_{l,l+1}$ is defined as $\frac{1}{2}(\hat{b}_l + \hat{b}_{l+1})$ for all $l = 1,\ldots,L+1$ and $\hat{v}_{0,1} = -\infty$ and $\hat{v}_{L,L+1} = \infty$. The use of the kernel representation is advantageous in case a linear rule does not suffice (see e.g. Fig. 4).

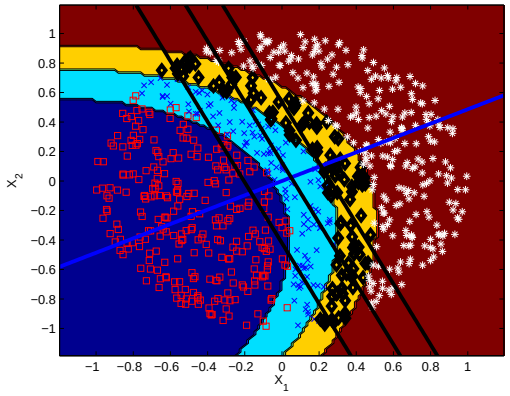

Fig. 4. *Example of the ordinal LS-SVM on a nonlinear toy example with quadratic boundaries. The linear LSC ordinal classifier is represented as the parallel solid lines, the nonlinear oLS-SVM result is visualized by the nonlinear contours.*

3.2. CONFUSION MATRIX AND ILLUSTRATIVE EXAMPLES

An important drawback of the method -at least in practice - is that multiple hyperparameters (γ and λ) are to be tuned. A qualitative assessment of the results of a ordinal discriminant rule can be given by the confusion matrix, a notion [12] well known in the literature on classical discriminant analysis. The confusion matrix $\Pi \in \mathbb{R}^{L\times L}$ can be defined in an ordinal setting as

$$\Pi_{kl} = \frac{1}{|\mathcal{S}_k|}\sum_{i\in\mathcal{S}_k}\mathcal{I}(c_l(x_i) = 1), \quad (21)$$

where $\mathcal{I}(z)$ equals one if z is true and zero otherwise. This notion is closely related to the ROC curve in the binary case, see [1] for a discussion in the bipartite ranking problem. If the confusion matrix is not diagonal dominant, the ordinal discriminant is not able to reflect the proper ordering of the labels well. Fig. 5 examplifies this tool.

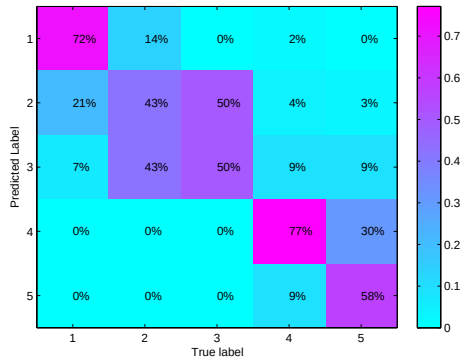

Fig. 5. *Confusion matrix of an ordinal discriminant rule discriminating between 5 classes. From this matrix, it becomes clear that the rule cannot reproduce well the class label of samples from the third class, while the rule clearly reproduces the class 1 labels.*

	LS	LDA	LS-SVM	OR	oMM	oLS-SVM
Abalone	52.34	51.43	**57.00**	52.34	55.02	56.12
Auto-mpg	50.12	55.88	53.67	50.13	52.20	**56.61**
Housing	44.50	50.28	51.44	57.22	**58.38**	57.86

Table 1. *Numerical performances (% accuracy) achieved on a number of UCI datasets, discretized in 8 equiprobable bins. LDA denotes a standard Linear Discriminant Analysis. The multiclass LS-SVMs are based on a one-versus-one encoding scheme, LS gives the result of a least squares regression on the labels. OR denotes the result of ordinal regression, while oMM gives results from an ordinal maximal approach [13]. For brevity, only linear tools (linear kernel) are applied, indicating the comparable performance of the methods. The last method denotes the ordinary LS-SVM with linear kernel which is equivalent to the primal ordinary Least Squares Classifier. Note that the ordering assumption may not yield optimal generalization performance as e.g. in case of the discretized Abalone dataset.*

An alternative model selection measure of a given rule is the risk as defined in (7). This continuous measure can be adviced to tune the parameter λ as it encourages a proper ordening, while avoiding the limit case of putting the thresholds $\{b_l\}_{l=1}^{L}$ equidistantly. It is however not recommended for tuning γ as it does not measure the generalization ability directly.

Six different methods are compared on a range of datasets taken from the UCI benchmark repository. At first, the LDA method equipped with a one-versus-one encoding scheme and a voting scheme are applied based on the LS-SVMlab toolbox respectively. Subsequently, the classical linear technique of ordinal regression (OR) is executed based on the StatBox 4.2 toolbox. Finally, the proposed ordinal linear least squares classifier and ordinal LS-SVM are applied. The tuning of the hyper-parameters is based on a 10-fold crossvalidation scheme. Table 1 reports some results on those datasets.

4. CONCLUSIONS

This paper discusses the ideas of linear discriminant analysis in a context of ordinal regression. These are implemented in a kernelk based context of ordinal LS-SVMs. [2]

5. REFERENCES

[1] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6:393–425, 2005.

[2] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifier. In *In Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–52. ACM, 1992.

[3] W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, 2006.

[4] W. Chu and S. S. Keerthi. New approaches to support vector ordinal regression. In *in Proc. of International Conference on Machine Learning*, pages 145–152. 2005.

[5] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[6] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.

[7] J. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84:165–175, 1989.

[8] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, Heidelberg, 2001.

[9] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Large Margin Classifiers*, pages 115–132, 2000. MIT Press, Cambridge, MA.

[10] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13:415–425, 2002.

[11] P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistics Society B.*, 42:109–142, 1980.

[12] G. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, 1946.

[13] A. Shashua and A. Levin. Ranking with large margin principle: Two approaches. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 937–944. MIT Press, Cambridge, MA, 2003.

[14] I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *IEEE Transactions on Information Theory*, 51:128–142, 2005.

[15] J.A.K. Suykens, T. van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

[16] J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.

[17] T. Van Gestel, J.A.K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. Vandewalle. A Bayesian framework for least squares support vector machine classifiers, gaussian processes and kernel Fisher discriminant analysis. *Neural Computation*, 14(5):1115–1147, 2002.

[18] V.N. Vapnik. *Statistical Learning Theory*. Wiley and Sons, 1998.

[2]- (KP): BOF PDM/05/161, FWO grant V4.090.05N; - (SCD:) GOA AMBioRICS, CoE EF/05/006, (FWO): G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0321.06, G.0553.06, (ICCoS, ANMMM, MLDM); (IWT): GBOU (McKnow), Eureka-Flite2 IUAP P5/22,PODO-II,FP5-Quprodis; ERNSI; - (JS) is an associate professor and (BDM) is a full professor at K.U.Leuven Belgium, respectively.