Math Meth Oper Res (2011) 74:381–407
DOI 10.1007/s00186-011-0368-z

ORIGINAL ARTICLE

Maximizing upgrading and downgrading margins for ordinal regression

Emilio Carrizosa · Belen Martin-Barragan

Received: 9 July 2010 / Accepted: 7 July 2011 / Published online: 20 September 2011

Abstract In ordinal regression, a score function and threshold values are sought to classify a set of objects into a set of ranked classes. Classifying an individual in a class with higher (respectively lower) rank than its actual rank is called an upgrading (respectively downgrading) error. Since upgrading and downgrading errors may not have the same importance, they should be considered as two different criteria to be taken into account when measuring the quality of a classifier. In Support Vector Machines, margin maximization is used as an effective and computationally tractable surrogate of the minimization of misclassification errors. As an extension, we consider in this paper the maximization of upgrading and downgrading margins as a surrogate of the minimization of upgrading and downgrading errors, and we address the biobjective problem of finding a classifier maximizing simultaneously the two margins. The whole set of Pareto-optimal solutions of such biobjective problem is described as translations of the optimal solutions of a scalar optimization problem. For the most popular case in which the Euclidean norm is considered, the scalar problem has a unique solution, yielding that all the Pareto-optimal solutions of the biobjective problem are translations of each other. Hence, the Pareto-optimal solutions can easily be

This research was partially supported by project MTM2009-14039, ECO2008-05080 of Ministerio de Educación y Ciencia (Spain), FQM-329 of Plan Andaluz de Investigación (Andalucía, Spain) and CCG07-UC3M/ESP-3389 of the Comunidad de Madrid (Spain).

E. Carrizosa
Facultad de Matemáticas, Universidad de Sevilla (Spain), Avda. Reina Mercedes, s/n, 41012 Sevilla, Spain
e-mail: ecarrizosa@us.es

B. Martin-Barragan (✉)
Facultad de CCSSJJ, Universidad Carlos III de Madrid (Spain), c/Madrid, 126, 28903 Getafe, Madrid, Spain
e-mail: belen.martin@uc3m.es

provided to the analyst, who, after inspection of the misclassification errors caused, should choose in a later stage the most convenient classifier. The consequence of this analysis is that it provides a theoretical foundation for a popular strategy among practitioners, based on the so-called ROC curve, which is shown here to equal the set of Pareto-optimal solutions of maximizing simultaneously the downgrading and upgrading margins.

Keywords Multi objective optimization · Support Vector Machines · Ordinal regression

1 Introduction

In many real-life situations one is interested in predicting, from a vector of variables, a variable which takes a finite number of ordered values. Examples of these situations appear in fields as diverse as Medicine (diseases are usually ranked from less severe to most severe), Education (students are sometimes graded on ordinal scales) or Collaborative Filtering (people are asked for their preferences, tastes or opinions categorized as excellent, good, . . ., horrible), (Ballarino et al. 2009; Cardoso et al. 2005; Grigoroudis et al. 2008; Lall et al. 2002 and the references therein).

These prediction problems are usually called *ordinal regression* problems, and can be seen as multi-class classification problems, where classes have a natural order, i.e., the variable representing the class is an ordinal variable instead of a categorical variable. In this sense, we use the term *rank* instead of class, to highlight that classes are ranked.

Classification (and, in particular, ordinal regression) problems can be solved by means of a good number of different procedures. We propose the use of Support Vector Machines (SVM), (Vapnik 1998), which have shown to be a very powerful classification tool. See Sect. 2 for a revision on the literature of SVM models for ordinal regression.

The basic SVM deals with a binary classification problem. The prediction is based on the value of a score function,

$$f(x) = \omega^\top x, \tag{1}$$

being below or above certain threshold β. Nonlinear score functions are also taken into account via a kernel that embeds the data into a higher dimensional space by replacing the coordinates x by the coordinates $\phi(x)$ in the embedded space. Mercer's theorem (1909) allows the representation of a kernel by the inner product in the embedded space, $k(x, y) = \langle \phi(x), \phi(y) \rangle$. The analysis developed in this paper analogously follows for the embedded space, just considering the embedded data $\phi(x)$ instead of the original x. Hence, under the Euclidean norm, the so-called kernel trick is as applicable in this analysis as in classical SVM. For simplicity in notation the analysis is derived in the original space.

For the multiclass SVM, several approaches have been proposed as in Allwein et al. (2000), Bredensteiner and Bennet (1999), Guermeur (2002), Platt et al. (2000) and

Weston and Watkins (1999). The most popular techniques are the one-against-one (o-a-o) (Cortes and Vapnik 1995) and the one-against-all (o-a-a) techniques (Hastie and Tibshirani 1998), where several two-class SVM problems are combined. Note that o-a-o and o-a-a approaches yield classifiers based on more than one score function. This has several disadvantages when applied to an ordinal regression problem. First, there is an instability concern, in the sense that similar objects might be predicted as belonging to very different ranks. Secondly, ordinal regression is often used to obtain, not only the prediction of the rank, but also an ordering of the objects. A typical example of this is collaborative filtering. In this kind of applications, a single score function can be directly applied to sort the objects. Finally, when interpretability is an issue, the contribution of every predictor variable to the rank is more simple if a single score function is used.

Two types of misclassification errors have especial relevance in ordinal regression: misclassifying an individual in a class with higher rank, called here *upgrading error*, and misclassifying an individual in a lower-ranked class, referred here to as *downgrading error*. These two types of misclassification errors may not have the same impact. For instance, misclassifying a patient with a severe disease as having an almost-harmless one is not as important as the other way round. Evidently, such difference of impact should be reflected somehow in the model. One might, for instance, introduce a cost matrix containing the different costs associated to the different types of misclassification errors, and seek a classification rule minimizing some aggregate cost measure (e.g. the expected misclassification cost). For instance, Li and Lin (2007) propose to transform an ordinal regression model with known costs into a binary classification problem with weighted samples. This transformation is used to derive generalization bounds based on the margin for the SVM-based ordinal regression. Lin and Li (2006) also use this transformation to derive generalization bounds for boosting in ordinal regression. In Lin and Li (2009), the reverse reduction is applied to design a boosting algorithm. In this latter method, the cost might be different for every object, but they are known for the objects in the training set.

The cost-sensitive approach for ordinal classification, see e.g. Li and Lin (2007), can be used in a wider range of situations, where not only upgrading and downgrading errors have different levels of importance, but also in situations where the cost of misclassification errors depends on the rank an object belongs to. Such misclassification costs are assumed to be known and the objective is to minimize the total cost. How to introduce these costs in the SVM framework is not so direct. SVM is mainly based on margin maximization, but the optimization problem also considers a loss term that penalizes misclassifications. The costs can be taken into account in this loss term, as proposed by Li and Lin (2007), but how the margin takes into account these costs is not so simple.

Giving precise values to such costs might not be an easy task in many practical applications (e.g. in Medicine, where costs measure how misclassifications damage the patient and relatives) (Adams and Hands 1999; Bradley 1997). Multiobjective optimization is then a useful approach to deal with this kind of imprecision or fuzziness affecting costs, as done for instance in Carrizosa and Martín-Barragán (2006) and Everson and Fieldsend (2006). The use of multiobjective optimization techniques in Machine Learning problems is not new, since much effort has been made in the

last decade. See Jin and Sendhoff (2008) for a recent overview of existing multiobjective Machine Learning approaches. For example, in Everson and Fieldsend (2006), a multiclass ROC analysis is proposed, where the ROC surface is defined by the Pareto-optimal solutions of a multiobjective optimization problem. Igel (2005) considers a multiobjective optimization problem where model complexity and training accuracy of an SVM are the two conflicting objectives. A multiobjective approach for the multi-class problem can be found in Tatsumi et al. (2007). In Nakayama et al. (2005) a Goal-Programming SVM model is proposed, and Carrizosa et al. (2008) addresses the biobjective problem of minimization of errors and measurement costs in an SVM. Approaches based on multicriteria decision making have also been proposed, as in Jiao et al. (2009), where SVM is combined with a multicriteria decision tool.

In this paper, we focus only on two types of errors: upgrading and downgrading errors. A model is proposed where the margin term takes into account the different cost/importance of these two types of errors. As we consider a biobjective problem, our model accommodates well the case in which these costs are fuzzy or unknown in advance.

The concept of margin, basic on SVM, needs to be extended to take into account the upgrading and downgrading errors. Hence, we define the upgrading margin and downgrading margin, first for the separable case, in which a solution correctly ranking all the objects exists, and we later extend the result obtained to the non-separable case. Our main result states that the set of all Pareto-optimal solutions is described as a set of solutions that are all translations of one of the solutions of the same optimization problem. In this way, this paper extends our previous result (Carrizosa and Martín-Barragán 2006) addressing a binary classification problem. In addition, the result obtained here is valid for any norm used to measure distances. When such norm is the Euclidean, the solutions are all translated of each others and the problem to solve is a convex quadratic problem with linear constraints.

The paper is organized as follows: In Sect. 2, the literature on SVM models for ordinal regression is reviewed. After this revision, in Sect. 3 we consider the separable case, where objects in the training set are not allowed to be misclassified. Since our aim is to take into account separately upgrading and downgrading errors, we extend the definition of margin given in Cristianini and Shawe-Taylor (2000) for the binary classification problem to the notion of upgrading margin and downgrading margin in ordinal regression. Then, following the SVM ideas, the simultaneous maximization of both the upgrading margin and the downgrading margin proceeds. We characterize the Pareto-optimal solutions of such biobjective problem. In Sect. 5, it is shown that a similar result is obtained for the non-separable case. Section 6 contains several illustrative examples that help to understand the main results of the paper and their applicability to real-world data. We finish with some conclusions and remarks for future lines of research.

2 SVM for ordinal regression. Review of the literature

Different proposals, briefly reviewed below, have been made to tackle ordinal regression with SVM-based tools. A training set I of objects is available. For every object

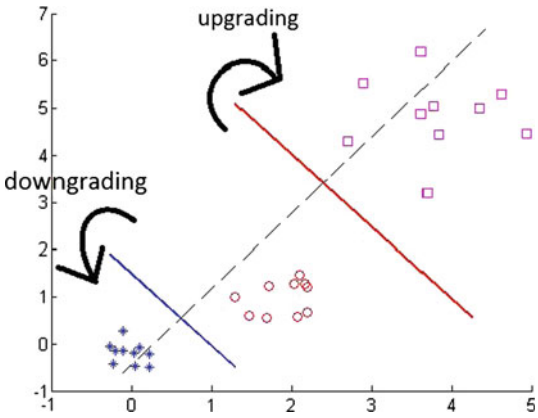

Fig. 1 Score function and thresholds

$u \in I$, its predictor vector $x^u \in I\!R^p$ and its rank $c^u \in \{1, 2, \ldots, R\}$ are known. The classes are ranked, thus the numbers associated with them are meaningful: class 1 is below class 2 and so on. The class set induces a partition in the training sample. In what follows, for $c \in \{1, 2, \ldots, R\}$, I^c denotes the set of objects having rank c, i.e., $I^c = \{u \in I : c^u = c\}$, and we assume throughout the paper that $I^c \neq \emptyset$, $\forall c \in \{1, 2, \ldots, R\}$.

Classification of future entries is made from the predictor vector as follows. First, a non-zero $\omega \in I\!R^p$ and $R-1$ thresholds $\beta^1 \leq \beta^2 \leq \ldots \leq \beta^{R-1}$ are determined. Then, an incoming individual with predictor vector x will be associated with rank $r(x)$,

$$r(x) = \begin{cases} 1, \text{ if } \omega^\top x < \beta^1 \\ 2, \text{ if } \beta^1 \leq \omega^\top x < \beta^2 \\ \vdots \qquad\qquad \vdots \\ R, \text{ if } \beta^{R-1} \leq \omega^\top x. \end{cases} \tag{2}$$

This way, given $\omega \neq 0$ and $\beta = \left(\beta^1, \ldots, \beta^{R-1}\right)$, $R-1$ parallel hyperplanes are obtained, splitting the space into R regions, associated with the R values of the rank.

Example 1 An example is shown in Fig. 1. A data set with 30 objects in dimension 2 and three ranks is generated. Objects in rank 1,2 and 3, represented respectively by stars, circles and squares, are generated following a gaussian distribution with mean $(0, 0)$, $(2, 1)$, $(4, 5)$ and covariance matrix $\frac{1}{16}I$, $\frac{1}{9}I$, $1.2I$, where I is the identity matrix.

Every threshold β^1, β^2 defines a hyperplane $\left\{x : \omega^\top x = \beta^c.\right\}$, represented as a solid line in the figure. Such hyperplanes split the space into three regions, associated with the three posible ranks: 1 (stars), 2 (circles) and 3 (squares).

The fact that the hyperplanes in Fig. 1 are parallel comes from the fact that only one score function (1) is used in the classifier. Other approaches are possible, as the ones existing for multi-class classification problems. We focus on a single score function because then the classifier is more stable, easier to interpret and provides an automatic ordering of all the objects.

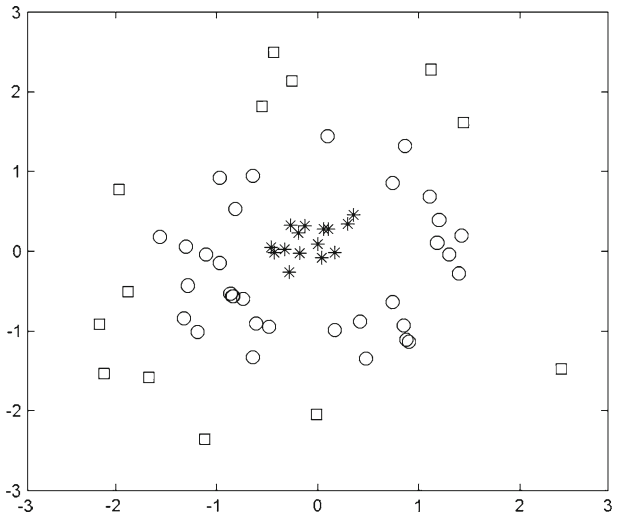

Fig. 2 Original data

As mentioned in the Introduction, we may perform a preprocessing step, transforming the predictor variables x via a nonlinear mapping ϕ, and use (2) with $\phi(x)$ instead of x. This strategy, common in the standard SVM is illustrated in Example 2 (Cristianini and Shawe-Taylor 2000; Shawe-Taylor and Cristianini 2004).

Example 2 In Fig. 2 a different example with three ranks (1, 2 and 3, represented as stars, circles and squares) is shown. No set of parallel hyperplanes exists correctly ranking the three classes. Figure 3 represents the same data after the mapping $\phi(x_1, x_2) = \left(x_1^2 + x_2^2, x_1\right)$, where it is obvious that such set of hyperplanes exists.

Several authors have proposed different SVM-based models for ordinal regression. Shashua and Levin introduce the so-called *fix-margin* model, where the parameters ω, β are obtained as optimal solution of the problem

$$\begin{array}{lll}
\min & \|\omega\|_2^2 + C\left(\sum\limits_{u\in I\setminus I^R} \xi^u + \sum\limits_{u\in I\setminus I^1} \tilde{\xi}^u\right) & \\
\text{s.t.:} & \left(\omega^\top x^u - \beta^c\right) \leq -1 + \xi^u & \forall u \in I^c,\ \forall c = 1, 2, \ldots, R-1 \\
& \left(\omega^\top x^u - \beta^{c-1}\right) \geq 1 - \tilde{\xi}^u & \forall u \in I^c,\ \forall c = 2, \ldots, R \\
& \xi^u \geq 0 & \forall u \in I^c,\ \forall c = 1, 2, \ldots, R-1 \\
& \tilde{\xi}^u \geq 0 & \forall u \in I^c,\ \forall c = 2, \ldots, R \\
& \omega \in I\!R^p,\ \beta \in I\!R^{R-1}. &
\end{array} \tag{3}$$

Note that nothing prevents in (3) the thresholds from being ordered inconsistently with the ordering in the ranks. To avoid this undesirable effect, Chu and Keerthi (2007) propose two different approaches. One way is to add the constraints $\beta^c \leq \beta^{c+1}$ for all $c = 1, 2, \ldots, R-2$. We refer to such modification as $(3)_c$.

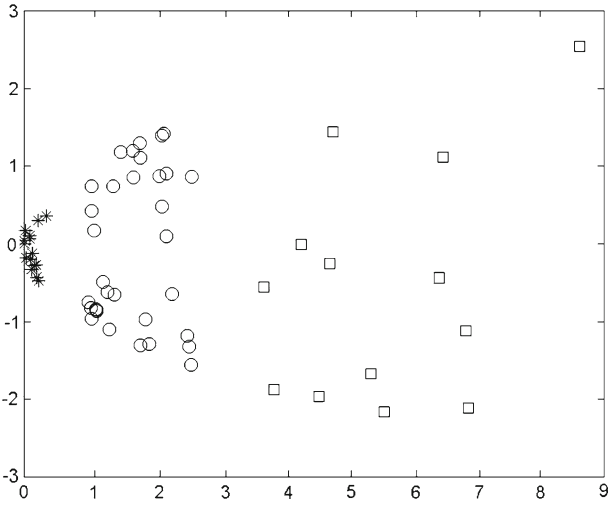

Fig. 3 Mapped data

Whereas problem (3) uses two slack variables ξ^u and $\tilde{\xi}^u$ for every object, the second variant proposed in Chu and Keerthi (2007) uses more slacks: one of the form ξ_j^u for every j with $c^u \leq j < R$, $u \in I \setminus I^R$, and another of the form $\tilde{\xi}_j^u$ for every $1 < j \leq c^u$, $u \in I \setminus I^1$. This yields the following formulation:

$$
\begin{array}{lll}
\min & \|\omega\|_2^2 + C\left(\displaystyle\sum_{j=1}^{R-1}\sum_{c=1}^{j}\sum_{u\in I^c}\xi_j^u + \sum_{j=2}^{R}\sum_{c=j}^{R}\sum_{u\in I^c}\tilde{\xi}_j^u\right) & \\
\text{s.t.:} & \left(\omega^\top x^u - \beta^j\right) \leq -1+\xi_j^u & \forall u \in I^c, \forall c = 1,2,\ldots,j, \forall j = 1,2,\ldots,R-1 \\
& \left(\omega^\top x^u - \beta^{j-1}\right) \geq 1-\tilde{\xi}_j^u & \forall u \in I^c, \forall c = j, j+1,\ldots,R, \forall j = 2,\ldots,R \\
& \xi_j^u \geq 0 & \forall u \in I^c, \forall c = 1,2,\ldots,j, \forall j = 1,2,\ldots,R-1 \\
& \tilde{\xi}_j^u \geq 0 & \forall u \in I^c, \forall c = j, j+1,\ldots,R, \forall j = 2,\ldots,R \\
& \omega \in I\!R^p, \beta \in I\!R^{R-1}. &
\end{array}
\tag{4}
$$

Note that, for an object belonging to rank c, if it is upgraded to rank $k > c$, then the error with respect to rank j, with $c \leq j < k$ would be $\xi_j^u = \omega^\top x^u - \beta^j + 1$. Similarly, when downgrading it into rank $k < c$, the error with respect to rank j, with $k < j \leq c$ is $\tilde{\xi}_j^u = \beta^{j-1} + 1 - \omega^\top x^u$. This means that, when an object is misclassified, the errors are computed as the sum of the errors with respect to all the ranks between the rank it has been allocated to and the rank it actually belongs to. Therefore, error is considered higher when an object is misranked several ranks away from its true rank, in other words, when multiple thresholds are crossed. Thus, solutions are encouraged to minimize the number of crossed thresholds (Rennie and Srebro 2005).

This model has been recently generalized in Cardoso et al. (2005), by fixing a parameter s (in practice to be chosen by cross-validation), and considering only the s

slacks ξ_j^u (respectively $\tilde{\xi}_j^u$) for rank j among the s closest ranks above c^u (respectively below c^u). This is written in Cardoso et al. (2005) as the optimization problem

$$\begin{array}{ll}
\min \|\omega\|_2^2 + C\left(\displaystyle\sum_{c=1}^{R-1}\sum_{j=c}^{\min\{c+s-1,R-1\}}\sum_{u\in I^c}\xi_j^u + \sum_{c=2}^{R}\sum_{\max\{2,c-s+1\}}^{R}\sum_{u\in I^c}\tilde{\xi}_j^u\right) & \\
\text{s.t.: } \left(\omega^\top x^u - \beta^j\right) \le -1+\xi_j^u & \forall j=c,\ldots,\min\{c+s-1,R-1\}, \forall u\in I^c, \forall c=1,2,\ldots,R-1 \\
\left(\omega^\top x^u - \beta^{j-1}\right) \ge 1-\tilde{\xi}_j^u & \forall j=\max\{2,c-s+1\},\ldots,c, \forall u\in I^c, \forall c=2,3,\ldots,R \\
\xi_j^u \ge 0 & \forall j=c,\ldots,\min\{c+s-1,R-1\}, \forall u\in I^c, \forall c=1,2,\ldots,R-1 \\
\tilde{\xi}_j^u \ge 0 & \forall j=\max\{2,c-s+1\},\ldots,c, \forall u\in I^c, \forall c=2,3,\ldots,R \\
\omega\in I\!R^p,\ \beta\in I\!R^{R-1}. &
\end{array} \tag{5}$$

A related approach, called *sum-of-margins* model, can be found in Shashua and Levin (2003).

The notion of *inversions* has been used in Carrizosa (2006) and Herbrich et al. (1999) to address ordinal regression using SVM. An inversion occurs when, given two objects $u \in I^c$ and $v \in I^d$, with $c < d$, object u is allocated to a rank greater than or equal to the rank where v is allocated. From the set I we generate the set of pairs $P = \{(u,v) \in I \times I : c^u \neq c^v\}$. A pair $(u,v) \in P$ is considered to be misclassified if it yields an inversion. We have then a 2-class classification problem, with P as training set and class labels $z^{(u,v)}$,

$$z^{(u,v)} = \begin{cases} 1, & \text{if } c^u > c^v \\ -1, & \text{otherwise.} \end{cases} \tag{6}$$

The classifier proposed is obtained from the SVM classifier on P : ω is the optimal solution of the problem

$$\begin{array}{ll}
\min \|\omega\|_2^2 + C\displaystyle\sum_{(u,v)\in P}\xi^{(u,v)} & \\
\text{s.t.: } z^{(u,v)}\left(\omega^\top x^u - \omega^\top x^v\right) \ge 1-\xi^{(u,v)} & \forall (u,v)\in P \\
\xi^{(u,v)} \ge 0 & \forall (u,v)\in P \\
\omega \in I\!R^p. &
\end{array} \tag{7}$$

In this case, thresholds β^c are considered as the midpoint of the pair x^u and x^v, having closest score values $\omega^\top x^u$ and $\omega^\top x^v$ among all $u \in I^c$ and $v \in I^{c+1}$.

We finish this section with a final remark on the choice of the norm. In all models mentioned above [an exception being (Carrizosa 2006)], the norm affecting ω is chosen to be the Euclidean norm, $\|\cdot\|_2$. Using $\|\cdot\|_2$ has two convenient properties: the resulting optimization problems are linearly-constrained convex quadratic optimization problems, solvable by a number of numerical procedures and packages, and they may allow the use of the so-called *kernel trick* (Cristianini and Shawe-Taylor 2000). However, other authors (Carrizosa 2008; Mangasarian 1965; Pedroso and Murata 2001) have proposed in SVM the use of other norms, such as the L_1 norm or L_∞, which yield Linear Programs, usually much easier to solve than quadratic problems.

In Pedroso and Murata (2001), empirical results show that 'in terms of separation performance, L_1, L_∞ and Euclidean norm-based SVMs tend to be quite similar'. The theoretical results that will be shown in this paper are true for any choice of this norm, hence they include the particular case of the Euclidean norm which allows us to use kernels.

3 Margins for separable classes

A key concept in the SVM is the margin. This section is devoted to extend the concept of *margin*, e.g., Cristianini and Shawe-Taylor (2000), to upgrading and downgrading margins. In Sect. 2 different models for ordinal regression with SVM found in the literature have been written as optimization problems. However, it is not direct to see how margin maximization is related to these models. The geometrical margin is much easier to understand when classes are *separable*, i.e., (ω, β) exists such that the classification rule (2) correctly ranks all objects in the training set I. The SVM approach that assumes that classes are separable is known in the literature as the *hard-margin* approach, e.g., Vapnik (1995). Note that although separability is a rather restrictive condition, it is usually fulfilled in practice after mapping the data into a vector space of higher dimension, in such a way that the classes become separable. See e.g., Cristianini and Shawe-Taylor (2000), Herbrich (2002).

The model for the separable case can be extended to the general case, known as *soft-margin* approach, in which (ω, β) is not forced to classify correctly all objects in I and misclassification is penalized via a loss term in the objective function of the corresponding optimization problem. Such case will be discussed in Sect. 5.

Under the assumption that the classes are separable, the search of a classifier is thus restricted to those (ω, β) satisfying the constraints

$$\begin{array}{c} \omega^\top x^u < \beta^1, \ \forall u \in I^1 \\ \beta^{c-1} < \omega^\top x^u < \beta^c, \ \forall c = 2, \ldots, R-1, \ \forall u \in I^c \\ \beta^{R-1} < \omega^\top x^u, \ \forall u \in I^R. \end{array} \tag{8}$$

Since, by assumption, $I^c \neq \emptyset$, $\forall c \in \{1, 2, \ldots, R\}$, one has, for any (ω, β) satisfying (8), that the threshold values β^c, with $c \in \{1, 2, \ldots, R-1\}$, are sorted, i.e.

$$\beta^1 < \beta^2 < \ldots < \beta^{R-1}. \tag{9}$$

Given (ω, β) satisfying (8) and $c \in \{1, 2, \ldots, R-1\}$ (respectively $c \in \{2, \ldots, R\}$), an object $u \in I^c$ satisfies $\omega^\top x^u \leq \beta^c$ (respectively $\omega^\top x^u \geq \beta^{c-1}$). Hence, the distance from x^u to the halfspace $\{x : \omega^\top x \geq \beta^c\}$ (respectively $\{x : \omega^\top x \leq \beta^{c-1}\}$) measures how far u is from being upgraded (respectively downgraded). This distance is basic for the concept of margin.

We recall that, given a norm ν in $\mathbb{R}^p$ and x^* such that $\omega^\top x^* \leq \beta$ (respectively $\omega^\top x^* \geq \beta$), the ν-distance from x^* to the halfspace $\{x : \omega^\top x \geq \beta\}$ (respectively

$\{x : \omega^\top x \leq \beta\}$) is given by $\frac{-\omega^\top x^* + \beta}{\nu^\circ(\omega)}$ (respectively $\frac{\omega^\top x^* - \beta}{\nu^\circ(\omega)}$), where ν° stands for the norm dual to ν. See Plastria (2009) for further details.

Definition 3 Given an arbitrary norm $\|\cdot\|$ in $\mathbb{R}^p$ and a classifier (ω, β) satisfying (8), the *upgrading margin* of (ω, β), is defined as

$$\rho_+(\omega, \beta) = \min_{c=1,2,\ldots,R-1} \min_{u \in I^c} \frac{\left(-\omega^\top x^u + \beta^c\right)}{\|\omega\|}. \tag{10}$$

Analogously, the *downgrading margin* of (ω, β) is defined as

$$\rho_-(\omega, \beta) = \min_{c=2,\ldots,R} \min_{u \in I^c} \frac{\left(\omega^\top x^u - \beta^{c-1}\right)}{\|\omega\|}. \tag{11}$$

Moreover, the *margin* $\rho(\omega, \beta)$ is defined as the smallest of the two margins above,

$$\rho(\omega, \beta) = \min\left\{\rho_+(\omega, \beta), \rho_-(\omega, \beta)\right\}. \tag{12}$$

Note that Definition 3 provides us with a new notion of margin for an ordinal regression problem. Although this notion of margin is not explicitly mentioned in any of the different models proposed in the literature, it turns out that it provides a general framework for all such models, in the sense that for the separable case, maximizing (12) is shown to be equivalent to any of such models already proposed. In order to prove this equivalence, let us first rephrase the problem of finding (ω, β) satisfying (8) yielding the largest margin ρ.

$$\begin{array}{ll} \min & \|\omega\| \\ \text{s.t.:} & \omega^\top x^u - \beta^c \leq -1 \qquad \forall u \in I^c,\ \forall c = 1, 2, \ldots, R-1 \\ & \omega^\top x^u - \beta^{c-1} \geq 1 \qquad \forall u \in I^c,\ \forall c = 2, \ldots, R \\ & \omega \in \mathbb{R}^p,\ \beta \in \mathbb{R}^{R-1}. \end{array} \tag{13}$$

Under the assumption that the classes are separable, we can force all slack variables $\xi, \tilde{\xi}$ in all problems in Sect. 2 to be equal to 0. For any such a problem (P), let $(\mathrm{P})^S$ denote the optimization problem (P) without slacks $\xi, \tilde{\xi}$. All such problems $(\mathrm{P})^S$ turn out to be equivalent to Problem (13), as stated in the following result.

Proposition 4 *Let $\|\cdot\|$ be the Euclidean norm. Problems $(3)^S$, $(3)^S_c$, $(4)^S$ and $(5)^S$ are identical to Problem (13), and they all are equivalent to Problem $(7)^S$ in the sense that*

(i) *if ω^* is an optimal solution of Problem $(7)^S$, then an optimal solution of Problem (13) is given by $(\hat{\omega}, \hat{\beta})$, with $\hat{\omega} = 2\omega^*$ and $\hat{\beta}^c = \max_{v \in I^c} \omega^{*\top} x^v + \min_{v \in I^{c+1}} \omega^{*\top} x^v$, $\forall c = 1, 2, \ldots, R-1$,*

(ii) *if $(\hat{\omega}, \hat{\beta})$, is an optimal solution of Problem (13), then $\omega^* = \frac{1}{2}\hat{\omega}$ is optimal for Problem $(7)^S$.*

For the proof, see the Appendix A.1.

This result links the SVM-based models available in the literature, revised in Sect. 2, with the margin maximization problem.

The different models proposed in the literature only differ in the soft-margin case, in the way they penalize the deviations. We can see two parts in every SVM-based model for ordinal regression: the margin part and the loss part. Different losses yield the different models proposed in the literature. For instance, as pointed out in Dembczyński and Kotłowski (2009) and Dembczyński et al. (2008), the loss in Problem (3) is tailored for minimization of 0/1 errors, whereas the loss in Problem (4) is for absolute deviations. In this sense, it can be seen that the loss used in Problem (5) is tailored for the minimization of the *truncated absolute deviations*, where the penalty is equal to the deviation when this deviation is not greater than s, and it is constant elsewhere.

Thanks to Proposition 4, we have a framework to extend the concept of margin to downgrading and upgrading margins.

4 Simultaneous maximization of both margins. The separable case

Problem (13), and thus also $(3)^S$, $(3)^S_c$, $(4)^S$, $(5)^S$ and $(7)^S$, maximize the margin ρ giving equal weight to upgrading and downgrading errors. In order to take into account the possible different importance of such errors, now we consider the biobjective optimization problem of simultaneously maximizing both margins ρ_+ and ρ_- under the constraint that (ω, β) separates the classes, i.e. satisfies (8):

$$\begin{array}{ll} \max & \{\rho_+(\omega,\beta), \rho_-(\omega,\beta)\} \\ \text{s.t.:} & \omega^\top x^u - \beta^c < 0 \qquad\qquad \forall u \in I^c,\ c = 1, 2, \ldots, R-1 \\ & \omega^\top x^u - \beta^{c-1} > 0 \qquad\qquad \forall u \in I^c,\ c = 2, \ldots, R \\ & \omega \in I\!R^p,\ \beta \in I\!R^{R-1},\ \omega \neq 0. \end{array} \tag{14}$$

Generalizing the results in Carrizosa and Martín-Barragán (2006), we seek the set of *Pareto-optimal* solutions to (14), i.e., the set of feasible (ω, β) such that no feasible (ω', β') exists satisfying

$$\begin{array}{l} \rho_+(\omega', \beta') \geq \rho_+(\omega, \beta), \\ \rho_-(\omega', \beta') \geq \rho_-(\omega, \beta), \end{array} \tag{15}$$

with at least one of those inequalities strict.

To characterize such a set, we first characterize the set of *weakly efficient* solutions, i.e., the set of feasible solutions that cannot be improved at the same time in both objectives $\rho_+(\omega, \beta)$ and $\rho_-(\omega, \beta)$. The reader is referred to Ehrgott and Gandibleaux (2002) for further details on multiobjective optimization.

Since all feasible solutions (ω, β) satisfy that $\rho_-(\omega, \beta) > 0$ and $\rho_+(\omega, \beta) > 0$, one can generate all weakly efficient solutions by solving max-min type scalarizations problems (Ehrgott and Gandibleaux 2002), as stated below for the sake of completeness.

Lemma 5 *The set of weakly efficient solutions of Problem (14) is obtained as the set of optimal solutions of problems of the form*

$$\begin{array}{ll} \max \min \{\rho_+(\omega,\beta), \theta\rho_-(\omega,\beta)\} & \\ s.t.: \ \left(\omega^\top x^u - \beta^c\right) < 0 & \forall u \in I^c,\ c = 1,2,\ldots,R-1 \\ \quad \left(\omega^\top x^u - \beta^{c-1}\right) > 0 & \forall u \in I^c,\ c = 2,\ldots,R \\ \quad \omega \in I\!R^p,\ \beta \in I\!R^{R-1},\ \omega \neq 0, & \end{array} \tag{16}$$

for $\theta \in (0, +\infty)$, in the sense that

(i) *any optimal solution of a problem of the form (16) is weakly efficient for Problem (14),*
(ii) *for every weakly efficient solution (ω, β) of Problem (14), (ω, β) is optimal for a problem of the form (16) for some $\theta \in (0, +\infty)$.*

For $\theta \in (0, +\infty)$, consider the convex minimization problem

$$\begin{array}{ll} \min \|\omega\| & \\ \text{s.t.: } \left(\omega^\top x^u - \beta^c\right) \leq -1 & \forall u \in I^c,\ c = 1,2,\ldots,R-1 \\ \quad \theta\left(\omega^\top x^u - \beta^{c-1}\right) \geq 1 & \forall u \in I^c,\ c = 2,\ldots,R \\ \quad \omega \in I\!R^p,\ \beta \in I\!R^{R-1}. & \end{array} \tag{P_θ}$$

Observe that Problem (P_1), the particular case of $\theta = 1$, corresponds to Problem (13). The following lemma relates the solution of Problem (P_θ) for the different values of θ.

Lemma 6 *Given $(\omega, \beta) \in I\!R^p \times I\!R^{R-1}$, the following statements are equivalent:*

(i) *There exists $\theta \in (0, +\infty)$ such that (ω, β) is optimal for Problem (P_θ).*
(ii) *There exists $\Lambda \in (-1, 1)$ such that $((1-\Lambda)\omega, \bar{\beta})$ with $\bar{\beta}^c = (1-\Lambda)\beta^c + \Lambda$, for all $c = 1, 2, \ldots, R-1$, is an optimal solution for Problem (P_1).*

For the proof, see the Appendix A.2.

Now we are in conditions to characterize the set of weakly efficient solutions to Problem (14).

Theorem 7 *Let $(\omega, \beta) \in I\!R^p \times I\!R^{R-1}$. The following statements are equivalent:*

(i) *(ω, β) is a weakly efficient solution of the biobjective Problem (14),*
(ii) *there exist $\gamma > 0$ and $\Lambda \in (-1, 1)$ such that $(\gamma\omega, \bar{\beta})$, with $\bar{\beta}^c = \gamma\beta^c + \Lambda$, $\forall c, = 1, 2, \ldots, R-1$, is an optimal solution of Problem (P_1).*

For the proof, see the Appendix A.3.

Corollary 8 (Characterization of Pareto-optimal solutions) *Let $(\omega, \beta) \in I\!R^p \times I\!R^{R-1}$. The following statements are equivalent:*

(i) *(ω, β) is a Pareto-optimal solution of the biobjective Problem (14),*
(ii) *there exist $\gamma > 0$ and $\Lambda \in (-1, 1)$ such that $(\gamma\omega, \bar{\beta})$, with $\bar{\beta}^c = \gamma\beta^c + \Lambda$, $\forall c, = 1, 2, \ldots, R-1$, is an optimal solution of Problem (P_1).*

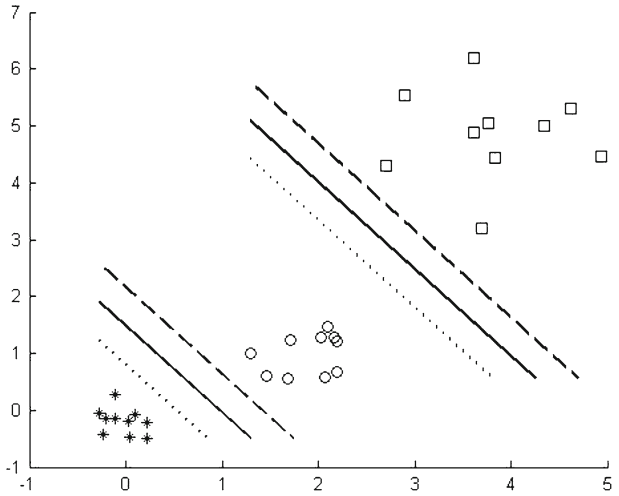

Fig. 4 Score function and thresholds

For the proof, see the Appendix A.4.

Observe that γ is a scaling parameter in the sense that (ω, β) and $(\gamma\omega, \gamma\beta)$ define the same ranking region. This Corollary characterizes the set of Pareto-optimal solutions. Indeed, every Pareto-optimal solution of the biobjective Problem (14) defines a region that is a translation of the region defined by one of the solutions of Problem (P_1). Remind that, in the particular case of the Euclidean norm, such solution is unique. Thus, for obtaining the whole set of Pareto-optimal solutions to (14), one just has to solve one problem, namely, Problem (P_1), and then let the threshold vector β vary in the appropriate range. Note that the different components of the threshold vector β should be moved all at the same time. Thus, there is an only free parameter, called Λ in Theorem 7, in order to get all the different Pareto-optimal solutions.

Example 9 A graphical illustration of this result can be seen in Fig. 4. The data in Example 1 are considered. Problem (P_1) is solved to find the coefficients (ω, β) of the score function. The problem is solved in Matlab 7.5. Remind that the classifier given by coefficients (ω, β) defines $R-1$ hyperplanes $\{x : \omega^\top x = \beta^c.\}$, separating rank c from rank $c-1$ for every $c = 1, 2, \ldots, R-1$. Such classifier is one Pareto-optimal soltuion of Problem (14). Corollary 8 characterizes all the Pareto-optimal solutions by the classifiers with coefficients $(\omega, \bar{\beta})$ with $\bar{\beta}^c = \beta^c + \Lambda,\ \forall \Lambda \in (-1, 1)$ and $c = 1, 2, \ldots, R-1$. The hiperplanes corresponding to several of these Pareto-optimal solutions are represented in Fig. 4. In particular, the solution with $\Lambda = 0$ is represented by a solid line, $\Lambda = 0.5$ by a dashed line, and $\Lambda = -0.5$ by a dotted line. The set of hyperplanes representing any Pareto-optimal solution is a translation of the set of hyperplanes representing another Pareto-optimal solution.

5 The general case: the soft-margin approach

In this section we address the general case and we propose a soft-margin approach to the problem.

In the geometrical interpretation for classical binary SVM, the hard-margin case is first analyzed. Afterwards the problem is perturbed by adding a loss term that penalizes such perturbation. This is called the soft-margin approach. A feasible optimization problem is obtained, but the meaning of margin is lost.

This has also been suggested in the literature of ordinal regression, as reviewed in Sect. 2. Different ways of defining these slack variables (one for each object, one for each object and possible rank, one for each pair of objects) yield the different formulations described in Sect. 2, namely, Problems (3), $(3)_c$, (4), (5) and (7) proposed in Cardoso et al. (2005), Chu and Keerthi (2007), Herbrich et al. (1999) and Shashua and Levin (2003).

For simplicity, we focus on one of these formulations, namely Problem $(3)_c$, proposed in Chu and Keerthi (2007). However, all the results in the section can be directly extended to the other slack variables proposed in Cardoso et al. (2005), Chu and Keerthi (2007) and Shashua and Levin (2003).

In Chu and Keerthi (2007) two slacks $\xi^u, \tilde{\xi}^u$, are associated to each object $u \in I$. Moreover, the constraint $\beta^c \leq \beta^{c+1}$ for all $c = 1, 2, \ldots, R-2$ is added to ensure that the thresholds are nondecreasing. Our aim is to perturb problem (14) for the nonlinearly separable case, keeping the geometrical interpretation of margin. This is done by defining the upgrading and downgrading margin by perturbing (10–11) via the slacks $\xi^u, \tilde{\xi}^u$.

Definition 10 Given an arbitrary norm $\|\cdot\|$ in $\mathbb{R}^p \times \mathbb{R}^{|I \setminus I^R|} \times \mathbb{R}^{|I \setminus I^1|}$, and a classifier $(\omega, \beta, \xi, \tilde{\xi})$, the *upgrading soft margin* of $(\omega, \beta, \xi, \tilde{\xi})$, is defined as

$$\rho_+(\omega, \beta, \xi, \tilde{\xi}) = \min_{c=1,2,\ldots,R-1} \min_{u \in I^c} \frac{\left(-\omega^\top x^u + \beta^c - \xi^u\right)}{\|(\omega, \xi, \tilde{\xi})\|}. \tag{17}$$

Analogously, the *downgrading soft margin* of $(\omega, \beta, \xi, \tilde{\xi})$ is defined as

$$\rho_-(\omega, \beta, \xi, \tilde{\xi}) = \min_{c=2,\ldots,R} \min_{u \in I^c} \frac{\left(\omega^\top x^u - \beta^{c-1} + \tilde{\xi}^u\right)}{\|(\omega, \xi, \tilde{\xi})\|}. \tag{18}$$

Moreover, the *soft margin* $\rho(\omega, \beta, \xi, \tilde{\xi})$ is defined as the smallest of the two margins above,

$$\rho(\omega, \beta, \xi, \tilde{\xi}) = \min\left\{\rho_+(\omega, \beta, \xi, \tilde{\xi}), \rho_-(\omega, \beta, \xi, \tilde{\xi})\right\}. \tag{19}$$

A convenient choice for $\|\cdot\|$ in Definition 10 is a weighted Euclidean norm,

$$\|(\omega,\xi,\tilde{\xi})\| = \sqrt{\|\omega\|_2^2 + C\left(\sum_{u\in I\setminus I^R}(\xi^u)^2 + \sum_{u\in I\setminus I^1}(\tilde{\xi}^u)^2\right)}, \tag{20}$$

where C is a constant trading off margin and generalization capability, Vapnik (1998).

The problem of finding the vector $(\omega,\beta,\xi,\tilde{\xi})$ yielding maximal soft margin, as given in Definition 10, can be formulated as the quadratic convex problem with linear constraints:

$$\begin{array}{lll}
\min & \|\omega\|_2^2 + C\left(\sum_{u\in I\setminus I^R}(\xi^u)^2 + \sum_{u\in I\setminus I^1}(\tilde{\xi}^u)^2\right) & \\
s.t.: & \left(\omega^\top x^u - \beta^c\right) - \xi^u \le -1 & \forall u\in I^c,\ \forall c = 1,2,\ldots,R-1 \\
& \left(\omega^\top x^u - \beta^{c-1}\right) + \tilde{\xi}^u \ge 1 & \forall u\in I^c,\ \forall c = 2,\ldots,R \\
& \beta^c \le \beta^{c+1} & \forall c = 1,2,\ldots,R-2 \\
& \xi^u \ge 0 & \forall u\in I^c,\ \forall c = 1,2,\ldots,R-1 \\
& \tilde{\xi}^u \ge 0 & \forall u\in I^c,\ \forall c = 2,\ldots,R \\
& \omega^\in I\!R^p,\ \beta\in I\!R^{R-1}. &
\end{array} \tag{21}$$

Observe that Problem (21) differs from Problem $(3)_c$, proposed by Chu and Keerthi (2007), only in that the squared slacks are penalized, instead of simply penalizing the slacks. This way, we retain the interpretation of margin, which is lost in their approach. Following the interpretation of the loss term in Dembczyński et al. (2008), Problem (21) is tailored for the minimization of 0/1 errors, since the loss does not depend on how many ranks away from the true rank an object is assigned to.

The soft-margin counterpart of Problem (14) is then

$$\begin{array}{lll}
\max & \left\{\rho_+(\omega,\beta,\xi,\tilde{\xi}), \rho_-(\omega,\beta,\xi,\tilde{\xi})\right\} & \\
s.t.: & \left(\omega^\top x^u - \beta^c\right) - \xi^u < 0 & \forall u\in I^c,\ \forall c = 1,2,\ldots,R-1 \\
& \left(\omega^\top x^u - \beta^{c-1}\right) + \tilde{\xi}^u > 0 & \forall u\in I^c,\ \forall c = 2,\ldots,R \\
& \beta^c \le \beta^{c+1} & \forall c = 1,2,\ldots,R-2 \\
& \xi^u \ge 0 & \forall u\in I^c,\ \forall c = 1,2,\ldots,R-1 \\
& \tilde{\xi}^u \ge 0 & \forall u\in I^c,\ \forall c = 2,\ldots,R \\
& \omega^\in I\!R^p,\ \beta\in I\!R^{R-1}. &
\end{array} \tag{22}$$

Results similar to those presented in Sect. 3 can be obtained for Problem (22). Indeed, the set of Pareto-optimal solutions of (22) can be characterized as translated solutions from the optimum of Problem (21). However, an intuitive interpretation of the range obtained under this translation is not straightforward. The characterization is stated in the following Theorem and Corollary.

Theorem 11 *Let* $(\omega,\beta,\xi,\tilde{\xi}) \in I\!R^p \times I\!R^{R-1} \times I\!R^{|I\setminus I^R|} \times I\!R^{|I\setminus I^1|}$. *The following statements are equivalent:*

(i) $(\omega, \beta, \xi, \tilde{\xi})$ *is a weakly efficient solution of the biobjective Problem (22),*
(ii) *there exist* $\Lambda \in (-1, 1)$, $\gamma > 0$, *such that* $(\gamma\omega, \bar{\beta}, \gamma\xi, \gamma\tilde{\xi})$, *with* $\bar{\beta}^c = \gamma\beta^c + \Lambda$, $\forall c, = 1, 2, \ldots, R-1$, *is an optimal solution of Problem (21).*

Corollary 12 (Characterization of Pareto-optimal solutions) *Let* $(\omega, \beta, \xi, \tilde{\xi}) \in \mathbb{R}^p \times \mathbb{R}^{R-1} \times \mathbb{R}^{|I \setminus I^R|} \times \mathbb{R}^{|I \setminus I^1|}$. *The following statements are equivalent:*

(i) $(\omega, \beta, \xi, \tilde{\xi})$ *is a Pareto-optimal solution of the biobjective Problem (22),*
(ii) *there exist* $\Lambda \in (-1, 1)$, $\gamma > 0$, *such that* $(\gamma\omega, \bar{\beta}, \gamma\xi, \gamma\tilde{\xi})$, *with* $\bar{\beta}^c = \gamma\beta^c + \Lambda$, $\forall c, = 1, 2, \ldots, R-1$, *is an optimal solution of Problem (21).*

6 Illustrative examples in real-world datasets

In binary classification, when the importance of the errors is different for the two different classes, the behavior of a classifier is usually analyzed using the ROC curve (Kupinski and Anastasio 1999). The ROC curve shows the sensitivity, i.e., the proportion of correctly classified objects of the positive class, against the specificity, proportion of correctly classified objects of the negative class, for different values of the threshold. For ordinary regression, as well as for multigroup classification and ordinal regression, some extensions to the ROC curve have been proposed, see e.g. Hand and Till (2001) and Waegeman et al. (2008). For our model, we first define two different error measurements. For every $u \in I$, let us denote by $\hat{c}^u$ the rank predicted by the classifier for the object u. Remind that c^u denotes the actual rank of object u, and let $\sharp$ denote the cardinality of a set.

- *Empirical downgrading error* is the proportion of downgraded objects in I among those whose rank is greater than one.

$$D = \frac{\sharp\{u \in I : \hat{c}^u < c^u,\ c^u \neq 1\}}{\sharp\{u \in S : c^u \neq 1\}}.$$

- *Empirical upgrading error* is the proportion of upgraded objects in I among those whose rank is lower than R.

$$U = \frac{\sharp\{u \in I : \hat{c}^u > c^u,\ c^u \neq R\}}{\sharp\{u \in S : c^u \neq R\}}.$$

Consider the classifiers with coefficients $(\omega, \bar{\beta})$ with $\bar{\beta}^c = \beta^c + \Lambda$, $\forall \Lambda \in (-1, 1)$. Let $D(\Lambda)$ (respectively, $U(\Lambda)$) be the empirical downgrading (resp. upgrading) error for the classifier with coefficients $(\omega, \bar{\beta})$. Consider the curve $(1 - D(\Lambda), 1 - U(\Lambda))$ for different values $\Lambda \in (-1, 1)$. we will call such plot the *DU curve* (from downgrading/upgrading). Every point in the DU curve represents the trade-off between downgrading and upgrading errors for a specific value of Λ. Observe that, thanks to the characterization given by Corollary 12, all the Pareto-optimal solutions of the biobjective optimization of the downgrading and upgrading margins are represented in such curve. This result reduces the resolution of the biobjective problem to solving just one single-objective problem. A property rather unusual in multiobjective optimization.

As illustrative examples, the DU curves are represented in real-world datasets (see Fig. 5). The datasets were previously used by Chu and Keerthi in (2007).[1] These datasets are modified versions whose originals had been previously used for metric regression problems. The metric regression task was converted into an ordinal regression task by discretizing the response variable into 5 or 10 ordinal scales, using equal-frequency binning. As representative example we choose the 5-classes case. Each dataset is randomly partitioned into training and testing set. In our experiments, we always use the first out of the 20 repetitions recollected by Chu and Keerthi. No other preprocessing apart from the one used by Chu and Keerthi has been done. More information about the datasets is given in Table 1.

In real-world applications, the soft-margin is usually considered either because data are not separable or to avoid overfitting. We focus on the soft-margin model inspired in Chu and Keerthi (2007), whose biobjective counterpart is Problem (22). Hence, optimization of Problem (21) proceeds. In our experiments it has been solved using Matlab 7.5. The parameter C which allows to control overfitting is chosen by 10-fold crossvalidation in the training set. A linear kernel is used in all the experiments, but other kernels can be used, as discussed in Sect. 2.

The DU curve is the thick curve represented in Fig. 5. The interpretation of this curve is similar to the interpretation of the ROC curve. For $\Lambda \in (-1, 1)$, $(1 - D(\Lambda))$ plays the role of the specificity while $(1 - U(\Lambda))$ plays the role of the sensitivity. The DU curve shows the tradeoff between downgrading and upgrading error (any decrease in the downgrading error will be accompanied by an increase in the upgrading error). The closer the curve follows the right-hand border and then the top border, the more accurate the prediction. The closer the curve comes to the 45-degree diagonal, the less accurate the prediction. In order to understand the behavior of the classifiers in forthcoming data, we computed the downgrading and upgrading errors in a test dataset. The corresponding downgrading/upgrading curve on the test data is represented as a thin curve.

Consider for instance the dataset `auto.data`. The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of eight attributes. The classification task 'mpg' (miles per gallon) has been discretized into 5 classes (very low, low, moderate, high, very high). Suppose that an environmental tax is fixed using this rank. Misclassifying a car as having 'moderate' consumption when it has 'high' consumption, would cause a damage in the environment because it indirectly increases the consumption of fuel. However, if the opposite error is incurred (misclassifying it as 'very high'), the car's company would make an extra effort to prove that the actual fuel consumption is lower. In this setting, it seems that the downgrading error is less important than the upgrading error, but, in any case, it does not seem to be easy to quantify how many times one error is more important than the other. Looking at the DU curve depicted in Fig. 5, one observes that in order to get a downgrading error of 5% (the 0.95 in the horizontal axis), one should admit an upgrading error of nearly 40%. But being less demanding in the downgrading error (for instance, allowing 10% error), the upgrading error would be less than 25%. This way, the decision maker can use the DU curve to trade off the importance of the two kinds of errors, and how being more demanding with one of them affects being more tolerant with the other.

[1] Datasets are freely available at http://www.liaad.up.pt/~ltorgo/Regression/DataSets.html.

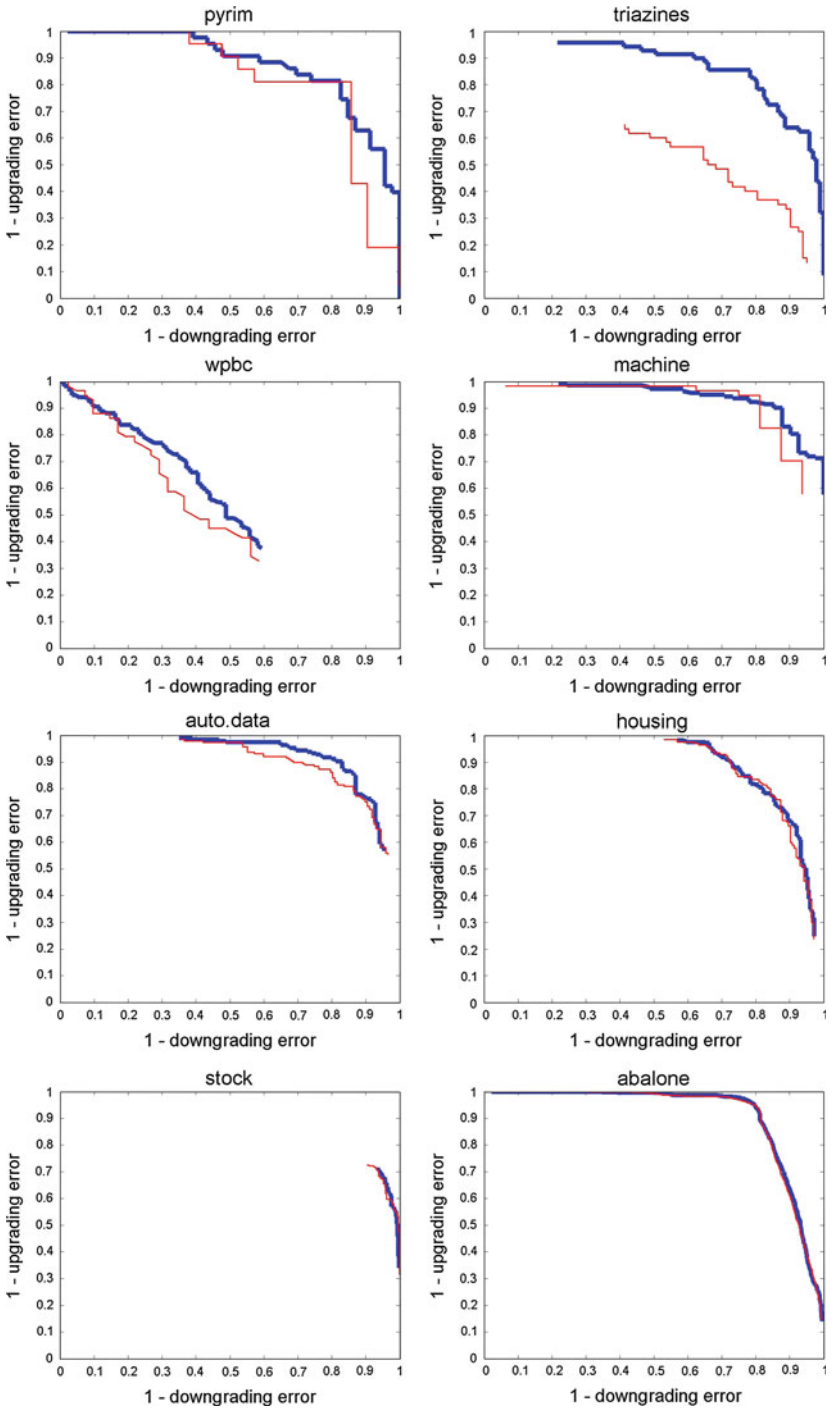

Fig. 5 DU curves. Pareto

Table 1 Information of the datasets

Database	p	Training size	Testing size
Pyrim	27	50	24
Triazines	60	100	86
wpbc	32	130	54
Machine	6	150	59
Autodata	7	200	192
Housing	13	300	206
Stock	9	600	350
Abalone	10	1,000	3,177

Note that the downgrading/upgrading curve can be drawn for any classifier that is based on a single score function. The contribution of this paper is that it provides a theoretical foundation and an interpretation of the points in this curve as the Pareto-optimal solutions of maximizing simultaneously the downgrading and upgrading margins.

7 Conclusions

The use of multiobjective optimization techniques for Machine Learning and Data Mining problems has gained much popularity in recent years. We have addressed in this paper the simultaneous minimization of upgrading and downgrading errors, via, following an SVM-based approach, upgrading and downgrading margins. Our main result states that the whole set of Pareto-optimal solutions can be obtained by solving one quadratic optimization problem, and then let the thresholds vary in an appropriate range. This type of characterization results are rather unusual in multiobjective optimization. In the hard-margin case, such quadratic optimization problem coincides with the hard-margin version of the problems reviewed from the literature.

Upgrading and downgrading errors in ordinal regression are only two kinds of errors that might be taken into account. An error cost that depends on the rank a misclassified object belongs to is another possibility that has been addressed in Tatsumi et al. (2007) for multiclass classification. In such paper, error costs are symmetric, in the sense that misclassifying an object from class c into a class c' is considered as important as the other way round. Considering these two costs as different and unknown deserves further research.

Acknowledgments The authors are grateful to the Editor and anonymous referees for their constructive suggestions.

Appendix: Proofs of the results

A.1 Proof of Proposition 4

Problems $(3)^S$, $(4)^S$ and $(5)^S$ are identical to Problem (13). Problem $(3)_c^S$ is a modification of $(3)^S$, where constraints $\beta^c \leq \beta^{c+1}$ for all $c = 1, 2, \ldots, R-1$ are added. Such constraints are trivially satisfied at any (ω, β) feasible for $(3)^S$, thus $(3)_c^S$ is

also equivalent to (13). Showing the equivalence between these four problems with Problem $(7)^S$ deserves further details.

We only show Part (*i*), since Part (*ii*) can be shown using the very same arguments. Let ω^* be an optimal solution of Problem $(7)^S$. We need to prove that the $(\hat{\omega}, \hat{\beta})$ given by the proposition is an optimal solution of Problem (13).

First we prove that $(\hat{\omega}, \hat{\beta})$ is feasible. Indeed, let $u \in I^c$ and $c = 1, 2, \ldots, R-1$. One has that

$$\begin{aligned}\hat{\omega}^\top x^u - \hat{\beta}^c &= 2\omega^{*\top} x^u - \max_{v \in I^c} \omega^{*\top} x^v - \min_{v \in I^{c+1}} \omega^{*\top} x^v \\ &= \left(\omega^{*\top} x^u - \max_{v \in I^c} \omega^{*\top} x^v\right) + \left(\omega^{*\top} x^u - \min_{v \in I^{c+1}} \omega^{*\top} x^v\right) \end{aligned} \tag{23}$$

Since $u \in I^c$, then

$$\omega^{*\top} x^u - \max_{v \in I^c} \omega^{*\top} x^v \leq 0. \tag{24}$$

Now we focus in the last two terms in (23). For $v \in I^{c+1}$, one has, by definition of $z^{(u,v)}$ in (6), that $z^{(u,v)} = -1$. Since ω^* is feasible for Problem $(7)^S$, we have that $z^{(u,v)}(\omega^{*\top} x^u - \omega^{*\top} x^v) \geq 1$, i.e.,

$$\left(\omega^{*\top} x^u - \omega^{*\top} x^v\right) \leq -1.$$

Since this is true for all $v \in I^{c+1}$, we get

$$\left(\omega^{*\top} x^u - \min_{v \in I^{c+1}} \omega^{*\top} x^v\right) \leq -1. \tag{25}$$

Joining (23) with (24) and (25), we conclude that $\hat{\omega}^\top x^u - \hat{\beta}^c \leq -1$. It is analogous to prove that for $u \in I^c$ we have $\hat{\omega}^\top x^u - \hat{\beta}^{c-1} \geq 1$. Hence, $(\hat{\omega}, \hat{\beta})$ is feasible for Problem (13).

Now we show the optimality of $(\hat{\omega}, \hat{\beta})$ for (13). Let (ω, β) be a feasible solution of Problem (13), i.e.,

$$\omega^\top x^u - \beta^c \leq -1, \ \forall u \in I^c, \ \forall c = 1, 2, \ldots, R-1, \tag{26}$$

$$\omega^\top x^u - \beta^{c-1} \geq 1, \ \forall u \in I^c, \ \forall c = 2, \ldots, R. \tag{27}$$

Then, $\frac{1}{2}\omega$ is feasible for Problem $(7)^S$. Indeed, let $u, v \in I$ with $c^u > c^v$ and thus $z^{(u,v)} = 1$. (The case with $c^u < c^v$, and thus $z^{(u,v)} = -1$, is completely analogous).

By (9), $\beta^{c^u-1} \geq \beta^{c^v}$. Hence,

$$z^{(u,v)}\left(\frac{1}{2}\omega^\top x^u - \frac{1}{2}\omega^\top x^v\right) = \frac{1}{2}\omega^\top x^u - \frac{1}{2}\omega^\top x^v$$
$$= \frac{1}{2}\left(\omega^\top x^u - \beta^{c^u-1}\right) + \frac{1}{2}\left(\beta^{c^u-1} - \beta^{c^v}\right)$$
$$+\frac{1}{2}\left(\beta^{c^v} - \omega^\top x^v\right) \geq \frac{1}{2} + 0 + \frac{1}{2} = 1. \tag{28}$$

Hence, $z^{(u,v)}\left(\frac{1}{2}\omega^\top x^u - \frac{1}{2}\omega^\top x^v\right) \geq 1\ \forall u, v \in I$ with $c^u \neq c^v$, and this shows that $\frac{1}{2}\omega$ is feasible for Problem $(7)^S$. By the optimality of ω^* with respect to Problem $(7)^S$, ω satisfies $\|\omega^*\| \leq \frac{1}{2}\|\omega\|$. In summary, we have that

$$\|\hat{\omega}\| = \|2\omega^*\| \leq \|2\frac{1}{2}\omega\| = \|\omega\|.$$

Since (ω, β) is an arbitrary feasible solution of Problem (13), we have shown that $(\hat{\omega}, \hat{\beta})$ is feasible for Problem (13) with better or equal objective value, $\|\hat{\omega}\| \leq \|\omega\|$. We conclude that $(\hat{\omega}, \hat{\beta})$ is an optimal solution of Problem (13), what proves the first part of the result.

A.2 Proof of Lemma 6

First suppose that there exists $\theta \in (0, +\infty)$ such that (ω, β) is optimal for Problem (P_θ). Let $\Lambda = \frac{1-\theta}{1+\theta} \in (-1, 1)$. We now prove that $((1-\Lambda)\omega, \bar{\beta})$ with $\bar{\beta}^c = (1-\Lambda)\beta^c + \Lambda$ for all $c = 1, 2, \ldots, R-1$, is feasible for Problem (P_1). Denote $\bar{\omega} = (1-\Lambda)\omega$. Since (ω, β) is optimal for Problem (P_θ), then for all $u \in I^c$, with $c = 1, 2, \ldots, R-1$, the inequality $\left(\omega^\top x^u - \beta^c\right) \leq -1$ holds. Hence, we have, for $(\bar{\omega}, \bar{\beta})$,

$$\left(\bar{\omega}^\top x^u - \bar{\beta}^c\right) = \left((1-\Lambda)\omega^\top x^u - (1-\Lambda)\beta^c - \Lambda\right)$$
$$= (1-\Lambda)\left(\omega^\top x^u - \beta^c\right) - \Lambda \leq (1-\Lambda)(-1) - \Lambda = -1. \tag{29}$$

Analogously, for all $u \in I^c$, with $c = 2, \ldots, R$, since for (ω, β) the inequality $\theta\left(\omega^\top x^u - \beta^{c-1}\right) \geq 1$ holds, we easily derive the following inequality for $(\bar{\omega}, \bar{\beta})$,

$$\left(\bar{\omega}^\top x^u - \bar{\beta}^{c-1}\right) = \left((1-\Lambda)\omega^\top x^u - (1-\Lambda)\beta^{c-1} - \Lambda\right) = (1-\Lambda)\left(\omega^\top x^u - \beta^{c-1}\right)$$
$$-\Lambda \geq (1-\Lambda)\frac{1}{\theta} - \Lambda = \left(1 - \frac{1-\theta}{1+\theta}\right)\frac{1}{\theta} - \frac{1-\theta}{1+\theta} = 1. \tag{30}$$

Hence, we have that $(\bar{\omega}, \bar{\beta})$ is feasible for Problem (P_1). Now, we need to prove that it is also optimal for Problem (P_1).

Let $(\hat{\omega}, \hat{\beta})$ be a feasible solution of Problem (P_1). Taking $\omega^* = \frac{1}{1-\Lambda}\hat{\omega}$, and $\beta^{*c} = \frac{1}{1-\Lambda}(\hat{\beta}^c - \Lambda)$, we can see that (ω^*, β^*) is feasible for Problem (P_θ). Indeed, for all $u \in I^c$, with $c = 1, 2, \ldots, R-1$, one has

$$\left(\omega^{*\top} x^u - \beta^{*c}\right) = \left(\frac{1}{1-\Lambda}\hat{\omega}^\top x^u - \frac{1}{1-\Lambda}(\hat{\beta}^c - \Lambda)\right)$$
$$= \frac{1}{1-\Lambda}\left(\hat{\omega}^\top x^u - \hat{\beta}^c + \Lambda\right) \leq \frac{1}{1-\Lambda}(-1+\Lambda) = -1. \quad (31)$$

Analogously, for all $u \in I^c$, with $c = 2, \ldots, R$, the following inequality holds

$$\theta\left(\omega^{*\top} x^u - \beta^{*c-1}\right) = \theta\left(\frac{1}{1-\Lambda}\hat{\omega}^\top x^u - \frac{1}{1-\Lambda}(\hat{\beta}^{c-1} - \Lambda)\right)$$
$$= \theta\frac{1}{1-\Lambda}\left(\hat{\omega}^\top x^u - \hat{\beta}^{c-1} + \Lambda\right) \geq \theta\frac{1}{1-\Lambda}(1+\Lambda) = \theta\frac{1}{1-\frac{1-\theta}{1+\theta}}\left(1 + \frac{1-\theta}{1+\theta}\right)$$
$$= 1. \quad (32)$$

Since, by assumption (ω, β), is an optimum of Problem (P_θ), we have that $\|\omega^*\| \geq \|\omega\|$, which yields

$$\|\bar{\omega}\| = (1-\Lambda)\|\omega\| \leq (1-\Lambda)\|\omega^*\| = (1-\Lambda)\frac{1}{1-\Lambda}\|\hat{\omega}\| = \|\hat{\omega}\|.$$

In summary we have that $(\bar{\omega}, \bar{\beta})$ is feasible for Problem (P_1) and with objective value better than or equal to an arbitrary feasible solution $(\hat{\omega}, \hat{\beta})$ of Problem (P_1). We conclude that $(\bar{\omega}, \bar{\beta})$ is an optimal solution of Problem (P_1).

Now we prove the converse. Let $\Lambda \in (-1, 1)$ be such that $((1-\Lambda)\omega, \bar{\beta})$, with $\bar{\beta}^c = (1-\Lambda)\beta^c + \Lambda$ is an optimal solution of Problem (P_1) for an arbitrary $(\omega, \beta) \in \mathbb{R}^p \times \mathbb{R}^{R-1}$. We have to prove that (ω, β) is an optimal solution of Problem (P_θ) for some $\theta \in (0, +\infty)$.

Denote $\bar{\omega} = (1-\Lambda)\omega$. Then, $\omega = \frac{1}{1-\Lambda}\bar{\omega}$ and $\beta^c = \frac{1}{1-\Lambda}(\bar{\beta}^c - \Lambda)$. Using the very same arguments than used in (31) for (ω^*, β^*), one gets that $(\omega^\top x^u - \beta^c) \leq -1$ for all $u \in I^c$, $c = 1, 2, \ldots, R-1$. Analogously, following the very same arguments than in (32), one gets that

$$\left(\omega^\top x^u - \beta^{c-1}\right) \geq \theta\frac{1}{1-\Lambda}(1+\Lambda),$$

which, taking $\theta = \frac{1-\Lambda}{1+\Lambda} \in (0, +\infty)$, is equal to 1 for all $u \in I^c$, $c = 2, \ldots, R$. Hence we have that (ω, β) is feasible for (P_θ).

Now we prove optimality of (ω, β) for (P_θ). Let $(\hat{\omega}, \hat{\beta})$ an arbitrary feasible solution for Problem (P_θ). Following the very same arguments than in (29) and (30), one gets that (ω^*, β^*), with $\omega^* = (1-\Lambda)\hat{\omega}$ and $\beta^{*c} = (1-\Lambda)\hat{\beta}^c + \Lambda$, $\forall\, c = 1, 2, \ldots, R-1$, is a feasible solution for Problem (P_1) and thus, the optimal solution of Problem (P_1), namely $(\bar{\omega}, \bar{\beta})$, is better than or equal to it, i.e., $\|\bar{\omega}\| \leq \|\omega^*\|$. Then, we have that

$$\|\omega\| = \frac{1}{1-\Lambda}\|\bar{\omega}\| \leq \frac{1}{1-\Lambda}\|\omega^*\| = \frac{1}{1-\Lambda}(1-\Lambda)\|\hat{\omega}\| = \|\hat{\omega}\|.$$

In summary, we get that (ω, β) is feasible for Problem (P_θ) and better than or equal to any arbitrary feasible solution $(\hat{\omega}, \hat{\beta})$. Hence we conclude that (ω, β) is an optimal solution for Problem (P_θ).

A.3 Proof of Theorem 7

Let $(\bar{\omega}, \bar{\beta}) \in I\!R^p \times I\!R$. By Lemma 5, $(\bar{\omega}, \bar{\beta})$ is weakly efficient for Problem (14) if and only if there exists $\theta \in (0, +\infty)$ such that $(\bar{\omega}, \bar{\beta})$ is an optimal solution of Problem (16). This is equivalent to $(\bar{\omega}, \bar{\beta})$ being optimal for

$$\begin{array}{ll} \min & \frac{\|\omega\|}{D(\omega,\beta)} \\ \text{s.t.:} & \left(\omega^\top x^u - \beta^c\right) < 0 \qquad\qquad \forall u \in I^c,\ c = 1, 2, \ldots, R-1 \\ & \left(\omega^\top x^u - \beta^{c-1}\right) > 0 \qquad\qquad \forall u \in I^c,\ c = 2, \ldots, R \\ & \omega \in I\!R^p,\ \beta \in I\!R^{R-1},\ \omega \neq 0. \end{array} \tag{33}$$

where $D(\omega, \beta)$ denotes

$$\min\left\{\min_{c=1,2,\ldots,R-1}\min_{u\in I^c}\left(-\omega^\top x^u + \beta^c\right), \theta \min_{c=2,\ldots,R}\min_{u\in I^c}\left(\omega^\top x^u - \beta^{c-1}\right)\right\}.$$

Observe that any (ω, β) is optimal for (33) if and only if $(\mu\omega, \mu\beta)$ is optimal for (33) for any $\mu > 0$. Hence, by normalizing the denominator in the objective of (33) we have that $(\bar{\omega}, \bar{\beta})$ is optimal for (33) if and only if there exists and $\mu > 0$ such that $(\mu\bar{\omega}, \mu\bar{\beta})$ is optimal for the following problem:

$$\begin{array}{ll} \min & \|\omega\| \\ \text{st:} & D(\omega, \beta) = 1, \\ & \omega \in I\!R^p,\ \beta \in I\!R^{R-1}. \end{array} \tag{34}$$

Such a problem is equivalent to the following one

$$\begin{array}{ll} \min & \|\omega\| \\ \text{st:} & D(\omega, \beta) \geq 1, \\ & \omega \in I\!R^p,\ \beta \in I\!R^{R-1}. \end{array} \tag{35}$$

Indeed, given a solution of Problem (35) with $D(\omega, \beta) > 1$, a better solution can be built by dividing ω and β by $D(\omega, \beta)$.

Problem (35) can be directly rephrased as Problem (P_θ). Hence, $(\bar{\omega}, \bar{\beta})$ is weakly efficient for Problem (14) iff there exists $\mu > 0$ such that $(\mu\bar{\omega}, \mu\bar{\beta})$ solves (P_θ) for some $\theta \in (0, +\infty)$.

In order to finish the proof, let (ω, β) be weakly efficient for Problem (14). Then, as shown above, there exist $\mu > 0$ and $\theta \in (0, +\infty)$ such that $(\mu\omega, \mu\beta)$ solves Problem (P_θ). By Lemma 6, there exists $\Lambda \in (-1, 1)$ such that $((1-\Lambda)\mu\omega, \bar{\beta})$ with

$\bar{\beta}^c = (1-\Lambda)\mu\beta^c + \Lambda$, for all $c = 1, 2, \ldots, R-1$, is an optimal solution for Problem (P_1). Hence, taking $\gamma = (1-\Lambda)\mu$, we have that $(\gamma\omega, \bar{\beta})$, with $\bar{\beta}^c = \gamma\beta^c + \Lambda$ for all $c = 1, 2, \ldots, R-1$, is an optimal solution for Problem (P_1). This proves that (i) implies (ii) in the Theorem.

To show that (ii) implies (i) let $\gamma > 0$ and $\Lambda \in (-1, 1)$ such that $(\gamma\omega, \bar{\beta})$ with $\bar{\beta}^c = \gamma\beta^c + \Lambda$ for all $c = 1, 2, \ldots, R-1$, is an optimal solution for Problem (P_1). Note that $(\gamma\omega, \bar{\beta})$ can be rephrased as $((1-\Lambda)\frac{\gamma}{(1-\Lambda)}\omega, \bar{\beta})$ with $\bar{\beta}^c = (1-\Lambda)\frac{\gamma}{(1-\Lambda)}\beta^c + \Lambda$. Applying Lemma (6), we get that $(\frac{\gamma}{(1-\Lambda)}\omega, \frac{\gamma}{(1-\Lambda)}\beta)$ solves Problem (P_θ) for some $\theta \in (0, +\infty)$. As stated above, taking $\mu = \frac{\gamma}{(1-\Lambda)} > 0$, this is equivalent to (ω, β) being weakly efficient for Problem (14).

A.4 Proof of Corollary 8

Any Pareto-optimal solution is, by definition weakly efficient. Let us show that in our problem the converse also holds. Let (ω, β) be weakly efficient to (14). By contradiction, suppose it is not Pareto-optimal. Then there would exist a feasible solution (ω', β') satisfying (15) with at least one of those inequalities strict.

Without loss of generality we can suppose that $\|\omega'\| = \|\omega\| = 1$. Note that both inequalities cannot be strict because it would contradict the fact that (ω, β) is weakly efficient. Suppose $\rho_-(\omega', \beta') = \rho_-(\omega, \beta)$, thus $\rho_+(\omega', \beta') > \rho_+(\omega, \beta)$. An analogous reasoning will follow in case $\rho_+(\omega', \beta') = \rho_+(\omega, \beta)$ and $\rho_-(\omega', \beta') > \rho_-(\omega, \beta)$. We are going to build a feasible solution which is strictly better than (ω, β) in both objectives.

Let Λ be such that $0 < \Lambda < \rho_+(\omega', \beta') - \rho_+(\omega, \beta)$. Consider $\hat{\beta}$ with $\hat{\beta}^c = \beta'^c - \frac{\Lambda}{2}$ for all $c = 1, 2, \ldots, R-1$. We have that $(\omega', \hat{\beta})$ is feasible for Problem (14):

$$\begin{aligned}
&\min_{c=1,2,\ldots,R-1} \min_{u\in I^c} \left(-\omega'^{\top} x^u + \hat{\beta}^c\right) \\
&\quad = \min_{c=1,2,\ldots,R-1} \min_{u\in I^c} \left(-\omega'^{\top} x^u + \beta'^c - \frac{\Lambda}{2}\right) = \rho_+\left(\omega', \beta'\right) \\
&\quad -\frac{\Lambda}{2} > \rho_+(\omega', \beta') - \frac{\rho_+(\omega', \beta') - \rho_+(\omega, \beta)}{2} = \frac{\rho_+(\omega', \beta') + \rho_+(\omega, \beta)}{2} > 0,
\end{aligned} \tag{36}$$

and

$$\begin{aligned}
&\min_{c=2,\ldots,R} \min_{u\in I^c} \left(\omega'^{\top} x^u - \hat{\beta}^{c-1}\right) \\
&\quad = \min_{c=2,\ldots,R} \min_{u\in I^c} \left(\omega'^{\top} x^u - \beta'^{c-1}\right) + \frac{\Lambda}{2} > 0.
\end{aligned} \tag{37}$$

We now prove that $(\omega', \hat{\beta})$ is better than (ω, β) in both objectives, which contradicts the efficiency of (ω, β). First, since $\rho_+(\omega', \beta') - \rho_+(\omega, \beta) > \Lambda$, one obtains

$$\begin{aligned}\rho_+(\omega', \hat{\beta}) &= \min_{c=1,2,\ldots,R-1} \min_{u\in I^c} \left(-\omega'^{\top} x^u + \hat{\beta}^c\right)\\ &= \min_{c=1,2,\ldots,R-1} \min_{u\in I^c} \left(-\omega'^{\top} x^u + \beta'^c - \frac{\Lambda}{2}\right)\\ &= \rho_+(\omega', \beta') - \frac{\Lambda}{2} > \rho_+(\omega', \beta')\\ &-\frac{\rho_+(\omega', \beta') - \rho_+(\omega, \beta)}{2} = \frac{\rho_+(\omega', \beta') + \rho_+(\omega, \beta)}{2}. \end{aligned} \tag{38}$$

Moreover, $\rho_+(\omega', \beta') > \rho_+(\omega, \beta)$, which yields

$$\frac{\rho_+(\omega', \beta') + \rho_+(\omega, \beta)}{2} > \frac{\rho_+(\omega, \beta) + \rho_+(\omega, \beta)}{2} = \rho_+(\omega, \beta). \tag{39}$$

By putting together inequalities (38) and (39), one gets that $\rho_+(\omega', \hat{\beta}) > \rho_+(\omega, \beta)$.

Secondly, since $\Lambda > 0$, one gets that

$$\begin{aligned}\rho_-(\omega', \hat{\beta}) &= \min_{c=2,\ldots,R} \min_{u\in I^c} \left(\omega'^{\top} x^u - \hat{\beta}^{c-1}\right)\\ &= \min_{c=2,\ldots,R} \min_{u\in I^c} \left(\omega'^{\top} x^u - \beta'^{c-1} + \frac{\Lambda}{2}\right)\\ &= \rho_-(\omega', \beta') + \frac{\Lambda}{2} > \rho_-(\omega', \beta') = \rho_-(\omega, \beta). \end{aligned} \tag{40}$$

In summary, we have found a feasible solution, better than (ω, β) in both objectives, which contradicts (ω, β) being weakly efficient. Hence, we conclude that (ω, β) is Pareto-optimal for Problem (14)

References

Adams NM, Hands DJ (1999) Comparing classifiers when the miallocation costs are uncertain. Pattern Recognit 32:1139–1147

Allwein EL, Schapire RE, Singer Y (2000) Reducing multiclass to binary: a unifying approach for margin classifiers. J Mach Learn Res 1:113–141

Ballarino G, Bernardi F, Requena M, Schadee H (2009) Persistent inequalities? expansion of education and class inequality in Italy and Spain. Eur Sociol Rev 25(1):123–138

Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 30:1145–1159

Bredensteiner E, Bennet K (1999) Multicategory classification by support vector machines. Comput Opt Appl 12:53–79

Cardoso JS, da Costa JF Pinto, Cardoso MJ (2005) Modelling ordinal relations with SVMs: an application to objective aesthetic evalutaion of breast cancer conservative treatment. Neural Netw 18:808–817

Carrizosa E (2006) Deriving weights in multiple-criteria decision making with support vector machines. TOP 14(2):399–424

Carrizosa E (2008) Support vector machines and distance minimization. In: Pardalos PM, Hansen P (eds) Data mining and mathematical programming. AMS, New York, pp 2–20

Carrizosa E, Martín-Barragán B (2006) Two-group classification via a biobjective margin maximization model. Eur J Oper Res 173(3):746–761

Carrizosa E, Martín-Barragán B, Morales D Romero (2008) Multi-group support vector machines with measurement costs: a biobjective approach. Discret Appl Math 156(6):950–966
Chu W, Keerthi SS (2007) Support vector ordinal regression. Neural Comput 19(3):792–815
Cortes C, Vapnik V (1995) Support-vector network. Mach Learn 20:273–297
Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines. Cambridge University Press, Cambridge
Dembczyński K, Kotłowski W (2009) Decision rule-based algorithm for ordinal classification based on rank loss minimization. In: Preference learning, ECML/PKDD workshop
Dembczyński K, Kotłowski W, Słowiński R (2008) Ordinal classification with decision rules. In: Proceedings of the 3rd ECML/PKDD international conference on Mining complex data. MCD'07. Springer, Berlin, pp 169–181
Ehrgott M, Gandibleaux X (eds) (2002) Multiple criteria optimization. State of the art annotated bibliographic surveys, volume 52 of international series in operations research and management science. Kluwer Academic Publishers, Boston
Everson RM, Fieldsend JE (2006) Multi-class ROC analysis from a multi-objective optimisation perspective. Pattern Recogn Lett 27(8):918–927
Grigoroudis E, Nikolopoulou G, Zopounidis C (2008) Customer satisfaction barometers and economic development: An explorative ordinal regression analysis. Total Qual Manag Bus Excell 19(5):441–460
Guermeur Y (2002) Combining discriminant models with multi-class SVMs. Pattern Anal Appl 5:168–179
Hand DJ, Till RJ (2001) A simple generalisation of the area under the roc curve for multiple class classification problems. Mach Learn 45(2):171–186
Hastie T, Tibshirani R (1998) Classification by pairwise coupling. Ann Stat 26(2):451–471
Herbrich R (2002) Learning theory classifiers. Theory and algorithms. MIT Press, Cambridge
Herbrich R, Graepel T, Obermayer K (1999) Support vector learning for ordinal regression. In: In Ninth international conference on artificial neural networks ICANN, vol. 17, pp 97–102
Igel C (2005) Multi-objective model selection for supprot vector machines. In: Evolution multi-criterion optimization. Lecture notes in computer sciences, vol. 3410, pp 534–546
Jiao T, Peng J, Terlaky T (2009) A confidence voting process for ranking problems based on support vector machines. Ann Oper Res 166:23–38
Jin Y, Sendhoff B (2008) Pareto-based multiobjective machine learning: an overview and case studies. IEEE Trans Syst Man Cybern Part C Appl Rev 38(3):397–415
Kupinski MA, Anastasio MA (1999) Multiobjective genetic optimization of diagnostic classifiers with implications for generating receiver operating characteristic curves. IEEE Trans Med Imaging 18(8):675–685
Lall R, Campbell MJ, Walters SJ, Morgan K (2002) A review of ordinal regression models applied on health-related quality of life assessments. Stat Methods Med Res 11(1):49–67
Li L, Lin HT (2007) Ordinal regression by extended binary classification. In: Schölkopf B, Platt J, Hoffman T (eds) Advances in neural information processing systems, vol. 19. MIT Press, Cambridge, pp 865–872
Lin HT, Li L (2006) Large-margin thresholded ensembles for ordinal regression: theory and practice. In: Algorithmic learning theory: ALT 2006. Lecture notes in computer sciences, vol. 4264, Springer, Berlin, pp 319–333
Lin HT, Li L (2009) Combining ordinal preferences by boosting. In: Second preference learning workshop at ECML/PKDD'09
Mangasarian OL (1965) Linear and nonlinear separation of patterns by linear programming. Oper Res 13:444–452
Mercer J (1909) Functions of positive and negative type and their connection with the theory of integral equations. Philos Trans Royal Soc Lond A 209:415–446
Nakayama H, Yun YB, Asada T, Yoon M (2005) MOP/GP models for machine learning. Eur J Oper Res 166:756–768
Pedroso JP, Murata N (2001) Support vector machines with different norms: motivation, formulations and results. Pattern Recognit Lett 22:1263–1272
Plastria F (2009) Asymmetric distances, semidirected networks and majority in Fermat-Weber problems. Ann Oper Res 167(1):121–155
Platt JC, Cristianini N, Shawe-Taylor J (2000) Large margin DAGs for multiclass classification. Adv Neural Inform Process Syst 12:547–553

Rennie JDM, Srebro N (2005) Loss functions for preference levels: regression with discrete ordered labels. In: Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling

Shashua A, Levin A (2003) Ranking with large margin principle: two approaches. In: Thrun S, Becker S, Obermayer K (eds) Advances in Neural Information Processing Systems, volume 15. MIT Press, Cambridge pp 937–944

Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, Cambridge

Tatsumi K, Hayashida K, Higashi H, Tanino T (2007) Multi-objective multiclass support vector machine for pattern recognition. SICE, 2007. Annual Conference, pp 1095–1098

Vapnik V (1995) The nature of statistical learning theory. Springer, Berlin

Vapnik V (1998) Statistical learning theory. Wiley, New York

Waegeman W, De Baets B, Boullart L (2008) Roc analysis in ordinal regression learning. Pattern Reognit Lett 29(1):1–9

Weston J, Watkins C (1999) Multi-class support vector machines. In: Proceedings of ESANN99. D. Facto Press, Brussels