Taxonomy of Large Margin Principle Algorithms for Ordinal Regression Problems*

Amnon Shashua
Computer Science Department
Stanford University
Stanford, CA 94305
email: shashua@cs.stanford.edu

Anat Levin
School of Computer Science and Engineering
Hebrew University of Jerusalem
Jerusalem 91904, Israel
email: alevin@cs.huji.ac.il

Abstract

We discuss the problem of ranking instances where an instance is associated with an integer from 1 to k. In other words, the specialization of the general multi-class learning problem when there exists an ordering among the instances — a problem known as "ordinal regression" or "ranking learning". This problem arises in various settings both in visual recognition and other information retrieval tasks. In the context of applying a large margin principle to this learning problem, we introduce two main approaches for implementing the large margin optimization criteria for $k-1$ margins. The first is the "fixed margin" policy in which the margin of the closest neighboring classes is being maximized — which turns out to be a direct generalization of SVM to ranking learning. The second approach allows for $k-1$ different margins where the sum of margins is maximized, thus effectively having the solution biased towards the pairs of neighboring classes which are the farthest apart from each other. This approach is shown to reduce to νSVM when the number of classes $k = 2$. Both approaches are optimal in size (of the dual functional) of $2l$ where l is the total number of training examples. Experiments performed on visual classification and "collaborative filtering" show that both approaches outperform existing ordinal regression algorithms applied for ranking and multi-class SVM applied to general multi-class classification.

1 Introduction

In this paper we investigate the problem of inductive learning from the point of view of predicting variables of ordinal scale [3, 7, 5], a setting referred to as *ranking learning* or *ordinal regression*. We consider the problem of applying the large margin principle used in Support Vector methods [11, 2] to the ordinal regression problem while maintaining an (optimal) problem size linear in the number of training examples.

Ordinal regression may be viewed as a problem bridging between the two standard machine

*This manuscript should be referenced as "Technical Report 2002-39, Leibniz Center for Research, School of Computer Science and Eng., the Hebrew University of Jerusalem."

learning tasks of classification and (metric) regression. Let $\mathbf{x}_i \in R^n$, $i = 1, ...l$, be the input vectors (the information upon which prediction takes place) drawn from some unknown probability distribution $D(\mathbf{x})$; let $y_i \in Y$ be the output of the prediction process according to a unknown conditional distribution function $D(y|\mathbf{x})$. The *training* set, on which the selection of the best predictor would be made, consists of $(\mathbf{x}_i, y_i)$ independent and identically distributed observations drawn from the joint distribution $D(\mathbf{x}, y) = D(\mathbf{x})D(y|\mathbf{x})$.

The learning task is to select a prediction function $f(\boldsymbol{x})$ from a family of possible functions $\mathcal{F}$ that minimizes the expected *loss* over the training set weighted by the joint distribution $D(\mathbf{x}, y)$ (also known as *risk functional*). The loss function $c : Y \times Y \rightarrow R$ represents the discrepancy between $f(\mathbf{x})$ and y. Since the joint distribution is unknown, the risk functional is replaced by the so-called *empirical risk functional*[11] which is simply the average of the loss function over the training set: $(1/l) \sum_i c(f(\mathbf{x}_i), y_i)$.

In a standard *classification* problem the input vectors are associated with one of k classes, thus $y_i \in Y = \{1, ..., k\}$ belongs to an *unordered* set of labels denoting the class membership. Since Y is unordered and since the metric distance between the prediction $f(\mathbf{x})$ and the correct output y is of no particular value, the loss function relevant for classification is the non-metric 0-1 indicator function $c(f(\mathbf{x}), y) = 0$ if $f(\mathbf{x}) = y$ and $c(f(\mathbf{x}), y) = 1$ if $f(\mathbf{x}) \neq y$. In a standard *regression* problem y ranges over the reals therefore the loss function can take into account the full metric structure — for example, $c(f(\mathbf{x}), y) = (f(\mathbf{x}) - y)^2$.

In ordinal regression, Y is a finite set (like in classification) but there is an *ordering* among the elements of Y (like in regression, but unlike classification). On the other hand, the ordering of the labels does not justify a metric loss function, thus casting the ranking learning problem as an ordinary regression (by treating the continuous variable with a coarse scale) may not be realistic [1]. Settings in which it is natural to rank or rate instances arise in many fields such as information retrieval, visual recognition, collaborative filtering, econometric models and classical statistics. We will later use some applications from collaborative filtering and visual recognition as our running examples in this paper. In collaborative filtering for example, the goal is to predict a person's rating on new items such as movies given the person's past ratings on similar items and the ratings of other people of all the items (including the new item). The ratings are ordered, such as "highly recommended", "good",..., "very bad" thus collaborative filtering falls naturally under the domain of ordinal regression.

In this paper we approach the ordinal regression problem within a classification problem framework, and in order to take advantage of the non-metric nature of the loss function we wish to embed the problem within a large margin principle used in Support Vector methods [11]. The Support Vector method (SVM) was introduced originally in the context of 2-class classification. The SVM paradigm has a nice geometric interpretation of discriminating one class from the other by a separating plane with maximum margin. The large-margin principle gives rise to the representation of the decision boundary by a small subset of the training examples called Support Vectors. The SVM approach is advantageous for representing the ordinal regression problem for two reasons. First, the computational machinery for finding the optimal classifier $f(\mathbf{x})$ is based on the non-metric 0-1 loss function. Therefore, by adopting the large-margin principle for ordinal regression we would be implementing an appropriate non-metric loss function as well. Second, the SVM approach is not limited to linear classifiers where through the mechanism of Kernel inner-products one can draw upon a rich family of learning functions applicable to non-linear decision boundaries.

To tackle the problem of using an SVM framework for regression learning, one may take the approach proposed in [7], which is to reduce the total order into a set of preferences over pairs which in effect increases the training set by from l to l^2. Another approach, inherited from the one-versus-many classifiers used for extending binary SVM to multi-class SVM,

is to solve $k-1$ binary classification problems. The disadvantage of this approach is that it ignores the total ordering of the class labels (and also the effective size of the training set is kl whereas we will show that regression learning can be performed with an effective training set of size $2l$). Likewise, the multi-class SVMs proposed in [4, 11, 12, 8] would also ignore the ordering of the class labels and use a training set of size kl.

In this paper we adopt the notion of maintaining a totally ordered set via projections in the sense of projecting the instances $\mathbf{x}_i$ onto the reals $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$ [7, 5] and show how this could be implemented within a large margin principle with an effective training size of $2l$. In fact, we show there is more than one way to implement the large margin principle as there $k-1$ possible margins. Essentially, we show, there are two strategies in general: a "fixed margin" strategy where the large margin principle is applied to the *closest* neighboring pairs of classes, or a multi-margin strategy where the sum of the $k-1$ margins is maximized.

2 The Ordinal Regression Problem

Let $\mathbf{x}_i^j$ be the set of training examples where $j = 1, ..., k$ denotes the class number, and $i = 1, ..., i_j$ is the index within each class. Let $l = \sum_j i_j$ be the total number of training examples. A straight-forward generalization of the 2-class separating hyperplane problem, where a single hyperplane determines the classification rule, is to define $k-1$ separating hyperplanes which would separate the training data into k ordered classes by modeling the ranks as intervals on the real line — an idea whose origins are with the classical cumulative model [9], see also [7, 5]. The geometric interpretation of this approach is to look for $k-1$ parallel hyperplanes represented by vector $\mathbf{w} \in R^n$ (the dimension of the input vectors) and scalars $b_1 \leq ... \leq b_{k-1}$ defining the hyperplanes $(\mathbf{w}, b_1), ..., (\mathbf{w}, b_{k-1})$, such that the data are *separated* by dividing the space into equally ranked regions by the decision rule

$$f(\mathbf{x}) = \min_{r \in \{1,...,k\}} \{r : \mathbf{w} \cdot \mathbf{x} - b_r < 0\}. \tag{1}$$

In other words, all input vectors $\mathbf{x}$ satisfying $b_{r-1} < \mathbf{w} \cdot \mathbf{x} < b_r$ are assigned the rank r (using the convention that $b_k = \infty$). For instance, recently [5] proposed an "on-line" algorithm (with similar principles to the classic "perceptron" used for 2-class separation) for finding the set of parallel hyperplanes which would comply with the separation rule above.

To continue the analogy to 2-class learning, in addition to the separability constraints on the variables $\alpha = \{\mathbf{w}, b_1 \leq ... \leq b_{k-1}\}$ one would like to control the tradeoff between lowering the "empirical risk" $R_{emp}(\alpha)$ (error measure on the training set) and lowering the "confidence interval" $\Phi(\alpha, h)$ controlled by the VC-dimension h of the set of loss functions. The "structural risk minimization" (SRM) principle [11] controls the "actual" risk $R(\alpha)$ (error measured on the "test" data) by keeping $R_{emp}(\alpha)$ fixed (in the ideal separable case it would be zero) while minimizing the confidence interval. The geometric interpretation for 2-class learning is to *maximize* the margin between the boundaries of the two sets [11, 2].

In our setting of ranking learning, there are $k-1$ margins to consider, thus there are two possible approaches to take on the "large margin" principle for ranking learning:

- "fixed margin" strategy: the margin to be maximized is the one defined by the *closest* (neighboring) pair of classes. Formally, let $\mathbf{w}, b_q$ be the hyperplane separating the two pairs of classes which are the closest among all the neighboring pairs of classes. Let $\mathbf{w}, b_q$ be scaled such the distance of the boundary points from the hyperplane is 1, i.e., the margin between the classes $q, q+1$ is $2/|\mathbf{w}|$ (see Fig. 1). Thus, the fixed margin policy for ranking learning is to find the direction

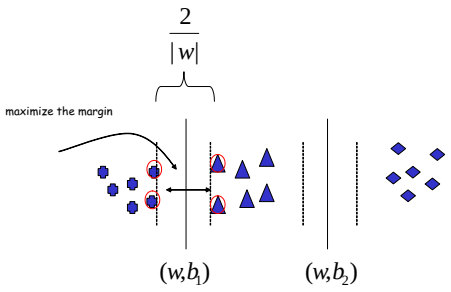


Figure 1: Fixed-margin policy for ranking learning. The margin to be maximized is associated with the two closest neighboring classes. As in conventional SVM, the margin is pre-scaled to be equal to $2/|\mathbf{w}|$ thus maximizing the margin is achieved by minimizing $\mathbf{w} \cdot \mathbf{w}$. The support vectors lie on the boundaries between the two closest classes.

$\mathbf{w}$ and the scalars $b_1, ..., b_{k-1}$ such that $\mathbf{w} \cdot \mathbf{w}$ is minimized (i.e., the margin between classes $q, q+1$ is maximized) subject to the separability constraints (modulo margin errors in the non-separable case).

- "sum of margins" strategy: the sum of all $k-1$ margins are to be maximized. In this case, the margins are not necessarily equal (see Fig. 2). Formally, the ranking rule employs a vector $\mathbf{w}$, $|\mathbf{w}| = 1$, and a set of $2(k-1)$ thresholds $a_1 \leq b_1 \leq a_2 \leq b_2 \leq ... \leq a_{k-1} \leq b_{k-1}$ such that $\mathbf{w} \cdot \mathbf{x}_i^j \leq a_j$ and $\mathbf{w} \cdot \mathbf{x}_i^{j+1} \geq b_j$ for $j = 1, ..., k-1$. In other words, all the examples of class $1 \leq j \leq k$ are "sandwiched" between two parallel hyperplanes $(\mathbf{w}, a_j)$ and $(\mathbf{w}, b_{j-1})$, where $b_0 = -\infty$ and $a_k = \infty$. The $k-1$ margins are therefore $(b_j - a_j)$ and the large margin principle is to maximize $\sum_j (b_j - a_j)$ subject to the separability constraints above.

It is also fairly straightforward to apply the SRM principle and derive the bounds on the actual risk functional by following [11] and making substitutions where necessary. Let the empirical risk be defined as:

$$R_{emp}(\alpha) = \frac{1}{l} \sum_{i=1}^{i_j} \sum_{j=1}^{k} |f(\mathbf{x}_i^j) - y_i^j| = \frac{m}{l},$$

where $f(\mathbf{x}_i^j)$ is the decision rule (1), i_j is the number of training examples of class j and l is the total number of training examples. The empirical risk is the average of the number of "mistakes" where the magnitude of a mistake is related to the total ordering, i.e., the loss function $Q(\mathbf{z}, \alpha) = |f(\mathbf{x}) - y|$, where $\mathbf{z} = (\mathbf{x}, y)$, is an integer between 0 and $k-1$ (unlike the 0/1 loss function associated with classification learning). Since the loss function is totally bounded, the VC-dimension of the class of loss functions $0 \leq Q(\mathbf{z}, \alpha) \leq k-1$ is equal to the VC-dimension h of the class of indicator (0/1) functions

$$I(\mathbf{z}, \alpha, \beta) = \left\{ \begin{array}{ll} 0, & Q(\mathbf{z}, \alpha) - \beta < 0 \\ 1, & Q(\mathbf{z}, \alpha) - \beta \geq 0 \end{array} \right\}$$

where $\beta \in (0, k-1)$. Let $\triangle$-margin k-separating hyperplanes be defined when $|\mathbf{w}| = 1$

and

$$y = \left\{ \begin{array}{ll} 1, & \mathbf{w} \cdot \mathbf{x} \leq a_1 \\ r, & b_{j-1} \leq \mathbf{w} \cdot \mathbf{x} \leq a_j \\ . & . \\ . & . \\ k, & b_{k-1} \leq \mathbf{w} \cdot \mathbf{x} \end{array} \right\}$$

and where $b_j - a_j = \triangle$ (fixed margin policy), and $\triangle$ is the margin between the closest pair of classes. From the arguments above, the VC-dimension of the set of $\triangle$-margin k-separating hyperplanes is bounded by the inequality (following [11]):

$$h \leq \min \left\{ \left[\frac{R^2}{\triangle^2} \right], n \right\} + 1,$$

where R is the radius of the sphere containing all the examples. Thus we arrive to the bound on the probability that a test example will not be separated correctly (following [[11], pp. 77,133]):

With probability $1 - \mu$ one can assert that the probability that a test example will not be separated correctly by the $\triangle$-margin k-separating hyperplanes has the bound

$$P_{error} \leq \frac{m}{l} + \frac{\epsilon(k-1)}{2} \left(1 + \sqrt{1 + \frac{4m}{l\epsilon(k-1)}} \right),$$

where

$$\epsilon = 4 \frac{h(\ln \frac{2l}{h} + 1) - \ln \mu / 4}{l}.$$

Therefore, the larger the "fixed" margin is the better bounds we obtain on the generalization performance of the ranking learning problem with the fixed-margin policy. Likewise, we obtain the same bound under the sum-of-margins principle, where $\triangle$ is defined by the sum of the $k-1$ margins.

In the remainder of this paper we will introduce the algorithmic implications of these two strategies for implementing the large margin principle for ranking learning. The fixed-margin principle will turn out to be a direct generalization of the Support Vector Machine (SVM) algorithm — in the sense that substituting $k = 2$ in our proposed algorithm would produce the dual functional underlying conventional SVM. It is interesting to note that the sum-of-margins principle reduces to νSVM (introduced by [10]) when $k = 2$.

3 Fixed Margin Strategy

Recall that in the fixed margin policy $(\mathbf{w}, b_q)$ is a "canonical" hyperplane normalized such that the margin between the closest classes $q, q+1$ is $2/|\mathbf{w}|$. The index q is of course unknown. The unknown variables $\mathbf{w}, b_1 \leq ... \leq b_{k-1}$ (and the index q) could be solved in a two-stage optimization problem: a Quadratic Linear Programming (QLP) formulation followed by a Linear Programming (LP) formulation.

The (primal) QLP formulation of the ("soft margin") fixed-margin policy for ranking learning takes the form:

$$\min_{w, b_j, \epsilon_i^j, \epsilon_i^{*j+1}} \quad \frac{1}{2} \mathbf{w} \cdot \mathbf{w} + C \sum_i \sum_j \left(\epsilon_i^j + \epsilon_i^{*j+1} \right) \tag{2}$$

$$subject\ to$$

$$\mathbf{w} \cdot \mathbf{x}_i^j - b_j \leq -1 + \epsilon_i^j, \tag{3}$$

$$\mathbf{w} \cdot \mathbf{x}_i^{j+1} - b_j \geq 1 - \epsilon_i^{*j+1}, \tag{4}$$

$$\epsilon_i^j \geq 0, \epsilon_i^{*j} \geq 0 \tag{5}$$

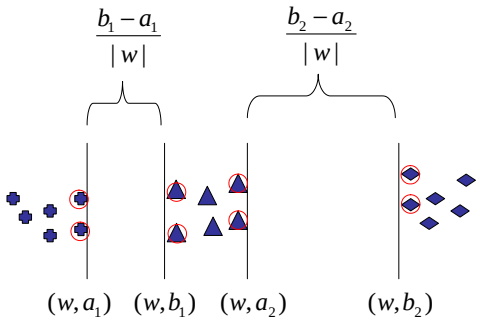


Figure 2: Sum-of-margins policy for ranking learning. The objective is to maximize the sum of $k-1$ margins. Each class is sandwiched between two hyperplanes, the norm of $\mathbf{w}$ is set to unity as a constraint in the optimization problem and as a result the objective is to maximize $\sum_j (b_j - a_j)$. In this case, the support vectors lie on the boundaries among all neighboring classes (unlike the fixed-margin policy). When the number of classes $k = 2$, the dual functional is equivalent to νSVM.

where $j = 1, ..., k-1$ and $i = 1, ..., i_j$, and C is some predefined constant. The scalars ϵ_i^j and ϵ_i^{*j+1} are positive for data points which are inside the margins or placed on the wrong side of the respective hyperplane — if the training data is linearly separable on all the k (ordered) classes then we wouldn't need those ("slack") variables.

The primal functional implements the fixed-margin principle even though we do not know in advance the index q. In the case of "hard" margin (the primal functional above when $\epsilon_i^j, \epsilon_i^{*j}$ are set to zero) the margin is maximized while maintaining separability, thus the margin will be governed by the closest pair of classes because otherwise the separability conditions would cease to hold. The situation may be slightly different and would depend on the choice of C in the soft margin implementation — but qualitatively the same type of behavior holds.

The solution to this optimization problem is given by the saddle point of the Lagrange functional (Lagrangian):

$$\begin{aligned} L(\cdot) &= \frac{1}{2}\mathbf{w}\cdot\mathbf{w} + C\sum_{i,j}\left(\epsilon_i^j + \epsilon_i^{*j+1}\right) \\ &+ \sum_{i,j}\lambda_i^j(\mathbf{w}\cdot\mathbf{x}_i^j - b_j + 1 - \epsilon_i^j) \\ &+ \sum_{i,j}\delta_i^j(1 - \epsilon_i^{*j+1} + b_j - \mathbf{w}\cdot\mathbf{x}_i^{j+1}) \\ &- \sum_{i,j}\zeta_i^j\epsilon_i^j - \sum_{i,j}\zeta_i^{*j+1}\epsilon_i^{*j+1} \end{aligned}$$

where $j = 1, ..., k-1, i = 1, ..., i_j$, and $\zeta_i^j, \zeta_i^{*j+1}, \lambda_i^j, \delta_i^j$ are all *non-negative* Lagrange multipliers. Since the primal problem is convex, there exists a strong duality between the primal and dual optimization functions. By first minimizing the Lagrangian with respect to $\mathbf{w}, b_j, \epsilon_i^j, \epsilon_i^{*j+1}$ we obtain the dual optimization function which then must be maximized with respect to the Lagrange multipliers. From the minimization of the Lagrangian with

respect to $\mathbf{w}$ we obtain:

$$\mathbf{w} = -\sum_{i,j} \lambda_i^j \mathbf{x}_i^j + \sum_{i,j} \delta_i^j \mathbf{x}_i^{j+1} \tag{6}$$

That is, the direction $\mathbf{w}$ of the parallel hyperplanes is described by a linear combination of the support vectors $\mathbf{x}$ associated with the non-vanishing Lagrange multipliers. From the Kuhn-Tucker theorem the support vectors are those vectors for which equality is achieved in the inequalities (3,4). These vectors lie on the two boundaries between the adjacent classes $q, q+1$ (and other adjacent classes which have the same margin). From the minimization of the Lagrangian with respect to b_j we obtain the constraint:

$$\sum_i \lambda_i^j = \sum_i \delta_i^j \quad j = 1, ..., k-1 \tag{7}$$

and the minimization with respect to ϵ_i^j and ϵ_i^{*j+1} yields the constraints:

$$C - \lambda_i^j - \zeta_i^j = 0 \tag{8}$$

$$C - \delta_i^j - \zeta_i^{*j+1} = 0 \tag{9}$$

which in turn gives rise to the constraints $0 \leq \lambda_i^j \leq C$ where $\lambda_i^j = C$ if the corresponding data point is a margin error ($\zeta_i^j = 0$, thus from the Kuhn-Tucker theorem $\epsilon_i^j > 0$), and likewise $0 \leq \delta_i^j \leq C$ where equality $\delta_i^j = C$ holds when the data point is a margin error. Note that a data point can count *twice* as a margin error — once with respect to the class on its "left" and once with respect to the class on its "right".

For the sake of presenting the dual functional in a compact form, we will introduce some new notations. Let X^j be the $n \times i_j$ matrix whose columns are the data points $\mathbf{x}_i^j$, $i = 1, ..., i_j$:

$$X^j = \left[\mathbf{x}_1^j, ..., \mathbf{x}_{i_j}^j\right]_{n \times i_j}.$$

Let $\lambda^j = (\lambda_1^j, ..., \lambda_{i_j}^j)^\top$ be the vector whose components are the Lagrange multipliers λ_i^j corresponding to class j. Likewise, let $\delta^j = (\delta_1^j, ..., \delta_{i_j}^j)^\top$ be the Lagrange multipliers δ_i^j corresponding to class $j+1$. Let $\mu = (\lambda^1, ..., \lambda^{k-1}, \delta^1, ..., \delta^{k-1})^\top$ be the vector holding all the λ_i^j and δ_i^j Lagrange multipliers, and let $\mu^1 = (\mu_1^1, ..., \mu_{k-1}^1)^\top = (\lambda^1, ..., \lambda^{k-1})^\top$ and $\mu^2 = (\mu_1^2, ..., \mu_{k-1}^2)^\top = (\delta^1, ..., \delta^{k-1})^\top$ the first and second halves of μ. Note that $\mu_j^1 = \lambda^j$ is a vector, and likewise so is $\mu_j^2 = \delta^j$. Let $\mathbf{1}$ be the vector of 1's, and finally, let Q be the matrix holding two copies of the training data:

$$Q = \left[-X^1, ..., -X^{k-1}, X^2, ..., X^k\right]_{n \times N}, \tag{10}$$

where $N = 2l - i_1 - i_k$. For example, (6) becomes in the new notations $\mathbf{w} = Q\mu$. By substituting the expression for $\mathbf{w} = Q\mu$ back into the Lagrangian and taking into account the constraints (7,8,9) one obtains the dual functional which should be maximized with respect to the Lagrange multipliers μ_i:

$$\max_{\mu} \quad \sum_{i=1}^{N} \mu_i - \mu^\top (Q^\top Q) \mu \tag{11}$$

$$subject\ to$$

$$0 \leq \mu_i \leq C \quad i = 1, ..., N \tag{12}$$

$$\mathbf{1} \cdot \mu_j^1 = \mathbf{1} \cdot \mu_j^2 \quad j = 1, ..., k-1 \tag{13}$$

There are several points worth noting at this stage. First, when $k = 2$, i.e., we have only two classes thus the ranking learning problem is equivalent to the 2-class classification problem, the dual functional reduces and becomes equivalent to the dual form of the conventional SVM. In that case $(Q^\top Q)_{ij} = y_i y_j \mathbf{x}_i \cdot \mathbf{x_j}$ where $y_i, y_j = \pm 1$ denoting the class membership.

Second, the dual problem is a function of the Lagrange multipliers λ_i^j and δ_i^j alone, that is, all the remaining Lagrange multipliers have dropped out. Therefore the size of the dual QLP problem (the number of unknown variables) is proportional to twice the number of training examples — precisely $N = 2l - i_1 - i_k$ where l is the number of training examples. This favorably compares to the $O(l^2)$ required by the recent SVM approach to ordinal regression introduced in [7] or the kl required by the general multi-class approach to SVM [4]. In fact, the problem size of $N = 2l - i_1 - i_k$ is the smallest possible for the ordinal regression problem since each training example is flanked by a class on each side (except examples of the first and last class), therefore the minimal number of constraints for describing an ordinal regression problem using separating hyperplanes is N.

Third, the criteria function involves only inner-products of the training examples, thereby making it possible to work with kernel-based inner-products. In other words, the entries $Q^\top Q$ are the inner-products of the training examples which can be represented by the kernel inner-product in the input space dimension rather than by inner-products in the feature space dimension. The decision rule, in this case, given a new instance vector $\mathbf{x}$ would be the rank r corresponding to the first smallest threshold b_r for which

$$\sum_{support\ vectors} \delta_i^j K(\mathbf{x}_i^{j+1}, \mathbf{x}) - \sum_{support\ vectors} \lambda_i^j K(\mathbf{x}_i^j, \mathbf{x}) \leq b_r,$$

where $K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{y})$ replaces the inner-products in the higher-dimensional "feature" space $\phi(\mathbf{x})$.

Finally, from the dual form one can solve for the Lagrange multipliers μ_i and in turn obtain $\mathbf{w} = Q\mu$ the direction of the parallel hyperplanes. The scalar b_q (separating the adjacent classes $q, q+1$ which are the closest apart) can be obtained from the support vectors, but the remaining scalars b_j cannot. Therefore an additional stage is required which amounts to a Linear Programming problem on the original primal functional (2) but this time $\mathbf{w}$ is already known (thus making this a linear problem instead of a quadratic one).

4 Sum-of-Margins Strategy

In the fixed margin policy for ranking learning the direction $\mathbf{w}$ of the $k-1$ parallel hyperplanes was determined such as to maximize the margin of the *closest* adjacent pair of classes. In other words, viewed as an extension to conventional SVM, the criteria function remained essentially a 2-class representation (maximizing the margin between two classes) while the linear constraints represented the admissibility constraints necessary for making sure that all classes are properly separable (modulo margin errors).

In this section we propose an alternative large-margin policy which allows for $k-1$ margins where the criteria function maximizes the *sum* of the $k-1$ margins. The challenge in formulating the appropriate optimization functional is that one cannot adopt the "prescaling" of $\mathbf{w}$ approach which is at the center of conventional SVM formulation and of the fixed-margin policy for ranking learning described in the previous section.

The approach we take is to represent the primal functional using $2(k-1)$ parallel hyperplanes instead of $k-1$. Each class would be "sandwiched" between two hyperplanes (except the first and last classes). This may appear superfluous, but in fact all the extra variables (having $2(k-1)$ thresholds instead of $k-1$) drop out in the dual functional —

therefore this approach has no detrimental effect in terms of computational efficiency. Formally, we seek a ranking rule which employs a vector $\mathbf{w}$ and a set of $2(k-1)$ thresholds $a_1 \leq b_1 \leq a_2 \leq b_2 \leq ... \leq a_{k-1} \leq b_{k-1}$ such that $\mathbf{w} \cdot \mathbf{x}_i^j \leq a_j$ and $\mathbf{w} \cdot \mathbf{x}_i^{j+1} \geq b_j$ for $j = 1, ..., k-1$. In other words, all the examples of class $1 \leq j \leq k$ are "sandwiched" between two parallel hyperplanes $(\mathbf{w}, a_j)$ and $(\mathbf{w}, b_{j-1})$, where $b_0 = -\infty$ and $a_k = \infty$.

The margin between two hyperplanes separating class j and $j+1$ is:

$$\frac{b_j - a_j}{\sqrt{(}\mathbf{w} \cdot \mathbf{w})}.$$

Thus, by setting the magnitude of $\mathbf{w}$ to be of unit length (as a constraint in the optimization problem), the margin which we would like to maximize is $\sum_j (b_j - a_j)$ for $j = 1, ..., k-1$ which we can formulate in the following *primal* Quadratic Linear Programming (QLP) problem (see also Fig. 2):

$$\min_{w, a_j, b_j} \quad \sum_{j=1}^{k-1} (a_j - b_j) + C \sum_i \sum_j \left(\epsilon_i^j + \epsilon_i^{*j+1} \right) \tag{14}$$

$$subject\ to$$

$$a_j \leq b_j, \tag{15}$$

$$b_j \leq a_{j+1}, \quad j = 1, ..., k-2 \tag{16}$$

$$\mathbf{w} \cdot \mathbf{x}_i^j \leq a_j + \epsilon_i^j, \tag{17}$$

$$b_j - \epsilon_i^{*j+1} \leq \mathbf{w} \cdot \mathbf{x}_i^{j+1}, \tag{18}$$

$$\mathbf{w} \cdot \mathbf{w} \leq 1, \tag{19}$$

$$\epsilon_i^j \geq 0, \epsilon_i^{*j+1} \geq 0 \tag{20}$$

where $j = 1, ..., k-1$ (unless otherwise specified) and $i = 1, ..., i_j$, and C is some predefined constant (whose physical role would be explained later). There are several points to note about the primal problem. First, the constraints $a_j \leq b_j$ and $b_j \leq a_{j+1}$ are necessary and sufficient to enforce the ordering constraint $a_1 \leq b_1 \leq a_2 \leq b_2 \leq ... \leq a_{k-1} \leq b_{k-1}$. Second, the (non-convex) constraint $\mathbf{w} \cdot \mathbf{w} = 1$ is replaced by the convex constraint $\mathbf{w} \cdot \mathbf{w} \leq 1$ since the optimal solution $\mathbf{w}^*$ would have unit magnitude in order to optimize the objective function. To see why this is so, consider first the case of $k = 2$ where we have a single (hard) margin:

$$\min_{\mathbf{w}, a, b} \quad (a - b)$$

$$subject\ to$$

$$a \leq b$$

$$\mathbf{w} \cdot \mathbf{x}_i \leq a \quad i = 1, ..., i_1$$

$$b \leq \mathbf{w} \cdot \mathbf{x}_i \quad i = i_1 + 1, ..., N$$

$$\mathbf{w} \cdot \mathbf{w} \leq 1$$

We would like to show that for the optimal solution (given that the data is linearly separable) $\mathbf{w}$ must be of unit norm. Let $\mathbf{w}, a, b$ be the optimal solution and $|\mathbf{w}| = \beta \leq 1$. Let $\mathbf{x}^+$ and $\mathbf{x}^-$ be points (support vectors) on the left and right boundary planes, i.e., $\mathbf{w} \cdot \mathbf{x}^- = a$ and $\mathbf{w} \cdot \mathbf{x}^+ = b$. Let $\mathbf{w}^* = (1/\beta)\mathbf{w}$ (thus $|\mathbf{w}^*| = 1$). We have therefore,

$$\begin{aligned} \mathbf{w}^* \cdot \mathbf{x}^- &= \frac{1}{\beta} a \\ \mathbf{w}^* \cdot \mathbf{x}^+ &= \frac{1}{\beta} b \end{aligned}$$

Therefore, the new solution $\mathbf{w}^*, (1/\beta)a, (1/\beta)b$ has a lower energy value (larger margin) of $(1/\beta)(a-b)$ when $\beta < 1$. As a result, $\beta = 1$ since the original solution was assumed to be optimal. This line of reasoning readily extends to multiple margins as the factor $1/\beta$ would apply to all the margins uniformly thus the sum $\sum_j (a_j - b_j)$ would decrease (larger sum of margins) by a factor of $1/\beta$ — thus $\beta = 1$. The introduction of the "soft" margin component (the second term in 14) does not affect this line of reasoning *as long as* the constant C is consistent with the existence of a solution with *negative* energy — otherwise there would be a *duality gap* between the primal and dual functionals. This consistency is related to the number of margin errors which we will discuss in more details later in this section and the following section. We will proceed to derive the dual functional below.

The Lagrangian takes the following form:

$$\begin{aligned} L(\cdot) &= \sum_j (a_j - b_j) + C \sum_{i,j} \left(\epsilon_i^j + \epsilon_i^{*j+1} \right) + \sum_j \psi_j (a_j - b_j) + \sum_{j=1}^{k-2} \eta_j (b_j - a_{j+1}) \\ &+ \sum_{i,j} \lambda_i^j (\mathbf{w} \cdot \mathbf{x}_i^j - a_j - \epsilon_i^j) + \sum_{i,j} \delta_i^j (b_j - \epsilon_i^{*j+1} - \mathbf{w} \cdot \mathbf{x}_i^{j+1}) \\ &+ \alpha (\mathbf{w} \cdot \mathbf{w} - 1) - \sum_{i,j} \zeta_i^j \epsilon_i^j - \sum_{i,j} \zeta_i^{*j+1} \epsilon_i^{*j} \end{aligned}$$

where $j = 1, ..., k-1$ (unless otherwise specified), $i = 1, ..., i_j$, and $\psi_j, \eta_j, \alpha, \zeta_i^j, \zeta_i^{*j}, \lambda_i^j, \delta_i^j$ are all *non-negative* Lagrange multipliers. From the minimization of the Lagrangian with respect to $\mathbf{w}$ we obtain:

$$\mathbf{w} = \frac{1}{2\alpha} Q \mu,$$

where the matrix Q was defined in (10) and the vector μ holds the Lagrange multipliers λ_i^j and δ_i^j as defined in the previous section. From the minimization with respect to b_j for $j = 1, ..., k-2$ we obtain:

$$\frac{\partial L}{\partial b_j} = -1 - \psi_j + \eta_j + \sum_i \delta_i^j = 0.$$

For $j = k-1$ we obtain,

$$\frac{\partial L}{\partial b_{k-1}} = -1 - \psi_{k-1} + \sum_i \delta_i^{k-1} = 0,$$

from which it follows that,

$$\sum_i \delta_i^{k-1} \geq 1. \tag{21}$$

Likewise, the minimization with respect to a_1 provides the constraint,

$$\sum_i \lambda_i^1 = 1 + \psi_1,$$

from which it follows (since $\psi_1 \geq 0$) that

$$\sum_i \lambda_i^1 \geq 1, \tag{22}$$

and with respect to a_j, $j = 2, ..., k-1$, we get the expression,

$$\frac{\partial L}{\partial a_j} = 1 + \psi_j - \eta_{j-1} - \sum_i \lambda_i^j = 0.$$

Summing up the Lagrange multiplier gives rise to another constraint (beyond (21) and (22)), as follows:

$$\sum_i \lambda_i^1 + \sum_{j=2}^{k-1} \sum_i \lambda_i^j = (k-1) + \sum_{j=1}^{k-1} \psi_j + \sum_{j=1}^{k-2} \eta_j,$$

and

$$\sum_{j=1}^{k-2} \sum_i \delta_i^j + \sum_i \delta_i^{k-1} = (k-1) + \sum_{j=1}^{k-1} \psi_j + \sum_{j=1}^{k-2} \eta_j,$$

Therefore, as a result we obtain the constraint:

$$\sum_{i,j} \lambda_i^j = \sum_{i,j} \delta_i^j. \tag{23}$$

Finally, the minimization with respect to ϵ_i^j and ϵ_i^{*j+1} yields the expressions (8) and (9) from which we obtain the constraints

$$0 \leq \lambda_i^j \leq C, \tag{24}$$

$$0 \leq \delta_i^j \leq C, \tag{25}$$

where $\lambda_i^j = C$ and/or $\delta_i^j = C$ if the corresponding data point $\mathbf{x}_i^j$ is a margin error (as mentioned before, a data point can count *twice* as a margin error — once with respect to the class in its "left" and once with respect to the class on its "right").

After substituting the expression for $\mathbf{w}$ back into the Lagrangian and considering the constraints borne out of the partial derivatives with respect to a_j, b_j we obtain the dual functional as a function of $\alpha, \lambda_i^j, \delta_i^j$ only (all the remaining variables drop out):

$$\max_{\alpha,\mu} \left\{ L'(\alpha, \mu) = -\alpha - \frac{1}{4\alpha} \mu^\top (Q^\top Q) \mu \right\},$$

subject to the constraints (21,22,24,25) and $\alpha \geq 0$. Note that $\alpha = 0$ cannot occur if there is an optimal solution with negative energy in the primal functional (otherwise we have a duality-gap, see later) since we have shown above that the $|\mathbf{w}| = 1$ in the optimal solution thus form the Kuhn-Tucker theorem $\alpha \neq 0$. We can eliminate α as follows:

$$\frac{\partial L'}{\partial \alpha} = -1 + \frac{1}{4\alpha^2} C = 0.$$

Substituting the expression for $\alpha = (1/2)\sqrt{C}$ back to $L'()$ provides a new dual functional $L''(\mu) = -\sqrt{\mu^\top Q^\top Q \mu}$ and maximization of $L''(\mu)$ is equivalent to maximization of the expression $-\mu^\top (Q^\top Q \mu$ since $Q^\top Q$ is positive definite. To conclude, the dual functional takes the following form:

$$\max_{\mu} \quad -\mu^\top (Q^\top Q) \mu \tag{26}$$

$$subject\ to$$

$$0 \leq \mu_i \leq C \quad i = 1, ..., N \tag{27}$$

$$\mathbf{1} \cdot \mu_1^1 \geq 1 \tag{28}$$

$$\mathbf{1} \cdot \mu_{k-1}^2 \geq 1 \tag{29}$$

$$\mathbf{1} \cdot \mu^1 = \mathbf{1} \cdot \mu^2 \tag{30}$$

where Q and μ are defined in the previous section. The direction $\mathbf{w}$ is represented by the linear combination of the support vectors:

$$\mathbf{w} = \frac{Q\mu}{|Q\mu|},$$

where, following the Kuhn-Tucker theorem, $\mu_i > 0$ for all vectors on the boundaries between the adjacent pairs of classes and margin errors. In other words, the vectors $\mathbf{x}$ associated with non-vanishing μ_i are those which lie on the hyperplanes, i.e., satisfy $a_j = \mathbf{w} \cdot \mathbf{x}_i^j$ or $b_j = \mathbf{w} \cdot \mathbf{x}_i^{j+1}$ or vectors tagged as margin errors ($\epsilon_i^j > 0$ or $\epsilon_i^{*j+1} > 0$). Therefore, all the thresholds a_j, b_j can be recovered from the support vectors — unlike the fixed-margin scheme which required another LP pass.

The dual functional (26) is similar to the dual functional (11) but with some crucial differences: (i) the quadratic criteria functional is homogeneous, and (ii) constraints (28,29) leads to the constraint $\sum_i \mu_i \geq 2$. From the Kuhn-Tucker theorem, $\psi_j = 0$ when $a_j < b_j$, and $\eta_j = 0$ when $b_j < a_{j+1}$ thus when the data is linearly separable the optimal solution we would have $\sum_i \mu_i = 2(k-1)$. Since a margin error implies that the corresponding Lagrange multiplier $\mu_i = C$, the number of margin errors is bounded since $\sum_i \mu_i$ is bounded. These two differences are also what distinguishes between conventional SVM and νSVM for 2-class learning proposed recently by [10]. Indeed, if we set $k = 2$ in the dual functional (26) we would be able to conclude that the two dual functionals are identical. The primal and dual functionals of νSVM and the sum-of-margins policy for ranking learning for $k = 2$ classes are summarized below:

$\nu\text{SVM} : \text{primal}$	$\nu\text{SVM} : \text{Dual}$
$\min_{\mathbf{w},b,\delta,\epsilon_i} \frac{1}{2}\mathbf{w} \cdot \mathbf{w} - \rho\nu + \frac{1}{N}\sum_{i=1}^N \epsilon_i$	$\max_{\mu_i} -\frac{1}{2}\mu^\top M \mu$
subject to	subject to
$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq \delta - \epsilon_i$	$0 \leq \mu_i \leq \frac{1}{N}$
$\epsilon_i \geq 0$	$\sum_i y_i \mu_i = 0$
$\rho \geq 0$	$\sum_i \mu_i \geq \nu$

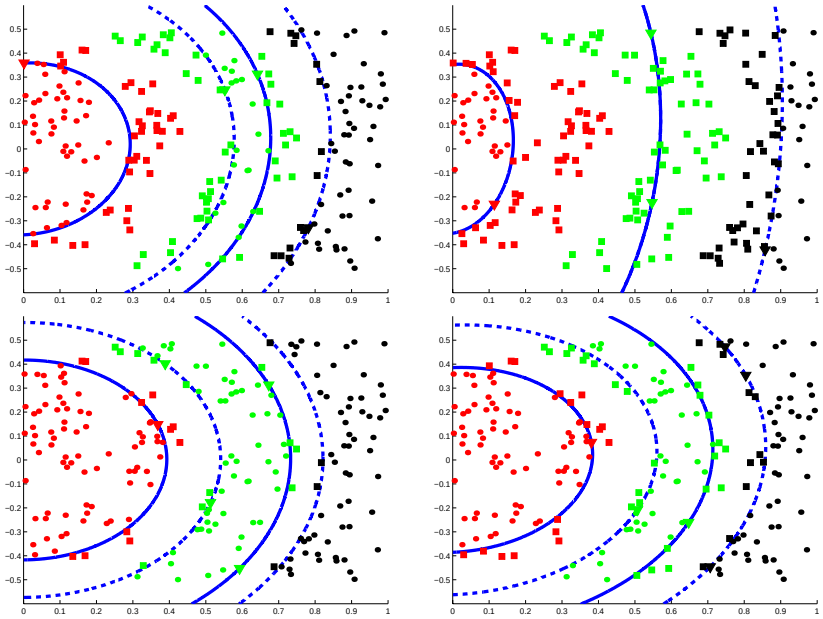
$k = 2 \text{ sum} - \text{of} - \text{margins} : \text{primal}$	$k = 2 \text{ sum} - \text{of} - \text{margins} : \text{dual}$
$\min_{\mathbf{w},a,b,\epsilon_i,\epsilon_i^*} (a-b) + C\left(\sum_{i=1}^{i_1} \epsilon_i + \sum_{i=i_1+1}^{N} \epsilon_i^*\right)$	$\max_{\mu_i} -\mu^\top M \mu$
subject to	subject to
$\mathbf{w} \cdot \mathbf{x}_i \leq a - \epsilon_i \quad i = 1, ..., i_1$	$0 \leq \mu_i \leq C$
$b - \epsilon_i^* \leq \mathbf{w} \cdot \mathbf{x}_i \quad i = i_1 + 1, ..., N$	$\sum_i y_i \mu_i = 0$
$\mathbf{w} \cdot \mathbf{w} \leq 1$	$\sum_i \mu_i \geq 2$
$a \leq b,\ \epsilon_i \geq 0,\ \epsilon_i^* \geq 0$	

where $M = Q^\top Q$ and $M_{ij} = y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$ where $y_i = \pm 1$ depending on the class membership. Although the primal functionals appear different, the dual functionals are similar and in fact can be made equivalent by the following change of variables. Scale the Lagrange multipliers μ_i associated with νSVM such that $\mu_i \to \frac{2\mu_i}{\nu}$. Then, $C = \frac{2}{\nu N}$ and equivalence between the two dual forms is established. Appendix A provides a more detailed analysis of the role of C in the case of $k = 2$.

In the general case of $k > 2$ classes (in the context of ranking learning) the role of the constant C carries the same meaning: $C \leq \frac{2(k-1)}{\#\text{m.e.}}$ where $\#\text{m.e.}$ stand for "total number of margin errors", thus

$$\frac{2(k-1)}{N} \leq C \leq 2(k-1).$$

Recall that in the worst case a data point can count twice for a margin error — being both a margin error in the context of its class and the class on its "left" and in the context of its



fixed-margin algorithm sum-of-margins algorithm

Figure 3: Synthetic data experiments for $k = 3$ classes with 2D data points using second order kernel inner-products. The solid lines correspond to a_1, a_2 and the dashed lines to b_1, b_2 (from left to right). Support vectors are marked as squares in the display. The left column illustrates fixed-margin (dual functional (35)) and the right column illustrates sum-of-margins (dual functional (26)). When the value of C is small (top row) the number of margin errors (and support vectors) is large in order to enable large margins, i.e. $b_j - a_j$ are large. In the case of sum-of-margins (top right display) a small value of C makes $b_1 = a_2$ in order to maximize the margins. When the value of C is large (bottom row) the number of margin errors (and support vectors) is small and as a result the margins are tight.

class and the class on its "right". Therefore the total number of margin errors in the worst case is $N = 2l - i_1 - i_k$ where l is the total number of data points.

The last point of interest to make is that, unlike the fixed margin policy, all the thresholds a_j, b_j are determined from the support vectors — the second Linear Programming optimization stage is not necessary in this case. In other words, there must be support vectors on each hyperplane $(\mathbf{w}, a_j)$ and $(\mathbf{w}, b_j)$, otherwise a better solution exists with larger margins.

To conclude, the multiple margin policy maximizes the sum of the $k-1$ margins allowing the margins to differ in size — thus effectively rewarding larger margins between neighboring classes which are spaced far apart from each other. This is opposite to the fixed margin policy in which the direction of the hyperplanes is dominated by the closest neighboring classes. We saw that the fixed margin policy reduces to conventional SVM when the number of classes $k = 2$ and the multiple margin policy reduces to νSVM. Other differences between the two policies of using the large margin principle is that the multiple margin policy requires a single optimization sweep for recovering both the direction $\mathbf{w}$ and the thresholds a_j, b_j whereas the fixed margin policy requires two sweeps: a QLP for recovering $\mathbf{w}$ and a Linear Programming problem for recovering the $k-1$ thresholds b_j.

5 Fixed Margin Policy Revisited: Generalization of νSVM

We have seen that the sum-of-margins policy reduces to νSVM when the number of classes $k = 2$. However, one cannot make the assertion in the other direction that the dual functional (26) is a generalization of νSVM. In fact, the fixed margin policy applied to νSVM for ranking learning would have the following form:

$$\min_{w,b_j,\epsilon_i^j,\epsilon_i^{*j+1}} \quad \frac{1}{2}\mathbf{w}\cdot\mathbf{w} - \rho\nu + \frac{1}{l}\sum_i\sum_j\left(\epsilon_i^j + \epsilon_i^{*j+1}\right) \tag{31}$$

$$subject\ \ to$$

$$\mathbf{w}\cdot\mathbf{x}_i^j - b_j \leq -\rho + \epsilon_i^j,$$

$$\mathbf{w}\cdot\mathbf{x}_i^{j+1} - b_j \geq \rho - \epsilon_i^{*j+1},$$

$$\rho \geq 0, \epsilon_i^j \geq 0, \epsilon_i^{*j} \geq 0,$$

and the resulting dual functional would have the form:

$$\max_{\mu} \quad -\frac{1}{2}\mu^\top(Q^\top Q)\mu \tag{32}$$

$$subject\ \ to$$

$$0 \leq \mu_i \leq \frac{1}{l} \quad i = 1,...,N \tag{33}$$

$$\sum_i \mu_i \geq \nu$$

$$\mathbf{1}\cdot\mu_j^1 = \mathbf{1}\cdot\mu_j^2, \tag{34}$$

which is not equivalent to the dual functional (26) of the multiple-margin policy (nor to the dual functional (11) of the fixed-margin policy).

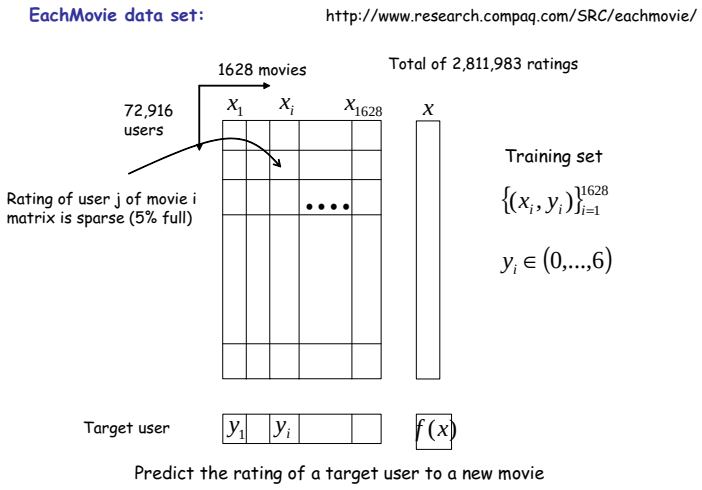


Figure 4: EachMovie dataset used for predicting a person's rating on a new movie given the past ratings on similar movies and the ratings of other people on all the movies. See text for details.

We saw that νSVM could be rederived using the principle of two parallel hyperplanes (primal functional (14) in the case $k = 2$). We will show next that the generalization of

νSVM to ranking learning (dual functional (32) above) can be derived using the $2(k-1)$ parallel hyperplanes approach. The primal functional takes the following form:

$$\begin{aligned}
\min_{w,a_j,b_j} \quad & t + C \sum_i \sum_j \left(\epsilon_i^j + \epsilon_i^{*j+1} \right) \\
& subject\ \ to \\
& a_j - b_j = t, \\
& \mathbf{w} \cdot \mathbf{x}_i^j \leq a_j + \epsilon_i^j, \\
& b_j - \epsilon_i^{*j+1} \leq \mathbf{w} \cdot \mathbf{x}_i^{j+1}, \\
& \mathbf{w} \cdot \mathbf{w} \leq 1, \\
& \epsilon_i^j \geq 0, \epsilon_i^{*j} \geq 0.
\end{aligned}$$

Note that the objective function $\min t$ subject to the constraint $a_j - b_j = t$ captures the fixed margin policy. The resulting dual functional takes the following form:

$$\max_{\mu} \quad -\mu^\top (Q^\top Q) \mu \tag{35}$$

$$subject\ \ to$$

$$0 \leq \mu_i \leq C \quad i = 1, ..., N \tag{36}$$

$$\sum_i \mu_i = 2 \tag{37}$$

$$\mathbf{1} \cdot \mu_j^1 = \mathbf{1} \cdot \mu_j^2 \tag{38}$$

which is equivalent (via change of variables) to the dual functional (32).

Thus to conclude, there are two fixed-margin implementations for ranking learning, one is a direct generalization of conventional SVM (dual functional (11)), and the second is a direct generalization of νSVM (dual functional (35)).

6 Experiments

We have conducted experiments on synthetic data in order to visualize the behavior of the new ranking algorithms, experiments on "collaborative filtering" problems, and experiments on ranking visual data of vehicles.

Fig. 3 shows the performance of the two types of algorithms on synthetic 2D data of a three class ($k = 3$) ordinal regression problem using second order kernel inner-products (thus the separating surfaces are conics). The value of the constant C changes the sensitivity to the number of margin errors and the number of support vectors and as a result the margins themselves (more margin errors allow larger margins). The left column illustrates fixed-margin (dual functional (35)) and the right column illustrates sum-of-margins (dual functional (26)). When the value of C is small (top row) the number of margin errors (and support vectors) is large in order to enable large margins, i.e. $b_j - a_j$ are large. In the case of sum-of-margins (top right display) a small value of C makes $b_1 = a_2$ in order to maximize the margins — as a result the center class completely vanishes (the decision rule will never make a classification in favor of the center class). When the value of C is large (bottom row) the number of margin errors (and support vectors) is small and as a result the margins are tight.

Fig. 4 shows the data structure of "EachMovie" dataset [6] which is used for collaborative filtering tasks. In general, the goal in collaborative filtering is to predict a person's rating on new items such as movies given the person's past ratings on similar items and the ratings of other people of all the items (including the new item). The ratings are ordered, such

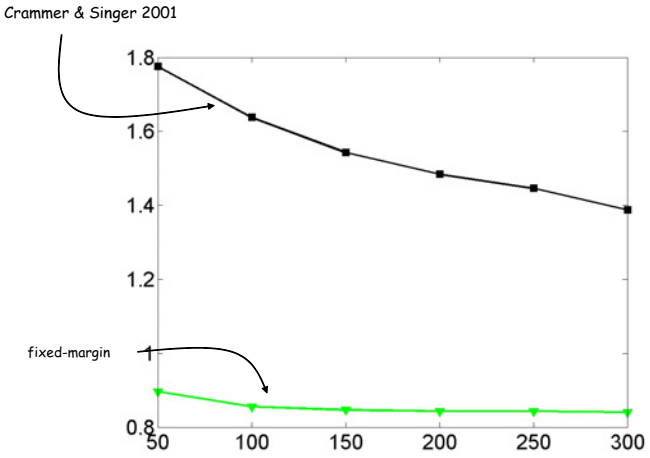


Figure 5: The results of the fixed-margin principle plotted against the results obtained by using the on-line algorithm of [5] which does not use a large-margin principle. The average error between the predicted rating and the correct rating is much lower.

as "highly recommended", "good",..., "very bad" thus collaborative filtering fall naturally under the domain of ordinal regression (rather than general multi-class learning).

The EachMovie dataset contains 1628 movies rated by 72,916 people arranged as a 2D array whose columns represent the movies and the rows represent the users — about 5% of the entries of this array are filled-in with ratings between $0, \ldots, 6$ totaling 2,811,983 ratings. Given a new user, the ratings of the user on the 1628 movies (not all movies would be rated) form the y_i and the i'th column of the array forms the $\mathbf{x}_i$ which together form the training data (for that particular user). Given a new movie represented by the vector $\mathbf{x}$ of ratings of all the other 72,916 users (not all the users rated the new movie), the learning task is to predict the rating $f(\mathbf{x})$ of the new user. Since the array contains empty entries, the ratings were shifted by -3.5 to have the possible ratings $\{-2.5, -1.5, -0.5, 0.5, 1.5, 2.5\}$ which allows to assign the value of zero to the empty entries of the array (movies which were not rated).

For the training phase we chose users which ranked about 450 movies and selected a subset $\{50, 100, \ldots, 300\}$ of those movies for training and tested the prediction on the remaining movies. We compared our results (collected over 100 runs) — the average distance between the correct rating and the predicted rating — to the best "on-line" algorithm of [5] called "PRank" (there is no use of large margin principle). In their work, PRank was compared to other known on-line approaches and was found to be superior, thus we limited our comparison to PRank alone. Attempts to compare our algorithms to other known ranking algorithms which use a large-margin principle ([7], for example) were not successful since those square the training set size which made the experiment with the Eachmovie dataset untractable computationally.

The graph in Fig. 5 shows that the large margin principle (dual functional 35) makes a significant difference on the results compared to PRank. The results we obtained with PRank are consistent with the reported results of [5] (best average error of about 1.25), whereas our fixed-margin algorithm provided an average error of about 0.7).

Figure 6: Classification of vehicle type: Small, Medium and Large. On the left are typical examples of correct classifications and on the right are typical examples of incorrect classifications.

We also applied the ranking learning algorithms to a visual classification problem where we consider images of vehicles taken from the rear where the task is to classify each picture to one of three classes: "small" (passenger cars), "medium" (SUVs, minivans) and "large" (buses, trucks). There is a natural order Small, Medium, Large since making a mistake between Small and Large is worse than confusing Small and Medium, for example. The ordering Small, Medium, Large makes it natural for applying ranking learning (rather than general multi-class). The problem of classifying vehicle types is relevant for applications in the area of "Intelligent Traffic Transportation" (ITS) where on-board sensors such as Visual and Radar would be responsible for a wide variety of "driving assistance" applications including active safety related to airbag deployment in which vehicle classification data is one important piece of information.

The training data included 1500 examples from each class where the input vector was simply the raw pixel values down-sampled to 20x20 pixels per image. The testing phase included 8081 pictures of Small vehicles, 3453 pictures of Medium vehicles and 2395 pictures of Large vehicles. The classification error (counting the number of miss-classifications) with the fixed-margin policy using second-order kernel inner-products was 20% of all test data compared to 25% when performing the classification using three rounds of 2-class conventional SVM (which is the conventional approach for using large margin principle for general multi-class). We also examined the ranking error by averaging the difference between the true rank $\{1, 2, 3\}$ and the predicted rank

$$f(\mathbf{x}) = \sum_{support\ vectors} \delta_i^j K(\mathbf{x}_i^{j+1}, \mathbf{x}) - \sum_{support\ vectors} \lambda_i^j K(\mathbf{x}_i^j, \mathbf{x}),$$

over all test vectors $\mathbf{x}$. The average was 0.216 compared to 1.408 using PRank. Fig. 6 shows a typical collection of correctly classified and incorrectly classified pictures from the test set.

7 Summary

We have introduced a number of algorithms — of linear size with the number of training examples — for implementing a large margin principle for the task of ordinal regression. The first type of algorithms (dual functionals 11, 32, 35) introduces the constraint of a *single* margin determined by the *closest* adjacent pair of classes. That particular margin is maximized while preserving (modulo margin errors) the separability constraints. The support vectors lie on the boundaries of the closest adjacent pair of classes only, thus a complete solution requires first a QLP for finding the hyperplanes direction $\mathbf{w}$ and an LP for finding the thresholds. This type of algorithm comes in two flavors: the first is a direct extension of conventional SVM (dual functional 11) and the second is a direct extension of νSVM (dual functionals 32, 35).

The second type of algorithm (dual functional 26) allows for multiple different margins where the optimization criteria is the sum of the $k-1$ margins. The key observation with this approach is that in order to accommodate different margins the pre-scaling concept (canonical hyperplane) used in conventional SVM (and in fixed-margin algorithms above) is not appropriate and instead one must have $2(k-1)$ parallel hyperplanes where the margins are represented explicitly by the intervals $b_j - a_j$ (rather than by $\mathbf{w} \cdot \mathbf{w}$ as with conventional SVM and fixed margin algorithms). A byproduct of the sum-of-margin approach is that the LP phase is not necessary any more, and that the role of the constant C has a natural interpretation. In fact when $k = 2$ the sum-of-margins algorithm is identical to νSVM. The drawback of this approach (a drawback shared with νSVM) is that unfortunate choices of the constant C might lead to a "duality gap" with the QLP thus rendering the dual functional irrelevant or degenerate.

Experiments performed on visual classification and "collaborative filtering" show that both approaches outperform existing ordinal regression algorithms (on-line approach) applied for ranking and multi-class SVM (applied to the visual classification problem).

Acknowledgements

Thanks for MobilEye Ltd. for the use of the vehicle data set. This work was done while the authors were at the Computer Science department at Stanford University. A.S. especially thanks his host Leo Guibas for making his visit to Stanford possible.

References

[1] J. Anderson. Regression and ordered categorical variables. *Journal of the Royal Statistical Society — Series B*, 46:1–30, 1984.

[2] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifers. In *Proc. of the 5th ACM Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.

[3] W.W. Cohen, R.E. Schapire, and Y. Singer. Learning to order things. *Journal of Artificial Intelligence Research (JAIR)*, 10:243–270, 1999.

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

[5] K. Crammer and Y. Singer. Pranking with ranking. In *Proceedings of the conference on Neural Information Processing Systems (NIPS)*, 2001.

[6] http://www.research.compaq.com/SRC/eachmovie/.

[7] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. Advances in Large Margin Classifiers, 2000. pp. 115–132.

[8] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines. Technical Report 1043, Univ. of Wisconsin, Dept. of Statistics, Sep. 2001.

[9] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 2nd edition edition, 1989.

[10] B. Scholkopf, A. Smola, R.C. Williamson, and P.L. Bartless. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[11] V.N. Vapnik. *The nature of statistical learning*. Springer, 2nd edition edition, 1998.

[12] J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *Proc. of the 7th European Symposium on Artificial Neural Networks*, April 1999.

A A Closer Look at $k = 2$: the Role of the Constant C

In νSVM the constant $0 < \nu < 1$ sets the tradeoff between the fraction of allowable margin errors (at most νN data points could be margin errors) and the minimal number of support vectors (at least νN support vectors). Therefore, the constant C in the sum-of-margins ranking learning specialized to $k = 2$ has a similar interpretation: $2/N < C \leq 2$ is inversely proportional to the allowable number of margin errors $\nu N = 2/C$. Thus, when $C = 2$ only a single margin error is tolerated (otherwise the optimization problem will be in a "weak duality" state — to be discussed later), and when $C = 2/N$ all the points could be margin errors (and in turn all the points are support vectors).

The role of the constant C as a tradeoff between the minimal number of support vectors and the allowable number of margin errors can be directly observed through the primal problem, as follows. Let $\mathbf{w}, a, b, \epsilon_i, \epsilon_i^*$ be a feasible solution for the primal problem. Let ϵ_0 be the smallest of the non-vanishing ϵ_i, i.e., the distance of the nearest margin error associated with the negative training examples; and let ϵ_0^* be the smallest of the non-vanishing ϵ_i^*, i.e., the distance of the nearest margin error associated with the positive training examples. Consider translating the two hyperplanes such that $\hat{a} = a + \epsilon_0$ and $\hat{b} = b - \epsilon_0^*$. The new feasible solution consists of: $\hat{a}, \hat{b}, \mathbf{w}, \hat{\epsilon}_i^*, \hat{\epsilon}_i$ where,

$$\hat{\epsilon}_i = \left\{ \begin{array}{cc} \epsilon_i - \epsilon_0 & \epsilon_i > 0 \\ 0 & \text{otherwise} \end{array} \right\}$$

and $\hat{\epsilon}_i^*$ is defined similarly. The value of the criterion function becomes:

$$\begin{aligned} &\hat{a} - \hat{b} + C \left(\sum_i \hat{\epsilon}_i + \sum_i \hat{\epsilon}_i^* \right) \\ &= a - b + C \left(\sum_i \epsilon_i + \sum_i \epsilon_i^* \right) + \epsilon_0(1 - \alpha C) + \epsilon_0^*(1 - \alpha^* C), \end{aligned}$$

where α is the number of margin errors (where $\epsilon_i > 0$) associated with the negative training examples, and α^* the number of margin errors associated with the positive examples. In order that the original solution would be optimal we must have that $1 - \alpha C + 1 - \alpha^* C \geq 0$ (otherwise we could lower the criteria function and obtain a better solution). Therefore,

$$C \leq \frac{2}{\alpha + \alpha^*}.$$

We see that $C = 2$ when only a single margin error is allowed and $C = 2/N$ when all training data, positive and negative, are allowed to be margin errors. In other words the smaller $C \leq 2$ is, the more margin errors are allowed in the final solution.

To see the connection between C and the necessary number of support vectors consider:

$$\epsilon_0 = \min_i \left\{ a - \mathbf{w} \cdot \mathbf{x}_i \mid a - \mathbf{w} \cdot \mathbf{x}_i > 0 \quad i = 1, ..., i_1 \right\},$$

which is the smallest distance between a negative example (which is not a support vector) and the "left" hyperplane. Likewise,

$$\epsilon_0^* = \min_i \left\{ \mathbf{w} \cdot \mathbf{x}_i - b \mid \mathbf{w} \cdot \mathbf{x}_i - b > 0 \quad i = i_1 + 1, ..., N \right\}$$

which is the smallest distance between a positive example (which is not a support vector) and the "right" hyperplane. Starting with a feasible solution $\mathbf{w}, a, b, \epsilon_i, \epsilon_i^*$ we create a new feasible solution $\mathbf{w}, \hat{a}, \hat{b}, \hat{\epsilon}_i, \hat{\epsilon}_i^*$ as follows. Let $\hat{a} = a - \epsilon_0$, $\hat{b} = b + \epsilon_0^*$,

$$\hat{\epsilon}_i = \left\{ \begin{array}{cc} \epsilon_i + \epsilon_0 & \mu_i > 0 \quad i = 1, ..., i_1 \\ 0 & \text{otherwise} \end{array} \right\},$$

and

$$\hat{\epsilon}_i^* = \left\{ \begin{array}{cc} \epsilon_i^* + \epsilon_0^* & \mu_i > 0 \quad i = i_1 + 1, ..., N \\ 0 & \text{otherwise} \end{array} \right\}.$$

Note that the support vectors are associated with points on the hyperplanes and points labeled as margin errors ($\mu_i > 0$ covers both). Since in the new solution the hyperplanes are shifted, all the old support vectors become margin errors (thus $\hat{\epsilon}_i > 0$). The value of the criteria function becomes:

$$\begin{aligned} & \hat{a} - \hat{b} + C \left(\sum_i \hat{\epsilon}_i + \sum_i \hat{\epsilon}_i^* \right) \\ & = a - b + C \left(\sum_i \epsilon_i + \sum_i \epsilon_i^* \right) + \epsilon_0(\beta C - 1) + \epsilon_0^*(\beta^* C - 1), \end{aligned}$$

where β is the number of negative support vectors and β^* is the number of positive support vectors. In order that the original solution would be optimal we must have that $\beta C - 1 + \beta^* C - 1 \geq 0$ (otherwise we could lower the criteria function and obtain a better solution). Therefore,

$$\beta + \beta^* \geq \frac{2}{C}.$$

We see that when $C = 2$ (a single margin error is allowed), the number of support vectors is at least 1, and when $C = 2/N$ (all instances are allowed to become margin errors) then the number of support vectors is N (i.e., all instances are support vectors). Taken together, C forms a tradeoff: the more margin errors are allowed, the more support vectors one will have in the optimal solution.

Finally, it is worth noting that with a wrong selection of the constant C (when there are more margin errors than the value of C allows for) would make the problem non-feasible as the primal criteria function would be positive (otherwise the constraints would not be satisfied). Since the dual criteria function is non-positive, a "duality gap" would emerge. In other words, even in the presence of "slack variables" (soft margin), there could be an unfortunate situation where the optimization problem is not feasible – and this situation is related to the choice of the constant C.

To conclude, the 2-parallel hyperplanes formulation, or equivalently the νSVM formulation, carries with it a tradeoff. On one hand, the role of the constant C is clear and intuitively simple: there is a direct relationship between the value of C and the fraction of data points which are allowed to be marked as margin errors. On the other hand, unlike conventional SVM which exhibits strong duality under all choices of the regularization constant C, the 2-plane formulation exhibits strong duality only for values of C which are consistent with the worst case scenario of margin errors.