SUPPORT VECTOR MACHINES FOR RANKING LEARNING: THE FULL AND THE TRUNCATED FIXED MARGIN STRATEGIES

ALEXANDER TATARCHUK [1], ALEXEY KURAKIN [2], VADIM MOTTL [1]

[1] Computing Center of the Russian Academy of Sciences, Vavilov St. 40, Moscow, 119991, Russia
[2] Moscow Institute of Physics and Technology, Institutsky Per. 9, Dolgoprudny, Moscow Region, 141700, Russia
E-MAIL: aitech@yandex.ru, alekseyvk@yandex.ru, vmottl@yandex.ru

Abstract:

Two known SVM-based approaches to ranking learning (ordinal regression estimation, supervised pattern recognition with ordered classes) are scrutinized as different generalizations of the classical principle of finding the optimal discriminant hyperplane in a linear space. Easily verifiable natural conditions are found under which the training result obtained by the computationally much more attractive truncated technique is completely equivalent to the hypothetical strict solution. The numerical procedures are essentially simplified for both techniques.

Keywords:

Ordinal regression; ranking learning; large margin learning; support vector machine; computational complexity

1 Introduction

The supervised learning problem or, what is the same, the problem of finding empirical dependences in a set of objects is still one of the glowing problems of the modern informatics. Usually, it is required to build an estimate $\hat{y}(\omega)$ of an unknown function $y(\omega):\Omega\to Y$ that maps a set of real-world objects $\omega\in\Omega$ into a set of values of their hidden characteristic $y\in Y$. The only information on the sought-for function $y(\omega)$ is an accessible subset of objects (training set) within which the values of the goal characteristic are known [1]:

$$y(\tilde{\omega})\in Y,\ \tilde{\omega}\in\Omega^*=\{\omega_1,...,\omega_N\}\subset\Omega. \qquad (1)$$

We use here the notation $\tilde{\omega}$ for the objects of the training set $\tilde{\omega}\in\Omega^*$ to distinguish them from other real-world objects $\omega\in\Omega\setminus\Omega^*$.

It is typical for practice that the finite set of values of the goal characteristic possesses the properties of the ordinal scale $y(\omega)\in Y=\{0,...,m\}$. The learning problem of such a kind is called the *problem of ordinal regression estimation* and bridges regression and classification.

If $m=1$, i.e. the number of ranks equals two, the goal characteristic of objects loses its ordinal nature, and the problem of ordinal regression estimation degenerates into the pattern recognition problem with two classes. This problem has been intensively studied in the literature under the assumption that the real-world objects $\omega\in\Omega$ are represented as points in the respective Euclidean linear space, for instance, by a feature vector $\mathbf{x}(\omega)\in R^n$. The most adopted SVM (Support Vector Machine) training principle [1] is based on the notion of the optimal discriminant hyperplane

$$f(\mathbf{x}(\omega))=\mathbf{a}^T\mathbf{x}(\omega)>h\to\hat{y}(\omega)=1,\ <0\to\hat{y}(\omega)=0, \qquad (2)$$

and consists in maximization of the margin between two classes of the training-set objects via finding the most appropriate direction vector $\mathbf{a}\in R^n$ and threshold $b\in R$.

From the computational point of view, the critical step of the SVM-based two-class training technique is solving the dual quadratic programming problem with the number of variables equal to the size of training set N (1). The computational complexity of this problem is proportional to N^3.

It appears natural to underlay the structure of the commonly adopted linear decision rule (2) by the following legend: It is assumed that the actual hidden properties of the real-world objects are expressed by a real-valued characteristic $f(\omega):\Omega\to R$, but, in contrast to the problem of regression estimation, the training set (1) reveals its values only as results of comparison with some unknown threshold h.

Among the known generalizations of the two-class linear decision rule onto the problem of ordinal regression estimation [2,3,4,5], the majority [2,3,4] boil down to the exploitation of just this legend, however, with usage, in contrast to one threshold in (2), of several increasing thresholds $h^{(1)}<...<h^{(m)}$ which separate the vector feature space R^n into a succession of ordered layers defined by a set of parallel hyperplanes with the same direction vector $\mathbf{a}\in R^n$:

$$\begin{cases} f(\mathbf{x}(\omega))<h^{(1)} & \to\hat{y}(\omega)=0,\\ h^{(1)}\le f(\mathbf{x}(\omega))<h^{(2)} & \to\hat{y}(\omega)=1,\\ \dots & \dots\\ h^{(m)}\le f(\mathbf{x}(\omega)) & \to\hat{y}(\omega)=m, \end{cases} \quad f(\mathbf{x}(\omega))=\mathbf{a}^T\mathbf{x}(\omega). \qquad (3)$$

It should be noted that in the original formulation of the two-class SVM problem the value of the margin at the discriminant hyperplane between two classes of the training-set objects is conventional, because the size of the margin depends on the norm of the direction vector $\|\mathbf{a}\|$ in R^n.

But in the case of several discriminant hyperplanes with a common direction vector (3), the maximum margin may be

different at different hyperplanes. The idea of the so-called fixed-margin approach to ranking learning is maximization of the minimum margin between the adjacent ranks in the training set, and the alternative sum-of-margins approach is based on maximization of the sum of the margins between all the ranks [2]. For brevity sake, we restrict here our consideration only to the former of the two approaches.

Within the bounds of the fixed-margin approach, Shashua and Levin [2] proposed the ranking learning criterion which is aimed at finding the direction vector $\mathbf{w}$ and the thresholds $h^{(i)}$ (3) which correctly separate, in the feature space, the training-set objects $x(\tilde{\omega}) \in \Omega^*$ of immediately adjacent ranks $y(\tilde{\omega}) = i-1$ and $y(\tilde{\omega}) = i$. The computational realization of this strategy involves solving the dual quadratic programming problem with the number of variables which is smaller than the doubled size of training set $2|\Omega^*|$.

This technique seems to be reasonable and computationally attractive, but it may result in the illegal disordered thresholds $h^{(i-1)} \geq h^{(i)}$ for some training sets. We shall call this technique of ranking learning the Truncated Fixed Margin (TFM) strategy.

Another strategy proposed by Chu and Keerthi [3], for finding each threshold value $h^{(i)}$, takes into account all the training-set objects with lower $y(\tilde{\omega}) \leq i-1$ and greater ranks $y(\tilde{\omega}) \geq i$. Thereby, the thresholds are automatically enforced to be correctly ordered. We shall call this technique the Full Fixed Margin (FFM) strategy.

The cost of the complete correctness of the FFM ranking learning strategy is that it leads to the dual quadratic programming problem with $m|\Omega^*|$ variables, i.e. is up to $(m/2)^3$ times more computationally expensive than TFM.

However, the theoretical incorrectness of the TFM strategy exhibits itself rather seldom, only in the case if the ranking model is poorly suitable to the given training set.

In this paper, we study easily verifiable conditions under which the training result obtained by the TFM technique is completely correct and, so, equivalent to the result which would be obtained by the FFM strategy. We propose a new ranking learning strategy which consists in using first the computationally advantageous truncated technique, verifying the correctness of the training result, and then application of the full technique only in the case the conditions are not met.

In contrast to the two-class SVM training problem, the ranking learning problem inevitably requires solving an additional dual problem for finding the optimal values of several thresholds. In this paper, we show the way of essentially reducing the computational complexity of this additional optimization stage and propose new simple algorithms for both TFM and FFM strategies.

For lack of the paper volume, we don't present here the full reasoning of all the mathematical issues and restrict ourselves to brief sketches of the proofs.

2 A common mathematical formulation of the competitive ranking learning strategies

2.1 The fixed-margin approach to ranking learning

In this Section, we mathematically formulate both ranking learning strategies in a common form which lends itself to an easy comparison. We start with partitioning the training-set $\tilde{\omega} \in \Omega^*: \; y_{\tilde{\omega}} = y(\tilde{\omega}) \in Y = \{0,...,m\}$ into the ordered collection of subsets $\Omega^{(0)},...,\Omega^{(m)}$ each consisting of the objects with the same rank $\Omega^{(i)} = \{\tilde{\omega} \in \Omega^*: y_{\tilde{\omega}} = i\}$.

The fixed margin approach [2] consists in finding the common direction vector $\mathbf{a} \in R^n$ and the thresholds $h^{(i)} \in R$ which maximize the minimum margin between the feature vectors $\mathbf{x}_{\tilde{\omega}} = \mathbf{x}(\tilde{\omega})$ of the training-set objects $\tilde{\omega} \in \Omega^*$ of respective ranks $y_{\tilde{\omega}}$ at all the parallel discriminant hyperplanes. In terms of the SVM training principle, this intent is equivalent to minimizing the squared Euclidean norm of the direction vector $\mathbf{a}^T\mathbf{a}$ under the constraint that the minimum margin equals 1. The distinction between the Full Fixed Margin and the Truncated Fixed Margin strategies consists in two different understandings of which ranks of objects should be separated by each of the hyperplanes within the training set.

2.2 The Full Fixed Margin strategy

The Full Fixed Margin strategy [3] is a straightforward generalization of the classical SVM training principle applied to each of m possible rank-related dissections of the training set at once:

$$\begin{cases} \mathbf{a}^T\mathbf{a} + C\sum_{i=1}^{m}\sum_{\tilde{\omega}\in\Omega^*}\delta_{\tilde{\omega}}^{(i)} \to \min(\mathbf{a}, h^{(i)}, \delta_{\tilde{\omega}}^{(i)} \geq 0), \\ \mathbf{a}^T\mathbf{x}_{\tilde{\omega}} - h^{(i)} \leq -1 + \delta_{\tilde{\omega}}^{(i)}, \; \tilde{\omega} \in \Omega^{(0)},..., \tilde{\omega} \in \Omega^{(i-1)}, \\ \mathbf{a}^T\mathbf{x}_{\tilde{\omega}} - h^{(i)} \geq 1 - \delta_{\tilde{\omega}}^{(i)}, \; \tilde{\omega} \in \Omega^{(i)},..., \tilde{\omega} \in \Omega^{(m)}, \; i = 1,...,m. \end{cases} \quad (4)$$

What is important here is that each training-set object participates in m partitions of the training set, so, all the objects occur in each of $m|\Omega^*|$ inequality constraints associated with the thresholds $h^{(i)}$, $i = 1,...,m$.

The solution of the FFM training problem (4) yields, first of all, the common direction vector of the optimal discriminant hyperplanes $\hat{\mathbf{a}}$. Almost the same mathematical reasoning as in the classical two-class SVM [1] leads to the representation of this direction vector as a linear combination of the feature vectors of the training-set objects

$$\hat{\mathbf{a}} = \sum_{\tilde{\omega}\in\Omega^*}\left(\sum_{i=1}^{m} g_{\tilde{\omega}}^{(i)}\hat{\lambda}_{\tilde{\omega}}^{(i)}\right)\mathbf{x}_{\tilde{\omega}}. \quad (5)$$

Here

$$g_{\tilde{\omega}}^{(i)} = 1 \text{ if } y_{\tilde{\omega}} \geq i, \; g_{\tilde{\omega}}^{(i)} = -1 \text{ if } y_{\tilde{\omega}} < i, \quad (6)$$

and nonnegative coefficients $\hat{\lambda}_{\tilde{\omega}}^{(i)}$ are the Lagrange multipliers at the inequality constraints in (4). It is clear that only the objects for which at least one Lagrange multiplier is positive $\hat{\lambda}_{\tilde{\omega}}^{(i)} > 0$ will affect the direction vector. In accordance with the classical terminology of the SVM training principle, these training-set objects are called the support objects.

On the force of (5), there is no need to evaluate the direction vector in an explicit form. The linear decision rule (3) applicable to any new real-world object $\omega \in \Omega \setminus \Omega^*$ will depend only on the inner products of its feature vector with those of the support objects:

$$f(\mathbf{x}_\omega) < \hat{h}^{(1)}, \ \hat{h}^{(i-1)} \le f(\mathbf{x}_\omega) < \hat{h}^{(i)}, \ f(\mathbf{x}_\omega) \ge \hat{h}^{(m)}, \quad f(\mathbf{x}_\omega) = \sum_{\tilde{\omega} \in \Omega^*} \left(\sum_{i=1}^{m} g_{\tilde{\omega}}^{(i)} \hat{\lambda}_{\tilde{\omega}}^{(i)} \right) \mathbf{x}_\omega^T \mathbf{x}_{\tilde{\omega}}. \tag{7}$$

It remains to specify the values of Lagrange multipliers $\hat{\lambda}_{\tilde{\omega}}^{(i)} \ge 0$ and thresholds $\hat{h}^{(i)}$.

The Lagrange multipliers $\hat{\lambda}_{\tilde{\omega}}^{(i)}$ in (5) and (7) are to be found as the solution of the dual problem whose structure is analogous to that in the two-class SVM but more complicated:

$$\begin{cases} W(\lambda_\omega^{(i)}, \omega \in \Omega^*, i = 1, ..., m) = \sum_{\tilde{\omega} \in \Omega^*} \sum_{i=1}^{m} \lambda_{\tilde{\omega}}^{(i)} - \\ (1/2) \sum_{\tilde{\omega} \in \Omega^*} \sum_{\tilde{\vartheta} \in \Omega^*} \sum_{i=1}^{m} \sum_{l=1}^{m} g_{\tilde{\omega}}^{(i)} g_{\tilde{\vartheta}}^{(l)} \mathbf{x}_{\tilde{\omega}}^T \mathbf{x}_{\tilde{\vartheta}} \lambda_{\tilde{\omega}}^{(i)} \lambda_{\tilde{\vartheta}}^{(l)} \to \max, \\ \sum_{\tilde{\omega} \in \Omega^*} g_{\tilde{\omega}}^{(i)} \lambda_{\tilde{\omega}}^{(i)} = 0, \ 0 \le \lambda_{\tilde{\omega}}^{(i)} \le C/2, \ \tilde{\omega} \in \Omega^*, i = 1, ..., m, \end{cases} \tag{8}$$

where indices $g_{\tilde{\omega}}^{(i)}$ are defined by (6). The number of variables in this quadratic programming program equals $m|\Omega^*|$.

As to the thresholds $\hat{h}^{(i)}$, it is assumed in the source paper [3] that these values are to be found as the solution of the initial optimization problem (4). This means that, along with the goal variables $(\hat{\mathbf{a}}, \hat{h}^{(1)}, ..., \hat{h}^{(m)})$, all the auxiliary variables $(\hat{\delta}_{\tilde{\omega}}^{(i)} \ge 0, \tilde{\omega} \in \Omega^*, i = 1, ..., m)$ occurring in (4) have to be evaluated. Each of the latter variables shows how the respective object, if it inhibits the correct separation of the training set by the i th hyperplane, should be shifted in the feature space $\mathbf{x}_{\tilde{\omega}} \in R^n$ along the direction vector $\mathbf{a} \in R^n$ for the training set becomes linearly separable into the subsets with ranks $l < i$ and $l \ge i$. It is easy to show that $\hat{\delta}_{\tilde{\omega}}^{(i)} > 0$ if and only if $\hat{\lambda}_{\tilde{\omega}}^{(i)} = C/2$, This information may be of interest, in some cases, but there is no need to evaluate the shifts $\hat{\delta}_{\tilde{\omega}}^{(i)}$ if only the decision rule parameters $(\hat{\mathbf{a}}, \hat{h}^{(1)}, ..., \hat{h}^{(m)})$ are the goal of data processing.

In this paper, we present explicit formulas that immediately express the thresholds $\hat{h}^{(i)}$ through the Lagrange multipliers $\hat{\lambda}_{\tilde{\omega}}^{(i)}$ and feature vectors of the training-set objects $\mathbf{x}_{\tilde{\omega}}$.

Theorem 1. The optimum threshold values $\hat{h}^{(i)}$ being part of the solution $(\hat{\mathbf{a}}, \hat{h}^{(i)}, \hat{\delta}_{\tilde{\omega}}^{(i)} \ge 0)$ of the FFM training problem (4) are expressed by the following formulas.

1) If $\left\{ \tilde{\omega} \in \Omega^* : 0 < \lambda_{\tilde{\omega}}^{(i)} < C/2 \right\} \ne \varnothing$,

$$\hat{h}^{(i)} = \frac{\sum\limits_{\substack{\tilde{\omega} \in \Omega^*: \\ 0 < \lambda_{\tilde{\omega}}^{(i)} < C/2}} \sum\limits_{\tilde{\vartheta} \in \Omega^*} \left(\sum_{l=1}^{m} g_{\tilde{\vartheta}}^{(l)} \lambda_{\tilde{\omega}}^{(l)} \lambda_{\tilde{\vartheta}}^{(l)} \right) \mathbf{x}_{\tilde{\omega}}^T \mathbf{x}_{\tilde{\vartheta}} + (C/2) \sum\limits_{\substack{\tilde{\omega} \in \Omega^*: \\ \lambda_{\tilde{\omega}}^{(i)} = C/2}} g_{\tilde{\omega}}^{(i)}}{\sum_{\tilde{\omega} \in \Omega^*: 0 < \lambda_{\tilde{\omega}}^{(i)} < C/2} \lambda_{\tilde{\omega}}^{(i)}}. \tag{9}$$

2) If $\left\{ \tilde{\omega} \in \Omega^* : 0 < \lambda_{\tilde{\omega}}^{(i)} < C/2 \right\} = \varnothing$,

$$\hat{h}^{(i)} = \dot{h}^{(i)} \ \text{ in case } \ \sum_{\tilde{\omega} \in \Omega^*: \lambda_{\tilde{\omega}}^{(i)} = C/2} g_{\tilde{\omega}}^{(i)} > 0, \tag{10a}$$

$$\hat{h}^{(i)} = \ddot{h}^{(i)} \ \text{ in case } \ \sum_{\tilde{\omega} \in \Omega^*: \lambda_{\tilde{\omega}}^{(i)} = C/2} g_{\tilde{\omega}}^{(i)} < 0, \tag{10b}$$

$$\hat{h}^{(i)} = (1/2)\left(\dot{h}^{(i)} + \ddot{h}^{(i)} \right) \ \text{ in case } \ \sum_{\tilde{\omega} \in \Omega^*: \lambda_{\tilde{\omega}}^{(i)} = C/2} g_{\tilde{\omega}}^{(i)} = 0, \tag{10c}$$

where

$$\dot{h}^{(i)} = \max \begin{cases} \max\limits_{\tilde{\omega} \in \Omega^*: \hat{\lambda}_{\tilde{\omega}}^{(i)} = 0, \, g_{\tilde{\omega}}^{(i)} = -1} \left[\sum_{\tilde{\vartheta} \in \Omega^*} \left(\sum_{l=1}^{m} g_{\tilde{\vartheta}}^{(l)} \hat{\lambda}_{\tilde{\vartheta}}^{(l)} \right) \mathbf{x}_{\tilde{\vartheta}}^T \mathbf{x}_{\tilde{\omega}} + 1 \right], \\ \max\limits_{\tilde{\omega} \in \Omega^*: \hat{\lambda}_{\tilde{\omega}}^{(i)} = C/2, \, g_{\tilde{\omega}}^{(i)} = 1} \left[\sum_{\tilde{\vartheta} \in \Omega^*} \left(\sum_{l=1}^{m} g_{\tilde{\vartheta}}^{(l)} \hat{\lambda}_{\tilde{\vartheta}}^{(l)} \right) \mathbf{x}_{\tilde{\vartheta}}^T \mathbf{x}_{\tilde{\omega}} - 1 \right], \end{cases}$$

$$\ddot{h}^{(i)} = \min \begin{cases} \min\limits_{\tilde{\omega} \in \Omega^*: \hat{\lambda}_{\tilde{\omega}}^{(i)} = C/2, \, g_{\tilde{\omega}}^{(i)} = -1} \left[\sum_{\tilde{\vartheta} \in \Omega^*} \left(\sum_{l=1}^{m} g_{\tilde{\vartheta}}^{(l)} \hat{\lambda}_{\tilde{\vartheta}}^{(l)} \right) \mathbf{x}_{\tilde{\vartheta}}^T \mathbf{x}_{\tilde{\omega}} + 1 \right], \\ \min\limits_{\tilde{\omega} \in \Omega^*: \hat{\lambda}_{\tilde{\omega}}^{(i)} = 0, \, g_{\tilde{\omega}}^{(i)} = 1} \left[\sum_{\tilde{\vartheta} \in \Omega^*} \left(\sum_{l=1}^{m} g_{\tilde{\vartheta}}^{(l)} \hat{\lambda}_{\tilde{\vartheta}}^{(l)} \right) \mathbf{x}_{\tilde{\vartheta}}^T \mathbf{x}_{\tilde{\omega}} - 1 \right]. \end{cases}$$

The proof consists in proving the expressions for two kinds of thresholds (9) and (10) which differ from each other by the validity of the condition $\left\{ \tilde{\omega} \in \Omega^* : 0 < \lambda_{\tilde{\omega}}^{(i)} < C/2 \right\} \ne \varnothing$ which means that there exist unmoved support objects related to the i th threshold.

In the particular case of only one threshold $m = 1$, namely, in the two-class SVM, the condition $\left\{ \tilde{\omega} \in \Omega^* : 0 < \lambda_{\tilde{\omega}} < C/2 \right\} \ne \varnothing$ is always met, and the sole threshold is uniquely defined. In the nontrivial ranking learning problem with $m > 1$, this condition is satisfied, as a rule, only for one of the thresholds, what is the consequence of the fixed margin approach, and the simple application of the classical reasoning to it determines the unique value (9).

This formula is inapplicable to other thresholds with $\left\{ \tilde{\omega} \in \Omega^* : 0 < \lambda_{\tilde{\omega}}^{(i)} < C/2 \right\} = \varnothing$, because its denominator is not defined. The way of evaluating their values is substituting the expression of the optimum direction vector (5) through the Lagrange multipliers and the feature vectors of the training-set objects into the initial criterion (4), which turns into inde-

pendent linear programming problems with respect to each threshold $h^{(i)}$ and the shifts $\delta_{\tilde{\omega}}^{(i)}$ associated with it.

Depending on the ratio between the numbers of objects with lower $y_{\tilde{\omega}} < i$ and greater ranks $y_{\tilde{\omega}} \geq i$, the respective linear programming problem has a unique decision $\hat{h}^{(i)}$ or an interval of decisions, which can be found analytically in both cases. The formulas (10a) and (10b) are just those analytical solutions for the former kind of thresholds, and (10c) points at the middle of the respective allowable interval for the thresholds of the latter kind. ■

So, to infer the decision rule from the training set in accordance with the FFM strategy, it is enough to solve the dual quadratic programming problem (8) and use the found Lagrange multipliers $\hat{\lambda}_{\tilde{\omega}}^{(i)}$ in (7), (9) and (10).

2.3 The Truncated Fixed Margin strategy

As distinct from the Full Fixed Margin strategy, the Truncated Fixed Margin strategy [2] stipulates adjusting each threshold only to the feature vectors of the objects indexed by two immediately adjacent ranks:

$$\begin{cases} \mathbf{a}^T\mathbf{a} + C\sum_{i=1}^{m}\left(\sum_{\tilde{\omega}\in\Omega^{(i-1)}}\delta'_{\tilde{\omega}} + \sum_{\tilde{\omega}\in\Omega^{(i)}}\delta''_{\tilde{\omega}}\right) \rightarrow \\ \qquad \min(\mathbf{a}, h^{(i)}, \delta'_{\tilde{\omega}} \geq 0, \delta''_{\tilde{\omega}} \geq 0), \\ \mathbf{a}^T\mathbf{x}_{\tilde{\omega}} - h^{(i)} \leq -1 + \delta'_{\tilde{\omega}},\ \tilde{\omega}\in\Omega^{(i-1)}, \\ \mathbf{a}^T\mathbf{x}_{\tilde{\omega}} - h^{(i)} \geq 1 - \delta''_{\tilde{\omega}},\ \tilde{\omega}\in\Omega^{(i)},\ i = 1,\ldots,m. \end{cases} \quad (11)$$

So, each i th inequality ignores the feature vectors of objects that not belong to the two respective ranks $i-1$ and i.. Each object of the minimum or maximum rank, i.e. $\tilde{\omega}\in\Omega^{(0)}$ or $\tilde{\omega}\in\Omega^{(m)}$, participates in only one partition of the training set and is associated with one unknown threshold, respectively, $h^{(1)}$ or $h^{(m)}$, whereas the objects of other ranks $\tilde{\omega}\in\Omega^{(1)},\ldots,\Omega^{(m-1)}$ participate in two partitions at once and each of them involves evaluation of two thresholds. All in all, the TFM training problem contains $|\Omega^*| + \sum_{i=1}^{m-1}|\Omega^{(i)}| < 2|\Omega^*|$ inequality constraints.

Using the standard mathematical reasoning of the classical SVM, it is not difficult to show that the direction vector $\hat{\mathbf{a}}$ found as the solution of the training problem (11) is a linear combination of the feature vectors of the training-set objects

$$\hat{\mathbf{a}} = \sum_{i=1}^{m}\left[\sum_{\tilde{\omega}\in\Omega^{(i)}}\hat{\lambda}''_{\tilde{\omega}}\mathbf{x}_{\tilde{\omega}} - \sum_{\tilde{\omega}\in\Omega^{(i-1)}}\hat{\lambda}'_{\tilde{\omega}}\mathbf{x}_{\tilde{\omega}}\right]. \quad (12)$$

The coefficients of the linear combination $\hat{\lambda}'_{\tilde{\omega}}$ for $\tilde{\omega}\in\Omega^{(0)}$, $\hat{\lambda}''_{\tilde{\omega}}$ for $\tilde{\omega}\in\Omega^{(m)}$, and both $(\hat{\lambda}'_{\tilde{\omega}}, \hat{\lambda}''_{\tilde{\omega}})$ for $\tilde{\omega}\in\Omega^{(i)}$, $i=1,\ldots,m-1$, are the nonegative Lagrange multipliers at the inequality constraints in (11).

In accordance with (12), the linear decision rule (3) obtains the following form, analogous to (7), in which the feature vectors of only the support objects occur:

$$\begin{gathered} f(\mathbf{x}_\omega) < \hat{h}^{(1)},\ \hat{h}^{(i-1)} \leq f(\mathbf{x}_\omega) < \hat{h}^{(i)},\ f(\mathbf{x}_\omega) \geq \hat{h}^{(m)}, \\ f(\mathbf{x}_\omega) = \sum_{i=1}^{m}\left[\sum_{\tilde{\omega}\in\Omega^{(i)}}\hat{\lambda}''_{\tilde{\omega}}\mathbf{x}_\omega^T\mathbf{x}_{\tilde{\omega}} - \sum_{\tilde{\vartheta}\in\Omega^{(i-1)}}\hat{\lambda}'_{\tilde{\vartheta}}\mathbf{x}_\omega^T\mathbf{x}_{\tilde{\vartheta}}\right]. \end{gathered} \quad (13)$$

The Lagrange multipliers $\hat{\lambda}_{\tilde{\omega}}^{(i)}$ in (12) and (13) are the solution of the dual problem:

$$\begin{cases} W(\lambda'_{\tilde{\omega}}, \tilde{\omega}\in\Omega^{(0)},\ldots,\Omega^{(m-1)}, \lambda''_{\tilde{\omega}}, \tilde{\omega}\in\Omega^{(1)},\ldots,\Omega^{(m)}) = \\ \sum_{i=0}^{m-1}\sum_{\tilde{\omega}\in\Omega^{(i)}}\lambda'_{\tilde{\omega}} + \sum_{l=1}^{m}\sum_{\tilde{\omega}\in\Omega^{(l)}}\lambda''_{\tilde{\omega}} - \\ (1/2)\left\{\sum_{i=0}^{m-1}\sum_{l=0}^{m-1}\sum_{\tilde{\omega}\in\Omega^{(i)}}\sum_{\tilde{\vartheta}\in\Omega^{(l)}}\mathbf{x}_{\tilde{\omega}}^T\mathbf{x}_{\tilde{\vartheta}}\lambda'_{\tilde{\omega}}\lambda'_{\tilde{\vartheta}} - \right. \\ 2\sum_{i=0}^{m-1}\sum_{l=1}^{m}\sum_{\tilde{\omega}\in\Omega^{(i)}}\sum_{\tilde{\vartheta}\in\Omega^{(l)}}\mathbf{x}_{\tilde{\omega}}^T\mathbf{x}_{\tilde{\vartheta}}\lambda'_{\tilde{\omega}}\lambda''_{\tilde{\vartheta}} + \\ \left.\sum_{i=1}^{m}\sum_{l=1}^{m}\sum_{\tilde{\omega}\in\Omega^{(i)}}\sum_{\tilde{\vartheta}\in\Omega^{(l)}}\mathbf{x}_{\tilde{\omega}}^T\mathbf{x}_{\tilde{\vartheta}}\lambda''_{\tilde{\omega}}\lambda''_{\tilde{\vartheta}}\right\} \rightarrow \max, \\ \sum_{\tilde{\omega}\in\Omega^{(i)}}\lambda''_{\tilde{\omega}} - \sum_{\tilde{\omega}\in\Omega^{(i-1)}}\lambda'_{\tilde{\omega}} = 0,\ i = 1,\ldots,m, \\ 0 \leq \lambda'_{\tilde{\omega}} \leq C/2,\ \tilde{\omega}\in\Omega^{(i)},\ i = 0,\ldots,m-1, \\ 0 \leq \lambda''_{\tilde{\omega}} \leq C/2,\ \tilde{\omega}\in\Omega^{(i)},\ i = 1,\ldots,m. \end{cases} \quad (14)$$

This quadratic programming problem contains $|\Omega^*| + \sum_{i=1}^{m-1}|\Omega^{(i)}|$ variables, i.e. less than $2|\Omega^*|$.

As we see, the number of variables in the dual formulation of the FFM training strategy exceeds that of the TFM strategy more than $(1/2)m$ times. It is well known that the computational complexity of the quadratic programming problem is proportional to the cube of the number of variables. Thus, the full training strategy is over $(1/8)m^3$ times more computationally expensive than the truncated strategy, and the loss in computational efficiency drastically grows as the number of ranks increases. For instance, $(1/8)m^3 = 42.875$ for $m = 7$.

The following theorem, which is analogous to Theorem 1, specifies explicit formulas for the thresholds $\hat{h}^{(i)}$ as expressed through the Lagrange multipliers $\hat{\lambda}_{\tilde{\omega}}^{(i)}$ and feature vectors of the training-set objects $\mathbf{x}_{\tilde{\omega}}$.

Theorem 2. The optimum threshold values $\hat{h}^{(i)}$ being part of the solution $(\hat{\mathbf{a}}, \hat{h}^{(i)}, \hat{\delta}_{\tilde{\omega}}^{(i)} \geq 0)$ of the TFM training problem (11) are expressed by the following formulas.

1) If $\left\{\tilde{\omega}\in\Omega^*: 0<\lambda'_{\tilde{\omega}}<C/2\right\} \neq \varnothing$ or $\left\{\tilde{\omega}\in\Omega^*: 0<\lambda''_{\tilde{\omega}}<C/2\right\} \neq \varnothing$,

$$\hat{h}^{(i)} = \frac{H'^{(i)} + H''^{(i)} + H'''^{(i)}}{\sum_{\tilde{\omega}\in\Omega^{(i-1)}:\,0<\hat{\lambda}'_{\tilde{\omega}}<C/2}\hat{\lambda}'_{\tilde{\omega}} + \sum_{\tilde{\omega}\in\Omega^{(i)}:\,0<\hat{\lambda}''_{\tilde{\omega}}<C/2}\hat{\lambda}''_{\tilde{\omega}}}, \quad (15)$$

where

$$H'^{(i)} = \sum_{\substack{\tilde{\omega}\in\Omega^{(i)}: \\ 0<\hat{\lambda}''_{\tilde{\omega}}<C/2}}\left[\sum_{i=1}^{m}\left(\sum_{\tilde{\vartheta}\in\Omega^{(i)}}\hat{\lambda}''_{\tilde{\vartheta}}\mathbf{x}_{\tilde{\vartheta}}^T\mathbf{x}_{\tilde{\omega}} - \sum_{\tilde{\vartheta}\in\Omega^{(i-1)}}\hat{\lambda}'_{\tilde{\vartheta}}\mathbf{x}_{\tilde{\vartheta}}^T\mathbf{x}_{\tilde{\omega}}\right)\right]\hat{\lambda}''_{\tilde{\omega}},$$

$$H'^{(i)}=\sum_{\substack{\tilde{\omega}\in\Omega^{(i-1)}:\\ 0<\hat{\lambda}'_{\tilde{\omega}}<C/2}}\left[\sum_{i=1}^{m}\left(\sum_{\tilde{\vartheta}\in\Omega^{(i)}}\hat{\lambda}''_{\tilde{\vartheta}}\mathbf{x}_{\tilde{\vartheta}}^{T}\mathbf{x}_{\tilde{\omega}}-\sum_{\tilde{\vartheta}\in\Omega^{(i-1)}}\hat{\lambda}'_{\tilde{\vartheta}}\mathbf{x}_{\tilde{\vartheta}}^{T}\mathbf{x}_{\tilde{\omega}}\right)\right]\hat{\lambda}'_{\tilde{\omega}},$$

$$H''^{(i)}=(C/2)\left(|\Omega''^{(i)}_{C/2}|-|\Omega'^{(i-1)}_{C/2}|\right),$$

$$\Omega'^{(i-1)}_{C/2}=\left\{\tilde{\omega}\in\Omega^{(i-1)}:\hat{\lambda}'_{\tilde{\omega}}=C/2\right\},\ \Omega''^{(i)}_{C/2}=\left\{\tilde{\omega}\in\Omega^{(i)}:\hat{\lambda}''_{\tilde{\omega}}=C/2\right\}.$$

2) If $\left\{\tilde{\omega}\in\Omega^{*}:0<\lambda'_{\tilde{\omega}}<C/2\right\}=\varnothing$ and $\left\{\tilde{\omega}\in\Omega^{*}:0<\lambda''_{\tilde{\omega}}<C/2\right\}=\varnothing$,

$$\hat{h}^{(i)}=\dot{h}^{(i)}\ \text{ in case }\ H''^{(i)}>0, \tag{16a}$$

$$\hat{h}^{(i)}=\ddot{h}^{(i)}\ \text{ in case }\ H''^{(i)}<0, \tag{16b}$$

$$\hat{h}^{(i)}=(1/2)\left(\dot{h}^{(i)}+\ddot{h}^{(i)}\right)\ \text{ in case }\ H''^{(i)}=0, \tag{16c}$$

where

$$\dot{h}^{(i)}=\max\left\{\max_{\tilde{\omega}\in\Omega^{(i-1)}:\lambda'_{\tilde{\omega}}=0}\left(B^{(i)}_{\tilde{\omega}}+1\right),\max_{\tilde{\omega}\in\Omega^{(i)}:\lambda''_{\tilde{\omega}}=C/2}\left(B^{(i)}_{\tilde{\omega}}-1\right)\right\}.$$

$$\ddot{h}^{(i)}=\min\left\{\min_{\tilde{\omega}\in\Omega^{(i-1)}:\lambda'_{\tilde{\omega}}=C/2}\left(B^{(i)}_{\tilde{\omega}}+1\right),\min_{\tilde{\omega}\in\Omega^{(i)}:\lambda''_{\tilde{\omega}}=0}\left(B^{(i)}_{\tilde{\omega}}-1\right)\right\},$$

$$B^{(i)}_{\tilde{\omega}}=\sum\nolimits_{i=1}^{m}\left(\sum\nolimits_{\tilde{\vartheta}\in\Omega^{(i)}}\hat{\lambda}''_{\tilde{\vartheta}}\mathbf{x}_{\tilde{\vartheta}}^{T}\mathbf{x}_{\tilde{\omega}}-\sum\nolimits_{\tilde{\vartheta}\in\Omega^{(i-1)}}\hat{\lambda}'_{\tilde{\vartheta}}\mathbf{x}_{\tilde{\vartheta}}^{T}\mathbf{x}_{\tilde{\omega}}\right).$$

The proof of this theorem is based on the same reasoning as that of Theorem 1. Just as in the FFM problem, the fixed margin approach results in that the set of unmoved support objects related to some threshold $\hat{h}^{(i)}$ is nonempty, as a rule, only for one i. However, in the TFM problem, each threshold is to be found as the value discriminating not all the training-set objects $\tilde{\omega}\in\Omega^{*}$ but only those of them which have the immediately adjacent ranks $\tilde{\omega}\in\Omega^{(i-1)}$ and $\tilde{\omega}\in\Omega^{(i)}$. As a consequence, the fact that the set of unmoved support objects related to $\hat{h}^{(i)}$ is nonempty is mathematically expressed as $\left[\left\{\tilde{\omega}\in\Omega^{*}:0<\lambda'_{\tilde{\omega}}<C/2\right\}\neq\varnothing\ \text{ or }\ \left\{\tilde{\omega}\in\Omega^{*}:0<\lambda''_{\tilde{\omega}}<C/2\right\}\neq\varnothing\right]$ instead of a more simple formula $\left\{\tilde{\omega}\in\Omega^{*}:0<\lambda^{(i)}_{\tilde{\omega}}<C/2\right\}\neq\varnothing$ in Theorem 1.

This distinction displays itself in the seemingly different appearance of formulas (15)-(16). In all other respects the essence of the proof remains the same. The formula (15) is result of application of the usual SVM logic to the thresholds with nonempty sets of unmoved support objects (as a rule, only one threshold). The other thresholds are the analytical solutions of the respective linear programming problems having a single minimum point (16a)-(16b) or are the middle points of the respective intervals of equivalent minima (16c). ■

There is no difference between the full and the truncated strategies in the classical SVM – it is easy to see that all the expressions in Sections 2.2 and 2.3 coincide if $m=1$. But these strategies may display essentially different behaviour in the nontrivial case $m>1$.

3 The correctness conditions for the training result obtained via the truncated technique

The very idea of ranking learning is based on the assumption that the sought-for linear function $\mathbf{a}^{T}\mathbf{x}_{\omega}$ is to be compared with increasing thresholds (3). It is shown in [3] that the fulfillment of the assumption $h^{(1)}<\ldots<h^{(m)}$ is guaranteed in the full formulation of the training problem (4), but it is not controlled in the truncated formulation (11) and may be broken. An attempt to include the constraints $h^{(1)}<\ldots<h^{(m)}$ into the truncated formulation [3] results in the loss of its computational advantage over the full one.

More over, the TFM strategy takes into account only a part of the relationships between the properties of the training-set objects, namely, those which are reflected by the truncated list of inequality constraints in (11), whereas the FFM strategy (4) stipulates that attention is paid to all the relationships. As a consequence, the parameters of the decision rule $(\hat{\mathbf{a}},\hat{h}^{(1)},\ldots,\hat{h}^{(m)})$ (3) estimated by the TFM strategy, even if the condition $h^{(1)}<\ldots<h^{(m)}$ is met, may differ from those $(\overline{\mathbf{a}},\overline{h}^{(1)},\ldots,\overline{h}^{(m)})$ which would be obtained by the FFM strategy. We shall say that the TFM-based training result is correct if it coincides with the FFM-based result:

$$(\hat{\mathbf{a}},\hat{h}^{(1)},\ldots,\hat{h}^{(m)})=(\overline{\mathbf{a}},\overline{h}^{(1)},\ldots,\overline{h}^{(m)}). \tag{17}$$

However, the TFM strategy may yield incorrect result only if the data set is drastically inconsistent with the ranking model forced upon it. Such situations are typical in the practice of data analysis when several models are tried before the most appropriate one is chosen, but they occur not unduly frequently. It is hardly reasonable always to use the computationally expensive FFM strategy only on the basis of quite vague apprehension that the TFM strategy may fail.

This reasoning leads to the idea of finding a way of verifying the correctness of the training result obtained via the computationally advantageous TFM strategy (17) without immediate comparing it with the FFM-based result.

Before we formulate the respective theorem, it should be remarked that, after the direction vector and thresholds $(\hat{\mathbf{a}},\hat{h}^{(1)},\ldots,\hat{h}^{(m)})$ are found as the essential part of the solution of the TFM training problem (11) in accordance with (12), (14), and Theorem 2, the shifts of the training-set objects in the feature space $\delta'_{\tilde{\omega}}\geq 0$ and $\delta''_{\tilde{\omega}}\geq 0$ can be easily evaluated by the formulas which evidently follow from the constraints in (11):

$$\hat{\delta}'_{\tilde{\omega}}=\max\left\{0,\hat{\mathbf{a}}^{T}\mathbf{x}_{\tilde{\omega}}-\hat{h}^{(1)}+1\right\},\ \tilde{\omega}\in\Omega^{(0)},$$

$$\begin{cases}\hat{\delta}'_{\tilde{\omega}}=\max\left\{0,\hat{\mathbf{a}}^{T}\mathbf{x}_{\tilde{\omega}}-\hat{h}^{(i+1)}+1\right\},\\ \hat{\delta}''_{\tilde{\omega}}=\max\left\{0,\hat{h}^{(i)}-\hat{\mathbf{a}}^{T}\mathbf{x}_{\tilde{\omega}}+1\right\},\end{cases}\quad \tilde{\omega}\in\Omega^{(i)},\ i=1,\ldots,m-1,$$

$$\hat{\delta}''_{\tilde{\omega}}=\max\left\{0,\hat{h}^{(m)}-\hat{\mathbf{a}}^{T}\mathbf{x}_{\tilde{\omega}}+1\right\},\ \tilde{\omega}\in\Omega^{(m)}.$$

The inner product $\hat{\mathbf{a}}^T\mathbf{x}_{\tilde{\omega}}$ in these formulas is completely represented through the feature vectors of the training set objects and the Lagrange multipliers in accordance with (12):

$$\hat{\mathbf{a}}^T\mathbf{x}_{\tilde{\omega}} = \sum_{i=1}^{m}\left[\sum_{\tilde{\vartheta}\in\Omega^{(i)}}\hat{\lambda}''_{\tilde{\vartheta}}\mathbf{x}_{\tilde{\vartheta}}^T\mathbf{x}_{\tilde{\omega}} - \sum_{\tilde{\vartheta}\in\Omega^{(i-1)}}\hat{\lambda}'_{\tilde{\vartheta}}\mathbf{x}_{\tilde{\vartheta}}^T\mathbf{x}_{\tilde{\omega}}\right].$$

Theorem 3. Let $(\hat{\mathbf{a}}, \hat{h}^{(1)},..., \hat{h}^{(m)}; \hat{\delta}'_{\tilde{\omega}}, \tilde{\omega}\in\Omega^{(0)}; \hat{\delta}'_{\tilde{\omega}}, \hat{\delta}''_{\tilde{\omega}}, \tilde{\omega}\in\Omega^{(i)}, i=1,...,m-1; \hat{\delta}''_{\tilde{\omega}}, \tilde{\omega}\in\Omega^{(m)})$ be the solution of the TFM training problem as the minimum point of the criterion (11). For the parameters of the decision rule $(\hat{\mathbf{a}}, \hat{h}^{(1)},..., \hat{h}^{(m)})$ would be also the solution of the FFM problem (4), it is necessary and sufficient that the following inequalities are met:

$$\begin{aligned}&\hat{\delta}'_{\tilde{\omega}} \le \hat{h}^{(i+2)} - \hat{h}^{(i+1)},\ \tilde{\omega}\in\Omega^{(i)},\ i=0,...,m-2,\\ &\hat{\delta}''_{\tilde{\omega}} \le \hat{h}^{(i)} - \hat{h}^{(i-1)},\ \tilde{\omega}\in\Omega^{(i)},\ i=2,...,m.\end{aligned} \tag{18}$$

The proof is based on rewriting the TFM training criterion (11) in an equivalent form which differs from the FFM training criterion (4) only by the list of inequality constraints:

$$\begin{cases}\mathbf{a}^T\mathbf{a} + C\sum_{i=1}^{m}\sum_{\tilde{\omega}\in\Omega^*}\delta_{\tilde{\omega}}^{(i)} \to \min(\mathbf{a}, h^{(i)}, \delta_{\tilde{\omega}}^{(i)} \ge 0),\\ \mathbf{a}^T\mathbf{x}_{\tilde{\omega}} - h^{(i)} \le -1 + \delta_{\tilde{\omega}}^{(i)},\ \tilde{\omega}\in\Omega^{(i-1)},\\ \mathbf{a}^T\mathbf{x}_{\tilde{\omega}} - h^{(i)} \ge 1 - \delta_{\tilde{\omega}}^{(i)},\ \tilde{\omega}\in\Omega^{(i)},\ i=1,...,m.\end{cases} \tag{19}$$

The "idle" variables in his formulation $\left(\delta_{\tilde{\omega}}^{(i)}, \tilde{\omega}\in\Omega^{(l)}, l<i-1\right)$ and $\left(\delta_{\tilde{\omega}}^{(i)}, \tilde{\omega}\in\Omega^{(k)}, k>i\right)$ don't occur in the constraints, therefore, they automatically take zero values at the minimum point of the criterion. Only one shift is actual for each training-set object of the minimum $\tilde{\omega}\in\Omega^{(0)}$ and maximum rank $\tilde{\omega}\in\Omega^{(m)}$, respectively, in accordance with denotations $\delta_{\tilde{\omega}}^{(1)} = \delta''_{\tilde{\omega}}$ and $\delta_{\tilde{\omega}}^{(m)} = \delta'_{\tilde{\omega}}$ accepted in (11), whereas the objects of the intermediate ranks $1 \le i < m$ are associated each with two active shifts $\delta_{\tilde{\omega}}^{(i)} = \delta'_{\tilde{\omega}}$ and $\delta_{\tilde{\omega}}^{(i+1)} = \delta''_{\tilde{\omega}}$. So, it is enough to consider the TFM problem in the form (19) instead of (11).

The objective function in both FFM (4) and TFM (19) formulations of the ranking learning problem remains the same

$$J_{FFM}(\mathbf{a}, h^{(i)}, \delta_{\tilde{\omega}}^{(i)} \ge 0) = J_{TFM}(\mathbf{a}, h^{(i)}, \delta_{\tilde{\omega}}^{(i)} \ge 0) = \mathbf{a}^T\mathbf{a} + C\sum_{i=1}^{m}\sum_{\tilde{\omega}\in\Omega^*}\delta_{\tilde{\omega}}^{(i)},$$

the only distinction is that the set of inequality constraints in the TFM problem (19) is a subset of those in (4). If we have found the solution $(\hat{\mathbf{a}}, \hat{h}^{(i)}, \hat{\delta}_{\tilde{\omega}}^{(i)} \ge 0)$ of the TFM problem, and it meets all the constraints of (4), then this solution is also the solution of the FFM problem and, so, is correct by definition.

The inequalities (18) are just the conditions, expressed in the initial denotations, which guarantee that all the constraints of the FFM formulation are met. ■

4 Conclusions

On the one hand, ranking learning is a generalization of pattern recognition under the additional assumption that the classes of objects to be distinguished are linearly ordered. On the other hand, if we view the ordinal variable as external manifestation of a hidden real variable, which can be observed only by comparison with a system of unknown thresholds, it will be suitable to consider the ranking model as a generalization of numerical regression.

The principle of two-class pattern recognition via finding the optimal discriminant hyperplane in a linear space, widely known under the name of SVM, is a natural prototype of both interpretations of the ranking learning problem in the trivial case of two ranks. However, there are many aspects with respect to which the classical SVM may be generalized onto a greater number of ranks and many ways of such generalization. We have scrutinized here only two ways, both within the bounds of the so-called fixed margin approach, the Full Fixed Margin and the Truncated Fixed Margin strategies. These strategies yield the same structure of the decision rule, but only the much more computationally expensive FFM strategy is completely correct from the viewpoint of its applicability to arbitrary data sets. At the same time, the incorrectness of the TFM strategy displays itself in relatively rare cases of "awkward" data sets.

In this paper, we essentially simplified both training procedures and have proposed an easily verifiable condition under which the result of application of the computationally advantage TFM strategy to the given data set is completely correct.

Acknowledgements

This work is supported by INTAS Grant 04-77-7347 and Grants of the Russian Foundation for Basic Research 05-01-00679, 06-01-08042, 06-07-89249.

References

[1] Vapnik V. Statistical Learning Theory. John-Wiley & Sons, Inc. 1998.

[2] Shashua A., Levin A. Ranking with large margin principle: two approaches. Advances in Neural Information Processing Systems, 15, 2003, pp. 937-944.

[3] Chu W., Keerthi S.S. New approaches to support vector ordinal regression. Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005.

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

[5] Waegeman W., Boullart L. An ensemble of weighted support vector machines for ordinal regression. Transactions on Engineering, Computing and Technology, Vol. 12, 2006.