INSTITUTE OF POLICY AND PLANNING SCIENCES

Discussion Paper Series

No. 1096

A Linear Classification Model Based on Conditional Geometric Score

by

Jun-ya Gotoh and Akiko Takeda

September 2004

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

A Linear Classification Model Based on Conditional Geometric Score

Jun-ya Gotoh* Akiko Takeda†

August 30, 2004
Revised: September 9, 2004

Abstract

We propose a two-class linear classification model by taking into account the Euclidean distance from each data point to the discriminant hyperplane and introducing a risk measure which is known as the conditional value-at-risk in financial risk management. It is formulated as a nonconvex programming problem and we present a solution method for obtaining either a globally or a locally optimal solution by examining the special structure of the problem. Also, this model is proved to be equivalent to the ν-support vector classification under some parameter setting, and numerical experiments show that the proposed model has better predictive accuracy in general.

The 2000 MSC: 68T10, 90C26, 90C90
Keywords: classification model, discriminant hyperplane, conditional value-at-risk, nonconvex programming

1 Introduction

Mathematical optimization approaches to the classification problem have attracted much attention of many researchers since the 1960s (e.g., [7]). Among such approaches are the support vector classification [15], the robust linear programming method [2], integer programming methods (e.g., [6]), the other kind of linear programming methods (e.g., [4, 5]). For the two-class classification problem in $\mathbb{R}^n$, these models seek to find a decision function $h : \mathbb{R}^n \to \{\pm 1\}$ using a set of training data $\{\boldsymbol{x}^i\} \subset \mathbb{R}^n$ which are labeled with binary values $\{y^i\} \subset \{\pm 1\}$ for $i \in I := \{1, ..., \ell\}$:

$$(\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^\ell, y^\ell) \in \mathbb{R}^n \times \{\pm 1\},$$

so that h will predict as accurately as possible the labels of new data points generated from the same probability distribution with the training data. In particular, the linear classification problem concerned in this paper is reduced to constructing a discriminant hyperplane:

$$H(\boldsymbol{w}, b) := \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0 \,\}, \quad \boldsymbol{w} \neq \boldsymbol{0}, \quad \boldsymbol{w} \in \mathbb{R}^n, \quad b \in \mathbb{R}, \tag{1}$$

*Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennoudai, Tsukuba, Ibaraki, 305-8573 Japan. e-mail: jgoto@sk.tsukuba.ac.jp

†Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, 2-12-1 Oh-Okayama, Meguro-ku, Tokyo, 152-8552 Japan. e-mail: takeda@is.titech.ac.jp

which corresponds to the decision function $h(\boldsymbol{x}) = \mathrm{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)$ where $\mathrm{sign}(\xi)$ is 1 if $\xi \geq 0$, -1, otherwise.

To this end, if given set of training data is linearly separable, i.e., there exists $(\boldsymbol{w}, b)$ such that $\boldsymbol{w} \neq \boldsymbol{0}$ and $y^i(\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle + b) > 0$ hold for $i \in I$, the hard margin support vector classification (HSVC) [15] provides a most reasonable hyperplane as an optimal solution of

$$\left| \underset{\boldsymbol{w} \neq \boldsymbol{0},\, b}{\text{maximize}} \ \min_{i \in I} \ g(\boldsymbol{w}, b; \boldsymbol{x}^i, y^i), \right. \tag{2}$$

where

$$g(\boldsymbol{w}, b; \boldsymbol{x}^i, y^i) := \frac{y^i(\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle + b)}{\|\boldsymbol{w}\|}, \qquad \text{for } i \in I. \tag{3}$$

In the paper, we call $g(\boldsymbol{w}, b; \boldsymbol{x}^i, y^i)$ the *geometric score* of data $(\boldsymbol{x}^i, y^i)$ with respect to $(\boldsymbol{w}, b)$. Formulation (2) implies that only a portion of training data which have the lowest geometric score contribute to determine the hyperplane.

On the other hand, for linearly inseparable data set, there is no hyperplane which correctly separates the given set of data points, and the criterion based on the lowest geometric score is less persuasive because the other misclassified data points are ignored. For overcoming this drawback, the other models such as the soft margin SVC and the robust linear programming method introduce some other misclassification risk measures to be minimized into their formulation.

In this paper, we propose a classification model which minimizes another misclassification risk. The risk measure applied in our model is based on a conditional expectation of the empirical distribution of the geometric score, which is known as the *conditional value-at-risk* (CVaR) in the financial risk management [12]. The CVaR measure is now highlighted as a tool of perceiving and controlling the financial downside risk since it has preferable properties such as coherency [1] and consistency with stochastic dominance [10, 9] and, in addition, many portfolio problems based on the CVaR minimization can be formulated as a convex programming problem [11].

Contrary to those portfolio problems, our model is basically formulated as a nonconvex quadratic optimization problem due to the nonconvexity of the geometric score g. Under some parameter setting, however, it can be transformed into an equivalent convex quadratic optimization problem. The properties of our model are

i) under some parameter setting, our model is equivalent to the ν-support vector classification (ν-SVC) model [13], and also equivalent to the HSVM model (2),

ii) the difficulty of the proposed problem depends on given parameters and data set $\{(\boldsymbol{x}^i, y^i) : i \in I\}$.

From i), the discussion in this paper provides another interpretation of the ν-SVC, and this similarity opens a gate for the use of the kernel functions in our linear model.

The structure of this paper is as follows. In the next section, we first define the conditional expectation of the geometric score as a risk measure by following Rockafellar & Uryasev [12], and

describe the formulation of the risk minimization. In section 3, we discuss the nonconvex structure of the formulation, and propose an algorithm for obtaining either an optimal solution when the problem is convex, or a locally optimal solution when the problem is essentially nonconvex. In section 4, we discuss a relationship between our model and other classification approaches: the HSVC and the ν-SVC, and show that under some parameter setting, our model is equivalent to them. Section 5 is devoted to some numerical experiments, showing that the nonconvexity in our model, which is eliminated in the ν-SVC, plays an important role in determining the hyperplane. Finally, we conclude the paper by adding some remarks and possible extension.

2 Conditional Geometric Score Optimization

We first introduce the several results of Rockafellar & Uryasev [12] and define the conditional expectation of the geometric score as misclassification risk measure in section 2.1. Then, in section 2.2, we describe the linear classification problem which minimizes the misclassification risk measure.

2.1 A Misclassification Risk Measure

We consider the two-class linear classification problem for a finite number of training data which are assumed to be generated from an unknown probability distribution.

We denote the given labeled data set by $\{(\boldsymbol{x}^i, y^i) : i \in I\} \subset \mathbb{R}^n \times \{\pm 1\}$ where $I := \{1, \ldots, \ell\}$ is the index set of the training data, and define two index sets I^+ and I^- by $I^+ := \{i \in I \,|\, y^i = +1\}$ and $I^- := \{i \in I \,|\, y^i = -1\}$, respectively. In addition, we suppose that the empirical probability assigned on the given data is given by $\mathrm{P}\{(\boldsymbol{x}, y) = (\boldsymbol{x}^i, y^i)\} = p_i > 0$ for $i \in I$, where $\sum_{i \in I} p_i = 1$. One simple choice for p_i is

$$p_i = \frac{1}{|I|} = \frac{1}{\ell} \text{ for } i \in I. \tag{4}$$

Another possibility is

$$p_i = \begin{cases} \frac{1}{2|I^+|} & \text{for } i \in I^+ \\ \frac{1}{2|I^-|} & \text{for } i \in I^- \end{cases} \tag{5}$$

which is often used to balance the two classes in the other classification models [2, 13].

For each data point $(\boldsymbol{x}^i, y^i)$, the geometric score is defined as in (3) with some given $(\boldsymbol{w}, b)$ and we sometimes use the notation $g(\boldsymbol{w}, b)$ in place of $g(\boldsymbol{w}, b; \boldsymbol{x}, y)$ for simplicity. $g(\boldsymbol{w}, b)$ is the Euclidean distance of a point $\boldsymbol{x} \in \mathbb{R}^n$ to the hyperplane $H(\boldsymbol{w}, b)$ if positive, and we can say that a data point with lower geometric score has higher risk to be misclassified. Denoting the minus value of the score g by f, i.e.,

$$f(\boldsymbol{w}, b) := -g(\boldsymbol{w}, b),$$

we interpret f as the magnitude of misclassification risk.

Let us define the distribution function $\Phi((\boldsymbol{w}, b), \alpha)$ of f by

$$\Phi((\boldsymbol{w}, b), \alpha) := \mathrm{P}\{\, f(\boldsymbol{w}, b) \leq \alpha \,\},$$

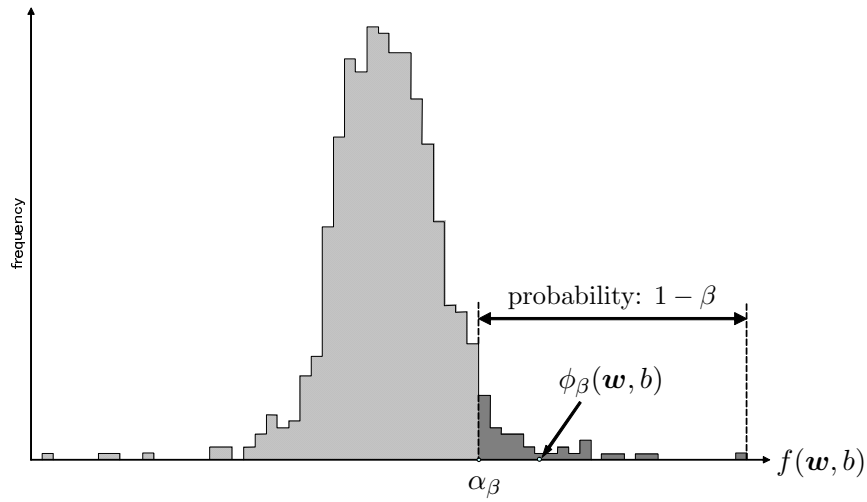


Figure 1: Illustration of the β-tail expectation of f

and a threshold α_β with some confidence level $\beta \in (0, 1)$ by

$$\alpha_\beta := \min\{\, \alpha \,|\, \Phi((\boldsymbol{w}, b), \alpha) \geq \beta \,\}.$$

We note that α_β is well-defined because $\Phi((\boldsymbol{w}, b), \alpha)$ is right continuous and nondecreasing with respect to α. The α_β is known as the *value-at-risk* (VaR) in the context of financial risk management. It is expected that $f(\boldsymbol{w}, b)$ exceeds α_β only in $(1 - \beta) \times 100\%$.

Following the results of Rockafellar & Uryasev [12], we introduce the β-tail distribution function to focus on the tail part of $\Phi((\boldsymbol{w}, b), \alpha)$ as

$$\Phi_\beta((\boldsymbol{w}, b), \alpha) := \begin{cases} 0 & \text{for } \alpha < \alpha_\beta, \\ \dfrac{\Phi((\boldsymbol{w}, b), \alpha) - \beta}{1 - \beta} & \text{for } \alpha \geq \alpha_\beta. \end{cases}$$

Using the expectation operator $\mathrm{E}_\beta[\cdot]$ under the β-tail distribution Φ_β, define the β-tail expectation of f as $\phi_\beta(\boldsymbol{w}, b) := \mathrm{E}_\beta[\, f(\boldsymbol{w}, b)\,]$, which is the risk measure known as the *conditional value-at-risk* (CVaR). Denoting the expectation under the original distribution Φ by $\mathrm{E}[\cdot]$, the following relation shown in [12]:

$$\mathrm{E}[\, f(\boldsymbol{w}, b) \,|\, f(\boldsymbol{w}, b) \geq \alpha_\beta \,] \;\leq\; \phi_\beta(\boldsymbol{w}, b) \;\leq\; \mathrm{E}[\, f(\boldsymbol{w}, b) \,|\, f(\boldsymbol{w}, b) > \alpha_\beta \,] \tag{6}$$

implies that $\phi_\beta(\boldsymbol{w}, b)$ is approximately equal to the conditional expectation of f which exceeds the threshold α_β. In other words, the risk measure $\phi_\beta(\boldsymbol{w}, b)$ is based on the distribution of the lower geometric score, and a hyperplane $H(\boldsymbol{w}, b)$ induced from the risk-measure minimization problem:

$$\underset{\boldsymbol{w} \neq \boldsymbol{0},\, b}{\text{minimize}} \;\; \phi_\beta(\boldsymbol{w}, b) \tag{7}$$

is expected to classify new data points generated from the same distribution correctly.

To solve the problem (7), Rockafellar & Uryasev [12] introduces a simpler auxiliary function $F_\beta : \mathbb{R}^{n+2} \to \mathbb{R}$:

$$F_\beta(\boldsymbol{w}, b, \alpha) := \alpha + \frac{1}{1 - \beta} \mathrm{E}\left[\, [f(\boldsymbol{w}, b) - \alpha]^+ \,\right],$$

where $[X]^+ := \max\{X, 0\}$, and provides a shortcut to minimizing $\phi_\beta(\boldsymbol{w}, b)$ as

$$\underset{(\boldsymbol{w},\, b)\, \in\, W}{\text{minimize}}\ \phi_\beta(\boldsymbol{w}, b) = \underset{((\boldsymbol{w},\, b),\, \alpha)\, \in\, W \times \mathbb{R}}{\text{minimize}}\ F_\beta(\boldsymbol{w}, b, \alpha), \tag{8}$$

where $W := (\mathbb{R}^n \setminus \{\mathbf{0}\}) \times \mathbb{R}$. This relation shows that the minimal value $\phi_\beta(\boldsymbol{w}, b)$ and an optimal solution can be achieved by minimizing the function $F_\beta(\boldsymbol{w}, b, \alpha)$ with respect to $(\boldsymbol{w}, b) \in W$ and $\alpha \in \mathbb{R}$ simultaneously. Furthermore, it is shown in [12] that an optimal α^* of the right-hand side optimization problem is almost equal to α_β.

2.2 Conditional Geometric Score (CGS) Optimization Problem

The optimization of the conditional lower geometric score given in (8) is described as

$$\left|\ \underset{\boldsymbol{w}\, \neq\, \mathbf{0},\, b,\, \alpha}{\text{minimize}}\ \alpha + \frac{1}{1-\beta} \sum_{i \in I} p_i \left[-\frac{y^i\,(\,\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle + b\,)}{\|\boldsymbol{w}\|} - \alpha \right]^+, \right. \tag{9}$$

where $\beta \in (0, 1)$ is a given confidence level. We call this problem the *Conditional Geometric Score* (CGS) optimization problem.

We show first that Problem (9) is equivalent to the following nonlinear programming problem with a nonconvex constraint:

$$(\mathrm{Q}(\beta))\ \left|\ \begin{array}{ll} \underset{\boldsymbol{w},\, b,\, \boldsymbol{z},\, \alpha}{\text{minimize}} & \alpha + \dfrac{1}{1-\beta} \displaystyle\sum_{i \in I} p_i\, z_i \\ \text{subject to} & z_i + y^i(\,\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle + b\,) + \alpha \geq 0, \quad i \in I, \\ & z_i \geq 0, \quad i \in I, \\ & \|\boldsymbol{w}\|^2 = 1. \end{array} \right. \tag{10}$$

The above two problems (9) and (10) have optimal solutions under the following assumption.

Assumption 2.1

- ▷ *Each class has at least one data, i.e.,* $|I^+|, |I^-| > 0$.
- ▷ *The parameter* β *is chosen so that*

$$1 - 2\min\{\sum_{i \in I^+} p_i, \sum_{i \in I^-} p_i\} \leq \beta < 1.$$

Lemma 2.2 *Under Assumption 2.1, Problem (10) has an optimal solution.*

Proof. We can find a feasible solution of (10) easily. Indeed, for arbitrary $\bar{b}$, $\bar{\alpha}$ and $\bar{\boldsymbol{w}}$ satisfying $\|\bar{\boldsymbol{w}}\|^2 = 1$, let $\bar{\boldsymbol{z}}$ satisfy

$$\bar{z}_i \geq \max\{-y_i\,\big(\,\langle \bar{\boldsymbol{w}}, \boldsymbol{x}^i \rangle + \bar{b}\,\big) - \bar{\alpha}, 0\}, \quad \forall i \in I,$$

then $(\bar{\boldsymbol{w}}, \bar{b}, \bar{\boldsymbol{z}}, \bar{\alpha})$ is a feasible solution of (10).

In the following, we will show that this problem never become unbounded under Assumption 2.1. Introducing the dual variables $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \delta) \in \mathbb{R}^{\ell}_{+} \times \mathbb{R}^{\ell}_{+} \times \mathbb{R}$ and constructing the Lagrangian function for Problem (10):

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}, b, \alpha, \boldsymbol{z}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \delta) \\
&\quad := \alpha + \frac{1}{1-\beta} \sum_{i \in I} p_i \, z_i - \sum_{i \in I} \lambda_i \{ z_i + y^i \left(\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle + b \right) + \alpha \} - \sum_{i \in I} \mu_i z_i + \frac{1}{2} \delta (\|\boldsymbol{w}\|^2 - 1),
\end{aligned}$$

the Lagrangian dual problem is represented as

$$\underset{\boldsymbol{\lambda} \geq \mathbf{0}, \, \boldsymbol{\mu} \geq \mathbf{0}, \, \delta}{\text{maximize}} \ \underset{\boldsymbol{w}, \, b, \, \alpha, \, \boldsymbol{z}}{\text{minimize}} \ \mathcal{L}(\boldsymbol{w}, b, \alpha, \boldsymbol{z}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \delta).$$

The stationarity conditions are

$$\sum_{i \in I} \lambda_i y^i \boldsymbol{x}^i = \delta \boldsymbol{w} \tag{11}$$

$$\sum_{i \in I} \lambda_i y^i = 0 \tag{12}$$

$$\sum_{i \in I} \lambda_i = 1 \tag{13}$$

$$0 \leq \lambda_i \leq \frac{p_i}{1-\beta}, \quad i \in I, \tag{14}$$

and substituting (11) to (13) into $\mathcal{L}$, we obtain the Lagrangian function:

$$\mathcal{L} = -\frac{1}{2} \delta \left(\|\boldsymbol{w}\|^2 + 1 \right). \tag{15}$$

Therefore, when $\delta \neq 0$, the Lagrangian dual problem is reformulated as

$$\underset{\boldsymbol{\lambda}, \, \delta}{\text{maximize}} \ -\frac{1}{2} \left(\delta + \frac{1}{\delta} \left\| \sum_{i \in I} y^i \lambda_i \boldsymbol{x}^i \right\|^2 \right) \qquad \text{subject to} \qquad (12) \text{ to } (14),$$

and the constraints (12) to (14) are equivalent to

$$\sum_{i \in I^+} \lambda_i = \sum_{i \in I^-} \lambda_i = \frac{1}{2}, \quad 0 \leq \lambda_i \leq \frac{p_i}{1-\beta}, \quad i \in I.$$

It is easy to verify that the Lagrangian dual is feasible under Assumption 2.1, which proves that the optimal value of (10) is bounded below. Considering that $\boldsymbol{w}$ is bounded in (10), we find that (10) has an optimal solution. □

Proposition 2.3 *Problem (9) is equivalent to Problem (10) in the following sense :*

- ▷ *Let $(\bar{\boldsymbol{w}}, \bar{b}, \bar{\alpha})$ be an optimal solution of (10). Then, $(\eta \bar{\boldsymbol{w}}, \eta \bar{b}, \bar{\alpha})$ for any $\eta > 0$ is an optimal solution of (9).*
- ▷ *Let $(\boldsymbol{w}^*, b^*, \alpha^*)$ be an optimal solution of (9). Then, $(\frac{\boldsymbol{w}^*}{\|\boldsymbol{w}^*\|}, \frac{b^*}{\|\boldsymbol{w}^*\|}, \alpha^*)$ is an optimal solution of (10).*

Proof. Now we show the proof only for the first statement since the second one can be shown similarly.

For any $\eta > 0$, $(\eta\bar{\boldsymbol{w}}, \eta\bar{b}, \bar{\alpha})$ is a feasible solution of (9). Suppose that there exists a feasible solution $(\boldsymbol{w}', b', \alpha')$ of (9) whose objective function value is less than that of $(\eta\bar{\boldsymbol{w}}, \eta\bar{b}, \bar{\alpha})$, i.e.,

$$\alpha' + \frac{1}{1-\beta}\sum_{i\in I} p_i \left[-\frac{y^i\,(\,\langle \boldsymbol{w}', \boldsymbol{x}^i\rangle + b'\,)}{\|\boldsymbol{w}'\|} - \alpha' \right]^+ < \bar{\alpha} + \frac{1}{1-\beta}\sum_{i\in I} p_i \left[-\frac{y^i\,(\,\langle \bar{\boldsymbol{w}}, \boldsymbol{x}^i\rangle + \bar{b}\,)}{\|\bar{\boldsymbol{w}}\|} - \bar{\alpha} \right]^+ .$$

Then, $(\frac{\boldsymbol{w}'}{\|\boldsymbol{w}'\|}, \frac{b'}{\|\boldsymbol{w}'\|}, \alpha')$ is feasible for (10) and it attains smaller objective function value than $(\bar{\boldsymbol{w}}, \bar{b}, \bar{\alpha})$, which contradicts the optimality of $(\bar{\boldsymbol{w}}, \bar{b}, \bar{\alpha})$ to (10). Therefore, for any $\eta > 0$, $(\eta\bar{\boldsymbol{w}}, \eta\bar{b}, \bar{\alpha})$ is an optimal solution of (9). □

From Lemma 2.2 and the first statement in Proposition 2.3, we see that Problem (9) has an optimal solution under Assumption 2.1.

3 Algorithm for the CGS Optimization

The conditional geometric score (CGS) optimization problem (10) includes a single nonconvex constraint. In this section, we examine the nonconvex structure of the problem, and then propose an algorithm for achieving either a globally optimal solution when the problem is convex or a locally optimal solution when the problem is essentially nonconvex.

3.1 2-Step Algorithm

The CGS optimization problem (10) is a nonconvex optimization problem and seems to be difficult to deal with in a direct manner. We consider three cases according to the optimal value of Problem (10) and show the way to solve the problem in each case.

Let us define a relaxation problem of Problem (10) as

$$(\mathrm{RQ}(\beta)) \left| \begin{array}{ll} \underset{\boldsymbol{w},\,b,\,\boldsymbol{z},\,\alpha}{\text{minimize}} & \alpha + \dfrac{1}{1-\beta}\displaystyle\sum_{i\in I} p_i\, z_i \\ \text{subject to} & z_i + y^i(\,\langle \boldsymbol{w}, \boldsymbol{x}^i\rangle + b\,) + \alpha \geq 0, \quad i \in I, \\ & z_i \geq 0, \quad i \in I, \\ & \|\boldsymbol{w}\|^2 \leq 1. \end{array} \right. \tag{16}$$

The existence of an optimal solution in (16) is shown by the similar proof to Lemma 2.2. Denoting the optimal value of Problem (10) by opt.(Q(β)) and that of Problem (16) by opt.(RQ(β)), then we have

$$\text{opt.}(\mathrm{RQ}(\beta)) \;\leq\; \text{opt.}(\mathrm{Q}(\beta)). \tag{17}$$

Proposition 3.1 *Let $(\hat{\boldsymbol{w}}, \hat{b}, \hat{\alpha}, \hat{\boldsymbol{z}})$ be an optimal solution of Problem (16). Then, under Assumption 2.1, the followings hold.*

1. *If* $\|\hat{\boldsymbol{w}}\| = 1$ *holds, then* $(\hat{\boldsymbol{w}}, \hat{b}, \hat{\alpha}, \hat{\boldsymbol{z}})$ *is an optimal solution of Problem (10). In this case,* $\mathrm{opt.}(Q(\beta)) \leq 0$.
2. *If* $0 < \|\hat{\boldsymbol{w}}\| < 1$ *holds, then* $\frac{1}{\|\hat{\boldsymbol{w}}\|}(\hat{\boldsymbol{w}}, \hat{b}, \hat{\alpha}, \hat{\boldsymbol{z}})$ *is an optimal solution of Problem (10) and its optimal value is 0.*
3. *If* $\|\hat{\boldsymbol{w}}\| = 0$ *holds,* $\mathrm{opt.}(Q(\beta)) \geq 0$.

Proof. Note that $(\boldsymbol{w}, b, \alpha, \boldsymbol{z}) = \mathbf{0}$ is a feasible solution of Problem (16), and therefore, $\mathrm{opt.}(\mathrm{RQ}(\beta)) \leq 0$. We use this to prove the statements.

1. We observe that $(\hat{\boldsymbol{w}}, \hat{b}, \hat{\alpha}, \hat{\boldsymbol{z}})$ is feasible and, accordingly, optimal to (10). Therefore, $\mathrm{opt.}(\mathrm{Q}(\beta)) = \mathrm{opt.}(\mathrm{RQ}(\beta)) \leq 0$ holds.
2. Suppose on contrary that $\mathrm{opt.}(\mathrm{RQ}(\beta)) < 0$, then $\frac{1}{\|\hat{\boldsymbol{w}}\|}(\hat{\boldsymbol{w}}, \hat{b}, \hat{\alpha}, \hat{\boldsymbol{z}})$ is a feasible solution of (16) and $\frac{1}{\|\hat{\boldsymbol{w}}\|}\mathrm{opt.}(\mathrm{RQ}(\beta)) < \mathrm{opt.}(\mathrm{RQ}(\beta))$. This contradicts that $(\hat{\boldsymbol{w}}, \hat{b}, \hat{\alpha}, \hat{\boldsymbol{z}})$ is an optimal solution of (16), and hence, $\mathrm{opt.}(\mathrm{RQ}(\beta)) = 0$ follows. Thus, $\mathrm{opt.}(\mathrm{Q}(\beta)) = 0$ and $\frac{1}{\|\hat{\boldsymbol{w}}\|}(\hat{\boldsymbol{w}}, \hat{b}, \hat{\alpha}, \hat{\boldsymbol{z}})$ is an optimal solution of Problem (10).
3. If $\mathrm{opt.}(\mathrm{Q}(\beta)) < 0$, the relation (17) indicates that $\mathrm{opt.}(\mathrm{RQ}(\beta)) < 0$. The dual problem of (16) with an additional constraint $\boldsymbol{w} = \mathbf{0}$ is reduced to

$$\underset{\boldsymbol{\lambda},\, \delta}{\text{maximize}} \quad 0 \quad \text{subject to (12) to (14).}$$

Since this problem has a feasible solution under Assumption 2.1, one has $\mathrm{opt.}(\mathrm{RQ}(\beta)) = 0$ which contradicts the assumption. □

In the third case of Proposition 3.1, any optimal solution of Problem (10) cannot be achieved from that of Problem (16) since the nonconvex constraint is essential. Accordingly, we have to solve a nonconvex programming problem in such a case.

Corollary 3.2 *Let* $(\hat{\boldsymbol{w}}, \hat{b}, \hat{\alpha}, \hat{\boldsymbol{z}})$ *be an optimal solution of Problem (16). Then, the followings hold.*

- ▷ *If* $\mathrm{opt.}(Q(\beta)) < 0$*, one has* $\|\hat{\boldsymbol{w}}\| = 1$.
- ▷ *If* $\mathrm{opt.}(Q(\beta)) > 0$*, one has* $\|\hat{\boldsymbol{w}}\| = 0$.

This corollary is proved by taking contraposition of the statements in Proposition 3.1. The corollary implies that when $\mathrm{opt.}(Q(\beta)) < 0$, Problem (10) turns into the convex problem (16), and when $\mathrm{opt.}(Q(\beta)) \geq 0$, it is equivalent to the nonconvex problem:

$$\left|\begin{array}{ll} \underset{\boldsymbol{w},\, b,\, \boldsymbol{z},\, \alpha}{\text{minimize}} & \alpha + \dfrac{1}{1-\beta}\displaystyle\sum_{i \in I} p_i\, z_i \\ \text{subject to} & z_i + y^i(\,\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle + b\,) + \alpha \geq 0, \quad i \in I, \\ & z_i \geq 0, \quad i \in I, \\ & \|\boldsymbol{w}\|^2 \geq 1. \end{array}\right. \tag{18}$$

Whether the convex constraint $\|\boldsymbol{w}\|^2 \leq 1$ in (16) is essential depends on the sign of the optimal value opt.$(Q(\beta))$ of (10) and we can roughly control the difficulty of Problem (10) by adjusting β since opt.$(Q(\beta))$ is non-decreasing with respect to β.

Now we outline our algorithm for solving the nonconvex program (10) to determine a discriminant hyperplane.

Algorithm 3.3 *(2-step framework)*

Step 1. At first, solve a convex optimization problem (16). For an optimal solution $(\hat{\boldsymbol{w}}, \hat{b}, \hat{\alpha}, \hat{\boldsymbol{z}})$,

- ▷ *if* $\|\hat{\boldsymbol{w}}\| = 1$, *terminate with an optimal solution* $(\hat{\boldsymbol{w}}, \hat{b}, \hat{\alpha})$ *of (10),*
- ▷ *if* $0 < \|\hat{\boldsymbol{w}}\| < 1$, *terminate with an optimal solution* $(\frac{\hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|}, \frac{\hat{b}}{\|\hat{\boldsymbol{w}}\|}, \frac{\hat{\alpha}}{\|\hat{\boldsymbol{w}}\|})$ *of (10),*
- ▷ *if* $\|\hat{\boldsymbol{w}}\| = 0$, *go to Step 2.*

Step 2. Solve the nonconvex program (10) approximately via the algorithm proposed in the next subsection.

3.2 Local Optimum Search Method with Approximation Accuracy

In this subsection, we focus on the case that an optimal solution of Problem (16) attains $\|\hat{\boldsymbol{w}}\| = 0$, that is, the third case of Proposition 3.1 where the optimal value of Problem (10) is nonnegative.

Lemma 3.4 *The following problem provides a lower bound for Problem (10) when the optimal value of (10) is nonnegative, i.e.,* opt.$(Q(\beta)) \geq 0$.

$$
(\mathrm{RP}_\infty(\beta)) \left| \begin{array}{ll} \underset{\boldsymbol{w},\, b,\, \boldsymbol{z},\, \alpha}{\textit{minimize}} & \alpha + \dfrac{1}{1-\beta} \displaystyle\sum_{i \in I} p_i\, z_i \\ \textit{subject to} & z_i + y^i(\,\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle + b\,) + \alpha \geq 0, \quad i \in I, \\ & z_i \geq 0, \quad i \in I, \\ & \|\boldsymbol{w}\|_\infty = \frac{1}{\sqrt{n}}. \end{array} \right. \tag{19}
$$

Proof. Obviously, the optimal value of (19) is nonnegative, that is, opt.$(\mathrm{RP}_\infty(\beta)) \geq 0$ and, in particular, when opt.$(Q(\beta)) = 0$, then opt.$(\mathrm{RP}_\infty(\beta)) = 0$ holds. Hence, we focus on the case that opt.$(Q(\beta)) > 0$ below. Consider a relaxation problem of (19):

$$
(\mathrm{RRP}_\infty(\beta)) \left| \begin{array}{ll} \underset{\boldsymbol{w},\, b,\, \boldsymbol{z},\, \alpha}{\text{minimize}} & \alpha + \dfrac{1}{1-\beta} \displaystyle\sum_{i \in I} p_i\, z_i \\ \text{subject to} & z_i + y^i(\,\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle + b\,) + \alpha \geq 0, \quad i \in I, \\ & z_i \geq 0, \quad i \in I, \\ & \|\boldsymbol{w}\|_\infty \geq \frac{1}{\sqrt{n}}, \end{array} \right. \tag{20}
$$

and note that (20) is also a relaxation problem of (10) because the feasible region of (10) is included in that of (20). Also, we can show that when opt.$(Q(\beta)) > 0$, any optimal solution of (20) satisfies $\|\boldsymbol{w}\|_\infty = \frac{1}{\sqrt{n}}$, which implies that (19) is also a relaxation problem for (10). Indeed,

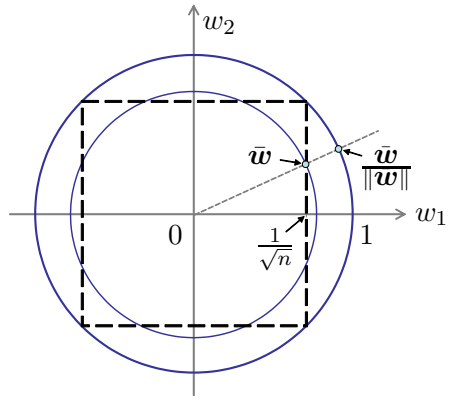


Figure 2: Constraints $\|\boldsymbol{w}\| = 1$ and $\|\boldsymbol{w}\|_\infty = 1/\sqrt{n}$

suppose on contrary that an optimal solution $\tilde{\boldsymbol{w}}$ of (20) satisfies $\|\tilde{\boldsymbol{w}}\|_\infty > \frac{1}{\sqrt{n}}$, then we obtain a feasible solution $\frac{1}{\sqrt{n}\|\tilde{\boldsymbol{w}}\|_\infty}(\tilde{\boldsymbol{w}}, \tilde{b}, \tilde{\alpha}, \tilde{\boldsymbol{z}})$ of (20) with smaller optimal value, which contradicts the optimality of $\tilde{\boldsymbol{w}}$. □

For all $(d, h) \in \{1, ..., n\} \times \{\pm\frac{1}{\sqrt{n}}\}$, solve the following linear programming problems:

$$(\mathrm{RLP}_\infty(d, h; \beta)) \left| \begin{array}{ll} \underset{\boldsymbol{w},\, b,\, \boldsymbol{z},\, \alpha}{\text{minimize}} & \alpha + \dfrac{1}{1-\beta} \displaystyle\sum_{i \in I} p_i\, z_i \\ \text{subject to} & z_i + y^i(\,\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle + b\,) + \alpha \geq 0, \quad i \in I, \\ & z_i \geq 0, \quad i \in I, \\ & -\frac{1}{\sqrt{n}} \leq w_j \leq \frac{1}{\sqrt{n}} \text{ for all } j = 1, ..., n, \\ & w_d = h. \end{array} \right. \tag{21}$$

Using the optimal values $\bar{f}(d, h)$ for all $(d, h) \in \{1, ..., n\} \times \{\pm\frac{1}{\sqrt{n}}\}$ and corresponding solutions, we obtain the optimal value $\bar{f}$ of (19) by setting

$$\bar{f} := \min\{\, \bar{f}(d, h) \,|\, (d, h) \in \{1, ..., n\} \times \{\pm\frac{1}{\sqrt{n}}\}\,\},$$

and its optimal solution $(\bar{\boldsymbol{w}}, \bar{b}, \bar{\alpha}, \bar{\boldsymbol{z}})$ by defining

$$(\bar{\boldsymbol{w}}, \bar{b}, \bar{\alpha}, \bar{\boldsymbol{z}}) : \text{an optimal solution of } (\mathrm{RLP}_\infty(d^*, h^*; \beta)), \tag{22}$$

where $(d^*, h^*) \in \arg\min\{\, \bar{f}(d, h) \,|\, (d, h) \in \{1, ..., n\} \times \{\pm\frac{1}{\sqrt{n}}\}\}$. Note that $0 < \|\bar{\boldsymbol{w}}\| < 1$, and $\frac{1}{\|\bar{\boldsymbol{w}}\|}(\bar{\boldsymbol{w}}, \bar{b}, \bar{\gamma}, \bar{\boldsymbol{z}})$ is feasible to (10) whose objective function value is $\frac{1}{\|\bar{\boldsymbol{w}}\|}\bar{f}$. One then has

$$0 \;\leq\; \bar{f} \;\leq\; \text{opt.}(Q(\beta)) \;\leq\; \frac{1}{\|\bar{\boldsymbol{w}}\|}\bar{f}.$$

Using the local optimum search technique proposed in [14], the upper bound $\frac{1}{\|\bar{\boldsymbol{w}}\|}\bar{f}$ might be improved. The procedure for local optimum search exploits the property of Problem (10), whose feasible region is reduced to a polyhedral cone by excluding the nonconvex constraint. We construct a polyhedral set by linearizing the nonconvex constraint as

$$\mathcal{F}(\boldsymbol{w}_k) := \left\{ (\boldsymbol{w}, b, \alpha, \boldsymbol{z}) \;\middle|\; \begin{array}{l} z_i + y_i\left(\,\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle + b\,\right) + \alpha \geq 0, \quad i \in I, \\ z_i \geq 0, \quad i \in I, \\ \langle \boldsymbol{w}_k, \boldsymbol{w} \rangle = 1 \end{array} \right\}$$

with some feasible solution $\boldsymbol{w}_k$ of (10), and then apply a pivoting technique for the linear programming problem:

$$\left| \underset{\boldsymbol{w},\, b,\, \boldsymbol{z},\, \alpha}{\text{minimize}} \;\; q(\boldsymbol{z}, \alpha) := \alpha + \frac{1}{1-\beta} \sum_{i \in I} p_i \, z_i \quad \text{subject to } (\boldsymbol{w}, b, \alpha, \boldsymbol{z}) \in \mathcal{F}(\boldsymbol{w}_k) \right. \tag{23}$$

to find a local optimum of Problem (10). Note that, when the optimal value of Problem (10) is nonnegative, it is equivalent to (18) which includes the constraint $\|\boldsymbol{w}\|^2 \geq 1$ instead of $\|\boldsymbol{w}\|^2 = 1$. Since $\mathcal{F}(\boldsymbol{w}_k)$ is included in the feasible region of (18), the optimal value of (23) is also nonnegative.

Here we introduce some additional definitions for the description of local optimum search algorithm. Let us define a polyhedral cone H by

$$H := \left\{ (\boldsymbol{w}, b, \alpha, \boldsymbol{z}) \;\middle|\; \begin{array}{l} z_i + y^i(\,\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle + b\,) + \alpha \geq 0, \quad i \in I \\ z_i \geq 0, \quad i \in I \end{array} \right\},$$

and a unit sphere G by

$$G := \left\{ (\boldsymbol{w}, b, \alpha, \boldsymbol{z}) \mid \|\boldsymbol{w}\|^2 = 1 \right\}.$$

Let ray$H$ denote the set of extreme rays of H, and we call a feasible solution lying on ray$H \cap G$ a *basic solution.*

Algorithm 3.5 *(Local optimum search)*

Step 0. Choose a feasible basic solution $(\boldsymbol{w}_0, b_0, \alpha_0, \boldsymbol{z}_0)$ *of (10). Let* $k = 0$.

Step 1. At $\boldsymbol{w}_k$*, construct the polyhedral set* $\mathcal{F}(\boldsymbol{w}_k)$.

Step 2. If there exists an extreme point $(\bar{\boldsymbol{w}}, \bar{b}, \bar{\alpha}, \bar{\boldsymbol{z}})$ *of* $\mathcal{F}(\boldsymbol{w}_k)$ *adjacent to* $(\boldsymbol{w}_k, b_k, \alpha_k, \boldsymbol{z}_k)$ *with* $q(\boldsymbol{z}_k, \alpha_k) \geq q(\bar{\boldsymbol{z}}, \bar{\alpha})$*, then compute the next basic solution*

$$(\boldsymbol{w}_{k+1}, b_{k+1}, \alpha_{k+1}, \boldsymbol{z}_{k+1}) := \frac{1}{\|\bar{\boldsymbol{w}}\|} (\bar{\boldsymbol{w}}, \bar{b}, \bar{\alpha}, \bar{\boldsymbol{z}}) \tag{24}$$

of (10), let $k = k + 1$ *and go to Step 1. Otherwise, terminate with a local minimizer* $(\boldsymbol{w}_k, b_k, \alpha_k, \boldsymbol{z}_k)$.

In Step 0, we can find a feasible basic solution of (10) easily. After obtaining a feasible solution $\frac{1}{\|\bar{\boldsymbol{w}}\|}(\bar{\boldsymbol{w}}, \bar{b}, \bar{\gamma}, \bar{\boldsymbol{z}})$ from (22), we construct $\mathcal{F}(\frac{\bar{\boldsymbol{w}}}{\|\bar{\boldsymbol{w}}\|})$. Then, we find an extreme point $(\boldsymbol{w}', b', \boldsymbol{z}', \alpha')$ of $\mathcal{F}(\frac{\bar{\boldsymbol{w}}}{\|\bar{\boldsymbol{w}}\|})$ after one pivot procedure, project the extreme point on G by (24), and achieve a feasible basic solution of (10).

The solution $(\boldsymbol{w}_k, b_k, \alpha_k, \boldsymbol{z}_k)$ obtained by this algorithm corresponds to a local minimizer of Problem (10), since this solution satisfies the *Karush-Kuhn-Tucker* (KKT) optimality conditions induced from (10). According to [14], it is easily shown that Algorithm 3.5 terminates within finite iterations when the optimal value of Problem (10) is nonnegative, i.e., opt.$(Q(\beta)) \geq 0$.

4 Relation to the Other Classification Problems

In this section, we discuss a relationship between the conditional geometric score (CGS) optimization problem and other classification problems such as the hard margin support vector classification (HSVC) and the ν-support vector classification (ν-SVC). Under some parameter setting, our model is proved to be equivalent to them.

4.1 The Hard Margin Support Vector Classification

We first explore the relation between the CGS optimization (9) and the hard margin support vector classification (2). Note that (2) can be represented as

$$\text{(HSVC2)} \quad \left| \underset{\boldsymbol{w} \neq \boldsymbol{0},\, b}{\text{minimize}} \ \max_{i \in I} \ -\frac{y^i\,(\,\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle + b\,)}{\|\boldsymbol{w}\|}. \right.$$

We show below that the formulation (9) coincides with (HSVC2) if $\beta > 1 - \min_i p_i$ holds.

Let $(\boldsymbol{w}^*, b^*, \alpha^*)$ be an optimal solution of (9) and let

$$k := \arg\max_{i \in I} \left\{ -\frac{y^i\,(\,\langle \boldsymbol{w}^*, \boldsymbol{x}^i \rangle + b^*\,)}{\|\boldsymbol{w}^*\|} \right\}.$$

Suppose that there exists $\gamma > 0$ such that

$$\alpha^* = -\frac{y^k\,(\,\langle \boldsymbol{w}^*, \boldsymbol{x}^k \rangle + b^*\,)}{\|\boldsymbol{w}^*\|} - \gamma.$$

Then,

$$\begin{aligned} \alpha^* + \frac{1}{1-\beta} \sum_{i \in I} p_i \left[-\frac{y^i\,(\,\langle \boldsymbol{w}^*, \boldsymbol{x}^i \rangle + b^*\,)}{\|\boldsymbol{w}^*\|} - \alpha^* \right]^+ &\geq -\frac{y^k\,(\,\langle \boldsymbol{w}^*, \boldsymbol{x}^k \rangle + b^*\,)}{\|\boldsymbol{w}^*\|} + (\frac{p_k}{1-\beta} - 1)\gamma \\ &> -\frac{y^k\,(\,\langle \boldsymbol{w}^*, \boldsymbol{x}^k \rangle + b^*\,)}{\|\boldsymbol{w}^*\|}, \end{aligned}$$

which contradicts the optimality of $(\boldsymbol{w}^*, b^*, \alpha^*)$. Accordingly, the optimal solution satisfies

$$\alpha^* \geq \max_{i \in I} \left\{ -\frac{y^i\,(\,\langle \boldsymbol{w}^*, \boldsymbol{x}^i \rangle + b^*\,)}{\|\boldsymbol{w}^*\|} \right\},$$

and Problem (9) results in (HSVC2).

From this viewpoint, the hard margin support vector classification (HSVC) corresponds to the CGS optimization with sufficiently large β. To attain more accurate classifier, it might be better to solve the CGS optimization problem (10) which takes into consideration the other misclassified data by adjusting the parameter β.

4.2 The ν-Support Vector Classification

We next show that the convex problem (16) solved at the first step of Algorithm 3.3 is equivalent to the ν-SVC when the optimal value of (10) is negative. Besides, it is shown that the ν-SVC results in a meaningless solution when the optimal value of (10) is positive.

Multiplying the objective function of (16) by $1-\beta > 0$ and introducing the Lagrangian multiplier $\theta \geq 0$ with respect to the convex constraint $\|\boldsymbol{w}\|^2 \leq 1$, we obtain a Lagrangian relaxation problem of (16):

$$(\mathrm{LRQ}(\theta;\beta)) \left| \begin{array}{ll} \underset{\boldsymbol{w},\, b,\, \boldsymbol{z},\, \alpha}{\text{minimize}} & \theta(\|\boldsymbol{w}\|^2 - 1) + (1-\beta)\alpha + \sum_{i \in I} p_i\, z_i \\ \text{subject to} & z_i + y^i(\,\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle + b\,) + \alpha \geq 0, \quad i \in I, \\ & z_i \geq 0, \quad i \in I. \end{array} \right. \tag{25}$$

Then, the optimal value of $(\mathrm{LRQ}(\theta;\beta))$ with any $\theta \geq 0$ is equal to or less than that of (16).

If we change the variable and parameters in Problem (25) as

$$\alpha = -\rho, \quad \beta = 1-\nu, \quad p_i = \frac{1}{|I|} \text{ for } i \in I, \tag{26}$$

add nonpositivity on $\alpha \leq 0$ and choose the Lagrangian multiplier as $\theta = \frac{1}{2}$, this problem is led into the primal formulation of the ν-SVC [13]:

$$(\nu\text{-SVC}(\nu)) \left| \begin{array}{ll} \underset{\boldsymbol{w},\, b,\, \boldsymbol{z},\, \rho}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{w}\|^2 - \nu\rho + \frac{1}{|I|}\sum_{i \in I} z_i \\ \text{subject to} & z_i + y^i(\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle + b) - \rho \geq 0, \quad i \in I, \\ & z_i \geq 0, \quad i \in I, \\ & \rho \geq 0. \end{array} \right. \tag{27}$$

In our classification model, α corresponds to a threshold of the distribution of $f(\boldsymbol{w}, b)$ with confidence level $\beta \in (0,1)$ and the nonpositivity is not imposed on α.

As shown in Corollary 3.2, when the optimal value of Problem (10) satisfies opt.$(Q(\beta)) < 0$, (10) is equivalent to the convex problem (16), and when opt.$(Q(\beta)) \geq 0$, it is equivalent to the nonconvex problem (18). From this point of view, the sign of Lagrangian multiplier θ should be determined according to the optimal value of (10). In addition, we can show the following theorem.

Theorem 4.1 *The ν-SVC (27) with parameters (26) provides a meaningless solution satisfying $\boldsymbol{w} = \boldsymbol{0}$ when the optimal value of (10) is positive.*

Proof. Corollary 3.2 shows that when opt.$(Q(\beta)) > 0$, Problem (16) provides a meaningless optimal solution satisfying $\boldsymbol{w} = 0$. In such a case, Problem (16) is equivalent to the following linear programming problem:

$$(\mathrm{RP2}(\beta)) \left| \begin{array}{ll} \underset{\boldsymbol{w},\, b,\, \boldsymbol{z},\, \alpha}{\text{minimize}} & (1-\beta)\alpha + \sum_{i \in I} p_i\, z_i \\ \text{subject to} & z_i + y^i(\,\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle + b\,) + \alpha \geq 0, \quad i \in I, \\ & z_i \geq 0, \quad i \in I. \end{array} \right. \tag{28}$$

By duality, the optimal value opt.$(\mathrm{RP2}(\beta))$ of (28) is 0, and the comparison between the ν-SVC (27) with parameters (26) and Problem (28) implies

$$0 = \text{opt.}(\mathrm{RP2}(\beta)) \leq \text{opt.}(\nu\text{-SVC}(\nu)).$$

Therefore, a feasible solution $(\boldsymbol{w}, b, \rho, \boldsymbol{z}) = \mathbf{0}$ is optimal in (27). We can show that there is no optimal solution satisfying $\|\boldsymbol{w}\| > 0$, since if (27) has such a solution and the optimal value is 0, then opt.(RP2(β)) < 0 holds. Thus, any optimal solution of (27) satisfies $\boldsymbol{w} = \mathbf{0}$, and the statement follows. □

Now we introduce a convex quadratic program:

$$(D^+) \left| \begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\sum_{i\in I}\sum_{j\in I} y^i y^j \langle \boldsymbol{x}^i, \boldsymbol{x}^j \rangle \lambda_i \lambda_j \\ \text{subject to} & \sum_{i\in I} \lambda_i y^i = 0, \\ & \sum_{i\in I} \lambda_i = 1, \\ & 0 \le \lambda_i \le \dfrac{1}{1-\beta} p_i, \quad i \in I. \end{array} \right.$$

Lemma 4.2 *When the optimal value of (10) is negative, (D^+) is the Lagrangian dual problem of (10) with zero gap.*

Proof. In the proof of Lemma 2.2, we have provided the Lagrangian dual problem of (10) as

$$\underset{\boldsymbol{\lambda},\,\delta}{\text{maximize}} \quad -\frac{1}{2}\,(\,\delta + \frac{1}{\delta}\left\|\sum_{i\in I} y^i \lambda_i \boldsymbol{x}^i\right\|^2\,) \quad \text{subject to (12) to (14)},$$

when $\delta \neq 0$. Since the optimal value of the primal problem (10) is negative, the optimal value of the above Lagrangian dual problem is less than 0, any optimal solution satisfies $\delta^* > 0$ and (11) shows $\left\|\sum_{i\in I} y^i \lambda_i \boldsymbol{x}^i\right\| > 0$. Then, the relation between the arithmetic and the geometric means implies that

$$-\frac{1}{2}\,(\,\delta + \frac{1}{\delta}\left\|\sum_{i\in I} y^i \lambda_i \boldsymbol{x}^i\right\|^2\,) \le -\left\|\sum_{i\in I} y^i \lambda_i \boldsymbol{x}^i\right\| = -\sqrt{\sum_{i\in I}\sum_{j\in I} y^i y^j \langle \boldsymbol{x}^i, \boldsymbol{x}^j \rangle \lambda_i \lambda_j},$$

and the Lagrangian dual problem results in (D^+), where the optimal solution δ^* is achieved at $\delta^* = \frac{1}{2}\sqrt{\sum_{i\in I}\sum_{j\in I} y^i y^j \langle \boldsymbol{x}^i, \boldsymbol{x}^j \rangle \lambda_i \lambda_j}$. Since Problem (10) is equivalent to Problem (16) when the optimal value of (10) is negative and both the Lagrangian dual (D^+) and (16) are convex programs, the result follows. □

Theorem 4.3 *The CGS optimization problem (10) and the ν-SVC (27) with parameters (26) provide the same discriminant hyperplane when the optimal value of (10) is negative.*

Proof. The dual of the ν-SVC (27) is known as

$$(\mathrm{D}\nu\mathrm{SVC}) \left| \begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\sum_{i\in I}\sum_{j\in I} y^i y^j \langle \boldsymbol{x}^i, \boldsymbol{x}^j \rangle \lambda_i \lambda_j \\ \text{subject to} & \sum_{i\in I} \lambda_i y^i = 0, \\ & \sum_{i\in I} \lambda_i \ge \nu, \\ & 0 \le \lambda_i \le \dfrac{1}{|I|}, \quad i \in I. \end{array} \right.$$

By replacing the variables $\boldsymbol{\lambda}$ by $\boldsymbol{\lambda}' := \frac{\boldsymbol{\lambda}}{\nu}$, the problem is proved to be equivalent to (D^+) except the inequality $\sum_{i \in I} \lambda'_i \geq 1$ which corresponds to $\rho \geq 0$ in (27). When the optimal value of (10) is negative, $\alpha^* < 0$ holds at any optimal solution α^*. Therefore, these two problems are equivalent when the optimal value of (10) is negative. □

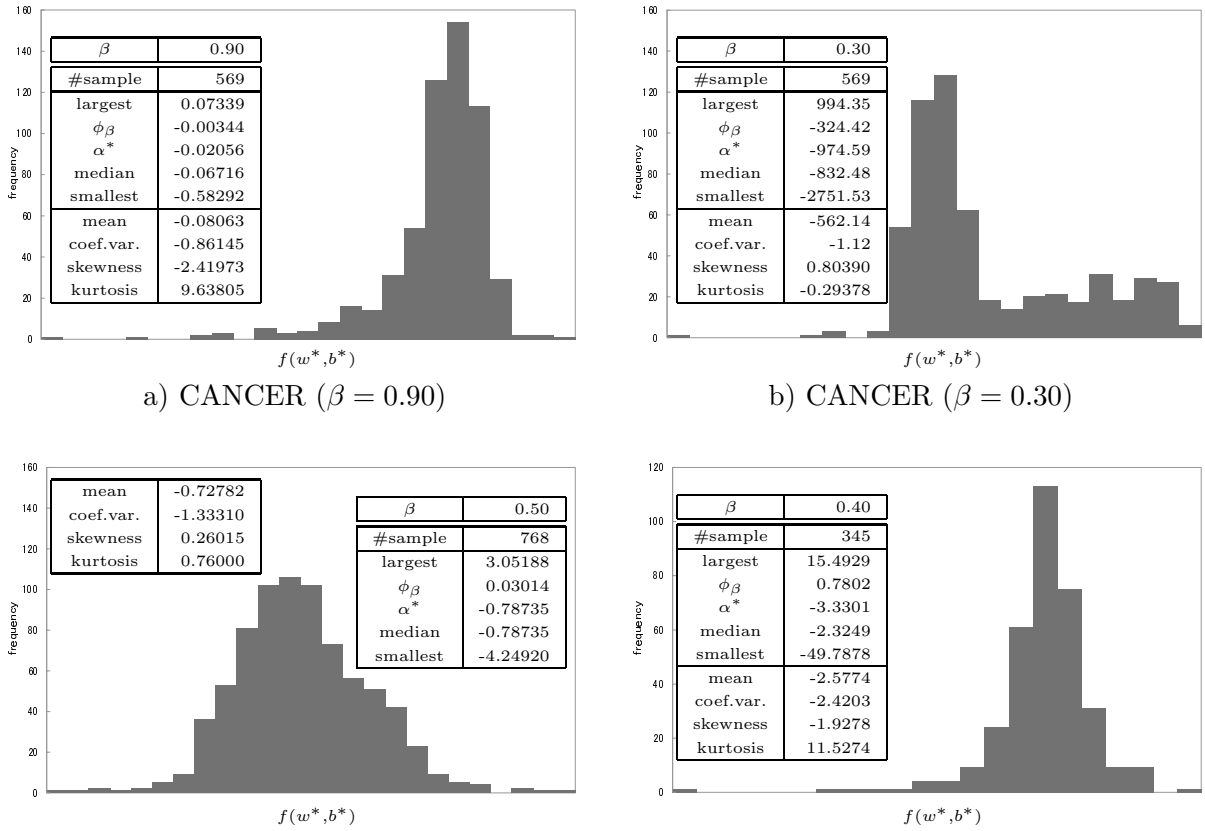
5 Computational Experiments

In this section, we examine our model by comparing with existing models such as the ν-SVC and the robust linear programming (RLP) approach [2] through numerical experiments. The instances to be used in the experiments are named 1) CANCER, 2) LIVER, and 3) DIABETES, all of which are obtained from the UCI repository of machine learning databases [3]. For the CANCER data, we use only three attributes which are shown to be highly predictable via RLP approach in [8], whereas we use all attributes provided in the database for the latter two instances, i.e., six attributes for LIVER, and eight for DIABETES.

Distribution of Optimal Geometric Score Figure 3 shows histograms and statistical characteristics for the optimal distribution of the geometric score $f(\boldsymbol{w}^*, b^*; \boldsymbol{x}, y) = -g(\boldsymbol{w}^*, b^*; \boldsymbol{x}, y)$, where $(\boldsymbol{w}^*, b^*, \alpha^*, \boldsymbol{z}^*)$ is an optimal solution of Problem (10) when the empirical probability p_i is given by (4). Figures a) and b) are the results of the CANCER data under the different parameter β. Figures a), c) and d) show the results of CANCER, DIABETES and LIVER, respectively, under the parameter β chosen so that the number of misclassification becomes very small. It is worth noting that the optimal values ϕ_β in Figures c) and d) are positive and the nonconvexity is essential. From these figures, we see that the shape of the optimal distribution depends on β and the given data sets.

Misclassification and Misprediction To show the potential superiority of our model over the ν-SVC, we performed ten-fold cross-validation for the three data sets. Figure 4 shows the misclassification (training error) and the misprediction (testing error) rates obtained by Problem (10) with $\beta = 0.4$ to 0.55 when the DIABETES data set is applied. The points highlighted by '$\circ$' and '$\times$' in the figure correspond to $\beta \geq 0.47$. For such β, at least one of the ten trials through the cross-validation proceeded to the Step 2 of Algorithm 3.3. It implies that when $\beta \geq 0.47$, that is, $\nu \leq 0.53$, the ν-SVC fails to attain any hyperplane at least one of the ten trials while the numerical results of our model and the ν-SVC are exactly the same when $\beta \leq 0.46$. Consequently, it is possible to find more predictable hyperplane by our model for larger β so that the ν-SVC returns a meaningless solution, i.e., $\boldsymbol{w}^* = \boldsymbol{0}$. This indicates that the nonconvexity plays an important role in the classification problem. Similar results are observed with the other data sets.

Tables 1 to 3 summarize the results of misclassification and misprediction of three data sets by using three models: the CGS, the ν-SVC, and the RLP. For the CGS and the ν-SVC, we compute discriminant hyperplanes under several values of β and report the best results. Each cell shows



a) CANCER ($\beta = 0.90$) b) CANCER ($\beta = 0.30$)

c) DIABETES ($\beta = 0.5$) d) LIVER ($\beta = 0.4$)

Figure 3: Histograms of f

the sample average and the sample standard deviation of the misclassification or misprediction rates. The column named '+1' shows the rates for the '+1'-labeled data set while the '-1' shows those for the other data set. The rows named 'CGS1' show results achieved by the model proposed in the paper when the empirical probability is given by (4), whereas 'CGS2' means the same model when the probability is given by (5). Similarly, 'RLP1' means the minimization of equally weighted misclassification error, i.e., $\frac{1}{\ell}\sum_{i\in I}\xi_i$, whereas 'RLP2' minimizes weighted error terms defined by $\frac{1}{|I^+|}\sum_{i\in I^+}\xi_i + \frac{1}{|I^-|}\sum_{i\in I^-}\xi_i$, where $\xi_i = [1 - y_i(\langle \boldsymbol{w}, \boldsymbol{x}^i\rangle + b)]^+$. Correspondingly, '$\nu$-SVC1' means (27), whereas 'ν-SVC2' means a variation of (27) whose objective function is replaced by $\frac{1}{2}\|\boldsymbol{w}\|^2 - \nu\rho + \frac{1}{2|I^+|}\sum_{i\in|I^+|} z_i + \frac{1}{2|I^-|}\sum_{i\in|I^-|} z_i$.

From these tables, our model achieves the best predictive accuracy in terms of the sample mean through ten-fold cross-validation. For the CANCER data, our model performs as well as the RLP2 which is shown to be highly predictable in [8] when the weighted empirical probability (5) is applied. However, the weighted probability did not always attain the better total predictability, but the misprediction rates of the two labels are more balanced than the equally weighted ones.

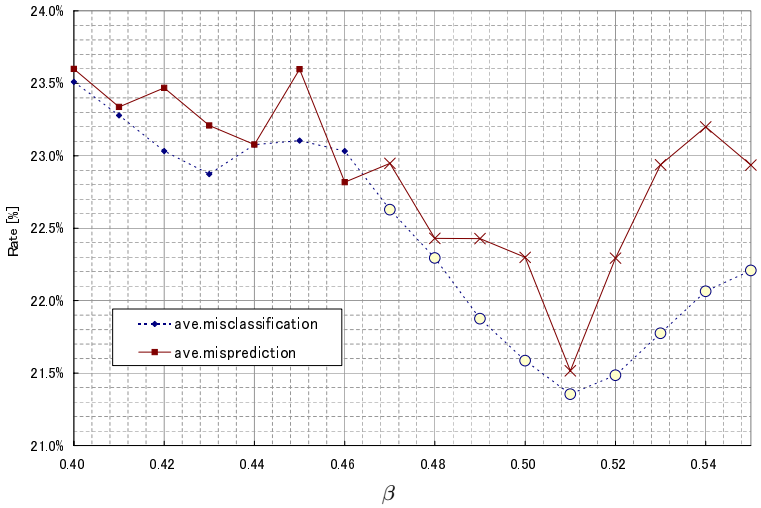


Figure 4: Misclassification vs. misprediction through Algorithm 3.3 (the DIABETES data)

Table 1: Misclassification and misprediction rates for the CANCER data

		misclassification			misprediction		
		total	+1	-1	total	+1	-1
CGS1	av.	3.03%	4.09%	2.40%	3.16%	4.26%	2.51%
($\beta = .97$)	s.d.	(0.53%)	(0.62%)	(0.49%)	(1.99%)	(4.29%)	(1.62%)
CGS2	av.	2.60%	3.30%	2.18%	**2.63%**	3.46%	2.21%
($\beta = .92$)	s.d.	(0.29%)	(0.37%)	(0.32%)	(1.70%)	(3.32%)	(1.73%)
ν-SVC1	av.	3.12%	6.29%	1.24%	3.16%	6.82%	1.10%
($\beta = .88$)	s.d.	(0.37%)	(0.88%)	(0.20%)	(2.45%)	(5.67%)	(1.92%)
ν-SVC2	av.	2.75%	4.19%	1.90%	2.81%	4.01%	2.25%
($\beta = .88$)	s.d.	(0.34%)	(0.92%)	(0.44%)	(1.69%)	(4.09%)	(1.76%)
RLP1	av.	3.14%	5.66%	1.65%	3.33%	6.32%	1.65%
	s.d.	(0.42%)	(0.92%)	(0.30%)	(2.54%)	(6.06%)	(1.92%)
RLP2	av.	2.54%	3.51%	1.96%	**2.63%**	3.46%	2.21%
	s.d.	(0.26%)	(0.56%)	(0.21%)	(1.70%)	(3.32%)	(1.73%)

Table 2: Misclassification and misprediction rates for the DIABETES data

		misclassification			misprediction		
		total	+1	-1	total	+1	-1
CGS1	av.	21.35%	11.11%	40.48%	**21.52%**	10.97%	40.36%
($\beta = .51$)	s.d.	(0.51%)	(0.68%)	(1.29%)	(7.43%)	(5.18%)	(10.74%)
CGS2	av.	24.07%	21.14%	29.57%	24.25%	21.77%	28.71%
($\beta = .42$)	s.d.	(0.45%)	(0.94%)	(1.19%)	(7.92%)	(5.98%)	(13.84%)
ν-SVC1	av.	23.03%	11.83%	43.95%	22.82%	11.80%	43.08%
($\beta = .46$)	s.d.	(0.97%)	(1.19%)	(1.04%)	(7.18%)	(4.89%)	(11.22%)
ν-SVC2	av.	24.07%	21.14%	29.57%	24.25%	21.77%	28.71%
($\beta = .42$)	s.d.	(0.45%)	(0.94%)	(1.19%)	(7.92%)	(5.98%)	(13.84%)
RLP1	av.	22.15%	11.09%	42.79%	22.56%	11.61%	42.77%
	s.d.	(0.84%)	(0.56%)	(1.43%)	(7.12%)	(4.89%)	(10.83%)
RLP2	av.	23.35%	21.03%	27.69%	23.20%	20.72%	27.55%
	s.d.	(0.81%)	(0.77%)	(1.44%)	(7.57%)	(5.60%)	(12.48%)

Table 3: Misclassification and Misprediction for the LIVER data

		misclassification			misprediction		
		total	+1	-1	total	+1	-1
CGS1	av.	29.66%	43.38%	19.73%	**28.64%**	42.88%	18.51%
($\beta = .41$)	s.d.	(1.85%)	(2.16%)	(2.28%)	(7.07%)	(13.77%)	(6.69%)
CGS2	av.	31.46%	34.33%	29.40%	31.55%	36.07%	28.51%
($\beta = .31$)	s.d.	(1.79%)	(2.26%)	(2.64%)	(7.72%)	(12.25%)	(8.30%)
ν-SVC1	av.	29.44%	49.09%	15.23%	30.69%	51.88%	15.46%
($\beta = .25$)	s.d.	(0.93%)	(3.54%)	(1.92%)	(5.63%)	(9.94%)	(5.75%)
ν-SVC2	av.	32.27%	30.73%	33.40%	33.91%	34.17%	34.03%
($\beta = .23$)	s.d.	(1.75%)	(2.10%)	(2.57%)	(6.64%)	(11.30%)	(6.04%)
RLP1	av.	31.69%	31.80%	31.62%	33.03%	34.80%	32.01%
	s.d.	(2.03%)	(1.95%)	(2.74%)	(7.87%)	(13.45%)	(6.32%)
RLP2	av.	28.86%	44.92%	17.23%	30.39%	48.46%	17.46%
	s.d.	(0.78%)	(1.96%)	(1.34%)	(6.07%)	(9.84%)	(6.58%)

6 Concluding Remarks

In this paper, we proposed a classification model based on the conditional expectation of the geometric score which is a simple extension of the Euclidean distance of data point to the discriminant hyperplane. The model is formulated as a nonconvex optimization problem, and we present a two-step algorithm for obtaining either a globally or a locally optimal solution of the problem. At the first step of the algorithm, a convex problem (16) is solved which is shown to be almost the same as the ν-SVC formulation. When the first step results in a meaningless one, i.e., the zero normal vector, the algorithm proceeds to the second step at which a local solution of Problem (10) is provided by solving a finite number of additional linear programming problems.

The numerical experiments indicate that the nonconvexity in our model plays an important role in achieving higher predictability. In fact, for any instance of CANCER, LIVER and DIABETES, the most predictable discriminant hyperplane was induced from an approximate solution of the nonconvex problem (10). In addition, our model with β satisfying Assumption 2.1 always provides a discriminant hyperplane while the ν-SVC results in a meaningless solution for large β.

One possible extension is to construct a nonlinear discriminant surface by incorporating the kernel functions into the dual problem (D^+) of Problem (16). Since (D^+) is equivalent to the ν-SVC formulation under some condition, it is easy to introduce the kernel functions into (D^+) in a similar manner to the ν-SVC. On the other hand, we need more discussion to incorporate the positive definite kernel functions into the nonconvex problem (18). The detailed exploration remains to be done in the future work.

Acknowledgement

Research of the first author is supported by MEXT Grant-in-Aid for Young Scientists (B) 14780343. Research of the second author is supported by MEXT Grant-in-Aid for Young Scientists (B) 16710110.

References

[1] P. Artzner, F. Delbaen, J.M. Eber and D. Heath, Coherent measures of risk, *Math. Finance* 9 (1999) 203-228.

[2] K.P. Bennett and O.L. Mangasarian, Robust linear programming discrimination of two linearly inseparable sets, *Optim. Methods Softw.* 1 (1992) 23-34.

[3] C.L. Blake and C.J. Merz, UCI Repository of machine learning databases [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[4] N. Freed and F. Glover, A linear programming approach to the discriminant problem, *Decis. Sci.* 12 (1981) 68-74.

[5] D.J. Hand, *Discrimination and classification*, John Wiley, Chichester U.K., 1981.

[6] G.J. Koehler and S.S. Erenguc, Minimizing misclassification in linear discriminant analysis, *Decis. Sci.* 21 (1990) 63-85.

[7] O.L. Mangasarian, Multisurface method of pattern separation, *IEEE Trans. Inform. Theory* IT-14 (1968) 801-807.

[8] O.L. Mangasarian, W.N. Street and W.H.Wolberg, Breast cancer diagnosis and prognosis via linear programming, *Oper. Res.* 43 (1995) 570-577.

[9] W. Ogryczak and A. Ruszczýnski, Dual stochastic dominance and related mean–risk models, *SIAM J. Optim.* 13 (2002) 60-78.

[10] G.Ch. Pflug, Some remarks on the value-at-risk and the conditional value-at-risk, in *Probabilistic Constrained Optimization: Methodology and Applications*, S. Uryasev (ed.), Kluwer Academic Publishers, Dordrecht, 2000, 272-281.

[11] R.T. Rockafellar and S. Uryasev, Optimization of conditional value-at-risk, *J. Risk* 2 (2000) 21-41.

[12] R.T. Rockafellar and S. Uryasev, Conditional value-at-risk for general loss distributions, *J. Bank. Finance* 26 (2002) 1443-1471.

[13] B. Schölkopf, A. J. Smola, R. C. Williamson and P. L. Bartlett, New support vector algorithms, *Neural Comput.* 12 (2000) 1207-1245.

[14] A. Takeda and H. Nishino, On measuring the inefficiency with the inner-product norm in date envelopment analysis, *European J. Oper. Res.* 133 (2001) 377-393.

[15] V. Vapnik, *The nature of statistical learning theory*, Springer Verlag, New York, 1995.