A Distributionally-robust Approach for Finding Support Vector Machines

Changhyeok Lee CHANGHYEOK.LEE@U.NORTHWESTERN.EDU
Department of IEMS
Northwestern University
Evanston, IL 60208, USA

Sanjay Mehrotra MEHROTRA@NORTHWESTERN.EDU
Department of IEMS
Northwestern University
Evanston, IL 60208, USA

Editor: TBD

Abstract

The classical SVM is an optimization problem minimizing the hinge losses of mis-classified samples with the regularization term. When the sample size is small or data has noise, it is possible that the classifier obtained with training data may not generalize well to population, since the samples may not accurately represent the true population distribution. We propose a distributionally-robust framework for Support Vector Machines (DR-SVMs). We build an ambiguity set for the population distribution based on samples using the Kantorovich metric. DR-SVMs search the classifier that minimizes the sum of regularization term and the hinge loss function for the worst-case population distribution among the ambiguity set. We provide semi-infinite programming formulation of the DR-SVMs and propose a cutting-plane algorithm to solve the problem. Computational results on simulated data and real data from University of California, Irvine Machine Learning Repository show that the DR-SVMs outperform the SVMs in terms of the Area Under Curve (AUC) measures on several test problems.

Keywords: Support Vector Machines (SVMs), Classification, Robust, Distributionally-robust, Kantrorovich metric

1. Introduction

Support Vector Machines (SVMs) (Vapnik, 1979; Boser et al., 1992) are one of the most popular and successful techniques for classification. The key idea of the SVMs is to find a hyperplane in the feature space that separates the data points with maximal margin. For the case of non-separable samples, the soft margin SVM was introduced in Cortes and Vapnik (1995) where a penalty term for each non-separable sample is added to the objective function.

Let us first review the soft margin SVM. We consider the binary classification problem, where a finite number of training data $\{\boldsymbol{x}_j, y_j\}_{j=1,\cdots,m} \subset \mathbb{R}^n \times \{-1, 1\}$ is given and we are to find a linear classifier, $f(\boldsymbol{x}; \boldsymbol{w}, b) = \operatorname{sgn}\left(\boldsymbol{w}^\top \boldsymbol{x} + b\right)$. Suppose that the data are not linearly-separable. The soft margin SVM finds a vector $\boldsymbol{w}$ and a scalar b by solving the following quadratic program:

$$\min_{\boldsymbol{w},b,\boldsymbol{s}} \; \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C\sum_{j=1}^{m} s_j \tag{1a}$$

$$\text{s.t.} \;\; s_j \geq 1 - y_j(\boldsymbol{w}^\top \boldsymbol{x}_j + b), \quad j = 1, \cdots, m, \tag{1b}$$

$$\boldsymbol{s} \geq \mathbf{0}. \tag{1c}$$

Minimizing the first term of the objective function relates to the maximization of the margin between two classes. The second term corresponds to the penalty for the mis-classified samples.

Let us define $h(\boldsymbol{w}, b; \boldsymbol{x}, y) := \max\{1 - y(\boldsymbol{w}^\top \boldsymbol{x} + b), 0\}$, $\widehat{C} := Cm$, and reformulate the above optimization model as:

$$\min_{\boldsymbol{w},b} \; \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + \widehat{C}\sum_{j=1}^{m} h(\boldsymbol{w}, b; \boldsymbol{x}_j, y_j)\frac{1}{m}. \tag{2}$$

The function $h(\boldsymbol{w}, b; \boldsymbol{x}_j, y_j)$ can be regarded as the mis-classification error of the SVM classifier for the training sample $(\boldsymbol{x}_j, y_j)$. The error is in a form of the hinge loss, which equals zero for the correctly classified sample, and is linear in the distance to the classifier for the mis-classified sample. The soft margin SVM, therefore, finds the linear classifier that minimizes the hinge losses of the training data with the L_2-norm regularization.

Let $\boldsymbol{\xi} := (\boldsymbol{x}, y)$ represent the pair of the random feature vector and class label with a probability distribution $\mathbb{P}$ and support $\Xi := \mathcal{X} \times \{-1, 1\} \subset \mathbb{R}^n \times \{-1, 1\}$. Consider the following convex optimization problem:

$$\min_{\boldsymbol{w},b} \; \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + \widehat{C}\int_{\Xi} h(\boldsymbol{w}, b; \boldsymbol{\xi})\mathbb{P}(d\boldsymbol{\xi}). \tag{3}$$

Here, the optimization problem (3) searches for the linear classifier with minimal generalization error of mis-classification with the L_2-norm regularization. Under the assumption that the samples are drawn independently from the distribution $\mathbb{P}$, equation (2) is SAA of (3). In the soft margin SVM (2), the empirical distribution, defined by the samples, may be considered as a proxy for the true unknown distribution $\mathbb{P}$ and we solve the approximated classification problem.

Our distributionally-robust framework is motivated for situations where only a small number of samples are available, as is often the case in health studies, or the collected data is noisy. It is possible that the classifier obtained with training data in such situation may not generalize well to population, since the samples may not accurately represent the true distribution $\mathbb{P}$. Alternatively, a more generalized model as presented here may provide an improved SVM. This motivates our work on the distributionally-robust approach to the SVMs. We consider the situation where only a small number of samples are available for training.

To improve the generalization capability of the SVMs, we propose a distributionally-robust version of the SVMs. We assume the unknown population distribution $\mathbb{P}$ is within a prescribed set $\mathcal{P}$ of probability distributions, called the ambiguity set, and consider the problem of minimizing the generalization error under the worst-case probability distribution among the ambiguity set:

$$\min_{\boldsymbol{w},b} \ \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + \widehat{C} \sup_{\mathbb{P}\in\mathcal{P}} \int_{\Xi} h(\boldsymbol{w}, b; \boldsymbol{\xi})\mathbb{P}(d\boldsymbol{\xi}). \tag{4}$$

We call the above minimax problem the distributionally-robust Support Vector Machine (DR-SVM).

Methods to robustify classification have been proposed in the literature. A robust optimization approach based on the uncertainty sets (Ben-Tal et al., 2009) has been sought in classification problem in Bhattacharyya (2004); Shivaswamy et al. (2006); Xu et al. (2009). The motivation for the robust optimization approach is the robustness to perturbation or measurement uncertainty. Xu et al. (2009) establishes the equivalence between a regularization and robust optimization in the context of SVMs. Lanckriet et al. (2002) proposes a distributionally-robust optimization approach as in our paper, but the ambiguity set is based on the mean and covariance of the underlying class-conditional distribution. Our distributionally-robust optimization approach for SVMs differs from Lanckriet et al. (2002) as we construct an ϵ-ball ambiguity set based on a statistical distance function. We will discuss in the following section in details on the motivation of such ambiguity set and choice of specific statistical distance function, Kantorovich metric.

This paper is organized as follows: Section 2 discusses the possible choices for ambiguity set in DR-SVMs and describes the ambiguity set based on Kantorovich metric. Section 3 introduces the detailed formulation of DR-SVMs under the Kantorovich ambiguity set and presents a semi-infinite programming formulation for the DR-SVMs. Section 4 describes a cutting-plane algorithm to solve the DR-SVMs and presents a linear programming formulation for the separation problem with L_1 and L_∞ distance function in the Kantorovich measure. Section 5 presents our computational experience with a simulated and 12 small real data sets from the literature. Several of these datasets are from health studies conducted in the past. Section 6 concludes with discussion and future works.

2. Ambiguity sets and Kantorovich metric

Let us assume that the support $\Xi \subset \mathbb{R}^n \times \{-1, 1\}$ is bounded. Let $\mathcal{F}$ be a σ-algebra over Ξ, and $\mathcal{M}(\Xi)$ be the set of all probability measures on the probability space $(\Xi, \mathcal{F})$. Suppose a set of probability measures $\mathcal{M} \subset \mathcal{M}(\Xi)$ is equipped with a metric ρ.

We consider the ambiguity set of the following types:

$$\mathcal{P} = \left\{ \mathbb{P} \in \mathcal{M} \,\middle|\, \rho(\mathbb{P}, \widehat{\mathbb{P}}) \leq \epsilon \right\}. \tag{5}$$

Here, $\widehat{\mathbb{P}}$ is some reference probability measure in $\mathcal{M}$, and $\epsilon > 0$. The ambiguity set includes all probability measures that are within an ϵ distance from $\widehat{\mathbb{P}}$. The parameter ϵ represents the *budget* of the ambiguity.

For the reference measure $\widehat{\mathbb{P}}$, we use the empirical distribution $\mathbb{P}_m$ of $\boldsymbol{\xi}$, which is the discrete probability distribution for the samples $\{\boldsymbol{\xi}_j\}_{j=1,\cdots,m}$, with equal weights of $1/m$. It is well-known that the empirical distribution $\mathbb{P}_m$ converges to $\mathbb{P}$ in various modes of convergence (Van der Vaart and Wellner, 1996).

2.1 Kantorovich metric

For the metric ρ, we adopt the Kantorovich metric (Villani, 2009) among many statistical distances. The Kantorovich metric, also called the Wasserstein metric of order 1, is a metric for a suitable space of probability measures that are defined on a metric space.

The definition of the Kantorovich metric in our context is as follows. Suppose the bounded support Ξ is equipped with a metric $d_{\boldsymbol{\xi}}$. The metric space $(\Xi, d_{\boldsymbol{\xi}})$ is Polish. Let $\mathcal{F}$ be the Borel σ-algebra over Ξ and $\mathcal{M}$ be the set of probability measures on the probability space $(\Xi, \mathcal{F})$ defined as

$$\mathcal{M} := \{\mathbb{P} \in \mathcal{M}(\Xi) \mid \int_{\Xi} d_{\boldsymbol{\xi}}(\boldsymbol{\xi}^0, \boldsymbol{\xi})\mathbb{P}(d\boldsymbol{\xi}) < \infty\}, \tag{6}$$

for any $\boldsymbol{\xi}^0 \in \Xi$. The Kantorovich metric between two probability measures $\mathbb{P}^1, \mathbb{P}^2 \in \mathcal{M}$ is defined as :

$$\begin{aligned}\rho(\mathbb{P}^1, \mathbb{P}^2) := \inf_{\mathbb{K}} \Big\{ \int_{\Xi \times \Xi} d_{\boldsymbol{\xi}}(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2)\mathbb{K}(d\boldsymbol{\xi}^1, d\boldsymbol{\xi}^2) \ \Big| \\ \int_{\Xi} \mathbb{K}(\boldsymbol{\xi}^1, d\boldsymbol{\xi}^2) = \mathbb{P}^1(\boldsymbol{\xi}^1), \int_{\Xi} \mathbb{K}(d\boldsymbol{\xi}^1, \boldsymbol{\xi}^2) = \mathbb{P}^2(\boldsymbol{\xi}^2), \forall \boldsymbol{\xi}^1, \boldsymbol{\xi}^2 \in \Xi \Big\},\end{aligned} \tag{7}$$

where $\mathbb{K}$ is a probability measure defined on $\Xi \times \Xi$ whose marginals are $\mathbb{P}^1$ and $\mathbb{P}^2$, respectively (Villani, 2009).

We further assume that the metric $d_{\boldsymbol{\xi}}$ on $\Xi = \mathcal{X} \times \{-1, 1\}$ to be L_1 norm product metric of the metric d on the feature space $\mathcal{X}$ and the discrete metric d_y on $\{-1, 1\}$. That is, for any $\boldsymbol{\xi}^1 = (\boldsymbol{x}^1, y^1), \boldsymbol{\xi}^2 = (\boldsymbol{x}^2, y^2) \in \Xi$, we have

$$d_{\boldsymbol{\xi}}(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2) := d(\boldsymbol{x}^1, \boldsymbol{x}^2) + d_y(y^1, y^2), \tag{8}$$

where d_y is defined as:

$$d_y(y^1, y^2) := \begin{cases} \delta & \text{if } y^1 \neq y^2, \\ 0 & \text{if } y^1 = y^2. \end{cases} \tag{9}$$

The function $d_{\boldsymbol{\xi}}(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2)$ may be taken a L_1, L_2, or L_∞ norms. The parameter $\delta > 0$ determines the relative importance of the distance between two classes compared to the distance in the feature space.

2.2 Alternative choices of metric in probability measures and ambiguity sets

Alternative definition of $\mathcal{P}$ with different choice of the metric function may be possible, as there are several definitions for the statistical distances studied in the literature. Among them are discrepancy metric, Hellinger distance, Kullback-Leibler divergence, Kolmogorov metric, Lévi metric, Prokhorov metric, separation distance, total variation distance, general Wasserstein metric, and χ^2-distance (Gibbs and Su, 2002).

Some of the popular statistical distances are less appropriate for ρ in the ambiguity set (5) for the DR-SVMs. Kullback-Leibler divergence and χ^2-distance are technically not metrics because they are not symmetric. Kolmogorov metric and Lévi metric are only defined on $\mathbb{R}$ and separation distance is only defined on a countable set. Other metrics, such as discrepancy metric, Prokhorov metric and total variation distance, are relatively difficult to compute as the definitions involve all closed balls or Borel sets of the probability space. Hellinger distance and the general Wasserstein metric with order $p > 1$ involve non-linearity in the definition, hence may not be practical for our purpose.

From the theoretical and practical standpoint, the Kantorovich distance is a good candidate for the choice of the metric for the following reasons. It is a metric and relatively easy to compute. For example, the Kantorovich distance between two discrete probability distributions on the same probability space can be computed by solving a transportation problem. As we show later, the Kantorovich metric leads to a tractable DR-SVM formulation. The Kantorovich metric is also defined on any metric space. This is particularly important as it allows us to extend the DR-SVMs with kernelization. The metric is also used in the previous works on distributionally-robust optimization (Pflug and Wozabal, 2007; Mehrotra and Zhang, 2014).

Other types of ambiguity sets are used for the distributionally-robust models in the literature. Most notably, the ambiguity sets based on moment conditions, especially in regards to the mean and covariance matrix, are studied in Scarf (1959); Bertsimas et al. (2010); Delage and Ye (2010); Ghaoui et al. (2003); Mehrotra and Zhang (2014). However, the ambiguity sets based on the moment conditions (Lanckriet et al., 2002) may be less practical for the machine learning problems as the dimension of the feature vector is typically large and the existing distributionally-robust optimization techniques based on the moment ambiguity sets can become computationally challenging. We show that under the given ϵ-ball ambiguity set (5) with a certain choice of metric ρ, the DR-SVMs remain computationally tractable.

2.3 Choice of the ambiguity parameter

A large value of ϵ may be suitable when we do not have many samples for training or the data is very noisy. A small value for ϵ, on the other hand, is suitable when we are more confident that the empirical distribution $\mathbb{P}_m$ is near the true distribution $\mathbb{P}$. But, the choice of the parameter ϵ in practice is rather subjective and it is difficult to know what would be the optimal value of the parameter.

A non-asymptotic convergence analysis can provide some guidance for a good choice of ϵ. It is shown that $\mathbb{P}_m$ converges to $\mathbb{P}$ as $m \to \infty$ when measured in the Wasserstien metric (Fournier and Guillin, 2013) and the reference also provides a detailed convergence analysis. For this purpose, we re-state the results in Fournier and Guillin (2013).

Theorem 1 *Let $\mathbb{P}_m$ be the empirical measure of m samples independently drawn from a unknown measure $\mathbb{P}$ and $n \geq 2$. Then, there exists a constant K such that, for all $m \geq 1$,*

$$\mathbb{E}[\rho(\mathbb{P}, \mathbb{P}_m)] \leq K\sqrt{2}^n m^{-1/(n+1)}. \tag{10}$$

Proof The Kantorovich metric is the Wasserstein metric of order $p = 1$. With $n + 1 \geq 3$, only the third case of the bound in Theorem 1 of Fournier and Guillin (2013) applies. Without loss of generality, we assume $\Xi \subset [-1, 1]^{n+1}$, since Ξ is compact. Then, we have $M_q(\mathbb{P}) := \int_\Xi |\boldsymbol{\xi}|^q \mathbb{P}(d\boldsymbol{\xi}) \leq 1$ for any $q > 1$. For sufficiently large q, the second term of the bound in Theorem 1 of Fournier and Guillin (2013) can be ignored. The exponential term

$\sqrt{2}^n$ follows from Lemma 5 in Fournier and Guillin (2013). ■

Equation (1) gives a bound on the expected Kantorovich distance between the unknown distribution $\mathbb{P}$ and the empirical distribution $\mathbb{P}_m$ that depends on the dimension of the feature vector n and the number of samples m.

In our empirical study, we use in-sample parameter grid search to heuristically select ϵ.

3. A reformulation of distributionally-robust Support Vector Machines with Kantorovich metric

The DR-SVM (4) can be reformulated as the following convex semi-infinite program by introducing a new variable z:

$$\min_{\boldsymbol{w},b,z} \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + z \tag{11a}$$

$$\text{s.t. } z \geq \widehat{C} \int_\Xi h(\boldsymbol{w}, b; \boldsymbol{\xi})\mathbb{P}(d\boldsymbol{\xi}), \quad \forall \mathbb{P} \in \mathcal{P}. \tag{11b}$$

We have a semi-infinite number of constraints in (11b), one for each $\mathbb{P} \in \mathcal{P}$.

An approach to solve (11) is to use the cutting plane method in Mehrotra and Papp (2014). The problem of finding a separating convex constraint defined by a distribution $\mathbb{P} \in \mathcal{P}$:

$$\sup_{\mathbb{P}\in\mathcal{P}} \int_\Xi h(\boldsymbol{w}, b; \boldsymbol{\xi})\mathbb{P}(d\boldsymbol{\xi}) \tag{12}$$

is a generalized moment problem with a finite number of moment conditions (See (15) below). For a given solution $\widehat{\boldsymbol{w}}$, we can obtain an optimal solution for the separation problem with a finite support and generate a cut which consists of a finite number of additional variables and linear constraints.

On the other hand, we can use the dual of the inner problem and obtain a monolithic formulation for the DR-SVMs. The following lemma gives a dual of (12) and the strong duality for this dual.

Theorem 2 *Suppose the ambiguity set $\mathcal{P}$ is defined as in Section 2. Then, the inner problem (12) of the DR-SVM is finite and it is equivalent to the following semi-infinite program:*

$$\min_{\boldsymbol{t},u} \frac{1}{m}\sum_{j=1}^m t_j + \epsilon u \tag{13a}$$

$$s.t.\ t_j \geq 1 - y(\boldsymbol{w}^\top \boldsymbol{x} + b) - [d(\boldsymbol{x}, \boldsymbol{x}_j) + d_y(y, y_j)]\, u, \quad (\boldsymbol{x}, y) \in \Xi,\ j = 1, \cdots, m, \tag{13b}$$

$$\boldsymbol{t} \geq \boldsymbol{0}, u \geq 0. \tag{13c}$$

Proof The ambiguity set (5) with the Kantorovich metric and the empirical distribution can be represented as:

$$\mathcal{P} = \Big\{\mathbb{P} \in \mathcal{M} \;\Big|\; \exists\mathbb{K} \text{ s.t.} \sum_{j=1}^m \mathbb{K}(\boldsymbol{\xi}, \boldsymbol{\xi}_j) = \mathbb{P}(\boldsymbol{\xi}), \forall \boldsymbol{\xi} \in \Xi, \\ \int_\Xi \mathbb{K}(d\boldsymbol{\xi}, \boldsymbol{\xi}_j) = \frac{1}{m}, \forall j,\ \int_\Xi \sum_{j=1}^m d_{\boldsymbol{\xi}}(\boldsymbol{\xi}, \boldsymbol{\xi}_j)\mathbb{K}(d\boldsymbol{\xi}, \boldsymbol{\xi}_j) \leq \epsilon\Big\}, \tag{14}$$

where $\mathbb{K}$ is a probability measure defined on $\Xi \times \widehat{\Xi}$. Then, the inner problem (12) is formulated as the following conic linear program:

$$\sup_{\mathbb{K}\in\mathcal{M}(\Xi\times\widehat{\Xi})} \int_{\Xi} \sum_{j=1}^{m} h(\boldsymbol{w}, b; \boldsymbol{\xi})\mathbb{K}(d\boldsymbol{\xi}, \boldsymbol{\xi}_j) \tag{15a}$$

$$\text{s.t.} \int_{\Xi} \mathbb{K}(d\boldsymbol{\xi}, \boldsymbol{\xi}_j) = \frac{1}{m}, \quad j = 1, \cdots, m, \tag{15b}$$

$$\int_{\Xi} \sum_{j=1}^{m} d_{\boldsymbol{\xi}}(\boldsymbol{\xi}, \boldsymbol{\xi}_j)\mathbb{K}(d\boldsymbol{\xi}, \boldsymbol{\xi}_j) \leq \epsilon, \tag{15c}$$

$$\mathbb{K} \geq 0. \tag{15d}$$

Since $\mathcal{F}$ is the Borel σ-algebra, every finite subset of Ξ is $\mathcal{F}$-measurable. Thus, we obtain the following sample-based dual problem of (15) (Shapiro, 2001):

$$\min_{\boldsymbol{t},u} \frac{1}{m}\sum_{j=1}^{m} t_j + \epsilon u \tag{16a}$$

$$\text{s.t. } t_j + d_{\boldsymbol{\xi}}(\boldsymbol{\xi}, \boldsymbol{\xi}_j)u \geq h(\boldsymbol{w}, b; \boldsymbol{\xi}), \quad \boldsymbol{\xi} \in \Xi,\ j = 1, \cdots, m, \tag{16b}$$

$$u \geq 0. \tag{16c}$$

Note that (16b) can be written as:

$$t_j + [d(\boldsymbol{x}, \boldsymbol{x}_j) + d_y(y, y_j)]\, u \geq 1 - y(\boldsymbol{w}^\top\boldsymbol{x} + b), \quad (\boldsymbol{x}, y) \in \Xi,\ j = 1, \cdots, m \tag{17a}$$

$$t_j + [d(\boldsymbol{x}, \boldsymbol{x}_j) + d_y(y, y_j)]\, u \geq 0, \quad (\boldsymbol{x}, y) \in \Xi,\ j = 1, \cdots, m. \tag{17b}$$

Equation (17b) implies that $\boldsymbol{t} \geq \boldsymbol{0}$, since $u \geq 0$ and $(\boldsymbol{x}_j, y_j) \in \Xi$. Therefore, (16) is equivalent to (13).

Note that $\mathbb{K} = \mathbb{P}^* \times \mathbb{P}^*$ is a strictly feasible solution of (15) as we assume $\epsilon > 0$. This satisfies the Slater condition and there is no duality gap between (13) and (15). Furthermore, we assume that $\mathcal{X}$ is bounded. Hence, the feasible set of (13) is non-empty. Therefore, the optimal objective value of (13) is finite. ■

We now give a monolithic formulation of distributionally-robust Support Vector Machines.

Corollary 3 *Suppose the ambiguity set is defined as in Section 2. Then, the DR-SVM (4) is equivalent to:*

$$\min_{\boldsymbol{w},b,\boldsymbol{t},u} \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} + \widehat{C}\left\{\frac{1}{m}\sum_{j=1}^{m} t_j + \epsilon u\right\} \tag{18a}$$

$$\textit{s.t. } t_j \geq 1 - y(\boldsymbol{w}^\top\boldsymbol{x} + b) - [d(\boldsymbol{x}, \boldsymbol{x}_j) + d_y(y, y_j)]\, u, \quad (\boldsymbol{x}, y) \in \Xi,\ j = 1, \cdots, m, \tag{18b}$$

$$\boldsymbol{t} \geq \boldsymbol{0}, u \geq 0. \tag{18c}$$

Proof The result follows by combing the dual of the inner problem from Theorem 2 with the outer minimization problem of (4). ∎

Note that the feasible set of (18) is non-empty and the optimal objective is bounded below. In Theorem 3, the inner problem of the DR-SVM is in its dual form and the DR-SVM is formulated as a convex semi-infinite program with infinitely-many linear constraints.

4. Algorithms to solve the distributionally-robust Support Vector Machines

In Section 3, we provided a monolithic formulation of the DR-SVM (18) which is a convex quadratic program with infinitely many linear constraints. We now provide a cutting-plane algorithm which is introduced below in detail.

4.1 Cutting-plane algorithm

The problem can be solved using a cutting-plane algorithm in which the following finite version of (18) is solved iteratively:

$$\min_{\boldsymbol{w},b,\boldsymbol{t},u} \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} + \widehat{C}\left\{\frac{1}{m}\sum_{j=1}^{m} t_j + \epsilon u\right\} \tag{19a}$$

$$\begin{aligned}\text{s.t.}\quad & t_j \geq 1 - y(\boldsymbol{w}^\top \boldsymbol{x}_{(j,k)} + b) - \left[d(\boldsymbol{x}_{(j,k)}, \boldsymbol{x}_j) + d_y(y_{(j,k)}, y_j)\right] u, \\ & \qquad j = 1, \cdots, m,\ k = 1, \cdots, K(j)\end{aligned} \tag{19b}$$

$$\boldsymbol{t} \geq \boldsymbol{0}, u \geq 0. \tag{19c}$$

Note that each row in (19b) is indexed by the tuple (j,k). We initialize the above problem with the training data, that is, $K(j) := 1$, $\boldsymbol{x}_{(j,1)} := \boldsymbol{x}_j$ and $y_{(j,1)} := y_j$, $j = 1, \cdots, m$. Then, the initial problem is identical to the soft margin SVM.

Let $\boldsymbol{w}^{(l)}, b^{(l)}, \boldsymbol{t}^{(l)}, u^{(l)}$ be an optimal solution of (19) at l-th iteration. Then, we have m separation problems to solve, one for each sample $(\boldsymbol{x}_j, y_j)$:

$$\theta_j(\boldsymbol{w}^{(l)}, b^{(l)}, \boldsymbol{t}^{(l)}, u^{(l)}) := \max_{\boldsymbol{x},y\in\Xi}\ 1 - y(\boldsymbol{w}^{(l)\top}\boldsymbol{x} + b^{(l)}) - t_j^{(l)} - \left[d(\boldsymbol{x}, \boldsymbol{x}_j) + d_y(y, y_j)\right] u^{(l)}. \tag{20}$$

For any j with $\theta_j > 0$, we increase $K(j)$ by one and take the optimal solution of the j-th separation problem as $(\boldsymbol{x}_{(j,K(j))}, y_{(j,K(j))})$. If $\boldsymbol{\theta} \leq \boldsymbol{0}$, then we stop and concluded that the optimal solution of the DR-SVM (18) is found. The cutting-plane algorithm is summarized in Algorithm 1. The convergence of the cutting-plane algorithm can be found in Gustafson and Kortanek (1973).

4.2 Separation problems

The separation problem (20) is equivalent to:

$$\max\{\theta_j^+, \theta_j^-\}, \tag{21}$$

where

$$\theta_j^+ := \max_{\boldsymbol{x}\in\mathcal{X}}\ 1 - y_j(\boldsymbol{w}^{(l)\top}\boldsymbol{x} + b^{(l)}) - t_j^{(l)} - u^{(l)} d(\boldsymbol{x}, \boldsymbol{x}_j), \tag{22}$$

Algorithm 1 Cutting-plane algorithm for the DR-SVM (18)

```
(Initialization) Set $K(j) \leftarrow 1$, $\boldsymbol{x}_{(j,1)} \leftarrow \boldsymbol{x}_j$ and $y_{(j,1)} \leftarrow y_j$, $j = 1, \cdots, m$. Set $l \leftarrow 0$.
Solve the problem (19) to obtain an optimal solution $\boldsymbol{w}^{(l)}, b^{(l)}, \boldsymbol{t}^{(l)}, u^{(l)}$.
for $j = 1, \cdots, m$ do
    Solve the $j$-th separation problem in (20).
    if $\theta_j > 0$ then
        Update $K(j) \leftarrow K(j) + 1$.
        Set $(\boldsymbol{x}_{(j,K(j))}, y_{(j,K(j))}) \leftarrow (\boldsymbol{x}, y)$, which is an optimal solution of $j$-th separation problem in (20).
    end if
end for
if no cut is added in the current iteration, then
    STOP.
end if
Update $l \leftarrow l + 1$ and go to Step 2.
```

and

$$\theta_j^- := \max_{\boldsymbol{x} \in \mathcal{X}} \ 1 + y_j(\boldsymbol{w}^{(l)\top}\boldsymbol{x} + b^{(l)}) - t_j^{(l)} - u^{(l)} \left[d(\boldsymbol{x}, \boldsymbol{x}_j) + \delta\right]. \tag{23}$$

They are both concave maximization problem over the feature space, since the metric d is convex. Therefore, the problem is tractable if we assume that the feature space is a convex set.

If we assume that the feature space is a polyhedron and the metric d is L_1 or L_∞ metric, then both (22) and (23) can be formulated as linear programs.

Proposition 4 *Assume that the metric d for $\mathcal{X}$ is L_1 metric and the support $\mathcal{X}$ is given by a polyhedron. Then, θ_j^+ and θ_j^- are the optimal objective values of the following linear programs:*

$$\theta_j^+ := \max_{\boldsymbol{x}, d} 1 - y_j(\boldsymbol{w}^{(l)\top}\boldsymbol{x} + b^{(l)}) - t_j^{(l)} - u^{(l)} \left(\sum_{i=1}^n d_i \right) \tag{24a}$$

$$s.t. \ -d_i \leq x_i - x_{ji} \leq d_i, \quad i = 1, \cdots, n, \tag{24b}$$

$$\boldsymbol{x} \in \mathcal{X} \tag{24c}$$

and

$$\theta_j^- := \max_{\boldsymbol{x}, d} 1 + y_j(\boldsymbol{w}^{(l)\top}\boldsymbol{x} + b^{(l)}) - t_j^{(l)} - u^{(l)} \left(\sum_{i=1}^n d_i + \delta \right) \tag{25a}$$

$$s.t. \ -d_i \leq x_i - x_{ji} \leq d_i, \quad i = 1, \cdots, n, \tag{25b}$$

$$\boldsymbol{x} \in \mathcal{X}. \tag{25c}$$

Proposition 5 *Assume that the metric d for $\mathcal{X}$ is L_∞ metric and the support $\mathcal{X}$ is given by a polyhedron. Then, θ_j^+ and θ_j^- are the optimal objective values of the following linear* *programs:*

$$
\begin{aligned}
\theta_j^+ := \max_{\boldsymbol{x},d}\ & 1 - y_j(\boldsymbol{w}^{(l)\top}\boldsymbol{x} + b^{(l)}) - t_j^{(l)} - u^{(l)}\,(d_0) && \text{(26a)}\\
s.t.\ & -d_i \le x_i - x_{ji} \le d_i, \quad i = 1, \cdots, n, && \text{(26b)}\\
& d_0 \ge d_i, \quad i = 1, \cdots, n, && \text{(26c)}\\
& \boldsymbol{x} \in \mathcal{X} && \text{(26d)}
\end{aligned}
$$

and

$$
\begin{aligned}
\theta_j^- := \max_{\boldsymbol{x},d}\ & 1 + y_j(\boldsymbol{w}^{(l)\top}\boldsymbol{x} + b^{(l)}) - t_j^{(l)} - u^{(l)}\,(d_0 + \delta) && \text{(27a)}\\
s.t.\ & -d_i \le x_i - x_{ji} \le d_i, \quad i = 1, \cdots, n, && \text{(27b)}\\
& d_0 \ge d_i, \quad i = 1, \cdots, n, && \text{(27c)}\\
& \boldsymbol{x} \in \mathcal{X}. && \text{(27d)}
\end{aligned}
$$

5. Computational experiments

We now present empirical findings on the performance of the proposed DR-SVMs and compare it to the soft margin SVMs. We also present the comparison to the R-SVMs proposed in Xu et al. (2009). We used hypothetical data sets based on the Gaussian distribution to study the performance of the DR-SVMs and to conduct sensitivity analysis. We also illustrate the effectiveness of the DR-SVMs using several real-world data from UCI Machine Learning Repository (Bache and Lichman, 2013).

The computational environment used in the study was a laptop with 2.53GHz CPU and 8GB RAM. The algorithms were implemented in Python 2.7 and CPLEX 12.5 was used to solve the quadratic and linear programs in the algorithms.

5.1 Test data set

We created a small test problem to illustrate the properties of the DR-SVMs compared to the SVMs. The hypothetical data was based on two 2-dimensional Gaussian distributions. We assumed that the data of the class with label 1 were distributed with the mean $\boldsymbol{\mu}_1 = (0.3, 0.2)$ and the covariance $\Sigma_1 = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}$ but truncated to be within $\{\boldsymbol{x} \in \mathbb{R}^2 \mid (-1, -1) \le \boldsymbol{x} \le (1, 1)\}$, and the data of the class with label -1 were distributed with the mean $\boldsymbol{\mu}_2 = (-0.3, -0.2)$ and the same covariance $\Sigma_2 = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}$ but truncated to be within the same box. The class distribution was the Bernoulli distribution with parameter $p = 0.5$.

5.2 Results for test data

For the initial test, we randomly generated 50 sample points for training data and 1,000 sample points for test data. We trained both the SVM and DR-SVM and computed the Area Under Curve (AUC) measure from the Receiver Operating Characteristic (ROC) curve generated from the trained model and the test data. The AUC is received as a better

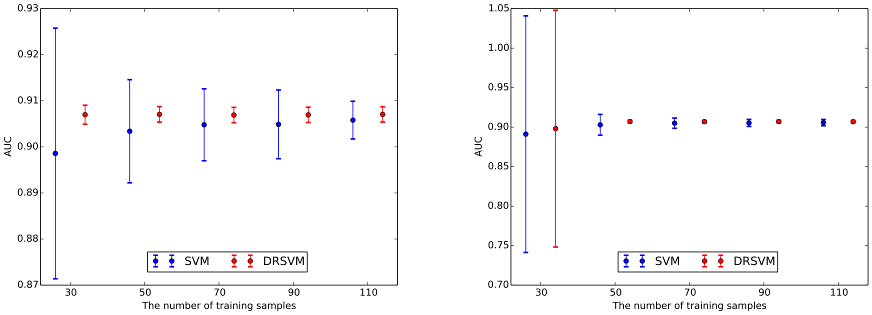
Table 1: Mean performance comparison between the SVMs and DR-SVMs for test data

	AUC	(S.E.)
SVM	0.9039	0.0046
DR-SVM with L_1-norm	0.9069	0.0008
DR-SVM with L_∞-norm	0.9069	0.0008

measure for the classification problem than the accuracy (Ling et al., 2003), particularly in the healthcare setting. We repeated the experiment 100 times to get the average and the standard errors of the AUC measure for SVMs and DR-SVMs. We used $\widehat{C} = 150, \epsilon = 0.1, \delta = 0.1$ and both L_1 and L_∞ norms were tested.

The performance comparison between the SVMs and DR-SVMs in Table 1 shows that the DR-SVMs and SVMs have similar performance in terms of the AUC measure, however, DR-SVMs have lower standard errors (denoted by S.E. in the table) for the mean performance which implies that the performance of DR-SVMs are more consistent.

To see how the number of training data m affects the performance of the classification models, we did the same experiments with different number of samples in the training data. The results in Figure 1 show that the confidence intervals of the AUC measures in DR-SVMs are significantly smaller than those in SVMs, especially when the number of training data m is small. The results, therefore, suggest that DR-SVMs are likely to provide a more robust classifier, without a significant trade-off in classification accuracy.



(a) L_1-norm

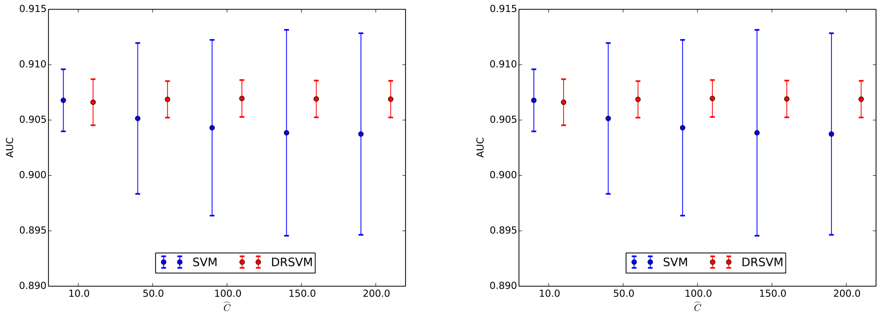
(b) L_∞-norm

Figure 1: Sensitivity analysis for the number of training samples m with 95% C.I.

5.3 Sensitivity analysis

We performed the sensitivity analysis for the parameters of the DR-SVMs as follows. First, we randomly generated the training data of 50 samples. For the parameter $\widehat{C}$, we trained both the soft margin SVM and the DR-SVMs with different values of the parameter. For the parameters ϵ and δ, we trained the DR-SVMs with different values of the parameters and additionally the soft margin SVM for a comparison. Then, we tested each trained model against the pre-generated test data of 1,000 samples. We repeated this process 100 times to estimate the mean and the standard errors for the AUC and accuracy measures. The experiment was conducted both with L_1 norm and L_∞ norm assumption.

Figure 2, when $\widehat{C}$ becomes too large, in other words, the mis-classification term is weighted more than the regularization terms, the performance of both the SVM and the DR-SVM deteriorate. This implies that the regularization still plays a role in the DR-SVM setting. However, DR-SVM is less sensitive to the values of $\widehat{C}$.

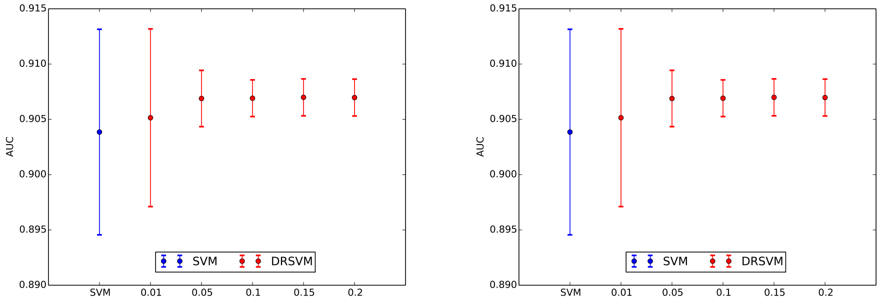


(a) L_1-norm

(b) L_∞-norm

Figure 2: Sensitivity analysis for $\widehat{C}$ ($\epsilon = 0.1$, $\delta = 0.1$) with 95% C.I.

The sensitivity analysis for the budget of ambiguity ϵ is shown in Figure 3. We find that when ϵ is small, the performance of the DR-SVMs are similar to SVMs, as expected. If ϵ is too large, the accuracy of the classifier from the DR-SVM worsens. The sensitivity analysis suggests that the DR-SVMs may provide an improvement in the SVMs after a proper tuning of the Kantorovich distance parameter ϵ.



(a) L_1-norm

(b) L_∞-norm

Figure 3: Sensitivity analysis for ϵ ($\widehat{C} = 150$, $\delta = 0.1$) with 95% C.I.

5.4 Results for real-world data

We compared the performance of DR-SVMs to SVMs and R-SVMs using several real-world data from UCI Machine Learning Repository (Bache and Lichman, 2013). Among the data

Table 2: Summary of data sets from UCI Machine Learning Repository

	Num of observations	Num of variables
Ionosphere	351	34
EEG eye state	14980	14
Statlog heart	270	12
SPECT heart	267	22
SPECTF heart	267	22
Pima Indians diabetes	768	8
Breast cancer Wisconsin	699	9
Banknote authentication	1372	4
Vertebral column	310	6
Connectionist bench	208	60
Climate model simulation crashes	540	18
Spambase	4601	57

sets in the repositories, we chose those for binary classification problem with a small feature space. The summary of 12 selected data set is shown in Table 2.

To compare the model performances on small samples, we used the repeated random sub-sampling validation approach. We randomly selected small subset, of size $n = 50, 75, 100, 150$, of available data for training and set aside the remaining data for testing. We selected the parameter $\widehat{C}$ for SVMs and parameter $\widehat{C}$ and ϵ for DR-SVMs and R-SVMs using a grid search. The candidate values for the parameter $\widehat{C}$ were $[100.0, 150.0, 200.0, 250.0]$, and for ϵ, $[0.0, 0.01, 0.05, 0.1, 0.15, 0.2, 0.5]$. When $\epsilon = 0$, the DR-SVMs and R-SVMs are the same as the classical SVMs. We did not allow the change in labels when perturbing the samples to find the worst-case probability distribution by setting the parameter δ arbitrary large $(1,000,000)$. The AUC measeure on training data was used as the criteria to select the best parameter. Once the best parameter setting was chosen, the models with the selected parameters applied to test data set and evaluated AUC measure. This entire process was repeated 100 times to find the average and standard errors of the AUC on test data set. The random sampling were *stratified*: the proportion of the class labels in the sub-samples were kept roughly equal to one in the original data set.

The average out-of-sample AUC measures between SVMs and DR-SVMs are summarized in Table 3. The comparison between SVMs and R-SVMs are shown in Table 4. We observe that DR-SVMs outperform the SVMs in terms of AUC for data sets such as Ionosphere, EEG eye state, Statlog heart and SPECT heart. They, however, give no statistically significant improvement in data sets such as SPECTF heart, Pima Indians diabetes, Breast cancer Wisconsin, Banknote authentication, Vertebral column and Connectionist bench. Only in one case of Climate model crashes data set, the performance of DR-SVMs is significantly worse and the results for Spambase dataset are mixed, depending on the distance function used in DR-SVMs. On the other hand, the performance of R-SVMs was found similar to that of the classical SVMs. This may be because the quadratic term in the objective function of the SVMs already serves as L_2-norm regularization and R-SVMs with L_1 or

L_∞ uncertainty set are equivalent to the classical SVMs with additional L_∞ or L_1-norm regularization term, respectively.

6. Conclusion and future work

We propose distributionally-robust SVMs, motivated to improve the generalization error. We find the semi-infinite convex formulation for the DR-SVMs with the convex quadratic objective function and infinitely-many linear constraints, which is solved through the cutting-plane algorithm. When L_1 or L_∞-norm is used in Kantorovich metric, we show that the separation problem can be formulated as linear programming. This makes the proposed DR-SVMs computationally tractable. We show that the DR-SVMs have improved generalization capabilities than the SVMs using simulated data as well as some real-world data in terms of the AUC measure.

In the next step, we will study te dual version of the DR-SVMs to incorporate the kernelization. Using the inner-product induced metric for the (higher-dimensional) mapped feature space, we can define the Kantorovich metric for the probability measures in the mapped feature space. Also, the inner-product induced metric can be computed using the kernel functions and so the kernel trick can be extended to the DR-SVMs. A sequential minimal optimization (SMO) type algorithm for the DR-SVMs will be developed.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (NSF CMMI 1100868).

Table 3: Mean performance comparison between the SVMs and DR-SVMs for real-world data

		SVM		DR-SVM with L_∞				DR-SVM with L_1			
	n	AUC	s.e.	AUC	s.e.	diff	p.val	AUC	s.e.	diff	p.val
Ionosphere	50	.8156	.0054	.8203	.0058	.0058	.2769	.8198	.0050	.0051	.2844
	75	.8388	.0037	.8597	.0063	.0249	.0023	.8391	.0037	.0004	.4772
	100	.8520	.0033	.8955	.0071	.0511	.0000	.8529	.0032	.0011	.4225
	150	.8712	.0029	.9284	.0068	.0657	.0000	.8718	.0028	.0007	.4409
EEG eye state	50	.7733	.0065	.7992	.0090	.0335	.0103	.8470	.0107	.0953	.0000
	75	.7710	.0066	.7892	.0090	.0236	.0523	.8295	.0112	.0759	.0000
	100	.7854	.0073	.7956	.0081	.0130	.1754	.8354	.0106	.0637	.0001
	150	.7923	.0063	.8076	.0079	.0193	.0658	.8455	.0105	.0671	.0000
Statlog heart	50	.8565	.0033	.8733	.0040	.0196	.0007	.8609	.0034	.0051	.1771
	75	.8836	.0021	.8941	.0026	.0119	.0010	.8858	.0020	.0025	.2245
	100	.8906	.0019	.9043	.0025	.0154	.0000	.8917	.0019	.0012	.3414
	150	.9007	.0022	.9112	.0026	.0117	.0012	.9009	.0022	.0002	.4744
SPECT heart	50	.7737	.0067	.7975	.0092	.0308	.0189	.7901	.0084	.0212	.0643
	75	.7878	.0033	.8119	.0064	.0306	.0005	.8121	.0068	.0308	.0008
	100	.8006	.0037	.8104	.0083	.0122	.1411	.8216	.0064	.0262	.0025
	150	.8183	.0071	.8306	.0089	.0150	.1406	.8343	.0093	.0196	.0865
SPECTF heart	50	.7771	.0037	.7776	.0037	.0006	.4620	.7767	.0038	-.0005	.4700
	75	.7957	.0030	.7963	.0030	.0008	.4438	.7961	.0031	.0005	.4631
	100	.8054	.0027	.8049	.0027	-.0006	.4480	.8064	.0029	.0012	.4005
	150	.8100	.0035	.8088	.0036	-.0015	.4057	.8126	.0033	.0032	.2947
Diabetes	50	.7894	.0033	.7918	.0032	.0030	.3011	.7858	.0034	-.0046	.2241
	75	.8016	.0027	.8029	.0025	.0016	.3621	.8002	.0025	-.0017	.3520
	100	.8095	.0016	.8094	.0016	-.0001	.4824	.8078	.0017	-.0021	.2337
	150	.8210	.0013	.8214	.0012	.0005	.4107	.8193	.0014	-.0021	.1873
Breast cancer	50	.9870	.0009	.9875	.0011	.0005	.3627	.9884	.0009	.0014	.1363
	75	.9898	.0005	.9890	.0008	-.0008	.1987	.9906	.0005	.0008	.1296
	100	.9906	.0003	.9907	.0003	.0001	.4070	.9913	.0003	.0007	.0503
	150	.9915	.0003	.9915	.0003	.0000	.5000	.9917	.0003	.0002	.3189
Banknote	50	.9981	.0001	.9973	.0006	-.0008	.0950	.9979	.0001	-.0002	.0794
	75	.9983	.0001	.9981	.0001	-.0002	.0794	.9984	.0001	.0001	.2402
	100	.9984	.0001	.9984	.0001	.0000	.5000	.9984	.0001	.0000	.5000
	150	.9985	.0001	.9984	.0001	-.0001	.2402	.9985	.0001	.0000	.5000
Vertebral column	50	.8674	.0027	.8664	.0030	-.0012	.4023	.8657	.0029	-.0020	.3342
	75	.8744	.0029	.8752	.0032	.0009	.4266	.8736	.0029	-.0009	.4228
	100	.8763	.0021	.8763	.0022	.0000	.5000	.8765	.0021	.0002	.4732
	150	.8827	.0028	.8831	.0028	.0005	.4598	.8827	.0028	.0000	.5000
Connectionist	50	.7887	.0035	.7890	.0034	.0004	.4755	.7902	.0035	.0019	.3811
	75	.8112	.0034	.8106	.0034	-.0007	.4504	.8114	.0033	.0002	.4832
	100	.8177	.0034	.8178	.0034	.0001	.4917	.8179	.0035	.0002	.4837
	150	.8243	.0051	.8243	.0051	.0000	.5000	.8240	.0051	-.0004	.4834
Climate model	50	.8352	.0068	.8415	.0062	.0075	.2472	.8291	.0064	-.0073	.2572
	75	.8444	.0075	.8517	.0072	.0086	.2417	.8407	.0080	-.0044	.3681
	100	.8720	.0083	.8677	.0084	-.0049	.3581	.8583	.0085	-.0157	.1251
	150	.7738	.0235	.7555	.0242	-.0236	.2940	.7607	.0242	-.0169	.3491
Spambase	50	.9022	.0018	.9009	.0018	-.0014	.3051	.9062	.0017	.0044	.0539
	75	.9180	.0014	.9103	.0093	-.0084	.2070	.9204	.0012	.0026	.0973
	100	.9235	.0011	.9109	.0132	-.0136	.1713	.9255	.0009	.0022	.0805
	150	.9298	.0009	.8999	.0183	-.0322	.0521	.9321	.0007	.0025	.0225

Table 4: Mean performance comparison between the SVMs and R-SVMs for real-world data

		SVM		R-SVM with L_∞				R-SVM with L_1			
	n	AUC	s.e.	AUC	s.e.	diff	p.val	AUC	s.e.	diff	p.val
Ionosphere	50	.8156	.0054	.8180	.0052	.0029	.3746	.8186	.0051	.0037	.3434
	75	.8388	.0037	.8383	.0037	-.0006	.4620	.8376	.0037	-.0014	.4094
	100	.8520	.0033	.8515	.0034	-.0006	.4580	.8502	.0035	-.0021	.3543
	150	.8712	.0029	.8691	.0027	-.0024	.2984	.8700	.0028	-.0014	.3831
EEG eye state	50	.7733	.0065	.7877	.0068	.0186	.0637	.7894	.0076	.0208	.0545
	75	.7710	.0066	.7807	.0068	.0126	.1536	.7802	.0069	.0119	.1682
	100	.7854	.0073	.7999	.0078	.0185	.0881	.7865	.0079	.0014	.4593
	150	.7923	.0063	.8203	.0070	.0353	.0017	.8032	.0078	.0138	.1392
Statlog heart	50	.8565	.0033	.8560	.0033	-.0006	.4574	.8568	.0034	.0004	.4748
	75	.8836	.0021	.8830	.0020	-.0007	.4182	.8837	.0021	.0001	.4866
	100	.8906	.0019	.8899	.0019	-.0008	.3974	.8899	.0019	-.0008	.3974
	150	.9007	.0022	.9021	.0021	.0016	.3229	.9020	.0022	.0014	.3383
SPECT heart	50	.7737	.0067	.7791	.0060	.0070	.2745	.7703	.0073	-.0044	.3659
	75	.7878	.0033	.7881	.0037	.0004	.4759	.7890	.0035	.0015	.4016
	100	.8006	.0037	.8038	.0036	.0040	.2680	.7991	.0037	-.0019	.3873
	150	.8183	.0071	.8219	.0070	.0044	.3592	.8218	.0063	.0043	.3564
SPECTF heart	50	.7771	.0037	.7779	.0037	.0010	.4393	.7775	.0038	.0005	.4700
	75	.7957	.0030	.7943	.0030	-.0018	.3709	.7939	.0030	-.0023	.3359
	100	.8054	.0027	.8048	.0029	-.0007	.4399	.8042	.0030	-.0015	.3833
	150	.8100	.0035	.8117	.0032	.0021	.3602	.8134	.0034	.0042	.2434
Diabetes	50	.7894	.0033	.7893	.0032	-.0001	.4913	.7887	.0033	-.0009	.4405
	75	.8016	.0027	.8024	.0025	.0010	.4141	.8014	.0026	-.0002	.4788
	100	.8095	.0016	.8102	.0016	.0009	.3787	.8097	.0016	.0002	.4648
	150	.8210	.0013	.8220	.0012	.0012	.2863	.8212	.0012	.0002	.4551
Breast cancer	50	.9870	.0009	.9881	.0009	.0011	.1943	.9882	.0009	.0012	.1735
	75	.9898	.0005	.9904	.0004	.0006	.1749	.9899	.0005	.0001	.4438
	100	.9906	.0003	.9910	.0003	.0004	.1735	.9909	.0003	.0003	.2402
	150	.9915	.0003	.9913	.0003	-.0002	.3189	.9914	.0003	-.0001	.4070
Banknote	50	.9981	.0001	.9980	.0001	-.0001	.2402	.9980	.0001	-.0001	.2402
	75	.9983	.0001	.9983	.0001	.0000	.5000	.9984	.0001	.0001	.2402
	100	.9984	.0001	.9983	.0001	-.0001	.2402	.9984	.0001	.0000	.5000
	150	.9985	.0001	.9985	.0001	.0000	.5000	.9986	.0001	.0001	.2402
Vertebral column	50	.8674	.0027	.8715	.0029	.0047	.1510	.8699	.0027	.0029	.2567
	75	.8744	.0029	.8771	.0028	.0031	.2519	.8762	.0028	.0021	.3279
	100	.8763	.0021	.8783	.0021	.0023	.2507	.8775	.0021	.0014	.3433
	150	.8827	.0028	.8831	.0027	.0005	.4591	.8836	.0027	.0010	.4086
Connectionist	50	.7887	.0035	.7895	.0034	.0010	.4350	.7897	.0036	.0013	.4212
	75	.8112	.0034	.8105	.0033	-.0009	.4414	.8112	.0034	.0000	.5000
	100	.8177	.0034	.8176	.0034	-.0001	.4917	.8174	.0036	-.0004	.4759
	150	.8243	.0051	.8199	.0053	-.0053	.2752	.8190	.0055	-.0064	.2403
Climate model	50	.8352	.0068	.8345	.0063	-.0008	.4699	.8360	.0071	.0010	.4676
	75	.8444	.0075	.8395	.0076	-.0058	.3234	.8432	.0074	-.0014	.4547
	100	.8720	.0083	.8581	.0083	-.0159	.1189	.8523	.0087	-.0226	.0515
	150	.7738	.0235	.7273	.0242	-.0601	.0848	.7371	.0245	-.0474	.1405
Spambase	50	.9022	.0018	.9023	.0018	.0001	.4844	.9025	.0018	.0003	.4532
	75	.9180	.0014	.9186	.0014	.0007	.3811	.9188	.0014	.0009	.3433
	100	.9235	.0011	.9234	.0010	-.0001	.4732	.9238	.0010	.0003	.4201
	150	.9298	.0009	.9305	.0009	.0008	.2915	.9311	.0009	.0014	.1542

References

K. Bache and M. Lichman. UCI machine learning repository, 2013.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton series in applied mathematics. Princeton University Press, Princeton, NJ, October 2009.

Dimitris Bertsimas, Xuan Vinh Doan, Karthik Natarajan, and Chung-Piaw Teo. Models for minimax stochastic linear optimization problems with risk aversion. *Mathematics of Operations Research*, 35(3):580–602, 2010.

C. Bhattacharyya. Robust classification of noisy data using second order cone programming approach. In *Intelligent Sensing and Information Processing, 2004. Proceedings of International Conference on*, pages 433–438, 2004.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, New York, NY, 1992. ACM. ISBN 0-89791-497-X.

Corinna Cortes and Vladimir N. Vapnik. Support-vector networks. *Machine Learning*, 20 (3):273–297, 1995.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. 2013. available on arXiv:1312.2128 [math.PR].

Laurent El Ghaoui, Maksim Oks, and Francois Oustry. Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations Research*, 51(4):543–556, 2003. ISSN 0030364X.

Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

S. Å. Gustafson and K. O. Kortanek. Numerical treatment of a class of semi-infinite programming problems. *Naval Research Logistics Quarterly*, 20(3):477–504, 1973.

Gert Lanckriet, Laurent E. Ghaoui, Chiranjib Bhattacharyya, and Michael I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3: 555–582, 2002.

Charles X. Ling, Jin Huang, and Harry Zhang. *AUC: A Better Measure than Accuracy in Comparing Learning Algorithms*, volume 2671, pages 329–341. Springer Berlin Heidelberg, 2003. ISBN 978-3-540-40300-5.

S. Mehrotra and D. Papp. A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization. *SIAM Journal on Optimization*, 24 (4):1670–1697, 2014.

Sanjay Mehrotra and He Zhang. Models and algorithms for distributionally robust least squares problems. *Mathematical Programming*, 146(1-2):123–141, August 2014. ISSN 0025-5610.

Georg Pflug and David Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7 (4):435–442, 2007.

Herbert Scarf. Bayes solutions of the statistical inventory problem. *The Annals of Mathematical Statistics*, 30(2):490–508, 1959. ISSN 00034851.

Alexander Shapiro. On duality theory of conic linear problems. In M.A. Goberna and M.A. Lopez, editors, *Semi-Infinite Programming: Recent Advances*, volume 57, pages 135–155. Kluwer Academic Publishers, 2001.

Pannagadatta K. Shivaswamy, Chiranjib Bhattacharyya, and Alexander J. Smola. Second order cone programming approaches for handling missing and uncertain data. *J. Mach. Learn. Res.*, 7:1283–1314, December 2006. ISSN 1532-4435.

AW Van der Vaart and JA Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

Vladimir N. Vapnik. *Estimation of dependences based on empirical data*. Nauka, Moscow, 1979. [in Russian].

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *J. Mach. Learn. Res.*, 10:1485–1510, December 2009. ISSN 1532-4435.