Robust Machine Learning Optimization Models with Data Uncertainties

Panos M. Pardalos

Center for Applied Optimization (CAO), University of Florida, USA
www.ise.ufl.edu/pardalos/

Laboratory of Algorithms and Technologies for Networks Analysis (LATNA)
National Research University, Higher School of Economics, Russia
https://nnov.hse.ru/en/latna/

Outline

Introduction

In recent years, machine learning and data mining have an explosive growth with new developments in science and technology.

- The essentials of most machine learning and data mining techniques are optimization problems.
- Traditional machine learning models are dealing with data when the exact values are known.
- This talk considers the case when uncertainties exist in data.

Introduction

Support Vector Machines (SVM) is one of the well known supervised classes of learning algorithms.

- It was proposed by Vapnik as a maximum-margin classifier.
- Basic SVM models are dealing with the situation that the exact values of the data points are known.
- When the data points are uncertain, different models have been proposed to formulate the SVM with uncertainties.

Introduction

- Robust SVM with Bounded Uncertainty
  - Trafalis et al. proposed a robust optimization model when the perturbation of the uncertain data is bounded by norm, where some efficient linear programming models are presented under certain conditions.
  - Ghaoui et al. derived a robust model when the uncertainty is expressed as intervals with support and extremum values.
  - Fan et al. studied a more general case for polyhedral uncertainties.

Introduction

- Chance Constrained SVM through Robust Optimization
  - The Chebyshev based model employs moment information of the uncertain training points.
  - The Bernstein bounds can be less conservative than the Chebyshev bounds since it employs both support and moment information, but it also makes a strong assumption that all the elements in the data set are independent.

Outline

Hard Margin SVM

Support Vector Machines (SVM) construct maximum-margin classifiers:

- A two-class dataset of m data points $\{\mathbf{x}_i, y_i\}_{i=1}^m$ with n-dimensional features $\mathbf{x}_i \in \mathbb{R}^n$ and class labels $y_i \in \{\pm 1\}$.
- For linearly separable datasets, there exists a hyperplane $\mathbf{w}^\top \mathbf{x} + b = 0$ to separate the two classes.
- The width between the margin lines $\mathbf{w}^\top \mathbf{x} + b = \pm 1$ is $\frac{2}{\|\mathbf{w}\|_2^2}$.

Hard Margin SVM (SVM-HardMargin)

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad i = 1, \ldots, m \end{aligned}$$

Soft Margin SVM

When two classes are not linearly separable:

- Soft margin SVM introduces non-negative slack variables ξ_i to measure the distance of data to the margin.
- $\xi_i = \max\{0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\}$
- When $0 < \xi_i < 1$, the data is within margine but correctly classified; when $\xi_i > 1$, the data is misclassified.

Soft Margin SVM (SVM-SoftMargin)

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, m \end{aligned}$$

Chance-Constrained SVM

When uncertainties exist in the data points:

- A two-class dataset of m uncertain training data points $\tilde{\mathbf{x}}_i \in \mathbb{R}^n$ and corresponding labels $y_i \in \{\pm 1\}$.
- The Chance-Constrained Program (CCP) is to ensure the small probability of misclassification for the uncertain data.

Chance-Constrained SVM (SVM-CCP)

$$\min_{\mathbf{w},b,\xi_i} \ \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \ \ \mathbb{P}\Big\{y_i(\mathbf{w}^\top\tilde{\mathbf{x}}_i + b) \le 1 - \xi_i\Big\} \le \varepsilon, \quad \xi_i \ge 0, \ \ i = 1,\ldots,m$$

Robust Chance-Constrained SVM

The exact probability distribution are often unknown:

- Only some properties of the distribution could be acquired, such as the first and second moments.
- The distributionally robust or ambiguous chance constraint is a conservative approximation of the original problem.
- Let $\mathcal{P}$ be the set of all probability distributions that have the known properties of $\mathbb{P}$.

Robust Chance-Constrained SVM (SVM-RCCP)

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \quad \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\Big\{ y_i(\mathbf{w}^\top \tilde{\mathbf{x}}_i + b) \leq 1 - \xi_i \Big\} \leq \varepsilon, \quad \xi_i \geq 0, \quad i = 1, \ldots, m$$

Outline

Moments Information

- Assume the first and second moment information of the random variables $\tilde{\mathbf{x}}_i$ are known.
- For random variable $\tilde{\mathbf{x}}_i$, let $\boldsymbol{\mu}_i = \mathbf{E}[\tilde{\mathbf{x}}_i] \in \mathbb{R}^n$ be the mean vector and $\boldsymbol{\Sigma}_i = \mathbf{E}\left[(\tilde{\mathbf{x}}_i - \boldsymbol{\mu}_i)(\tilde{\mathbf{x}}_i - \boldsymbol{\mu}_i)^\top\right] \in \mathbb{S}^n$ be the covariance matrix.
- Combine the first and second moments $\boldsymbol{\Sigma}_i, \boldsymbol{\mu}_i$ into one matrix Ω_i:

$$\Omega_i = \begin{bmatrix} \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top & \boldsymbol{\mu}_i \\ \boldsymbol{\mu}_i^\top & 1 \end{bmatrix}$$

- Let $\mathcal{P}$ be the set of all probability distributions that have the same first and second moments.

Reformulation of (SVM-RCCP) into SDP

SVM SDP Model (SVM-SDP)

$$\min_{\mathbf{w},b,\xi_i,\mathbf{N}_i,\alpha_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \quad \alpha_i - \frac{1}{\varepsilon}\text{Trace}(\Omega_i\mathbf{N}_i) \geq 0, \quad \xi_i \geq 0$$

$$\mathbf{N}_i \succeq 0, \quad \mathbf{N}_i + \begin{bmatrix} 0 & \frac{1}{2}y_i\mathbf{w} \\ \frac{1}{2}y_i\mathbf{w}^\top & y_ib + \xi_i - 1 - \alpha_i \end{bmatrix} \succeq 0$$

Theorem

(SVM-RCCP) is equivalent to (SVM-SDP).

Reformulation of (SVM-RCCP) into SDP

Proof Sketch

- For the $p = \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\Big\{ y_i(\mathbf{w}^\top \tilde{\mathbf{x}}_i + b) \leq 1 - \xi_i \Big\}$, define the indicator function

$$I(\mathbf{x}_i) = \begin{cases} 1, & \text{if } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq 1 - \xi_i \\ 0, & \text{otherwise} \end{cases}$$

- Then p can be expressed by the following program:

$$\begin{aligned} p = \sup_{\mathbb{P}} \quad & \int_{\mathbb{R}^n} I(\mathbf{x}_i)\mathbb{P}\{\mathbf{x}_i\}\mathrm{d}\mathbf{x}_i \\ \text{s.t.} \quad & \int_{\mathbb{R}^n} \mathbb{P}\{\mathbf{x}_i\}\mathrm{d}\mathbf{x}_i = 1 \\ & \int_{\mathbb{R}^n} \mathbf{x}_i\mathbb{P}\{\mathbf{x}_i\}\mathrm{d}\mathbf{x}_i = \boldsymbol{\mu}_i \\ & \int_{\mathbb{R}^n} \mathbf{x}_i\mathbf{x}_i^\top\mathbb{P}\{\mathbf{x}_i\}\mathrm{d}\mathbf{x}_i = \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top \\ & \mathbb{P}\{\mathbf{x}_i\} \geq 0 \end{aligned}$$

Reformulation of (SVM-RCCP) into SDP

Proof Sketch (Continued)

- The dual of the program is:

$$p = \inf_{\mathbf{Z}_i, \mathbf{z}_i, z_{0i}} \ (\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top) \cdot \mathbf{Z}_i + \boldsymbol{\mu}_i^\top \mathbf{z}_i + z_{0i}$$

$$\text{s.t. } \mathbf{x}_i^\top \mathbf{Z}_i \mathbf{x}_i + \mathbf{x}_i^\top \mathbf{z}_i + z_{0i} \geq I(\mathbf{x}_i), \quad \forall \mathbf{x}_i \in \mathbb{R}^n$$
$$\mathbf{Z}_i \in \mathbb{S}^n, \ \mathbf{z}_i \in \mathbb{R}^n, \ z_{0i} \in \mathbb{R}$$

- The first constraint can be expressed in two constraints:

$$\mathbf{x}_i^\top \mathbf{Z}_i \mathbf{x}_i + \mathbf{x}_i^\top \mathbf{z}_i + z_{0i} \geq 0, \quad \forall \mathbf{x}_i \in \mathbb{R}^n$$

$$\mathbf{x}_i^\top \mathbf{Z}_i \mathbf{x}_i + \mathbf{x}_i^\top \mathbf{z}_i + z_{0i} \geq 1, \quad \text{if } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq 1 - \xi_i$$

- Let $\mathbf{M}_i = \begin{bmatrix} \mathbf{Z}_i & \frac{1}{2}\mathbf{z}_i \\ \frac{1}{2}\mathbf{z}_i^\top & z_{0i} \end{bmatrix}$, $\Omega_i = \begin{bmatrix} \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top & \boldsymbol{\mu}_i \\ \boldsymbol{\mu}_i^\top & 1 \end{bmatrix}$
- The objective function becomes $\text{Trace}(\Omega_i \mathbf{M}_i)$

Reformulation of (SVM-RCCP) into SDP

Proof Sketch (Continued)

- $\mathbf{x}_i^\top \mathbf{Z}_i \mathbf{x}_i + \mathbf{x}_i^\top \mathbf{z}_i + z_{0i} \geq 0, \quad \forall \mathbf{x}_i \in \mathbb{R}^n$ becomes

$$[\mathbf{x}_i^\top 1]\mathbf{M}_i[\mathbf{x}_i^\top 1]^\top \geq 0, \forall \mathbf{x}_i \in \mathbb{R}^n \qquad \text{i.e.} \qquad \mathbf{M}_i \succeq 0$$

- $\mathbf{x}_i^\top \mathbf{Z}_i \mathbf{x}_i + \mathbf{x}_i^\top \mathbf{z}_i + z_{0i} \geq 1$, if $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq 1 - \xi_i$ is equivalent that the following system has no solution $\mathbf{x}_i \in \mathbb{R}^n$ such that

$$\mathbf{x}_i^\top \mathbf{Z}_i \mathbf{x}_i + \mathbf{x}_i^\top \mathbf{z}_i + z_{0i} - 1 < 0$$
$$y_i \mathbf{w}^\top \mathbf{x}_i + y_i b + \xi_i - 1 \leq 0$$

- According to S-lemma, there exists a nonnegative number $\beta_i \geq 0$ such that

$$\mathbf{x}_i^\top \mathbf{Z}_i \mathbf{x}_i + \mathbf{x}_i^\top \mathbf{z}_i + z_{0i} - 1 + \beta_i(y_i \mathbf{w}^\top \mathbf{x}_i + y_i b + \xi_i - 1) \geq 0 \quad \forall \mathbf{x}_i \in \mathbb{R}^n$$

Reformulation of (SVM-RCCP) into SDP

Proof Sketch (Continued)

- The dual program becomes:

$$p = \inf_{\mathbf{M}_i, \beta_i} \ \text{Trace}(\Omega_i \mathbf{M}_i)$$
$$\text{s.t.} \ \ \mathbf{M}_i \succeq 0, \ \beta_i \geq 0$$
$$[\mathbf{x}_i^\top 1]\mathbf{M}_i[\mathbf{x}_i^\top 1]^\top - 1 + \beta_i(y_i \mathbf{w}^\top \mathbf{x}_i + y_i b + \xi_i - 1) \geq 0 \ \ \forall \mathbf{x}_i \in \mathbb{R}^n$$

- The whole program becomes:

$$\min_{\mathbf{w}, b, \xi_i, \mathbf{M}_i, \beta_i} \ \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m} \xi_i$$
$$\text{s.t.} \ \ \mathbf{M}_i \succeq 0, \ \beta_i \geq 0, \ \xi_i \geq 0$$
$$\text{Trace}(\Omega_i \mathbf{M}_i) \leq \varepsilon$$
$$[\mathbf{x}_i^\top 1]\mathbf{M}_i[\mathbf{x}_i^\top 1]^\top - 1 + \beta_i(y_i \mathbf{w}^\top \mathbf{x}_i + y_i b + \xi_i - 1) \geq 0 \ \ \forall \mathbf{x}_i \in \mathbb{R}^n$$

Reformulation of (SVM-RCCP) into SDP

Proof Sketch (Continued)

- To get rid of the bilinear terms in $\beta_i(y_i\mathbf{w}^\top\mathbf{x}_i + y_i b + \xi_i - 1)$, first it could be verified that β_i cannot be zero since $\text{Trace}(\Omega_i\mathbf{M}_i) \leq \varepsilon$, and $0 < \varepsilon < 1$. If $\beta_i = 0$, then $[\mathbf{x}_i^\top 1]\mathbf{M}_i[\mathbf{x}_i^\top 1]^\top \geq 1 > \varepsilon,\ \forall \mathbf{x}_i \in \mathbb{R}^n$, a contradiction. Therefore, $\beta_i > 0$.
- Then the constraints become

$$\frac{1}{\varepsilon}\text{Trace}(\Omega_i\frac{\mathbf{M}_i}{\beta_i}) - \frac{1}{\beta_i} \leq 0$$

$$[\mathbf{x}_i^\top 1]\frac{\mathbf{M}_i}{\beta_i}[\mathbf{x}_i^\top 1]^\top - \frac{1}{\beta_i} + (y_i\mathbf{w}^\top\mathbf{x}_i + y_i b + \xi_i - 1) \geq 0 \ \ \forall \mathbf{x}_i \in \mathbb{R}^n$$

- Replace $\frac{\mathbf{M}_i}{\beta_i}$ with $\mathbf{N}_i \succeq 0$, and $\frac{1}{\beta_i}$ with $\alpha_i > 0$, the second constraint could further be expressed as a semidefinite constraint as:

$$\mathbf{N}_i + \begin{bmatrix} 0 & \frac{1}{2}y_i\mathbf{w} \\ \frac{1}{2}y_i\mathbf{w}^\top & y_i b + \xi_i - 1 - \alpha_i \end{bmatrix} \succeq 0$$

- $\alpha_i > 0$ is guaranteed since $\mathbf{N}_i \succeq 0$ and $\frac{1}{\varepsilon}\text{Trace}(\Omega_i\mathbf{N}_i) - \alpha_i \leq 0$.

Reformulation of (SVM-RCCP) into SDP

Proof Sketch (Continued)

- The whole program becomes:

$$\min_{\mathbf{w},b,\xi_i,\mathbf{N}_i,\alpha_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \quad \alpha_i - \frac{1}{\varepsilon}\text{Trace}(\Omega_i\mathbf{N}_i) \geq 0, \quad \xi_i \geq 0$$

$$\mathbf{N}_i \succeq 0, \quad \mathbf{N}_i + \begin{bmatrix} 0 & \frac{1}{2}y_i\mathbf{w} \\ \frac{1}{2}y_i\mathbf{w}^\top & y_i b + \xi_i - 1 - \alpha_i \end{bmatrix} \succeq 0$$

- This completes the proof of the (SVM-SDP) reformulation.

Reformulation of (SVM-RCCP) into SDP

The standard SDP formulation is

$$\min_{\mathbf{w},b,\xi_i,\mathbf{N}_i,\alpha_i} \quad \sum_{i=1}^{m} \xi_i$$

$$\text{s.t.} \quad \alpha_i - \frac{1}{\varepsilon}\text{Trace}(\Omega_i \mathbf{N}_i) \geq 0, \quad \xi_i \geq 0$$

$$\mathbf{N}_i \succeq 0, \quad \mathbf{N}_i + \begin{bmatrix} 0 & \frac{1}{2}y_i\mathbf{w} \\ \frac{1}{2}y_i\mathbf{w}^\top & y_i b + \xi_i - 1 - \alpha_i \end{bmatrix} \succeq 0$$

$$\begin{bmatrix} W\mathbf{I} & \mathbf{w} \\ \mathbf{w}^\top & W \end{bmatrix} \succeq 0$$

Reformulation of (SVM-RCCP) into SOCP

Theorem

The SDP constraints could yield the SOCP constraints:

$$\alpha_i - \frac{1}{\varepsilon} Trace(\Omega_i \mathbf{N}_i) \geq 0, \quad \mathbf{N}_i \succeq 0,$$

$$\mathbf{N}_i + \begin{bmatrix} 0 & \frac{1}{2} y_i \mathbf{w} \\ \frac{1}{2} y_i \mathbf{w}^\top & y_i b + \xi_i - 1 - \alpha_i \end{bmatrix} \succeq 0$$

$$\Longrightarrow \quad y_i(\mathbf{w}^\top \mu_i + b) \geq 1 - \xi_i + \sqrt{\frac{1-\varepsilon}{\varepsilon}} ||\mathbf{\Sigma}_i^{\frac{1}{2}} \mathbf{w}||_2$$

SDP Constraints into SOCP Constraints

Proof Sketch

- Consider the following problem:

$$\begin{aligned} \inf_{b,\xi_i,\mathbf{N}_i,\alpha_i} \quad & y_i b + \xi_i - 1 \\ \text{s.t.} \quad & \alpha_i - \frac{1}{\varepsilon}\text{Trace}(\Omega_i \mathbf{N}_i) \geq 0 \\ & \mathbf{N}_i \succeq 0 \\ & \mathbf{N}_i + \begin{bmatrix} 0 & \frac{1}{2} y_i \mathbf{w} \\ \frac{1}{2} y_i \mathbf{w}^\top & y_i b + \xi_i - 1 - \alpha_i \end{bmatrix} \succeq 0 \end{aligned}$$

- Let γ_i, $\mathbf{C}_i$, $\overline{\mathbf{D}}_i = \begin{bmatrix} \mathbf{D}_i & \mathbf{d}_i \\ \mathbf{d}_i^\top & d_{0i} \end{bmatrix}$ represent the dual variables of the constraints.

SDP Constraints into SOCP Constraints

Proof Sketch (Continued)

- The Lagrangian is:

$$\inf_{b,\xi_i,\mathbf{N}_i,\alpha_i} \sup_{\gamma_i\geq 0,\mathbf{C}_i\succeq 0,\overline{\mathbf{D}}_i\succeq 0} \mathscr{L}(\mathbf{w},b,\xi_i,\mathbf{N}_i,\alpha_i,\gamma_i,\mathbf{C}_i,\overline{\mathbf{D}}_i)$$

$$= y_i b + \xi_i - 1 - \gamma_i\Big(\alpha_i - \frac{1}{\varepsilon}\text{Trace}(\Omega_i\mathbf{N}_i)\Big) - \text{Trace}(\mathbf{C}_i\mathbf{N}_i)$$

$$- \text{Trace}\left(\overline{\mathbf{D}}_i, \mathbf{N}_i + \begin{bmatrix} 0 & \frac{1}{2}y_i\mathbf{w} \\ \frac{1}{2}y_i\mathbf{w}^\top & y_i b + \xi_i - 1 - \alpha_i \end{bmatrix}\right)$$

$$= y_i b + \xi_i - 1 - \gamma_i\alpha_i + \frac{\gamma_i}{\varepsilon}\text{Trace}(\Omega_i\mathbf{N}_i) - \text{Trace}(\mathbf{C}_i\mathbf{N}_i) - \text{Trace}(\overline{\mathbf{D}}_i,\mathbf{N}_i) - y_i\mathbf{w}^\top\mathbf{d}_i$$

$$- d_{0i}(y_i b + \xi_i - 1 - \alpha_i)$$

$$= (y_i b + \xi_i - 1)(1 - d_{0i}) - (\gamma_i - d_{0i})\alpha_i + \text{Trace}\Big(\frac{\gamma_i}{\varepsilon}\Omega_i - \mathbf{C}_i - \overline{\mathbf{D}}_i, \mathbf{N}_i\Big) - y_i\mathbf{w}^\top\mathbf{d}_i$$

SDP Constraints into SOCP Constraints

Proof Sketch (Continued)

- The dual function is finite if and only if
$$1 - d_{0i} = 0, \quad \gamma_i - d_{0i} = 0, \quad \frac{\gamma_i}{\varepsilon}\Omega_i - \mathbf{C}_i - \overline{\mathbf{D}}_i = 0$$
- Therefore, $\gamma_i = 1$ and $\frac{1}{\varepsilon}\Omega_i - \overline{\mathbf{D}}_i = \mathbf{C}_i \succeq 0$.
- Then the dual problem is:
$$\begin{aligned} \sup_{\overline{\mathbf{D}}_i} \quad & -y_i \mathbf{w}^\top \mathbf{d}_i \\ \text{s.t.} \quad & \frac{1}{\varepsilon}\Omega_i \succeq \overline{\mathbf{D}}_i \succeq 0 \end{aligned}$$

SDP Constraints into SOCP Constraints

Proof Sketch (Continued)

- Since $\Omega_i = \begin{bmatrix} \mathbf{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top & \boldsymbol{\mu}_i \\ \boldsymbol{\mu}_i^\top & 1 \end{bmatrix}$, $\overline{\mathbf{D}}_i = \begin{bmatrix} \mathbf{D}_i & \mathbf{d}_i \\ \mathbf{d}_i^\top & d_{0i} \end{bmatrix}$, $d_{0i} = 1$, and $\varepsilon > 0$, the constraint $\frac{1}{\varepsilon}\Omega_i \succeq \overline{\mathbf{D}}_i \succeq 0$ is equivalent to

$$\begin{bmatrix} \mathbf{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top - \varepsilon \mathbf{D}_i & \boldsymbol{\mu}_i - \varepsilon \mathbf{d}_i \\ \boldsymbol{\mu}_i^\top - \varepsilon \mathbf{d}_i^\top & 1 - \varepsilon \end{bmatrix} \succeq 0, \quad \begin{bmatrix} \varepsilon \mathbf{D}_i & \varepsilon \mathbf{d}_i \\ \varepsilon \mathbf{d}_i^\top & \varepsilon \end{bmatrix} \succeq 0$$

- According to Schur Complement Lemma, the above is equivalent to
$\mathbf{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top - \varepsilon \mathbf{D}_i - \frac{1}{1-\varepsilon}(\boldsymbol{\mu}_i - \varepsilon \mathbf{d}_i)(\boldsymbol{\mu}_i - \varepsilon \mathbf{d}_i)^\top \succeq 0, \quad \varepsilon \mathbf{D}_i - \frac{1}{\varepsilon}\varepsilon \mathbf{d}_i \varepsilon \mathbf{d}_i^\top \succeq 0$

$$\text{i.e.,} \quad \mathbf{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top - \frac{1}{1-\varepsilon}(\boldsymbol{\mu}_i - \varepsilon \mathbf{d}_i)(\boldsymbol{\mu}_i - \varepsilon \mathbf{d}_i)^\top \succeq \varepsilon \mathbf{D}_i \succeq \varepsilon \mathbf{d}_i \mathbf{d}_i^\top$$

- The above holds for some $\mathbf{D}_i$ if and only if

$$\mathbf{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top \succeq \frac{1}{1-\varepsilon}(\boldsymbol{\mu}_i - \varepsilon \mathbf{d}_i)(\boldsymbol{\mu}_i - \varepsilon \mathbf{d}_i)^\top + \varepsilon \mathbf{d}_i \mathbf{d}_i^\top$$

SDP Constraints into SOCP Constraints

Proof Sketch (Continued)

- Expand the constraint

$$\begin{aligned}\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top &\succeq \frac{1}{1-\varepsilon}\boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top - \frac{\varepsilon}{1-\varepsilon}\boldsymbol{\mu}_i\mathbf{d}_i^\top - \frac{\varepsilon}{1-\varepsilon}\mathbf{d}_i\boldsymbol{\mu}_i^\top + \frac{\varepsilon^2}{1-\varepsilon}\mathbf{d}_i\mathbf{d}_i^\top + \varepsilon\mathbf{d}_i\mathbf{d}_i^\top \\ &= \frac{1}{1-\varepsilon}\boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top - \frac{\varepsilon}{1-\varepsilon}(\boldsymbol{\mu}_i\mathbf{d}_i^\top + \mathbf{d}_i\boldsymbol{\mu}_i^\top) + \frac{\varepsilon}{1-\varepsilon}\mathbf{d}_i\mathbf{d}_i^\top\end{aligned}$$

- It is equivalent to

$$\boldsymbol{\Sigma}_i \succeq \frac{\varepsilon}{1-\varepsilon}(\boldsymbol{\mu}_i - \mathbf{d}_i)(\boldsymbol{\mu}_i - \mathbf{d}_i)^\top$$

- The dual problem becomes

$$\begin{aligned}\sup_{\mathbf{d}_i} \quad & -y_i\mathbf{w}^\top\mathbf{d}_i \\ \text{s.t.} \quad & \frac{1-\varepsilon}{\varepsilon}\boldsymbol{\Sigma}_i - (\boldsymbol{\mu}_i - \mathbf{d}_i)(\boldsymbol{\mu}_i - \mathbf{d}_i)^\top \succeq 0\end{aligned}$$

SDP Constraints into SOCP Constraints

Proof Sketch (Continued)

- From the constraint, there is
$$y_i \mathbf{w}^\top \left(\frac{1-\varepsilon}{\varepsilon} \mathbf{\Sigma}_i - (\boldsymbol{\mu}_i - \mathbf{d}_i)(\boldsymbol{\mu}_i - \mathbf{d}_i)^\top \right) y_i \mathbf{w} \geq 0$$
- Since $y_i \in \{+1, -1\}$, $y_i^2 = 1$. Then
$$(y_i \mathbf{w}^\top \boldsymbol{\mu}_i - y_i \mathbf{w}^\top \mathbf{d}_i)^2 \leq \frac{1-\varepsilon}{\varepsilon} \mathbf{w}^\top \mathbf{\Sigma}_i \mathbf{w}$$
- Therefore,
$$-y_i \mathbf{w}^\top \mathbf{d}_i \leq \sqrt{\frac{1-\varepsilon}{\varepsilon}} ||\mathbf{\Sigma}_i^{\frac{1}{2}} \mathbf{w}||_2 - y_i \mathbf{w}^\top \boldsymbol{\mu}_i$$
- The maximum value of $-y_i \mathbf{w}^\top \mathbf{d}_i$ is $\sqrt{\frac{1-\varepsilon}{\varepsilon}} ||\mathbf{\Sigma}_i^{\frac{1}{2}} \mathbf{w}||_2 - y_i \mathbf{w}^\top \boldsymbol{\mu}_i$ since $\mathbf{d}_i$ is the decision variable such that the equality could be obtained.

SDP Constraints into SOCP Constraints

Proof Sketch (Continued)

- Combine the primal problem and the result for the dual problem, it could yield that

$$\sqrt{\frac{1-\varepsilon}{\varepsilon}}||\mathbf{\Sigma}_i^{\frac{1}{2}}\mathbf{w}||_2 - y_i\mathbf{w}^\top\boldsymbol{\mu}_i \leq y_i b + \xi_i - 1$$

- Or equivalently,

$$y_i(\mathbf{w}^\top\boldsymbol{\mu}_i + b) \geq 1 - \xi_i + \sqrt{\frac{1-\varepsilon}{\varepsilon}}||\mathbf{\Sigma}_i^{\frac{1}{2}}\mathbf{w}||_2$$

- This completes the proof of the SDP-SOCP transformation.

Reformulation of (SVM-RCCP) into SOCP

Multivariate Chebyshev Inequality

- Let $\tilde{\mathbf{x}} \sim (\mu, \mathbf{\Sigma})$ denote random vector $\tilde{\mathbf{x}}$ with mean μ and convariance matrix $\mathbf{\Sigma}$.
- The multivariate Chebyshev inequality states that for an arbitrary closed convex set S, the supremum of the probability that $\tilde{\mathbf{x}}$ takes a value in S is

$$\sup_{\tilde{\mathbf{x}} \sim (\mu, \mathbf{\Sigma})} \mathbb{P}\{\tilde{\mathbf{x}} \in S\} = \frac{1}{1 + d^2}$$

$$d^2 = \inf_{\mathbf{x} \in S} (\mathbf{x} - \mu)^\top \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)$$

Reformulation of (SVM-RCCP) into SOCP

- For SVM constraint, the $S = \left\{ y(\mathbf{w}^\top \mathbf{x} + b) \leq 1 - \xi \right\}$ is a half-space produced by a hyperplane and therefore a closed convex set.
- Using multivariate Chebyshev inequality, the SOCP reformulation of (SVM-RCCP) is

SVM SOCP Model (SVM-SOCP)

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top \boldsymbol{\mu}_i + b) \geq 1 - \xi_i + \sqrt{\frac{1-\varepsilon}{\varepsilon}}\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}\mathbf{w}\|_2$$

$$\xi_i \geq 0, \quad i = 1, \ldots, m$$

Outline

Geometric Interpretation of (SVM-SOCP)

- For each point $\mathbf{x}_i$, it is no longer a single point, but an ellipsoid centered at $\boldsymbol{\mu}_i$, and shaped with the covariance matrix $\sqrt{\frac{1-\varepsilon}{\varepsilon}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}$:

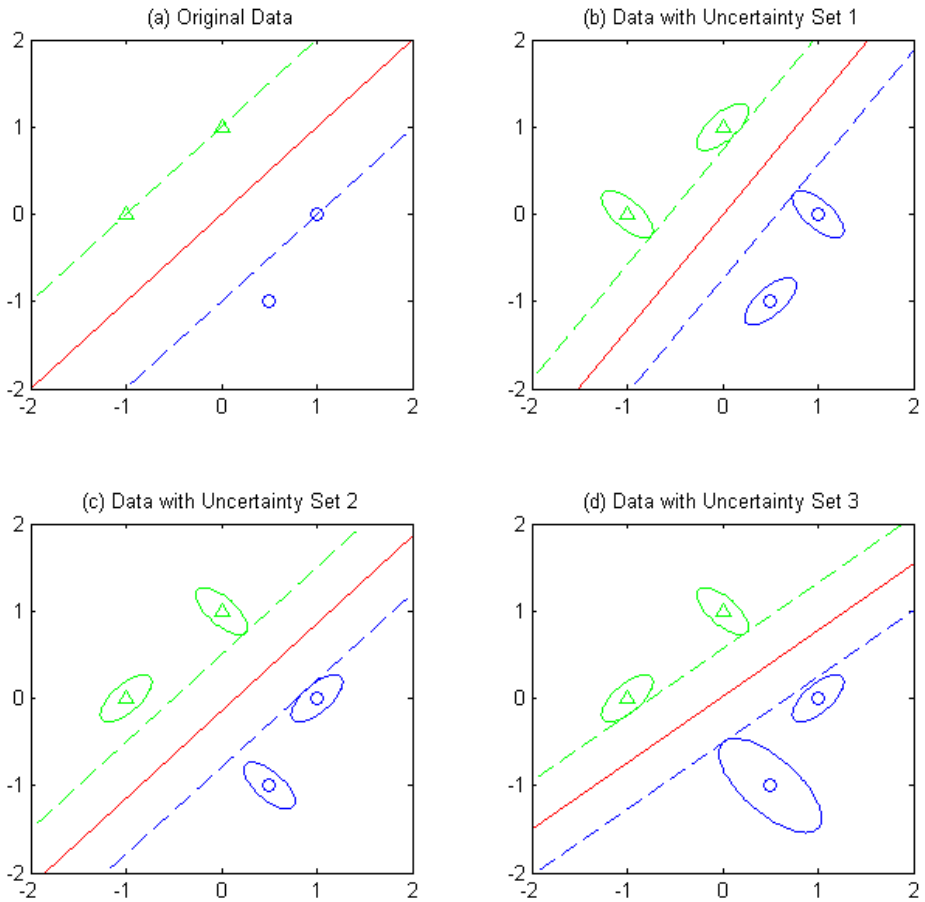
$$\mathscr{E}\left(\boldsymbol{\mu}_i, \sqrt{\frac{1-\varepsilon}{\varepsilon}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\right) = \left\{\mathbf{x} = \boldsymbol{\mu}_i + \sqrt{\frac{1-\varepsilon}{\varepsilon}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\mathbf{a} \ : \ ||\mathbf{a}||_2 \leq 1\right\}$$

- The SOCP constraint is satisfied if and only if

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \forall \mathbf{x}_i \in \mathscr{E}\left(\boldsymbol{\mu}_i, \sqrt{\frac{1-\varepsilon}{\varepsilon}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\right)$$

- This transforms the (SVM-RCCP) into a robust optimization problem over the uncertainty set $\mathscr{E}\left(\boldsymbol{\mu}_i, \sqrt{\frac{1-\varepsilon}{\varepsilon}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\right)$ for each uncertain training data point.

Geometric Interpretation of (SVM-SOCP)



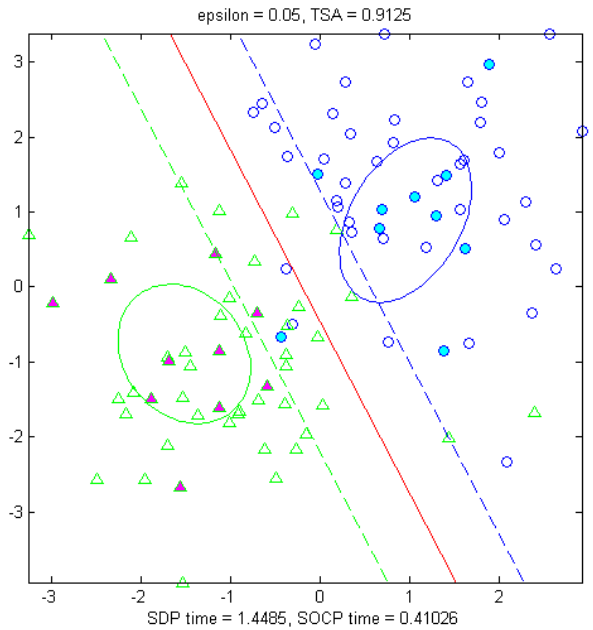
Outline

Numerical Experiments

- The model (SVM-SDP) and model (SVM-SOCP) are equivalent since they both use the exact supremum of the chance constraints $\sup \mathbb{P}\left\{ y_i(\mathbf{w}^\top \tilde{\mathbf{x}}_i + b) \leq 1 - \xi_i \right\}$ and both based on the exact means and covariance matrices of the random data points.
- Numerical experiments in MATLAB using SeDuMi solver for both models on YALMIP platform also show that these two formulations would get the same result.

Synthetic Data

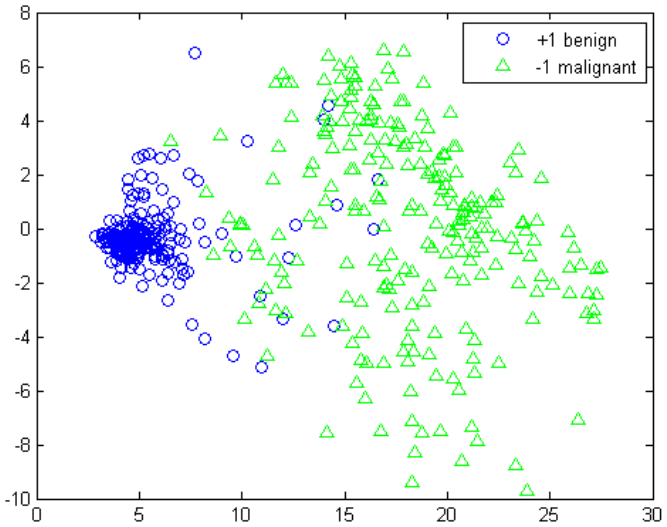
$+1$ class: 2-d normal distribution with $\boldsymbol{\mu}_+ = [1,1]^\top$, $\boldsymbol{\Sigma}_+ = \boldsymbol{I}$
-1 class: 2-d normal distribution with $\boldsymbol{\mu}_- = [-1,-1]^\top$, $\boldsymbol{\Sigma}_- = \boldsymbol{I}$
Each class has 50 points: 10 for training, 40 for test



Wisconsin Breast Cancer Data

Wisconsin breast cancer data from UCI dataset:

- 444 benign($+1$) class data, 239 malignant (-1) class data
- 9-dimensional features
- Use PCA to show the first 2 principle components



Wisconsin Breast Cancer Data Classification Result

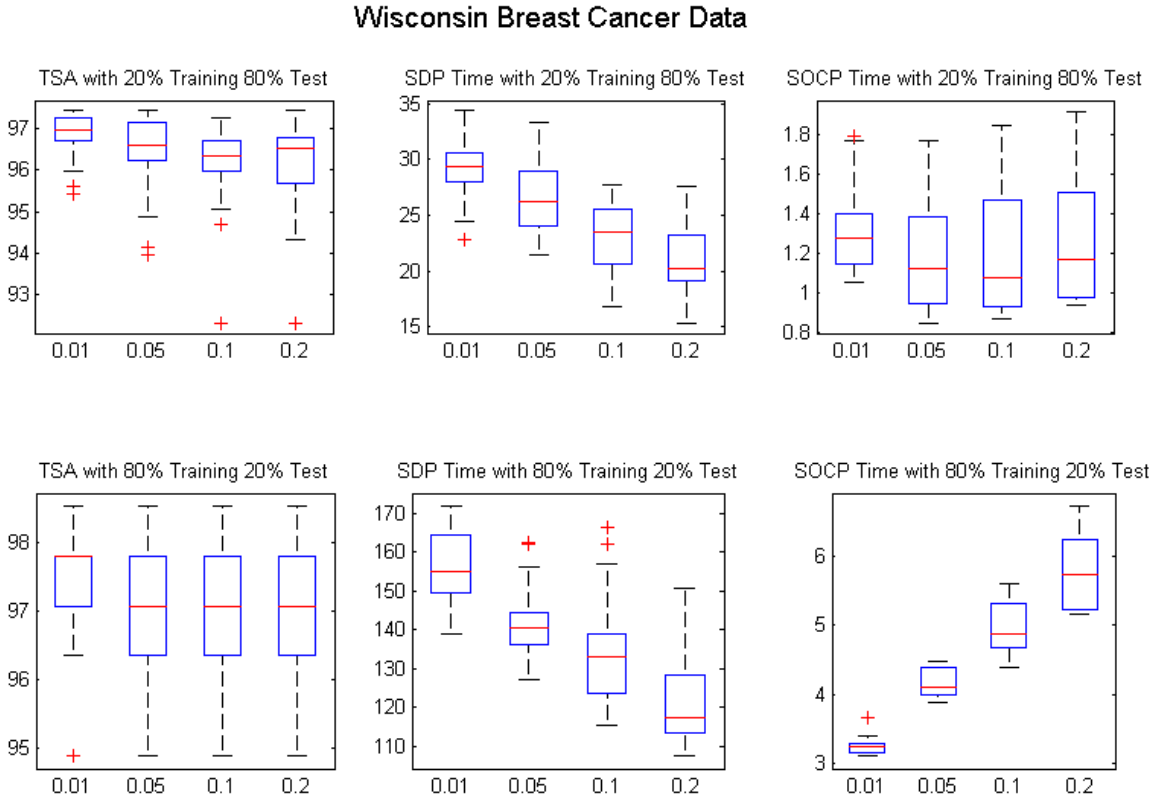
Table: 20% training, 80% test

	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$
Test Set Accuracy	96.8±0.6%	96.4±1.0%	96.1±1.1%	96.1±1.2%
SDP Running Time	29.0±2.5	26.5±3.2	23.1±3.2	21.1±2.9
SOCP Running Time	1.3±0.2	1.2±0.3	1.2±0.3	1.2±0.3

Table: 80% training, 20% test

	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$
Test Set Accuracy	97.4±0.9%	97.1±1.0%	97.0±1.1%	97.1±1.0%
SDP Running Time	155.6±9.7	141.8±9.1	134.2±13.7	121.4±11.7
SOCP Running Time	3.3±0.1	4.1±0.2	5.0±0.4	5.8±0.6

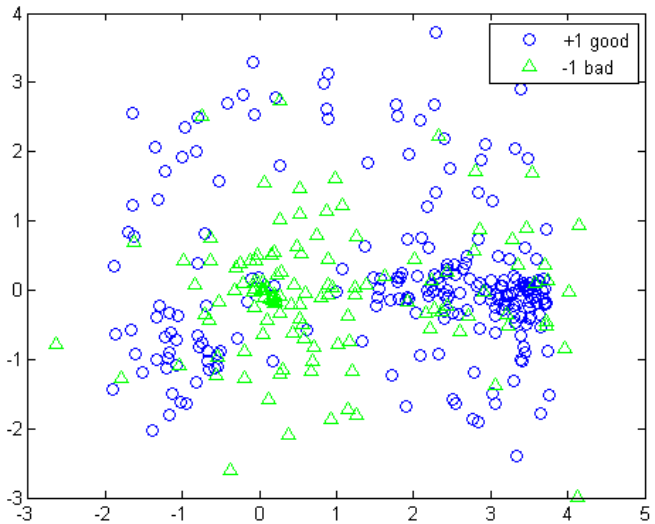
Wisconsin Breast Cancer Data Classification Result



Ionosphere Data

Ionosphere data from UCI dataset:

- 225 for $+1$ good class, 126 for -1 bad class
- 34-dimensional data
- Use PCA to show the first 2 principle components



Extracted Ionosphere Data Classification Result

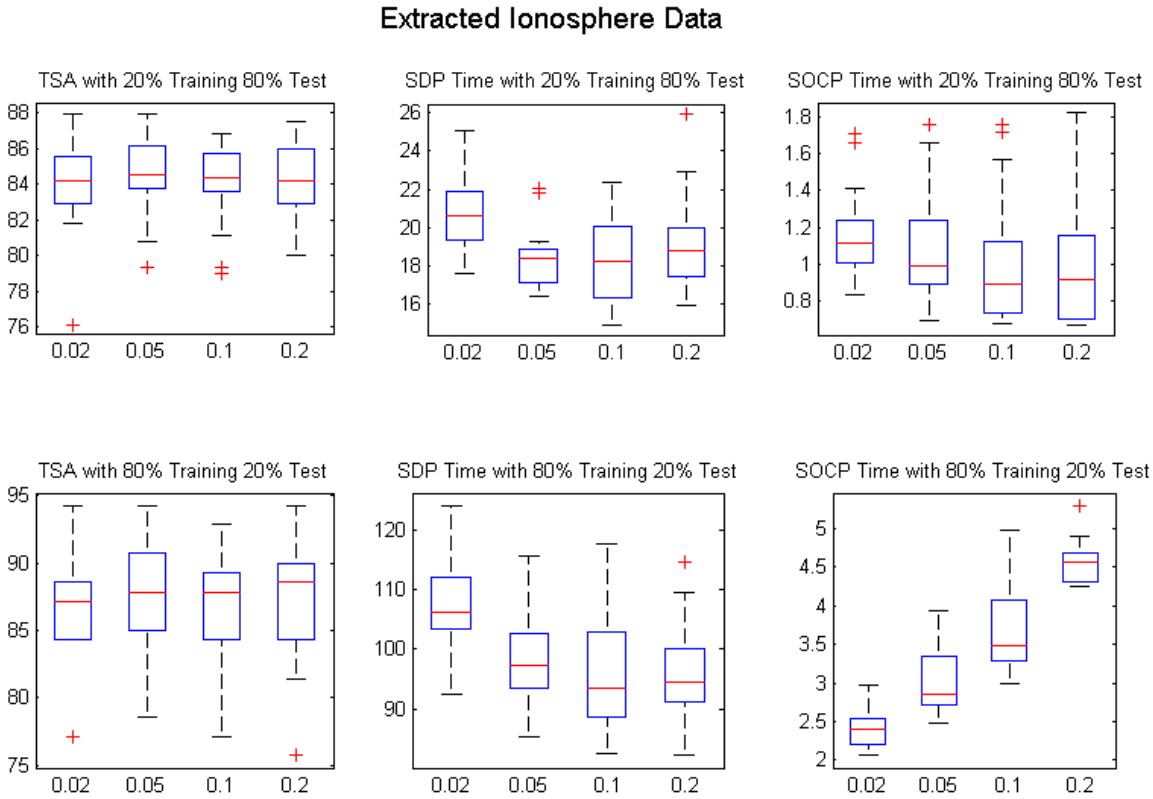
Table: 20% training, 80% test

	$\varepsilon = 0.02$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$
Test Set Accuracy	84.0±2.5%	84.4±2.1%	84.1±2.2%	84.2±2.2%
SDP Running Time	20.6±1.8	18.3±1.6	18.1±2.1	19.1±2.4
SOCP Running Time	1.1±0.2	1.1±0.3	1.0±0.3	1.0±0.4

Table: 80% training, 20% test

	$\varepsilon = 0.02$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$
Test Set Accuracy	86.9±3.6%	87.8±3.8%	87.2±3.9%	87.2±4.3%
SDP Running Time	107.4±7.6	97.9±7.3	96.0±10.0	95.7±8.3
SOCP Running Time	2.4±0.2	3.0±0.4	3.7±0.5	4.6±0.3

Extracted Ionosphere Data Classification Result



Outline

Estimation Errors

- In practice, the distribution properties are often unknown but need to be estimated from data.
- If an uncertain data point $\tilde{\mathbf{x}}_i = [\tilde{x}_{i1}, \ldots, \tilde{x}_{in}]^\top$ has N samples $\mathbf{x}_{i_k}, k = 1, \ldots, N$, then the sample mean $\bar{\mathbf{x}}_i = \frac{1}{N} \sum_{k=1}^{N} \mathbf{x}_{i_k}$ is used to estimate the mean vector $\boldsymbol{\mu}_i = \mathbf{E}[\tilde{\mathbf{x}}_i]$, and the sample covariance $\mathbf{S}_i = \frac{1}{N-1} \sum_{k=1}^{N} (\mathbf{x}_{i_k} - \bar{\mathbf{x}}_i)(\mathbf{x}_{i_k} - \bar{\mathbf{x}}_i)^\top$ is used to estimate the covariance matrix $\boldsymbol{\Sigma}_i = \mathbf{E}\left[(\tilde{\mathbf{x}}_i - \boldsymbol{\mu}_i)(\tilde{\mathbf{x}}_i - \boldsymbol{\mu}_i)^\top\right]$.
- However, these could cause possible estimation errors.
- Three special cases when the mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ are not exactly known are discussed here.

Estimation Error I: $\boldsymbol{\mu}_i \in [\boldsymbol{\mu}_i^-, \boldsymbol{\mu}_i^+]$, $\boldsymbol{\Sigma}_i = \mathbf{S}_i$

The interval of $\boldsymbol{\mu}_i$ works for each element in the feature vector, i.e. $\mu_{ij} \in [\mu_{ij}^-, \mu_{ij}^+]$, $j = 1, \ldots, n$.

(SVM-SOCP-Mu1)

$$\min_{\mathbf{w}, b, \xi_i, z_{ij}} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m} \xi_i$$

$$\text{s.t.} \quad \sum_j z_{ij} + y_i b \geq 1 - \xi_i + \sqrt{\frac{1-\varepsilon}{\varepsilon}}\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}\mathbf{w}\|_2$$

$$z_{ij} \leq y_i \mu_{ij}^- w_j, \; z_{ij} \leq y_i \mu_{ij}^+ w_j$$

$$\xi_i \geq 0, \quad i = 1, \ldots, m$$

Estimation Error I: $\boldsymbol{\mu}_i \in [\boldsymbol{\mu}_i^-, \boldsymbol{\mu}_i^+]$, $\boldsymbol{\Sigma}_i = \mathbf{S}_i$

- This case is applied when the confidence interval of μ_{ij} could be estimated.
- For a random variable $\tilde{x}_{ij}$ with normal distribution, and N samples $x_{ij_k}, k = 1, \ldots, N$, the sample mean $\bar{x}_{ij} = \frac{1}{N}\sum_{k=1}^{N} x_{ij_k}$, the unbiased sample variance $s_{ij}^2 = \frac{1}{N-1}\sum_{k=1}^{N}(x_{ij_k} - \bar{x}_{ij})^2$, then $\frac{\bar{x}_{ij}-\mu_{ij}}{s_{ij}/\sqrt{N}} \sim t_{N-1}$.
- The confidence interval of μ_{ij} is $\left[\bar{x}_{ij} - t_{crit} \cdot s_{ij}/\sqrt{N}, \bar{x}_{ij} + t_{crit} \cdot s_{ij}/\sqrt{N}\right]$, where t_{crit} is the coefficient corresponding to the confidence level $1-\alpha$ and the degree of freedom $N-1$.

Estimation Error I: $\boldsymbol{\mu}_i \in [\boldsymbol{\mu}_i^-, \boldsymbol{\mu}_i^+]$, $\boldsymbol{\Sigma}_i = \mathbf{S}_i$

- For n-dimensional vector $\tilde{\mathbf{x}}_i \in \mathbb{R}^n$, the Bonferroni correction factor uses α/n instead of α for each of the n univariate confidence interval.

- The geometric interpretation of this case is that, for each point $\mathbf{x}_i$, it is replaced by a union of ellipsoids, with center varies in the hyper-rectangle $[\boldsymbol{\mu}_i^-, \boldsymbol{\mu}_i^+]$, and shaped with the covariance matrix $\sqrt{\frac{1-\varepsilon}{\varepsilon}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}$.

Estimation Error II: $(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i) \leq \nu_i^2, \boldsymbol{\Sigma}_i = \mathbf{S}_i$

- Since the Bonferroni correction is for the case when the random variables $\tilde{x}_{ij}$ in $\tilde{\mathbf{x}}_i$ are independent, it would over-correct and result in lower α and larger robust region than it needs to be when they are not independent.
- The Hotelling's T-square test statistic $T^2 = N(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})$, and it has the property that $\frac{N-n}{n(N-1)} T^2 \sim F(n, N-n)$.
- Then the confidence region for $\boldsymbol{\mu}_i$ is $T^2 \leq \frac{n(N-1)}{N-n} F_{crit}$, i.e., $(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i)^\top \mathbf{S}_i^{-1} (\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i) \leq \frac{n(N-1)}{N(N-n)} F_{crit}$, where F_{crit} is the coefficient corresponding to the confidence level $1 - \alpha$ and the degree of freedom $(n, N-n)$.

Estimation Error II: $(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i) \leq \nu_i^2, \boldsymbol{\Sigma}_i = \mathbf{S}_i$

- Let $\nu_i^2 = \frac{n(N-1)}{N(N-n)} F_{crit}$, the geometric interpretation is that, the mean vector $\boldsymbol{\mu}_i$ varies in an ellipsoid centered at $\bar{\mathbf{x}}_i$ and shaped with $\nu_i \boldsymbol{\Sigma}_i^{\frac{1}{2}}$.
- Then the uncertainty set for each point $\mathbf{x}_i$ is a union of ellipsoids, with center varies in the ellipsoid $\mathscr{E}\left(\bar{\mathbf{x}}_i, \nu_i \boldsymbol{\Sigma}_i^{\frac{1}{2}}\right)$, and shaped with $\sqrt{\frac{1-\varepsilon}{\varepsilon}} \boldsymbol{\Sigma}_i^{\frac{1}{2}}$.
- This has a more concise form that the union of these ellipsoids is also an ellipsoid $\mathscr{E}\left(\bar{\mathbf{x}}_i, (\sqrt{\frac{1-\varepsilon}{\varepsilon}} + \nu_i) \boldsymbol{\Sigma}_i^{\frac{1}{2}}\right)$.

Estimation Error II: $(\mu_i - \bar{\mathbf{x}}_i)^\top \mathbf{\Sigma}_i^{-1}(\mu_i - \bar{\mathbf{x}}_i) \leq \nu_i^2, \mathbf{\Sigma}_i = \mathbf{S}_i$

The model in this case is:

(SVM-SOCP-Mu2)

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top \mu_i + b) \geq 1 - \xi_i + \left(\sqrt{\frac{1-\varepsilon}{\varepsilon}} + \nu_i\right)||\mathbf{\Sigma}_i^{\frac{1}{2}}\mathbf{w}||_2$$

$$\xi_i \geq 0, \quad i = 1, \ldots, m$$

Estimation Error III: $\mu_i = \bar{\mathbf{x}}_i$, $||\mathbf{\Sigma}_i - \mathbf{S}_i||_F \leq \rho_i$

The Frobenius norm is $||\mathbf{A}||_F^2 = \text{Trace}(\mathbf{A}^\top \mathbf{A}) = \sum_{ij} A_{ij}^2$. In this case, the uncertainty set becomes $\mathscr{E}\left(\boldsymbol{\mu}_i, \sqrt{\frac{1-\varepsilon}{\varepsilon}}\left(\mathbf{\Sigma}_i + \rho_i I_n\right)^{\frac{1}{2}}\right)$.

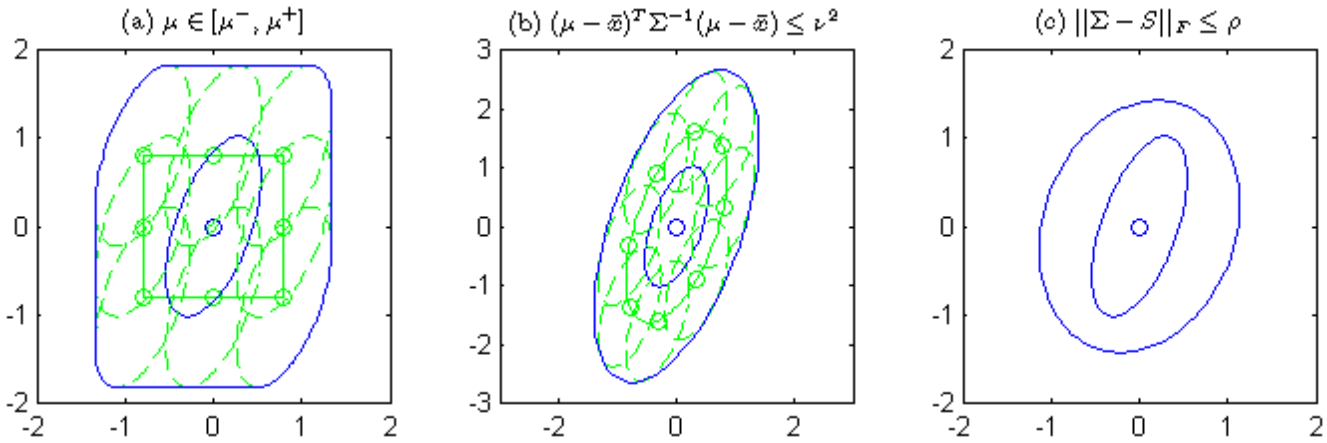
(SVM-SOCP-Cov)

$$\min_{\mathbf{w},b,\xi_i} \ \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \ \ y_i(\mathbf{w}^\top \boldsymbol{\mu}_i + b) \geq 1 - \xi_i + \sqrt{\frac{1-\varepsilon}{\varepsilon}}\left\|\left(\mathbf{\Sigma}_i + \rho_i I_n\right)^{\frac{1}{2}}\mathbf{w}\right\|_2$$

$$\xi_i \geq 0, \quad i = 1, \ldots, m$$

Estimation Errors



Performance Measures

- Test Set Accuracy (TSA) is a direct way to evaluate the model performance.
- TSA is computed by counting the number of correctly predicted labels in the test data set and divided by the size of the test set.
- The class label y_i is decided by the sign$(\mathbf{w}^\top \mathbf{x}_i + b)$.
- When there are replicates $\mathbf{x}_{i_k}$ for the test point $\mathbf{x}_i$, the class label y_i is decided by the majority label of the replicates sign$(\mathbf{w}^\top \mathbf{x}_{i_k} + b)$.

Performance Measures

- Nominal Error and Optimal Error can also be used to evaluate the model performance.
- The nominal error is similar to TSA but the opposite, i.e., TSA + NomErr = 1.
- The expression for NomErr is:

$$\text{NomErr} = \frac{\sum_i 1_{y_i^{pr} \neq y_i}}{\#\text{ test datapoints}} \times 100\%$$

Performance Measures

- The optimal error is based on the probability of misclassification.
- For $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\{y_i(\mathbf{w}^\top \tilde{\mathbf{x}}_i + b) \leq 0\} \leq \varepsilon$, it can be similarly transformed into $y_i(\mathbf{w}^\top \boldsymbol{\mu}_i + b) \geq \sqrt{\frac{1-\varepsilon}{\varepsilon}} ||\boldsymbol{\Sigma}_i^{\frac{1}{2}} \mathbf{w}||_2$.

$$\varepsilon_{opt} = \frac{\mathbf{w}^\top \boldsymbol{\Sigma}_i \mathbf{w}}{(\mathbf{w}^\top \boldsymbol{\mu}_i + b)^2 + \mathbf{w}^\top \boldsymbol{\Sigma}_i \mathbf{w}}$$

- The OptErr of data point $\mathbf{x}_i$ is

$$\text{OptErr}_i = \begin{cases} 1, & \text{if } y_i^{pr} \neq y_i \\ \varepsilon_{opt}, & \text{if } y_i^{pr} = y_i \end{cases}$$

- The OptErr of the whole test set is

$$\text{OptErr} = \frac{\sum_i \text{OptErr}_i}{\#\text{ test datapoints}} \times 100\%$$

Outline

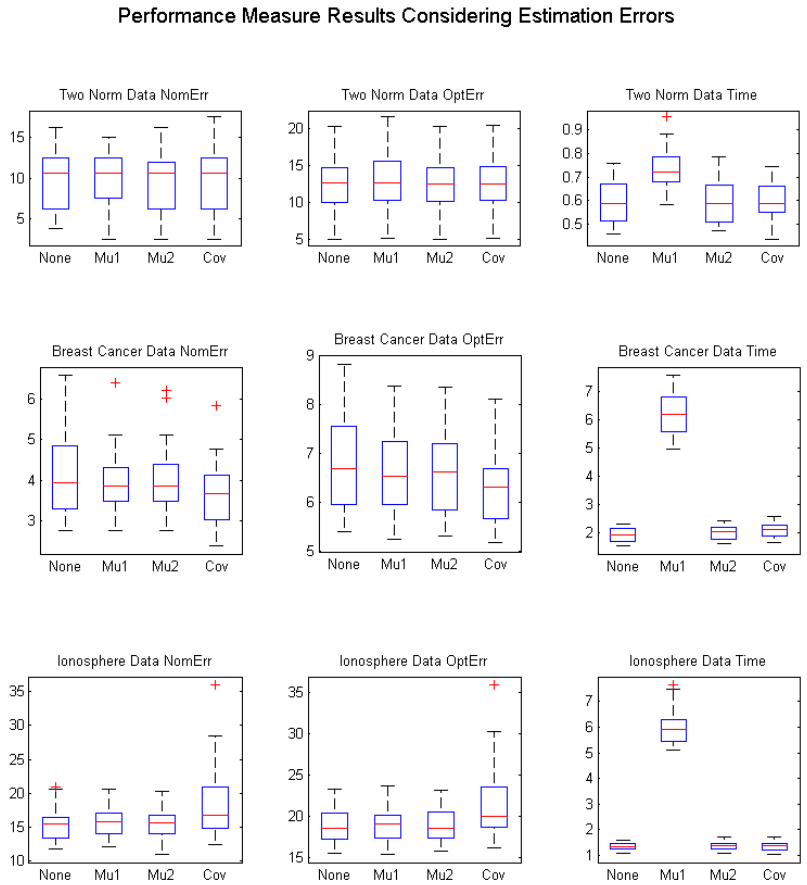
Numerical Experiments

- The three estimation error cases $\boldsymbol{\mu}_i \in [\boldsymbol{\mu}_i^-, \boldsymbol{\mu}_i^+]$, $(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i) \leq \nu_i^2$, and $||\boldsymbol{\Sigma}_i - \mathbf{S}_i||_F \leq \rho_i$ were experimented on the two norm data, the Wisconsin breast cancer data, and the Ionosphere data.
- For each data point $\mathbf{x}_i$, $N = 50$ replicates $\mathbf{x}_{i_k} (k = 1, \ldots, N)$ were generated with mean equal to the value of the data point $\mathbf{x}_i$, and covariance equal to 0.01 times the covariance of the training dataset.
- For the two norm data, since we generated the data using $\boldsymbol{\Sigma}_+ = \boldsymbol{\Sigma}_- = \mathbf{I}$, the replicates generation covariance used $0.01\mathbf{I}$.

Numerical Experiments

Two Norm	$\boldsymbol{\mu}_i = \bar{\mathbf{x}}_i$ $\boldsymbol{\Sigma}_i = \mathbf{S}_i$	$\boldsymbol{\mu}_i \in [\boldsymbol{\mu}_i^-, \boldsymbol{\mu}_i^+]$ $\boldsymbol{\Sigma}_i = \mathbf{S}_i$	$\boldsymbol{\mu}_i \in \mathscr{E}(\bar{\mathbf{x}}_i, \nu_i \boldsymbol{\Sigma}_i^{\frac{1}{2}})$ $\boldsymbol{\Sigma}_i = \mathbf{S}_i$	$\boldsymbol{\mu}_i = \bar{\mathbf{x}}_i$ $\|\|\boldsymbol{\Sigma}_i - \mathbf{S}_i\|\|_F \leq \rho_i$
NomErr	9.75±3.69%	9.63±3.42%	9.63±3.61%	9.69±3.82%
OptErr	12.44±3.92%	12.62±3.92%	12.46±3.84%	12.43±3.75%
Time	0.59±0.10	0.74±0.09	0.59±0.09	0.60±0.08
Breast Cancer	$\boldsymbol{\mu}_i = \bar{\mathbf{x}}_i$ $\boldsymbol{\Sigma}_i = \mathbf{S}_i$	$\boldsymbol{\mu}_i \in [\boldsymbol{\mu}_i^-, \boldsymbol{\mu}_i^+]$ $\boldsymbol{\Sigma}_i = \mathbf{S}_i$	$\boldsymbol{\mu}_i \in \mathscr{E}(\bar{\mathbf{x}}_i, \nu_i \boldsymbol{\Sigma}_i^{\frac{1}{2}})$ $\boldsymbol{\Sigma}_i = \mathbf{S}_i$	$\boldsymbol{\mu}_i = \bar{\mathbf{x}}_i$ $\|\|\boldsymbol{\Sigma}_i - \mathbf{S}_i\|\|_F \leq \rho_i$
NomErr	4.16±1.03%	3.99±0.83%	4.07±0.92%	3.66±0.82%
OptErr	6.82±0.99%	6.61±0.88%	6.64±0.89%	6.26±0.75%
Time	1.94±0.24	6.18±0.77	2.01±0.26	2.10±0.28
Ionosphere Data	$\boldsymbol{\mu}_i = \bar{\mathbf{x}}_i$ $\boldsymbol{\Sigma}_i = \mathbf{S}_i$	$\boldsymbol{\mu}_i \in [\boldsymbol{\mu}_i^-, \boldsymbol{\mu}_i^+]$ $\boldsymbol{\Sigma}_i = \mathbf{S}_i$	$\boldsymbol{\mu}_i \in \mathscr{E}(\bar{\mathbf{x}}_i, \nu_i \boldsymbol{\Sigma}_i^{\frac{1}{2}})$ $\boldsymbol{\Sigma}_i = \mathbf{S}_i$	$\boldsymbol{\mu}_i = \bar{\mathbf{x}}_i$ $\|\|\boldsymbol{\Sigma}_i - \mathbf{S}_i\|\|_F \leq \rho_i$
NomErr	15.21±2.58%	15.60±2.17%	15.55±2.41%	19.15±6.81%
OptErr	18.87±2.20%	18.92±2.15%	18.93±2.21%	22.02±5.76%
Time	1.35±0.13	5.98±0.69	1.38±0.16	1.37±0.18

Numerical Experiments



Outline

Primal-Dual Interior Point Method

- Primal-dual interior point methods were originally used for linear programming. These methods create the system of equations $\mathbf{Au} = \mathbf{b}$, $\mathbf{A}^\top \mathbf{v} + \mathbf{s} = \mathbf{c}$, and relax the complementarity conditions into $u_i s_i = \mu$, then apply Newton's method to solve the system.

- SeDuMi uses primal-dual interior-point method to solve conic linear program problems. In each iteration, a search direction $(\Delta\mathbf{u}, \Delta\mathbf{v}, \Delta\mathbf{s})$ is computed and added to the current feasible solution $(\mathbf{u}, \mathbf{v}, \mathbf{s})$ with the step length $t > 0$, and the next feasible solution $(\mathbf{u}^+, \mathbf{v}^+, \mathbf{s}^+)$ is

$$(\mathbf{u}^+, \mathbf{v}^+, \mathbf{s}^+) = (\mathbf{u}, \mathbf{v}, \mathbf{s}) + t(\Delta\mathbf{u}, \Delta\mathbf{v}, \Delta\mathbf{s})$$

Primal-Dual Interior Point Method

- The search direction $(\Delta\mathbf{u}, \Delta\mathbf{v}, \Delta\mathbf{s})$ is defined by the following equations:

$$\Delta\mathbf{u} + \mathbf{\Pi}\Delta\mathbf{s} = \mathbf{r}$$
$$\mathbf{A}\Delta\mathbf{u} = \mathbf{0}$$
$$\mathbf{A}^\top\Delta\mathbf{v} + \Delta\mathbf{s} = \mathbf{0}$$

  where $\mathbf{\Pi}$ is an invertible block diagonal matrix which satisfies $\mathbf{\Pi}^\top\mathbf{s} = \mathbf{u}$.

- When setting $\mathbf{r} = -\mathbf{u}$, then $(\mathbf{u}^+)^\top\mathbf{s}^+ = (1-t)\mathbf{u}^\top\mathbf{s}$. The duality gap is decreasing in each iteration.

- Instead of solving the system directly, $\Delta\mathbf{v}$ can be solved by a reduced system. When $\mathbf{r} = -\mathbf{u}$, the reduced system is

$$\mathbf{A}\mathbf{\Pi}\mathbf{A}^\top\Delta\mathbf{v} = \mathbf{b}$$

  After solving $\Delta\mathbf{v}$, then we can obtain $\Delta\mathbf{s} = -\mathbf{A}^\top\Delta\mathbf{v}$, $\Delta\mathbf{u} = \mathbf{r} - \mathbf{\Pi}\Delta\mathbf{s}$.

Primal-Dual Interior Point Method

Primal-Dual Interior Method in SeDuMi

Step 0: Initial solution $(\mathbf{u}, \mathbf{v}, \mathbf{s}) \in K \times \mathbb{R}^m \times K$ with
$\mathbf{A}\mathbf{u} = \mathbf{b}$ and $\mathbf{A}^\top \mathbf{v} + \mathbf{s} = \mathbf{c}$ such that $\lambda(P(\mathbf{u})^{1/2}\mathbf{s}) \in N$.
Step 1: If $\mathbf{u}^\top \mathbf{s} \leq \epsilon$ then STOP.
Step 2: Choose $\mathbf{\Pi}$ and $\mathbf{r}$ according to algorithmic settings.
Compute the search direction $(\Delta\mathbf{u}, \Delta\mathbf{v}, \Delta\mathbf{s})$.
Determine the step length $t > 0$ s.t. $\lambda(P(\mathbf{u} + t\Delta\mathbf{u})^{1/2}(\mathbf{s} + t\Delta\mathbf{s})) \in N$.
Step 3: Update $(\mathbf{u}, \mathbf{v}, \mathbf{s}) \leftarrow (\mathbf{u} + t\Delta\mathbf{u}, \mathbf{v} + t\Delta\mathbf{v}, \mathbf{s} + t\Delta\mathbf{s})$ and return to Step 1.

The worst case iteration bound for SeDuMi is $O(\sqrt{\nu(K)}|\log\epsilon|)$ where $\nu(K)$ is the order of the cone K.

Outline

Stochastic Gradient Descent Method

- For a two-class dataset of m data points $\{\mathbf{x}_i, y_i\}_{i=1}^m$ with n-dimensional features $\mathbf{x}_i \in \mathbb{R}^n$ and respective class labels $y_i \in \{+1, -1\}$, the soft margin SVM can be expressed in this particular form:

$$\min_{\mathbf{w},b} \quad f(\mathbf{w}, b) = \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C\sum_{i=1}^{m} \max\{0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\}$$

- The second term is also called penalty function. The function $L(\mathbf{x}_i, y_i) = \max\{0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\}$ decreases linearly for $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq 1$ and then remains 0, so it is a hinge function, its value is called the hinge loss.

Stochastic Gradient Descent Method

- For large scale data, gradient descent method has advantages. To minimize f, the gradient is computed, then the current $\mathbf{w}$ is moved in the direction opposite to the direction of the gradient.

- A constant η_t is chosen to be the fraction of the gradient that to be moved in each round. The gradient descent iteration is:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \nabla_{\mathbf{w}} f$$

- The gradient of f with respect to $\mathbf{w}$ can be computed as

$$\begin{aligned}\nabla_{\mathbf{w}} f &= \mathbf{w} + C \sum_{i=1}^{m} \nabla_{\mathbf{w}} L(\mathbf{x}_i, y_i) \\ &= \mathbf{w} + C \sum_{i=1}^{m} \begin{cases} 0, & \text{if } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \\ -y_i \mathbf{x}_i, & \text{otherwise} \end{cases}\end{aligned}$$

Stochastic Gradient Descent Method

- The gradient descent method is also called batch gradient descent because all the training points are considered as a batch at each round. The problem of the batch gradient descent method is that to compute $\nabla_{\mathbf{w}} f$, it needs to go over all the m training data points. When the data size is large, it can be too time-consuming to visit every training point and often iterates many times before convergence.

- The stochastic gradient descent, on the other hand, considers one training point at a time and adjusts the current solution in the direction evaluated by the only training point:

$$\nabla_{\mathbf{w}} f_t = \mathbf{w} + Cm\nabla_{\mathbf{w}} L(\mathbf{x}_{i_t}, y_{i_t})$$

  The training point $(\mathbf{x}_{i_t}, y_{i_t})$ can be selected randomly or according to some fixed strategy.

Stochastic Gradient Descent Method

Batch Gradient Descent Method

Iterate until convergence:

Evaluate: $\nabla_{\mathbf{w}} f = \mathbf{w} + C \sum_{i=1}^{m} \nabla_{\mathbf{w}} L(\mathbf{x}_i, y_i)$

Update: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \nabla_{\mathbf{w}} f$

Stochastic Gradient Descent Method

Iterate until convergence:

Evaluate: $\nabla_{\mathbf{w}} f_t = \mathbf{w} + Cm \nabla_{\mathbf{w}} L(\mathbf{x}_{i_t}, y_{i_t})$

Update: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \nabla_{\mathbf{w}} f_t$

Stochastic Gradient Descent Method

The batch gradient descent method improves the value of the objective function at every step. The stochastic gradient descent method improves the value in a noisy way since it only considers one point at each iteration. The batch gradient descent method takes fewer iterations to converge, but in each iteration, it takes much longer to compute. In practice, the stochastic gradient descent method is much faster.

The computational cost of batch gradient descent method and stochastic gradient descent method is:

	BGD	SGD
Time per iteration	mn	n
Iterations to ϵ-accuracy	$\log(1/\epsilon)$	$1/\epsilon$
Time to ϵ-accuracy	$mn\log(1/\epsilon)$	n/ϵ

Stochastic Gradient Descent Method

- As an application of the stochastic gradient descent method, Pegasos (Primal Estimated sub-GrAdient SOlver for SVM) studied the SVM problem in this form:

$$\min_{\mathbf{w}} \quad f(\mathbf{w}) = \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w} + \frac{1}{m}\sum_{i=1}^{m}\max\{0, 1 - y_i\mathbf{w}^\top\mathbf{x}_i\}$$

- On iteration t, a random training point $(\mathbf{x}_{i_t}, y_{i_t})$ is chosen uniformly with $i_t \in \{1, \ldots, m\}$. The objective function is approximated with the training point $(\mathbf{x}_{i_t}, y_{i_t})$:

$$f_t(\mathbf{w}) = \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w} + \max\{0, 1 - y_{i_t}\mathbf{w}^\top\mathbf{x}_{i_t}\}$$

- The step size is set to be $\eta_t = 1/(\lambda t)$. The update is:

$$\mathbf{w}_{t+1} \leftarrow (1 - \frac{1}{t})\mathbf{w}_t + \eta_t 1_{[y_{i_t}\mathbf{w}_t^\top\mathbf{x}_{i_t}<1]} y_{i_t}\mathbf{x}_{i_t}$$

Stochastic Gradient Descent Method

Pegasos

Initialize $\mathbf{w}_1 = \mathbf{0}$
For $t = 1, 2, \ldots, T$
  Choose $i_t \in \{1, \ldots, m\}$ uniformly at random
  Set $\eta_t = 1/(\lambda t)$
  If $y_{i_t} \mathbf{w}_t^\top \mathbf{x}_{i_t} < 1$, then
    Set $\mathbf{w}_{t+1} \leftarrow (1 - 1/t)\mathbf{w}_t + \eta_t y_{i_t} \mathbf{x}_{i_t}$
  Else $(y_{i_t} \mathbf{w}_t^\top \mathbf{x}_{i_t} \geq 1)$
    Set $\mathbf{w}_{t+1} \leftarrow (1 - 1/t)\mathbf{w}_t$

Pegasos can obtain an ϵ-accuracy solution for the primal problem in $\tilde{O}(1/\epsilon)$ iterations with the cost per iteration $O(n)$.

Outline

Solving Large Scale SVM-RCCP

- The SOCP reformulation of robust chance-constrained SVM is:

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \boldsymbol{\mu}_i + b) \geq 1 - \xi_i + \sqrt{\frac{1-\varepsilon}{\varepsilon}}||\boldsymbol{\Sigma}_i^{\frac{1}{2}}\mathbf{w}||_2 \\ & \xi_i \geq 0, \quad i = 1,\ldots,m \end{aligned}$$

- Write the model in the form similar to gradient descent method:

$$\min_{\mathbf{w},b} \; f(\mathbf{w},b) = \frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\sum_{i=1}^{m}\max\left\{0, 1 - y_i(\mathbf{w}^\top\boldsymbol{\mu}_i + b) + \sqrt{\frac{1-\varepsilon}{\varepsilon}\mathbf{w}^\top\boldsymbol{\Sigma}_i\mathbf{w}}\right\}$$

Solving Large Scale SVM-RCCP

- For $L(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, y_i) = \max\left\{0, 1 - y_i(\mathbf{w}^\top \boldsymbol{\mu}_i + b) + \sqrt{\frac{1-\varepsilon}{\varepsilon}\mathbf{w}^\top \boldsymbol{\Sigma}_i \mathbf{w}}\right\}$, the gradient

$$\nabla_{\mathbf{w}} L(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, y_i) = \begin{cases} 0, & \text{if } y_i(\mathbf{w}^\top \boldsymbol{\mu}_i + b) \geq 1 + \sqrt{\frac{1-\varepsilon}{\varepsilon}}||\boldsymbol{\Sigma}_i^{\frac{1}{2}}\mathbf{w}||_2 \\ -y_i\boldsymbol{\mu}_i + \sqrt{\frac{1-\varepsilon}{\varepsilon}}\frac{\boldsymbol{\Sigma}_i \mathbf{w}}{||\boldsymbol{\Sigma}_i^{\frac{1}{2}}\mathbf{w}||_2}, & \text{otherwise} \end{cases}$$

- For $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \nabla_{\mathbf{w}} f_t$, the gradient considered is $1/(Cm)$ of the original f:

$$\nabla_{\mathbf{w}} f_t = \frac{1}{Cm}\mathbf{w} + \nabla_{\mathbf{w}} L(\boldsymbol{\mu}_{i_t}, \boldsymbol{\Sigma}_{i_t}, y_{i_t})$$

- The step size $\eta_t = Cm/t$.

Solving Large Scale SVM-RCCP

SVM-RCCP SGD Method

Initialize $\mathbf{w}_1 = \mathbf{0}, b_1 = 0$

For $t = 1, 2, \ldots, T$

Choose $i_t \in \{1, \ldots, m\}$ uniformly at random

Set $\eta_t = Cm/t$

If $y_{i_t}(\mathbf{w}_t^\top \boldsymbol{\mu}_{i_t} + b_t) < 1 + \sqrt{\frac{1-\varepsilon}{\varepsilon}}||\boldsymbol{\Sigma}_{i_t}^{\frac{1}{2}}\mathbf{w}_t||_2$, then

$$\text{Set } \mathbf{w}_{t+1} \leftarrow (1 - 1/t)\mathbf{w}_t + \eta_t\left(y_{i_t}\boldsymbol{\mu}_{i_t} - \sqrt{\frac{1-\varepsilon}{\varepsilon}}\frac{\boldsymbol{\Sigma}_{i_t}\mathbf{w}_t}{||\boldsymbol{\Sigma}_{i_t}^{\frac{1}{2}}\mathbf{w}_t||_2}\right)$$

$$b_{t+1} \leftarrow b_t + \eta_t y_{i_t}$$

Else $(y_{i_t}(\mathbf{w}_t^\top \boldsymbol{\mu}_{i_t} + b_t) \geq 1 + \sqrt{\frac{1-\varepsilon}{\varepsilon}}||\boldsymbol{\Sigma}_{i_t}^{\frac{1}{2}}\mathbf{w}_t||_2)$

$$\text{Set } \mathbf{w}_{t+1} \leftarrow (1 - 1/t)\mathbf{w}_t$$

$$b_{t+1} \leftarrow b_t$$

Numerical Experiments

- Three sets of data are used in the numerical experiments.

  Wisconsin breast cancer data: 683 samples, 9-dimensional features
  Ionosphere data: 351 samples, 34-dimensional features
  MAGIC Gamma Telescope data: 19020 samples, 10-dimensional features

- The sampling procedure is chosen to be sampling without replacement and new permutation is generated every epoch. In the experiments, the iterations goes 2 epochs, 10 epochs, and 50 epochs over the training data.

- As comparison, SeDuMi is used to solve the SOCP reformulation of robust chance-constrained SVM directly.

Wisconsin Breast Cancer Data

Wisconsin Breast Cancer Data: 683 samples, 9-dimensional features

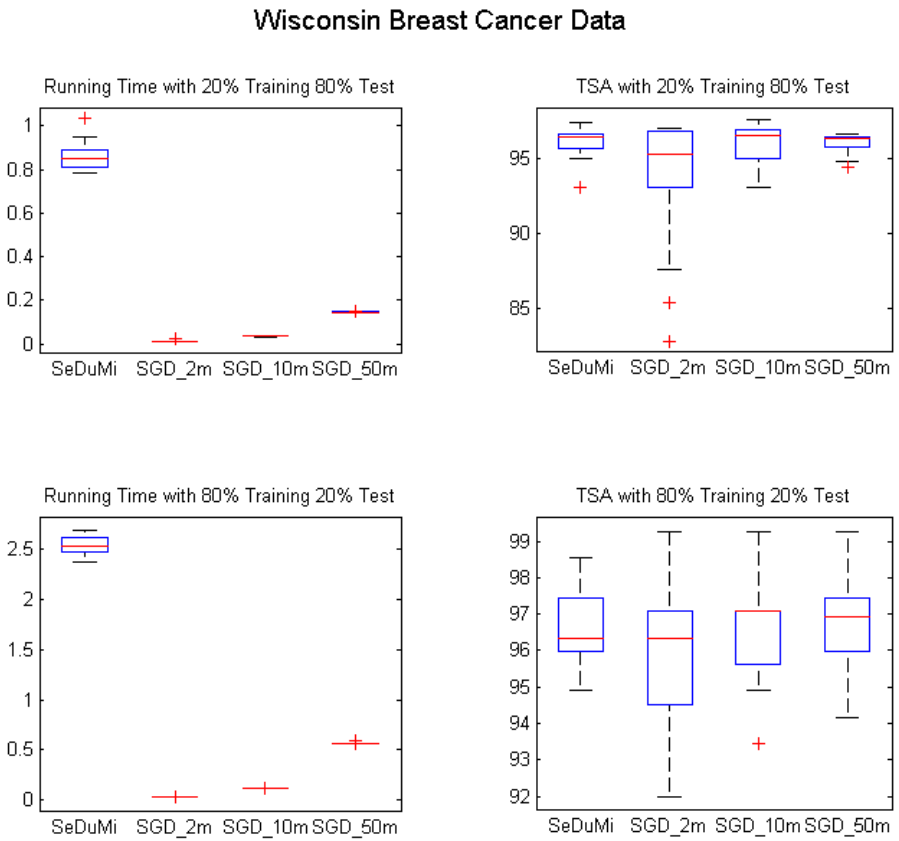
Table: 20% training, 80% test

	SeDuMi	SGD $2m$	SGD $10m$	SGD $50m$
Time	0.857±0.063	0.013±0.002	0.034±0.001	0.146±0.002
TSA	96.16±0.98%	93.90±4.23%	96.09±1.27%	96.08±0.64%

Table: 80% training, 20% test

	SeDuMi	SGD $2m$	SGD $10m$	SGD $50m$
Time	2.542±0.094	0.029±0.001	0.117±0.001	0.564±0.007
TSA	96.72±1.12%	95.84±2.12%	96.68±1.31%	96.70±1.28%

Wisconsin Breast Cancer Data



Ionosphere Data

Ionosphere Data: 351 samples, 34-dimensional features

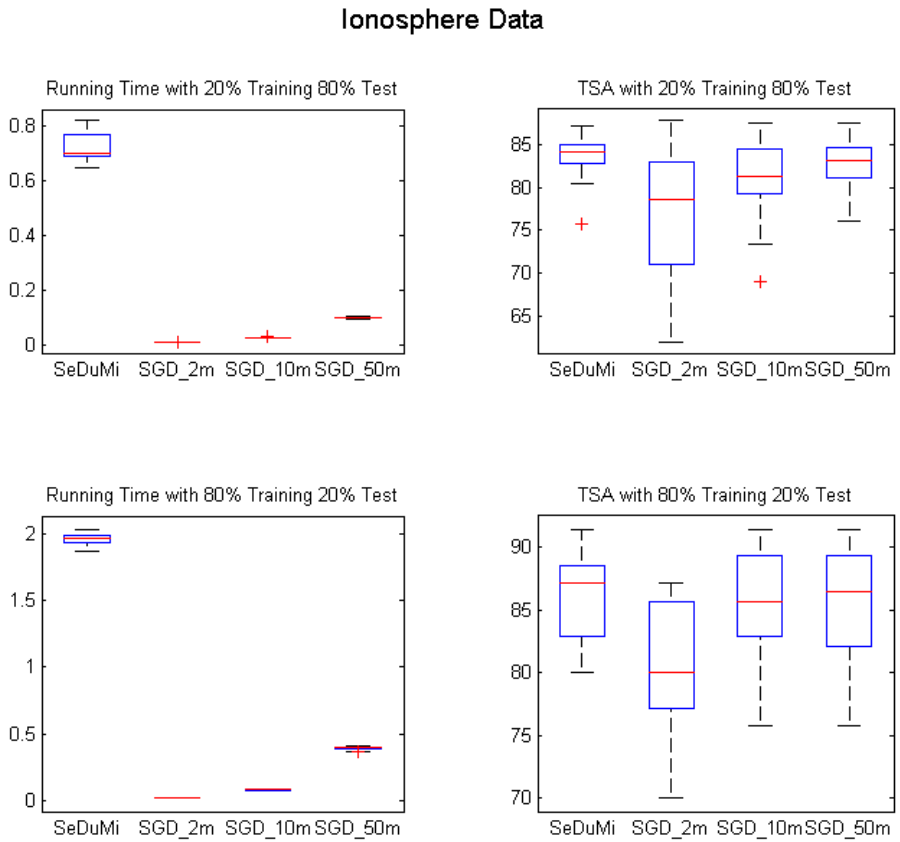
Table: 20% training, 80% test

	SeDuMi	SGD $2m$	SGD $10m$	SGD $50m$
Time	0.728±0.053	0.010±0.000	0.024±0.001	0.099±0.003
TSA	83.56±2.50%	76.89±7.96%	81.09±4.63%	82.85±2.93%

Table: 80% training, 20% test

	SeDuMi	SGD $2m$	SGD $10m$	SGD $50m$
Time	1.961±0.044	0.022±0.001	0.083±0.002	0.394±0.013
TSA	85.93±3.48%	80.71±5.07%	85.57±4.14%	85.64±4.33%

Ionosphere Data



MAGIC Gamma Telescope Data

MAGIC Gamma Telescope Data: 19020 samples, 10-dimensional features

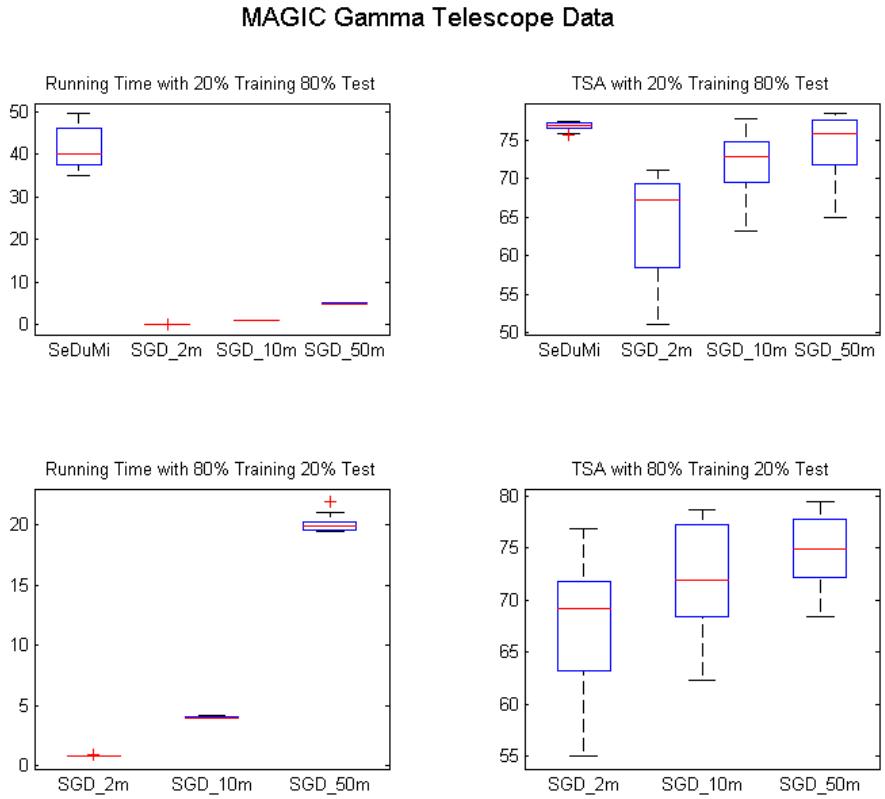
Table: 20% training, 80% test

	SeDuMi	SGD $2m$	SGD $10m$	SGD $50m$
Time	41.198±4.558	0.215±0.002	1.031±0.011	5.011±0.115
TSA	76.80±0.52%	64.29±6.61%	72.09±4.09%	74.35±3.87%

Table: 80% training, 20% test

	SeDuMi	SGD $2m$	SGD $10m$	SGD $50m$
Time	-	0.850±0.028	4.061±0.075	20.105±0.654
TSA	-	67.47±6.49%	72.14±4.71%	74.68±3.39%

MAGIC Gamma Telescope Data



Conclusions

- This talk presents robust chance-constrained SVM with second-order moment information and obtains equivalent SDP and SOCP reformulations.
- Three types of estimation errors for mean and covariance matrix are considered and the corresponding formulations and techniques to handle these types of errors are presented.
- The method to solve robust chance-constrained SVM with large scale data is proposed based on stochastic gradient descent method to process big data.

References


Vapnik, V. N. *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998.


Wang, X., and Pardalos, P. M. A Survey of Support Vector Machines with Uncertainties. *Annals of Data Science* (2014) 1: 293-309. doi:10.1007/s40745-014-0022-8


Wang, X., Fan, N., and Pardalos, P. M. Robust Chance-Constrained Support Vector Machines with Second-Order Moment Information. *Annals of Operations Research* (2015), 1-24. doi:10.1007/s10479-015-2039-6


Wang, X., Fan, N., and Pardalos, P. M. Stochastic Subgradient Descent Method for Large-Scale Robust Chance-Constrained Support Vector Machines. *Optimization Letters* (2017) 11: 1013-1024. doi:10.1007/s11590-016-1026-4


Wang, X., and Pardalos, P. M. A Modified Active Set Algorithm for Transportation Discrete Network Design Bi-Level Problem. *Journal of Global Optimization* (2017) 67: 325-342. doi:10.1007/s10898-015-0396-y

Thank You!