Support Vector Machines Classification with Robust Chance Constraints

Ximing Wang   Panos Pardalos

Department of Industrial and Systems Engineering
University of Florida

November 11, 2014

Outline

Outline

1. Robust Chance-Constrained SVM
   - Basic SVM Models
   - SVM with Robust Chance Constraints
2. Reformulation of RCC-SVM
   - RCC-SVM into SDP Models
   - RCC-SVM into SOCP Models
3. Preliminary Computational Results
   - Synthetic Data
   - Wisconsin Breast Cancer Data
4. Conclusions

Hard Margin SVM

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

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 $y_i \in \{\pm 1\}$.

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.

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

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

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

Soft Margin SVM

When two classes are not linearly separable:

Soft Margin SVM

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.

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.
- $\xi_i = \max\{0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\}$

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

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

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

$$\text{s.t.} \ \ y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, m$$

Outline

Chance-Constrained SVM

When uncertainties exist in the data points:

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 $y_i \in \{\pm 1\}$.

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

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

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

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

Robust Chance-Constrained SVM

The exact probability distribution are often unknown:

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.

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.

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

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

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

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

Outline

Moments Information

- Assume the first and second moment information of the random variables $\tilde{\mathbf{x}}_i$ are known.

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.

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:

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

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.

Theorem

RCC-SVM is equivalent to the following SDP formulation:

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

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

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

Outline

Multivariate Chebyshev Inequality

- Let $\tilde{\mathbf{x}} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote random vector $\tilde{\mathbf{x}}$ with mean $\boldsymbol{\mu}$ and convariance matrix $\boldsymbol{\Sigma}$.

Multivariate Chebyshev Inequality

- Let $\tilde{\mathbf{x}} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote random vector $\tilde{\mathbf{x}}$ with mean $\boldsymbol{\mu}$ and convariance matrix $\boldsymbol{\Sigma}$.
- 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

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

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

- 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.

- 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 RCC-SVM is:

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

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

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

Geometric Interpretation of the SOCP Model

- 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 $\boldsymbol{\Sigma}_i$:

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

Geometric Interpretation of the SOCP Model

- 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 $\boldsymbol{\Sigma}_i$:

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

- The SOCP constraint is satisfied if and only if

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

Geometric Interpretation of the SOCP Model

- 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 $\boldsymbol{\Sigma}_i$:

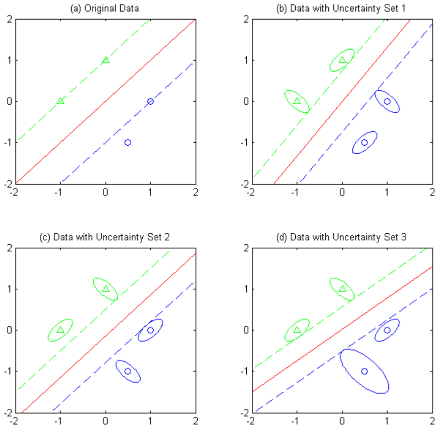
$$\mathscr{E}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \left\{ \mathbf{x} = \boldsymbol{\mu}_i + \sqrt{\frac{1-\varepsilon}{\varepsilon}} \boldsymbol{\Sigma}_i^{\frac{1}{2}} \mathbf{a} \ : \ ||\mathbf{a}||_2 \leq 1 \right\}$$

- The SOCP constraint is satisfied if and only if

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

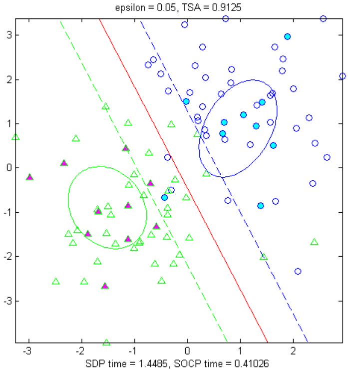
- This transforms the RCC-SVM into a robust optimization problem over the uncertainty set $\mathscr{E}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ for each uncertain training data point.

Geometric Interpretation of the SOCP Model

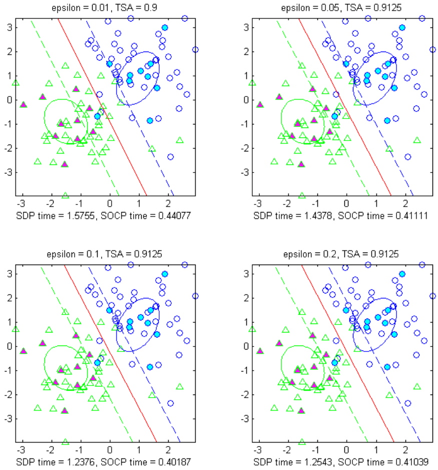


Outline

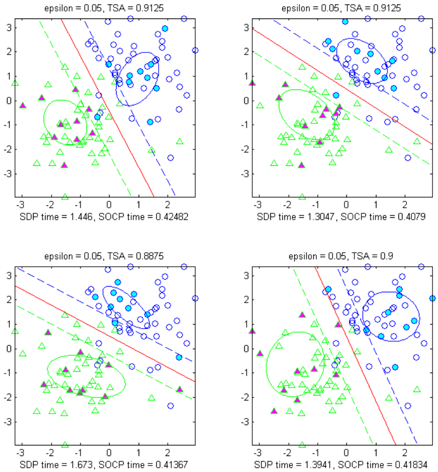
$+1$ class: 2-d normal distribution with $\boldsymbol{\mu}_+ = [1, 1]^\top$, $\boldsymbol{\Sigma}_+ = \boldsymbol{I}$
-1 class: 2-d normal distribution with $\boldsymbol{\mu}_- = [-1, -1]^\top$, $\boldsymbol{\Sigma}_- = \boldsymbol{I}$
Each class has 50 points: 10 for training, 40 for test



Classification Result with Different ε



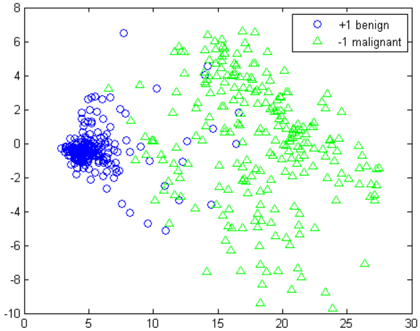
Classification Result with Different Training Set



Outline

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	96.52%	95.24%	95.24%	95.24%
SDP Running Time	35.1723	34.0854	28.6718	28.9409
SOCP Running Time	1.6846	1.6724	1.9918	2.1012

Table : 90% training, 10% test

	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$
Test Set Accuracy	98.53%	98.53%	97.06%	97.06%
SDP Running Time	216.1927	182.9951	189.6738	164.3278
SOCP Running Time	9.3391	10.9000	12.4002	13.8963

Conclusions

- The robust chance-constrained SVM is to ensure the small probability of misclassification for the uncertain data.
  - The exact probability distribution of the random variables are unknown.
  - Some properties of the distribution are known, for example, the moments information.

- When the mean and covariance of the data points are known, the RCC-SVM can be reformulated as both SDP and SOCP models.
  - The SDP and SOCP models are equivalent, which could be proved theoretically and by experiments.
  - The SOCP model runs more efficiently than SDP model.

References

Ben-Tal, A., Bhadra, S., Bhattacharyya, C., and Nath, J. S. Chance constrained uncertain classification via robust optimization. *Mathematical Programming* 127, 1 (2011),145-173.

Fan, N., Sadeghi, E., and Pardalos, P. M. Robust support vector machines with polyhedral uncertainty of the input data. In *Learning and Intelligent Optimization*. Springer, 2014, pp. 291-305.

Xanthopoulos, P., Guarracino, M. R., and Pardalos, P. M. Robust generalized eigenvalue classifier with ellipsoidal uncertainty. *Annals of Operations Research* 216, 1 (2014), 327-342.

Xanthopoulos, P., Pardalos, P. M., and Trafalis, T. B. *Robust Data Mining*. Springer, 2012.

Zymler, S., Kuhn, D., and Rustem, B. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming* 137, 1-2 (2013), 167-198.

Thank You!