Literature Review Presentation

A Distributionally-Robust Approach For Finding Support Vector Machines

Author: Changkyeok Lee and Sanjay Mehrotra at Northwestern University

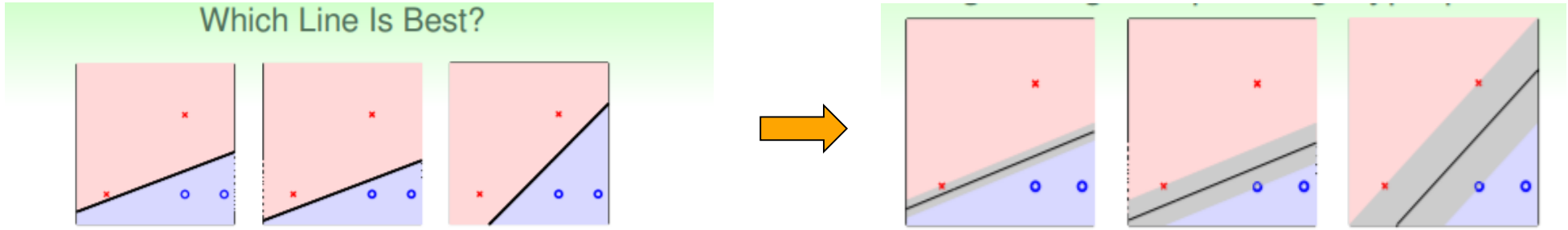
Wenbo Ma

Agenda

- Motivation
  - What is Support Vector Machine (SVM)
  - What motivates Distributionally-robust SVM (DR-SVM)
- Formulation and Reformulation
  - Kantorovich Metric
  - Monolithic Formulation
- Solution Algorithm
  - Cutting-plane Algorithm
  - Separation Problems
- Computational Experiments
  - Simulation Data
  - Real Data

Motivation: What is SVM

- What is SVM



$$\min_{w,b} \frac{1}{2} w^T w + \hat{C} \frac{1}{m} \sum_{i=1}^{m} h(w, b; x_j, y_j) \quad h(w, b; x, y) = \max\{1 - y(w^T x + b), 0\}$$

- A Toy Example
  - Credit Card Application
    - The predicted variable: default/not default
    - The predictors/features: credit score, number of credit cards,monthly income …

Figure sourced from Hsuan-Tien Lin's Machine Learning Techniques Course

Motivation: Why do we need DR-SVM

- Take a second look. What we are doing? Expectation

$$\min_{w,b} \ \frac{1}{2}w^T w + \hat{C}\sum_{i=1}^{m} h(w,b;x_j,y_j)\frac{1}{m} \quad h(w,b;x,y) = \max\{1 - y(w^T x + b), 0\}$$

- Empirical distribution. Is it representative for population distribution?

$$\{\frac{1}{m}, \frac{1}{m}, ..., \frac{1}{m}\} \quad \Rightarrow \quad \{\frac{1}{m} + \delta, \frac{1}{m} - \delta, ..., \frac{1}{m}\}$$

- Can we define a set of distributions/probablities similar to the empirical one?

$$\mathcal{P} = \{p | d(p, \hat{p}) \leq \epsilon\}$$

- If yes, we can optimize over the worst scenario(robust).

$$\min_{w,b} \ \frac{1}{2}w^T w + \hat{C} \sup_{p \in P} \int_{\Xi} h(w,b;\Xi) P(d\xi)$$

- Fortunately some smart minds in history said YES WE CAN….

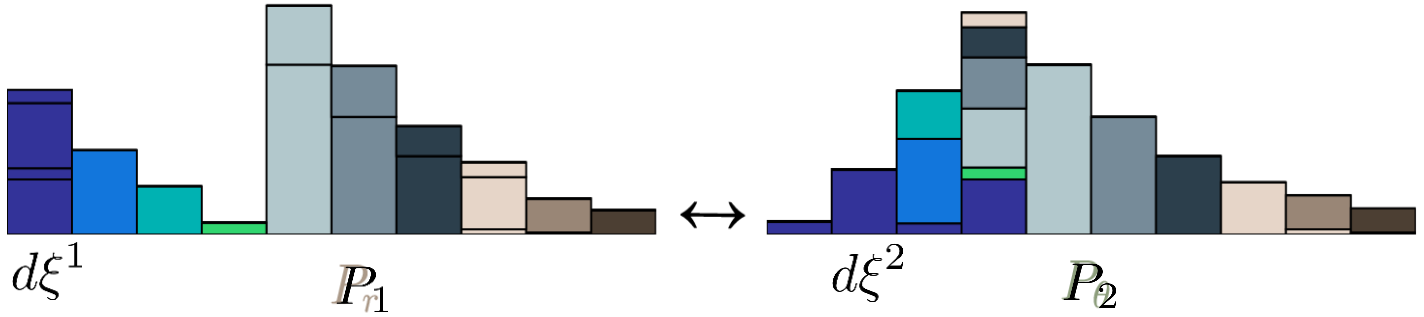
Kantorovich Metric – A Distance between Two Probability Measure

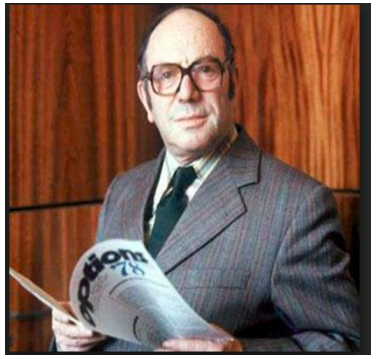
- Definition

$$d(p_1, p_2) := \inf_K \{ \int_{\Xi,\Xi} d_\xi(\xi^1, \xi^2) K(d\xi^1, d\xi^2) | \int_\Xi K(\xi^1, d\xi^2) = P^1(\xi^1), \int_\Xi K(d\xi^1, \xi^2) = P^2(\xi^2),$$

$$\forall \xi_1 \xi_2 \in \Xi \}$$

- Intution – Perspective from Optimal Transport Plan / Earth Mover Distance (CS Community)
  - How much work needed to convert P1 to P2?





Leonid Kantorovich

$K(d\xi^1, d\xi^2)$: How much earth I want to transport from $d\xi^1$ to $d\xi^2$

$d_\xi(\xi^1, \xi^2)$: Distance between ξ^1, ξ^2

Figure sourced from https://vincentherrmann.github.io/blog/wasserstein/ and modified by Wenbo Ma

Kantorovich Metric – A Distance between Two Probability Measures

- Definition

$$d(p_1, p_2) := \inf_K \{ \int_{\Xi,\Xi} d_\xi(\xi^1, \xi^2) K(d\xi^1, d\xi^2) | \int_\Xi K(\xi^1, d\xi^2) = P^1(\xi^1), \int_\Xi K(d\xi^1, \xi^2) = P^2(\xi^2),$$

$$\forall \xi_1 \xi_2 \in \Xi \}$$

- Intution – From Optimal Transport Plan to Probability

$$K(d\xi^1, d\xi^2) : \text{transport plan to joint density}$$

$$d_\xi(\xi^1, \xi^2) : \text{Distance between } \xi^1, \xi^2$$

- Other Distance Metrics
  - Kullback-Leibler(KL) Divergence
    - Deep Learning/Neural Network/Varational Inference
    - Not symmetric
    - Solomon Kullback

Ph.D. Math GW 1934
Chair of Dept. of Statistics

DR-SVM Formulation

- DR-SVM Formulation

$$\min_{w,b} \ \frac{1}{2} w^T w + \hat{C} \sup_{p \in P} \int_{\Xi} h(w, b; \Xi) P(d\xi)$$

$$d(p_1, p_2) := \inf_{K} \{ \int_{\Xi,\Xi} d_{\xi}(\xi^1, \xi^2) K(d\xi^1, d\xi^2) | \int_{\Xi} K(\xi^1, d\xi^2) = P^1(\xi^1), \int_{\Xi} K(d\xi^1, \xi^2) = P^2(\xi^2),$$

$$\forall \xi_1 \xi_2 \in \Xi \}$$

$$\mathcal{P} = \{ p | d(p, \hat{p}) \leq \epsilon \} \text{ and } \hat{p} = \frac{1}{m}$$

- Inner Problem - Reparameterization

$$\sup_{p \in P} \int_{\Xi} h(w, b; \Xi) P(d\xi) \qquad \mathcal{P} = \{ p | \exists \ K \text{satisfy } 1,2,3 \}$$

$$s.t.1) \sum_{j=1}^{m} K(\xi, \xi_j) = p(\xi), \forall \xi \in \Xi \qquad 3) \int_{\Xi} \sum_{i=1}^{m} d\xi(\xi, \xi j) K(d\xi, \xi_j) \leq \epsilon \}$$

$$2) \int_{\Xi} K(d\xi, \xi_j) = \frac{1}{m}, \forall j$$

DR-SVM Formulation – Conic Linear Program

- DR-SVM Formulation (from Last Slide)

$$\sup_{p \in P} \int_{\Xi} h(w, b; \Xi) P(d\xi) \qquad \mathcal{P} = \{p | \exists\ K \text{satisfy } 1,2,3\}$$

$$s.t.1) \sum_{j=1}^{m} K(\xi, \xi_j) = p(\xi), \forall \xi \in \Xi \quad 2) \int_{\Xi} K(d\xi, \xi_j) = \frac{1}{m}, \forall j \qquad 3) \int_{\Xi} \sum_{i=1}^{m} d\xi(\xi, \xi j) K(d\xi, \xi_j) \leq \epsilon\}$$

- Reformulation

$$\sup_{K \in M(\Xi \times \hat{\Xi})} \int_{\Xi} \sum_{j=1}^{m} h(w, b; \xi) K(d\xi, \xi_j)$$

$$\text{s.t. } \int_{\Xi} K(d\xi, \xi_j) = \frac{1}{m}, j = 1, ..., m \text{ (marginal is the emprical distribution)}$$

$$\int_{\Xi} \sum_{i=1}^{m} d\xi(\xi, \xi j) K(d\xi, \xi_j) \leq \epsilon \text{ (ambiguity constraint)}$$

$$K \geq 0$$

DR-SVM Formulation – Dual Problem of the Inner Problem

Primal Problem (from Last Slide)

$$\sup_{K \in M(\Xi \times \hat{\Xi})} \int_{\Xi} \sum_{j=1}^{m} h(w, b; \xi) K(d\xi, \xi_j)$$

$$\text{s.t. } \int_{\Xi} K(d\xi, \xi_j) = \frac{1}{m}, j = 1, ..., m$$

$$\int_{\Xi} \sum_{i=1}^{m} d_{\xi}(\xi, \xi j) K(d\xi, \xi_j) \leq \epsilon$$

$$K \geq 0$$

Dual Problem

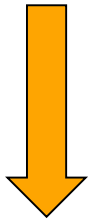
$$\min_{t,u} \frac{1}{m} \sum_{j=1}^{m} t_j + \epsilon\mu$$

$$\text{s.t. } t_j + d_{\xi}(\xi, \xi_j)\mu \geq h(w, b; \xi), \xi \in \Xi, j = 1, ..., m$$

$$\mu \geq 0$$

Complexity and Characteristic

- Decision Variable: K – probability function
- Target: Max over a linear function
- m+1 constraint: all linear



Linear Dual

Complexity and Characteristic

- Decision Variable: $t \in R^m, \mu$ (m+1)
- Target: Min over a linear function
- m linear constraint
- Implicitly infinite constraints since Ξ is infinite

DR-SVM Monolithic Formulation – Combine Inner and Outer Probelm

Original Problem

$$\min_{w,b} \ \frac{1}{2} w^T w + \hat{C} \sup_{p \in P} \int_{\Xi} h(w, b; \Xi) P(d\xi)$$

Dual Inner Problem

$$\min_{t,u} \frac{1}{m} \sum_{j=1}^{m} t_j + \epsilon\mu$$

$$\text{s.t. } t_j + d_\xi(\xi, \xi_j)\mu \geq h(w, b; \xi), \xi \in \Xi, j = 1, ..., m$$

$$\mu \geq 0$$

Monolithic Version

$$\min_{w,b,t,\mu} \ \frac{1}{2} w^T w + \hat{C}\{\frac{1}{m} \sum_{j=1}^{m} t_j + \epsilon\mu\}$$

$$\text{s.t. } t_j \ \geq h(w, b; \xi) - d_\xi(\xi, \xi_j)\mu, \xi \in \Xi, j = 1, ..., m$$

In which

$$h(w, b; x, y) := \max\{1 - y(w^T x + b), 0\}, \quad \xi := (x, y)$$

$$d_\xi(\xi^1, \xi^2) := d(x^1, x^2) + d(y^1, y^2)$$

DR-SVM Monolithic Formulation

Monolithic Version

$$\min_{w,b,t,\mu} \ \frac{1}{2}w^T w + \hat{C}\{\frac{1}{m}\sum_{j=1}^{m} t_j + \epsilon\mu\}$$

$$\text{s.t. } t_j \ \geq 1 - y(w^T x + b) - [d(x, x_j) + d_y(y, y_j)]\mu, \ (x, y) \ \in \ \Xi, \ j = 1, ..., m$$

$$t \geq 0, \mu \geq 0$$

Complexity and Characteristic

- Decision Variable: w,b,t,u
- Target Function: QP
- Constraint: infinite linear inequality

Solution

- Cutting Plane Algorithm (Approximation)

Solution Algorithm – Cutting Plane Algorithm

Semi-Infinite Programming (SIP) - Infinite Constraints)

$$\min_{w,b,t,\mu} \ \frac{1}{2}w^T w + \hat{C}\{\frac{1}{m}\sum_{j=1}^{m} t_j + \epsilon\mu\}$$

$$\text{s.t. } t_j \ \geq 1 - y(w^T x + b) - [d(x, x_j) + d_y(y, y_j)]\mu, \ (x, y) \ \in \ \Xi, \ j = 1, ..., m$$

$$t \geq 0, \mu \geq 0$$

Cutting Plane Algorithm

- Master Problem

$$\min_{w,b,t,\mu} \ \frac{1}{2}w^T w + \hat{C}\{\frac{1}{m}\sum_{j=1}^{m} t_j + \epsilon\mu\}$$

$$\text{s.t. } t_j \ \geq 1 - y(w^T x_{(j,k)} + b) - [d(x_{(j,k)}, x_j) + d_y(y_{(j,k)}, y_j)]\mu$$

$$j = 1, ..., m, k = 1, ..., K(j)$$

- Separation Problem(for each j)

$$\theta_j(w^l, b^l, t^l, \mu^l) := \max_{x,y\in\Xi} 1 - y(w^l x + b^l) - t_j^l - [d(x, x_j) + d_y(y, y_j)]u^l$$

Solution Algorithm – Cutting Plane Algorithm

Cutting Plane Algorithm

- Master Problem

$$\min_{w,b,t,\mu} \ \frac{1}{2}w^T w + \hat{C}\{\frac{1}{m}\sum_{j=1}^{m} t_j + \epsilon\mu\}$$

$$\text{s.t. } t_j \ \geq 1 - y(w^T x_{(j,k)} + b) - [d(x_{(j,k)}, x_j) + d_y(y_{(j,k)}, y_j)]\mu$$

$$j = 1, ..., m, k = 1, ..., K(j)$$

- Separation Problem(for each j)

$$\theta_j(w^l, b^l, t^l, \mu^l) := \max_{x,y\in\Xi} 1 - y(w^l x + b^l) - t_j^l - [d(x, x_j) + d_y(y, y_j)]u^l$$

Idea:

- At each iteration, solve a finite version (the master problem).
- Check if all constraints satisfied (is θ_j less than 0?).
- If yes, achieve the optimal solution for the DR-SVM problem.
- If not, put the optimal solution (x,y) into the master problem to cut the current solution of the master problem and do next iteration

Solution Algorithm – Complexity

Cutting Plane Algorithm

- Master Problem

$$\min_{w,b,t,\mu} \ \frac{1}{2}w^T w + \hat{C}\{\frac{1}{m}\sum_{j=1}^{m} t_j + \epsilon\mu\}$$

$$\text{s.t. } t_j \ \geq 1 - y(w^T x_{(j,k)} + b) - [d(x_{(j,k)}, x_j) + d_y(y_{(j,k)}, y_j)]\mu$$

$$j = 1, ..., m, k = 1, ..., K(j)$$

Complexity and Characteristic

- Decision Variable: w, b, t, μ
- Target Function: convex QP (PSD)
- Constraint: finite linear inequality

- Separation Problem(for each j)

$$\theta_j(w^l, b^l, t^l, \mu^l) := \max_{x,y\in\Xi} 1 - y(w^{lT} x + b^l) - t_j^l - [d(x, x_j) + d_y(y, y_j)]u^l$$

Complexity and Characteristic

- Decision Variable: x,y
- Target Function: convex if distance function is norm
- Norm can be further linearized $\min_x ||x - x_j||_1, \ x \in R^n \Leftrightarrow$

$$\min_{d\in R_+^n} \sum d_i$$

$$-d_i \leq x_{ji} - x_i \leq d_i$$

Computation Experiment - Simulation

- Simulation Procedure
  - Training Data: 50 samples

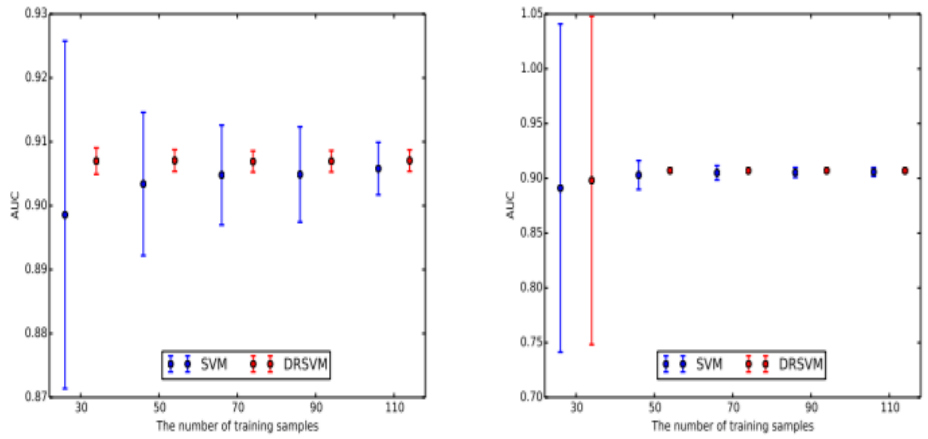
$$x \sim N((0.3, 0.2), \begin{bmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{bmatrix}) \text{ if } y = 1$$

$$x \sim N((-0.3, -0.2), \begin{bmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{bmatrix}) \text{ if } y = -1 \qquad y \sim \text{Bernoulli}(0.5)$$

  - Testing Data: 1000 samples
  - Process: fitting DR-SVM and SVM model 100 times on training samples and evaluate its performance on Testing Data

- Performance

	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



(a) L_1-norm (b) L_∞-norm

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

Computation Experiment – Real Data

- Data and Performance

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

Significant Improvement

Significant Worse

Depends on Norm

Conclusion

- DR-SVM Framework to Improve Generalization Error
- Semi-infinite Convex Formulation
- Cutting Plane Algorithm
- Evaluation on Simulation and Real Data