Maximum Margin Semi-supervised Learning With Irrelevant Data

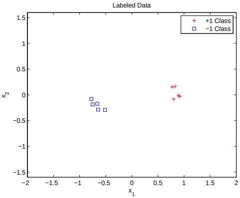
Haiqin Yang

Department of Computer Science & Engineering
The Chinese University of Hong Kong

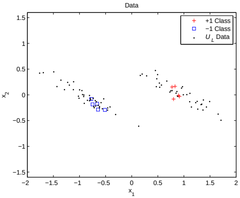
Outline

Data

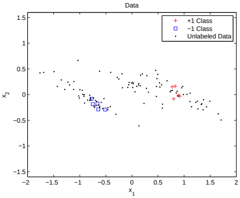
Labeled data



Clean unlabeled data



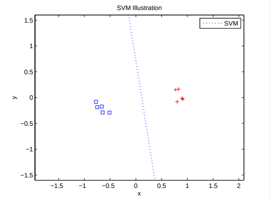
Mixed unlabeled data



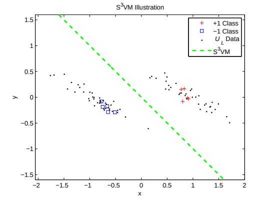
Models

Many models try to learn from both labeled and unlabeled data, e.g.,

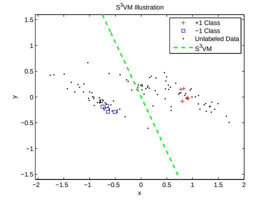
SVMs



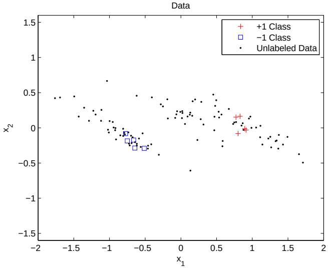
S^3VMs



S^3VMs



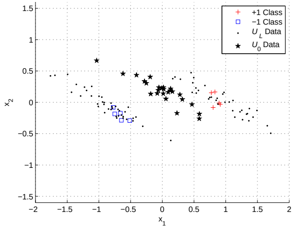
Problems



- Previous SSL assumption: unlabeled data are from the same distribution as the labeled data.
- Usual situation: unlabeled data may be a mixture of relevant and irrelevant data.
- Very common in web applications: unlabeled data are not well-prepared.

Setup

Data Illustration



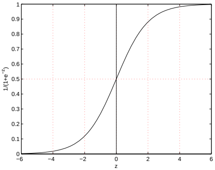
- $\mathcal{L} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{L}$
  $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$, $y_i \in \{-1, 0, 1\}$
- $\mathcal{U} = \mathcal{U}_R \cup \mathcal{U}_0 = \{\mathbf{x}_i\}_{i=1}^{U}$
- **Objective:** seek $f_\vartheta(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$, $\vartheta = (\mathbf{w}, b)$, to separate the binary class data correctly with the help of (mixed) unlabeled data

Definition

- **Objective function:**

$$\min_{\vartheta}. \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{\mathbf{x}_i \in \mathcal{L}} r_i \ell_L(f_{\vartheta}(\mathbf{x}_i), y_i) + \sum_{\mathbf{x}_i \in \mathcal{U}} r_i \ell_U(f_{\vartheta}(\mathbf{x}_i)),$$

- **Facts:** if $f_{\vartheta}(\mathbf{x}_i) \gg 0$, more confident on $+1$-class
  if $f_{\vartheta}(\mathbf{x}_i) \ll 0$, more confident on -1-class



- **Principle:** rely more on labeled and relevant data,
  risk measured by hinge loss, symmetrical hinge loss

  Principle : ignore irrelevant data,
  risk measured by ε-insensitive loss

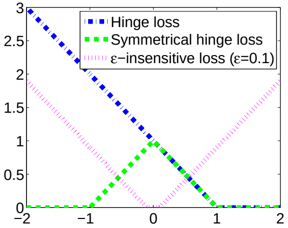
Definition

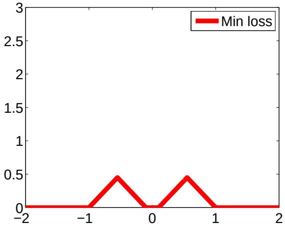
- **Objective function:**

$$\min_{\vartheta}. \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{\mathbf{x}_i \in \mathcal{L}_{\pm 1}} r_i H_1(y_i f_{\vartheta}(\mathbf{x}_i)) + \sum_{\mathbf{x}_i \in \mathcal{L}_0} r_i I_{\varepsilon}(f_{\vartheta}(\mathbf{x}_i)) + \sum_{\mathbf{x}_i \in \mathcal{U}} r_i \min\{H_1(|f_{\vartheta}(\mathbf{x}_i)|), I_{\varepsilon}(|f_{\vartheta}(\mathbf{x}_i)|)\}.$$

$$H_1(z) = \max\{0, 1 - z\}, \quad I_{\varepsilon}(z) = \max\{0, |z| - \varepsilon\}.$$

- **Loss functions illustration:**





Model Generalization

- **Objective function:**

$$\min_{\vartheta} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{\mathbf{x}_i\in\mathcal{L}_{\pm1}} r_i H_1(y_i f_\vartheta(\mathbf{x}_i)) + \sum_{\mathbf{x}_i\in\mathcal{L}_0} r_i I_\varepsilon(f_\vartheta(\mathbf{x}_i)) + \sum_{\mathbf{x}_i\in\mathcal{U}} r_i \min\{H_1(|f_\vartheta(\mathbf{x}_i)|), I_\varepsilon(|f_\vartheta(\mathbf{x}_i)|)\}.$$

- **Model relationship:**

3C-SVM

$\mathcal{L}$	-1	0	1
$\mathcal{U}$	-1	0	1

SVM

$\mathcal{L}$	-1		1
$\mathcal{U}$			

S^{3}VM

$\mathcal{L}$	-1		1
$\mathcal{U}$	-1		1

$\mathcal{U}$-SVM

$\mathcal{L}$	-1	0	1
$\mathcal{U}$			

Theorem

Objective function:

$$\min_{\vartheta} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{\mathbf{x}_i \in \mathcal{L}_{\pm 1}} r_i H_1(y_i f_{\vartheta}(\mathbf{x}_i)) + \sum_{\mathbf{x}_i \in \mathcal{L}_0} r_i I_{\varepsilon}(f_{\vartheta}(\mathbf{x}_i)) + \sum_{\mathbf{x}_i \in \mathcal{U}} r_i \min\{H_1(|f_{\vartheta}(\mathbf{x}_i)|), I_{\varepsilon}(|f_{\vartheta}(\mathbf{x}_i)|)\}.$$

3C-SVM with $r_i = \infty$ for unlabeled data and $\varepsilon = 0$

Unlabeled data $\mathbf{x}_j$ satisfies

(a) $|\mathbf{w}^T\phi(\mathbf{x}_j) + b| \geq 1 \Rightarrow$ data lie on or out of the margin gap, or

(b) $\mathbf{w}^T\phi(\mathbf{x}_j) + b = 0 \Rightarrow \mathbf{w}^T(\phi(\mathbf{x}_j) - \phi(\mathbf{x}_0)) = 0$, $\mathbf{x}_j, \mathbf{x}_0 \in \mathcal{U}_0$

Removing Min-Terms and Absolute Values

$$\min_{\boldsymbol{\vartheta},\mathbf{g}} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{\mathbf{x}_i\in\mathcal{L}_{\pm 1}} r_i H_1(y_i f_{\boldsymbol{\vartheta}}(\mathbf{x}_i)) + \sum_{\mathbf{x}_i\in\mathcal{L}_0} r_i I_\varepsilon(f_{\boldsymbol{\vartheta}}(\mathbf{x}_i))$$

$$+ \sum_{\mathbf{x}_{k+L}\in\mathcal{U}} r_{k+L}\left(\underbrace{H_1(|f_{\boldsymbol{\vartheta}}(\mathbf{x}_i)|+D(1-g_k))}_{Q_1} + \underbrace{I_\varepsilon(|f_{\boldsymbol{\vartheta}}(\mathbf{x}_i)|-Dg_k)}_{Q_2}\right),$$

- $g_k = 0 \Rightarrow Q_1 = 0$,
- $g_k = 1 \Rightarrow Q_2 = 0$,
- $H_1(|z|+a)$: non-convexity, approximated by ramploss, $H_{1-a}(z) - H_\kappa(z) + H_{1-a}(-z) - H_\kappa(-z)$,
- $I_\varepsilon(|z|-b) = H_{-\varepsilon-b}(-z) + H_{-\varepsilon-b}(z)$,
- $H_1(|z|+a)$ and $I_\varepsilon(|z|-b)$ are symmetrical loss.

Concave-Convex Procedure

- **Objective function:** $Q^{\kappa}(\boldsymbol{\vartheta},\mathbf{g})=Q_{vex}(\boldsymbol{\vartheta},\mathbf{g})+Q^{\kappa}_{cav}(\boldsymbol{\vartheta})$
- Each step

$$\boldsymbol{\vartheta}^{t+1} = \arg\min_{\boldsymbol{\vartheta}}\left(Q_{vex}(\boldsymbol{\vartheta},\mathbf{g}^t)+\frac{\partial Q^{\kappa}_{cav}(\boldsymbol{\vartheta}^t)}{\partial\boldsymbol{\vartheta}}\cdot\boldsymbol{\vartheta}\right),$$

$$\underset{\text{QP}}{\overset{\text{Dual}}{\Longleftrightarrow}}\begin{cases}\max_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*}. & -\frac{\lambda}{2}\|\mathbf{w}(\boldsymbol{\alpha},\boldsymbol{\alpha}^*)\|^2+\boldsymbol{\varrho}(\boldsymbol{\alpha},\boldsymbol{\alpha}^*)\\ \text{s.t.} & \mathbf{A}_e[\boldsymbol{\alpha};\boldsymbol{\alpha}^*]=\boldsymbol{\mu}^T\mathbf{Y}_{\bullet U}\,,\\ & \mathbf{A}[\boldsymbol{\alpha};\boldsymbol{\alpha}^*]\le\mathbf{0}\,,\\ & \mathbf{0}\le\boldsymbol{\alpha},\boldsymbol{\alpha}^*\le\mathbf{r}\,.\end{cases}$$

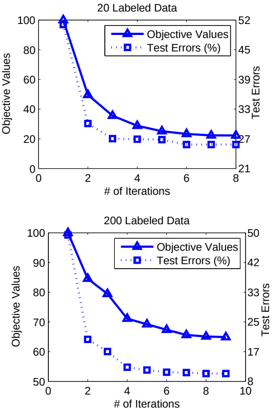
$$g_k = \begin{cases}1 & \text{if} \quad \xi_k\le\xi_k^*\\ 0 & \text{otherwise}\end{cases}, \quad \xi_k=H_1(|f_{\boldsymbol{\vartheta}}(\mathbf{x}_{k+L})|),\ \xi_k^*=I_{\varepsilon}(|f_{\boldsymbol{\vartheta}}(\mathbf{x}_{k+L})|),\ k=1,\ldots,U.$$

- **Solution:** w is linear combined by α and α^*,
  b is attained by KKT condition.

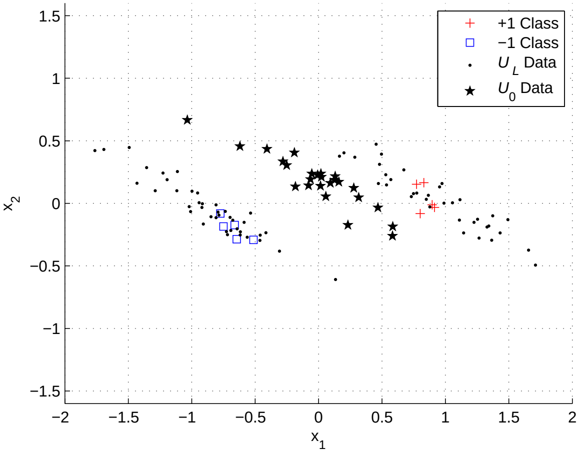
Algorithm

Algorithm 1 CCCP for 3C-SVMs

Initialization:
$t = 0$;
Calculate $\boldsymbol{\vartheta}^0 = (\mathbf{w}^0, b^0)$ from a $\mathcal{U}$-SVM solution on the labeled/unlabeled data;
Compute
$$\mu_i^0 = \begin{cases} r_i & \text{if } y_i f_{\boldsymbol{\vartheta}^0}(\mathbf{x}_i) < \kappa \text{ and } i \geq L+1 \\ 0 & \text{otherwise} \end{cases};$$
repeat
$t \leftarrow t + 1$;
Solve the optimization in (6) to obtain $\boldsymbol{\vartheta}^t$;
Update $\mathbf{g}^t$ from (4);
Update $\boldsymbol{\mu}^t$ from (5);
if $Q^\kappa(\boldsymbol{\vartheta}^t, \mathbf{g}^t) > Q^\kappa(\boldsymbol{\vartheta}^{t-1}, \mathbf{g}^{t-1})$ **then**
Let $\mathbf{g}^t = \mathbf{g}^{t-1}$;
Solve the optimization in (6) to obtain $\boldsymbol{\vartheta}^t$ by fixing $\mathbf{g}^{t-1}$;
Update $\boldsymbol{\mu}^t$ from (5);
end if
until $|\boldsymbol{\mu}^{t+1} - \boldsymbol{\mu}^t| \leq \epsilon$.

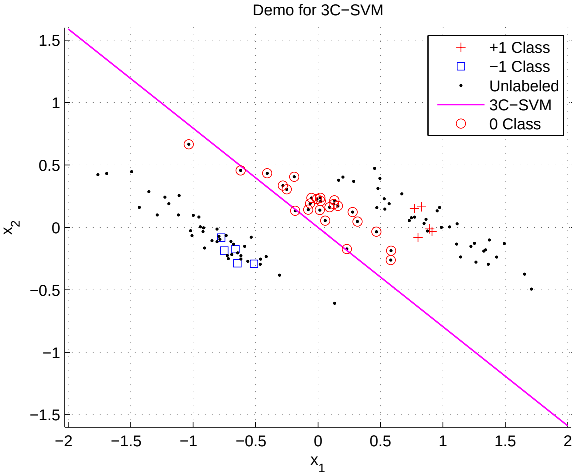


3CSVM Demo



Video

3CSVM Result



- **Comparing Algorithms:**
  - SVMs
  - S^3VMs
  - $\mathcal{U}$-SVMs
  - 3C-SVMs
- **Platform:**
  - Matlab 7.3
  - MOSEK 5.0

Data Generation

- Follow scheme from Sinz et al., 2008.
- ± 1-class: $c_i^{\pm} = \pm 0.3$, $i = 1, \ldots, 50$, $\sigma_{1,2}^2 = 0.08$, $\sigma_{3,\ldots,50}^2 = 10$.
- Two Gaussians with the Bayes risk being approximately 5%.
- First $\mathcal{U}_0$: zero mean, $\sigma_{1,2}^2 = 0.1$, $\sigma_{3,\ldots,50}^2 = 10$.
- Second $\mathcal{U}_0$: variance values are the same as ± 1-class data, mean is $t \cdot \mathbf{c}^{+}$, $t = 0.5$.

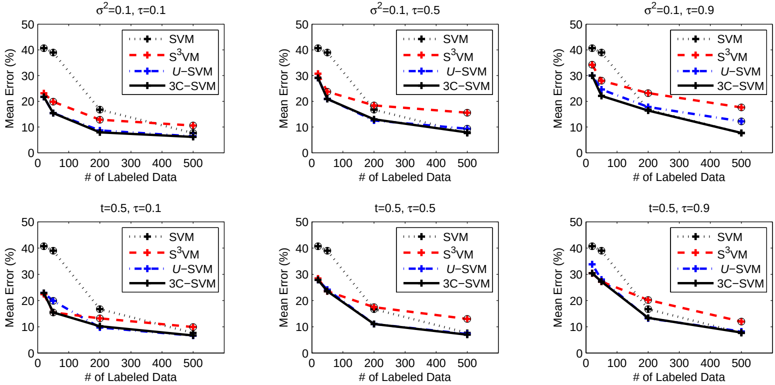
Test procedure

- $L = 20, 50, 200, 500$
- $U = 500 = (\tau U, (1-\tau)U)$, $\tau = 0.1, 0.5, 0.9$
- Labeled + Unlabeled/500 Test, ten-run average
- Hyperparameters
  - Linear kernel
  - Regularized parameters, forward tuning

	$C_{\mathcal{L}}$	$C_{\mathcal{U}}$	ε	κ
SVM	✓	×	×	×
$\mathcal{U}$-SVM	–	✓	✓	×

  - Further tune on S^3VM
  - 3C-SVM uses the same parameters of other models

Accuracy


σ²=0.1, τ=0.1
σ²=0.1, τ=0.5
σ²=0.1, τ=0.9
t=0.5, τ=0.1
t=0.5, τ=0.5
t=0.5, τ=0.9
Mean Error (%)
# of Labeled Data
SVM
S³VM
U−SVM
3C−SVM

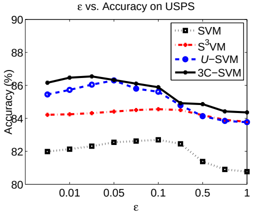
Description

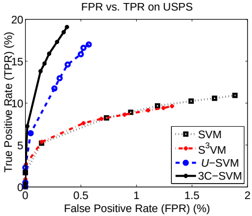
- Datasets:
  - Small size: USPS
  - Large size: MNIST
- Setup
  - ± 1-class: Digits "5" and "8"
  - $\mathcal{U}_0$: Other digits
  - $L = 20$
  - $U = 500 = (\tau U, (1-\tau)U)$, $\tau = 0.1, 0.5, 0.9$
  - RBF kernel: $K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$, $\gamma = \frac{1}{0.3d}$
  - Other hyperparameters are set similar to those in the synthetic datasets

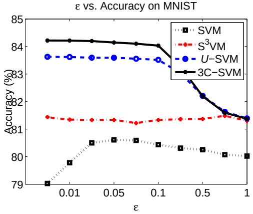
Accuracy Results

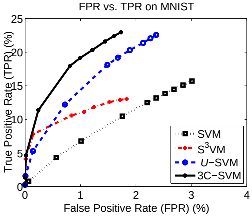
Dataset	Algorithm	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
USPS	SVM	72.4$\pm$ 15.9 (**0.7**)	72.4$\pm$ 15.9 (**9.5**)	72.4$\pm$ 15.9 (53.1)
	S^3VM	63.6 $\pm$ 8.9 (**0.0**)	68.2 $\pm$ 8.0 (**2.2**)	73.2 $\pm$ 7.0 (**9.5**)
	$\mathcal{U}$-SVM	83.1 $\pm$ 2.5 (**0.0**)	73.4 $\pm$ 4.4 (**0.0**)	64.2 $\pm$ 3.6 (**0.0**)
	3C-SVM	**87.2**$\pm$2.3	**80.6**$\pm$4.8	**75.4**$\pm$7.3
MNIST	SVM	70.9$\pm$ 11.4 (**0.3**)	70.9$\pm$ 11.4 (**0.8**)	70.9$\pm$ 11.4 (13.6)
	S^3VM	70.9 $\pm$ 10.5 (**0.7**)	72.4 $\pm$ 10.1 (**1.0**)	75.7 $\pm$ 9.1 (**9.8**)
	$\mathcal{U}$-SVM	84.2 $\pm$ 2.2 (**0.2**)	80.0 $\pm$ 4.6 (**0.9**)	75.0 $\pm$ 3.9 (**1.0**)
	3C-SVM	**85.3**$\pm$1.6	**82.8**$\pm$2.9	**77.6**$\pm$3.9

Accuracy on Detecting 0-class









Balance Constraint

- Ideally, $\frac{1}{U}\sum_{t=L+1}^{L+U} f_{\vartheta}(\mathbf{x}_t) = \frac{1}{L}\sum_{i=1}^{L} y_i$, but no improvement from experimental results;
- A possible better on, $\frac{1}{U}\sum_{t=L+1}^{L+U} f_{\vartheta}(\mathbf{x}_t) = c$,
  c: a user-specified constant, but need tuning.

Conclusions

Conclusions

- A novel maxi-margin classifier, 3C-SVM, can distinguish data into -1, $+1$, and 0, three categories.
- The model incorporates standard SVMs, S^3VMs, and $\mathcal{U}$-SVMs as specific cases.
- It is solved by the CCCP, in a high efficiency algorithm.
- Effectiveness and efficiency are demonstrated.

Future works

- Model speedup
- Multi-class extension
- Theoretical analysis, generalization bound

References

1. Chapelle, O., Schölkopf, B., and Zien, A. (Eds.). *Semi-supervised learning*. Cambridge, MA: MIT Press. 2006
2. Collobert, R., Sinz, F., Weston, J., and Bottou, L. Large scale transductive svms. *Journal of Machine Learning Research*, *7*, 1687–1712. 2006.
3. Weston, J., Collobert, R., Sinz, F. H., Bottou, L., and Vapnik, V. Inference with the universum. In ICML'06, 1009–1016. 2006.
4. Vapnik, V., and Kotz, S. *Estimation of dependences based on empirical data: Empirical inference science (information science and statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. 2nd edition. 2006.
5. Sinz, F. H., Chapelle, O., Agarwal, A., and Schölkopf, B. An analysis of inference with the universum. In *Proceedings of the 21th Neural Information Processing Systems Conference*, 1369–1376. Cambridge, MA, USA: MIT Press. 2008.

Questions ?

Haiqin Yang
www.cse.cuhk.edu.hk/~hqyang
hqyang@cse.cuhk.edu.hk