Regularization and Robustness of Support Vector Machines

Huan Xu

McGill University

Dec. 12nd, 2008

Joint work with Constantine Caramanis and Shie Mannor.

Outline

1. Introduction
2. SVM & Robust Classification
3. Robustness Implies Consistency

Outline

1. **Introduction**
2. SVM & Robust Classification
3. Robustness Implies Consistency

Statistical Learning

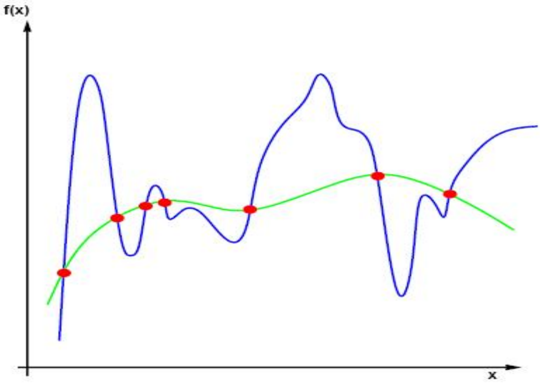
Supervised Learning Problem:

- Training Data: $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ generated according to unknown distribution.
- Goal: Find labelling rule $\mathcal{L}(\boldsymbol{x})$ to minimize generalization error:

$$\mathbb{E}[\ell(\mathbf{x}, \mathcal{L}(\mathbf{x}), y^{\text{true}})]$$

- Problems: Do not know distribution. Control overfitting.

Overfitting: An Example [1]



[1] Adapted from http://www.mit.edu/∼9.520/Classes/class02.pdf

Regularization

- Fact 1: Overfitting solutions are unnecessarily complicated.
- Approach 1: Penalizing the complexity of the solution.

$$\min_{\mathcal{L}}: \sum_{i=1}^{m} \ell(\mathbf{x}_i, \mathcal{L}(\mathbf{x}_i), y_i) + \rho(\mathcal{L}).$$

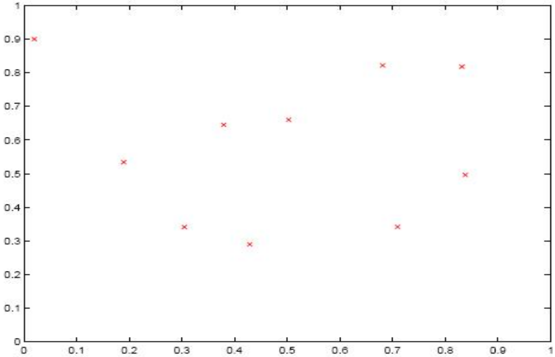
- $\rho(\mathcal{L})$ is the regularization term. Typically chosen as a norm function.
- Adding apples with oranges.

Robustness

- Fact 2: Overfitting solutions are sensitive to disturbance.

Robustness & Overfitting: an example[2]

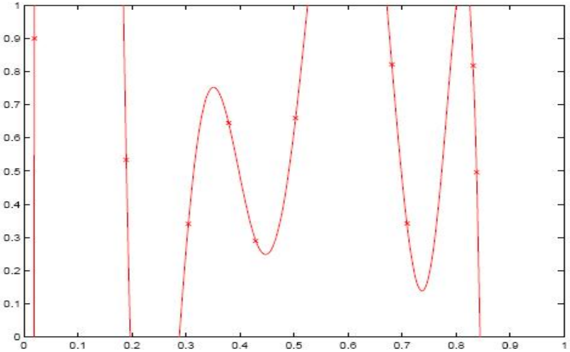
Consider the 10-sample example



[2] Adapted from http://www.mit.edu/∼9.520/Classes/class02.pdf

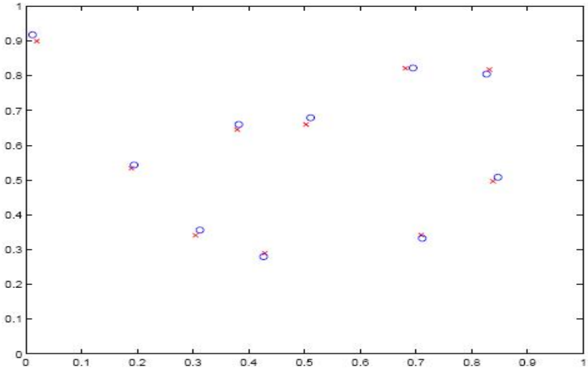
Robustness & Overfitting: an example (Cont.)

Fitting the samples with an arbitrary degree polynomial



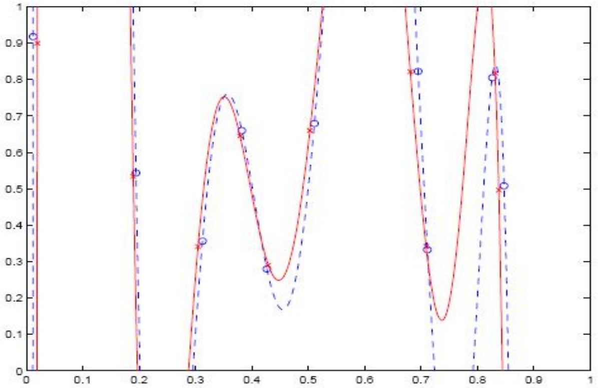
Robustness & Overfitting (Cont.)

Perturbing the sample slightly



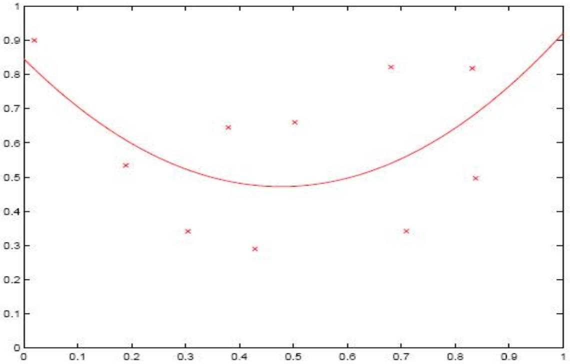
Robustness & Overfitting (Cont.)

The solution changes dramatically



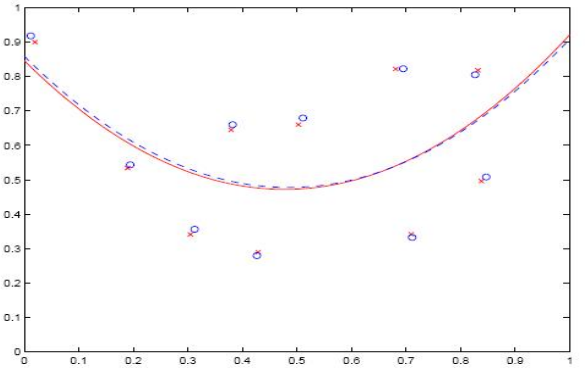
Robustness & Overfitting (Cont.)

Degree-2 polynomial fitting



Robustness & Overfitting (Cont.)

Not sensitive to perturbation



Robustness

- Fact 2: Overfitting solutions are sensitive to disturbance.
- Approach 2: Find a **robust** (w.r.t sample perturbation) solution.
- How? Robust Optimization.

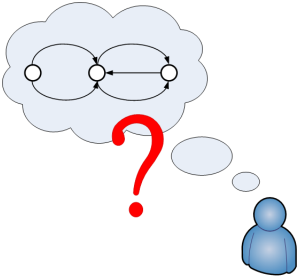
Robust Optimization

- General decision problem:

$$\max_{\mathbf{x}} \quad u(\mathbf{x}, \boldsymbol{\xi}).$$

- What if ξ is unknown?

  - noisy/incorrect observation
  - estimation from finite samples
  - simplification of the problem

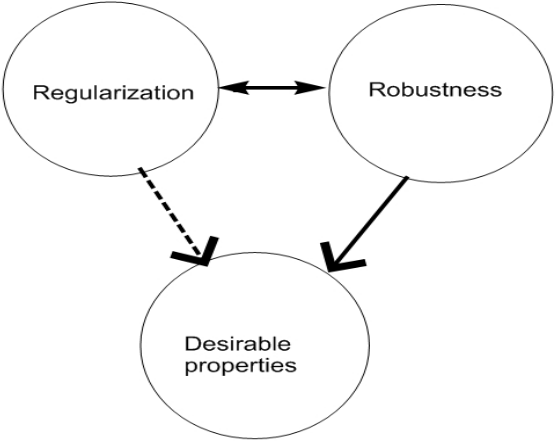


- Max-min solution.

$$\max_{\mathbf{x}} \min_{\xi \in \Delta} u(\mathbf{x}, \boldsymbol{\xi}).$$

Main Contribution: Regularization = Robustness

- Fact 3: Approach 1 and Approach 2 are equivalent!



Outline

1. Introduction
2. **SVM & Robust Classification**
3. Robustness Implies Consistency

Regularized SVM

- Support Vector Machine:
  Look for a **linear classifier** in the **feature space.**

$$\begin{aligned}\min_{\boldsymbol{w},b}: \quad & c\|\boldsymbol{w}\|_2 + \sum_{i=1}^{m} \xi_i \\ \text{s.t.}: \quad & \xi_i \geq 1 - y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \\ & \xi_i \geq 0\end{aligned}$$

Or equivalently:

$$\min_{\boldsymbol{w},b}: c\|\boldsymbol{w}\|_2 + \sum_{i=1}^{m} \max[1 - y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b), 0]$$

Robust SVM without Regularization

For some set $\mathcal{N}$, solve the following:

$$\min_{\boldsymbol{w},b} : \sup_{(\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_m)\in\mathcal{N}} \sum_{i=1}^{m} \max[1 - y_i(\langle \boldsymbol{w}, (\boldsymbol{x}_i - \boldsymbol{\delta}_i)\rangle + b), 0]$$

Here, the set $\mathcal{N}$ is called *Uncertainty Set*. In particular, we investigate *Sublinear Aggregated Uncertainty Set*.

Uncertainty Set/Allowed Disturbance: Formal definition

A set $\mathcal{N}_0 \subseteq \mathbb{R}^n$ is called an *Atomic Uncertainty Set* if

$$\text{(I)} \qquad \mathbf{0} \in \mathcal{N}_0;$$

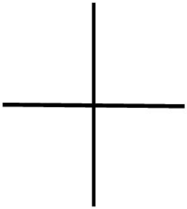
$$\text{(II)} \qquad \sup_{\boldsymbol{\delta} \in \mathcal{N}_0} \left[\mathbf{w}^\top \boldsymbol{\delta}\right] = \sup_{\boldsymbol{\delta}' \in \mathcal{N}_0} \left[-\mathbf{w}^\top \boldsymbol{\delta}'\right] < \infty, \ \forall \mathbf{w} \in \mathbb{R}^n.$$

Sublinear Aggregated Uncertainty set $\mathcal{N}$ for $\mathcal{N}_0$:

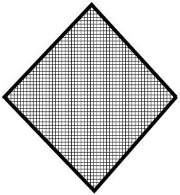
$$(i) \ \{(\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_m) \,|\, \boldsymbol{\delta}_t \in \mathcal{N}_0, \ \boldsymbol{\delta}_{i \neq t} = 0\} \subseteq \mathcal{N}, \quad t = 1, \dots, m$$

$$(ii) \ \mathcal{N} \subseteq \{(\alpha_1 \boldsymbol{\delta}_1, \dots, \alpha_m \boldsymbol{\delta}_m) \,|\, \sum_{i=1}^m \alpha_i = 1, \ \alpha_i \geq 0, \ \boldsymbol{\delta}_i \in \mathcal{N}_0, i = 1, \dots, m\}.$$

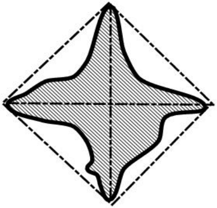
Sublinear Aggregated Uncertainty Set: Illustration



(a) Inner Set



(b) Outer Set



(c) An SAU Set

Sublinear Aggregated Uncertainty Set: Some Examples

- (1) $\{(\delta_i, \dots, \delta_m) \mid \sum_{i=1}^{m} ||\delta_i|| \leq c\}$.
- (2) $\{(\delta_1, \cdots, \delta_m) | \exists t \in [1:m];\ \|\delta_t\| \leq c;\ \ \delta_i = \mathbf{0}, \forall i \neq t\}$.
- (3) $\{(\delta_1, \cdots, \delta_m) | \sum_{i=1}^{m} \sqrt{c_i \|\delta_i\|} \leq c\}$.

Shocker: Regularization = Robustness

Proposition: Assume $\{\boldsymbol{x}_i, y_i\}_{i=1}^m$ are non-separable. Then

$$\min_{\boldsymbol{w},b} : \quad \sup_{(\boldsymbol{\delta}_1,\dots,\boldsymbol{\delta}_m)\in\mathcal{N}} \sum_{i=1}^m \max[1 - y_i(\langle \boldsymbol{w}, (\boldsymbol{x}_i - \boldsymbol{\delta}_i)\rangle + b), 0]$$

is equivalent to

$$\begin{aligned} \min_{\boldsymbol{w},b} : \quad & \sup_{\boldsymbol{\delta}\in\mathcal{N}_0} (\boldsymbol{w}^\top \boldsymbol{\delta}) + \sum_{i=1}^m \xi_i \\ \text{s.t.} : \quad & \xi_i \geq 1 - y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) \\ & \xi_i \geq 0 \end{aligned}$$

- This is a regularization term.

Regularization = Robustness (Cont.)

Corollary:
Consider $\mathcal{N} = \{(\boldsymbol{\delta}_i, \ldots, \boldsymbol{\delta}_m) \mid \sum_{i=1}^m ||\boldsymbol{\delta}_i||^* \leq c\}$. If the training sample $\{\mathbf{x}_i, y_i\}_{i=1}^m$ are non-separable, then the following two optimization problems on $(\mathbf{w}, b)$ are equivalent

$$\min: \quad \max_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m) \in \mathcal{N}} \sum_{i=1}^m \max\left[1 - y_i\big(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + b\big), 0\right],$$

$$\min: \quad c\|\mathbf{w}\| + \sum_{i=1}^m \max\left[1 - y_i\big(\langle \mathbf{w},\, \mathbf{x}_i \rangle + b\big), 0\right].$$

- Standard regularization essentially assumes that the disturbance is spherical
- A physical meaning to the regularization constant

Kernelization

Linear Classifier in abstract feature space:

$$\begin{aligned} \min_{\mathbf{w},b}: \quad & c\|\mathbf{w}\|_{\mathcal{H}} + \sum_{i=1}^{m} \xi_i \\ \text{s.t.}: \quad & \xi_i \geq \left[1 - y_i(\langle \mathbf{w}, \Phi(\mathbf{x}_i)\rangle + b)\right], \\ & \xi_i \geq 0. \end{aligned}$$

Here, $\|\mathbf{w}\|_{\mathcal{H}} = \sqrt{\langle \mathbf{w}, \mathbf{w}\rangle}$.

Regularization = Robustness still holds

Consider $\mathcal{N} = \{(\boldsymbol{\delta}_i, \ldots, \boldsymbol{\delta}_m) \mid \sum_{i=1}^m ||\boldsymbol{\delta}_i||_{\mathcal{H}} \leq c\}$. If $\{\Phi(\mathbf{x}_i), y_i\}_{i=1}^m$ are non-separable, then the following two optimization problems on $(\mathbf{w}, b)$ are equivalent

$$\min: \quad \max_{(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)\in\mathcal{N}} \sum_{i=1}^m \max\left[1 - y_i\big(\langle \mathbf{w},\ \Phi(\mathbf{x}_i) - \boldsymbol{\delta}_i\rangle + b\big), 0\right],$$

$$\min: \quad c||\mathbf{w}||_{\mathcal{H}} + \sum_{i=1}^m \max\left[1 - y_i\big(\langle \mathbf{w},\ \Phi(\mathbf{x}_i)\rangle + b\big), 0\right].$$

Conclusion: standard kernelized SVM is implicitly a robust classifier (without regularization) with noises lie in the feature-space.

Input Space Uncertainty

- Feature-space uncertainty $\Rrightarrow$ input-space uncertainty.

Lemma 1:
Suppose there exist $\mathcal{X} \subseteq \mathbb{R}^n$, $\rho > 0$, and a continuous non-decreasing function $f : \mathbb{R}^+ \to \mathbb{R}^+$ satisfying $f(0) = 0$, such that

$$k(\mathbf{x},\mathbf{x})+k(\mathbf{x}',\mathbf{x}')-2k(\mathbf{x},\mathbf{x}') \leq f(\|\mathbf{x}-\mathbf{x}'\|_2^2), \quad \forall \mathbf{x},\mathbf{x}' \in \mathcal{X}, \|\mathbf{x}-\mathbf{x}'\|_2 \leq \rho.$$

Then

$$\|\Phi(\hat{\mathbf{x}}+\boldsymbol{\delta}) - \Phi(\hat{\mathbf{x}})\|_{\mathcal{H}} \leq \sqrt{f(\|\boldsymbol{\delta}\|_2^2)}, \quad \forall \|\boldsymbol{\delta}\|_2 \leq \rho,\ \hat{\mathbf{x}}, \hat{\mathbf{x}}+\boldsymbol{\delta} \in \mathcal{X}.$$

Input Space Uncertainty (Cont.)

- Example: Degree-2 Polynomial for 2-d data,

$$\Phi(\mathbf{x}) = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}.$$

- The image of a small-ball in input space $\Phi(\mathcal{B}_I) \subseteq$ a small-ball in feature space $\mathcal{B}_F$.
- Robust to $\mathcal{B}_F \Rrightarrow$ robust to $\mathcal{B}_I$.

Outline

1. Introduction
2. SVM & Robust Classification
3. **Robustness Implies Consistency**

PAC Setup

- $\mathcal{X} \subseteq \mathbb{R}^n$ is bounded.
- The training samples $(\mathbf{x}_i, y_i)_{i=1}^{\infty}$ are generated i.i.d. according to an unknown distribution $\mathbb{P}$ supported on $\mathcal{X} \times \{-1, +1\}$.
- Kernel function $k(\cdot, \cdot)$ satisfies the condition of Lemma 1.
- Denote $K \triangleq \max_{\mathbf{x} \in \mathcal{X}} k(\mathbf{x}, \mathbf{x})$.

Consistency: Main result

Theorem:
There exists a random sequence $\{\gamma_{m,c}\}$ independent of $\mathbb{P}$ such that, $\forall c > 0$, $\lim_{m\to\infty} \gamma_{m,c} = 0$ almost surely, and the following bounds on the Bayes loss and the hinge loss hold uniformly $\forall(\boldsymbol{w}, b) \in \mathcal{H} \times \mathbb{R}$:

$$\mathbb{E}_{\mathbb{P}}(\mathbf{1}_{y\neq sgn(\langle \mathbf{w},\, \Phi(\mathbf{x})\rangle + b)}) \leq$$

$$\gamma_{m,c} + c\|\mathbf{w}\|_{\mathcal{H}} + \frac{1}{m}\sum_{i=1}^{m} \max\left[1 - y_i(\langle \mathbf{w},\ \Phi(\mathbf{x}_i)\rangle + b), 0\right];$$

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}}\Big(\max(1 - y(\langle \mathbf{w},\ \Phi(\mathbf{x})\rangle + b),\ 0)\Big) \leq$$

$$\gamma_{m,c}(1 + K\|\mathbf{w}\|_{\mathcal{H}} + |b|) + c\|\mathbf{w}\|_{\mathcal{H}} + \frac{1}{m}\sum_{i=1}^{m} \max\left[1 - y_i(\langle \mathbf{w},\ \Phi(\mathbf{x}_i)\rangle + b), 0\right].$$

Proof sketch: Linear case

- Regard testing samples as perturbed version of training samples.
- A testing sample $(\mathbf{x}', y')$ and a training sample $(\mathbf{x}, y)$ are called a **sample pair** if $y = y'$ and $\|\mathbf{x} - \mathbf{x}'\|_2 \leq c$.
- Given m training samples and m testing samples, M_m is the largest number of pairings.
- For paired samples, the testing error & hinge-loss is upper bounded by

$$\max_{(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)\in\mathcal{N}_0\times\cdots\times\mathcal{N}_0} \sum_{i=1}^{m} \max\left[1 - y_i\big(\langle\mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b\big), 0\right]$$

$$\leq cm\|\mathbf{w}\|_2 + \sum_{i=1}^{m} \max\left[1 - y_i(\langle\mathbf{w},\, \mathbf{x}_i\rangle + b),\, 0\right].$$

Proof sketch: Linear case (Cont.)

Lemma 2:
Given $c > 0$, $M_m/m \to 1$ almost surely as $m \to +\infty$, uniformly w.r.t. $\mathbb{P}$.

- Partition $\mathcal{X}$ into finite "small" sets.
- N_i^{tr} and N_i^{te} be the number of training samples and testing samples falling in the i^{th} set.
- $(N_1^{tr}, \cdots, N_T^{tr})$ and $(N_1^{te}, \cdots, N_T^{te})$ are multinomial r.v following a same distribution.
- $\sum_{i=1}^{T} \left|N_i^{tr} - N_i^{te}\right|/m \to 0$ with probability one.

Kernelized version

- For good kernels, robustness in the feature-space implies robustness in the input-space, which completes the proof.
- Bad kernels can be non-consistent. Eg., $k(\mathbf{x}, \mathbf{x}') = \mathbf{1}_{(\mathbf{x}=\mathbf{x}')}$. The result of SVM is $\mathrm{sign}(\sum_{i=1}^{m} \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b)$, and provides no meaningful prediction if $\mathbf{x}$ is not one of the training samples.

Conclusion

- Conclusion:
  1. Regularization is indeed Robustness, and Vice Versa.
  2. Consistency is the result of Robustness.

- Future works:
  1. New regularization schemes using Robustness.
  2. A general robust learning framework.

- Preprint available: http://www.cim.mcgill.ca/∼xuhuan/