Statistical performances of SVM regularization in classification

Sébastien Loustau
Université de Provence, LATP

December 5-6, 2008, Berlin

Contents

Classification

We observe $D_n = \{(X_1, Y_1), \ldots (X_n, Y_n)\}$ where

- (X_i, Y_i) i.i.d. from π **unknown** probability distribution
- $X_i \in \mathcal{X}$ $(= \mathbb{R}^d)$ input variable
- $Y_i \in \{-1, +1\}$ corresponding class

Goal: $X \longrightarrow Y$? $\Leftrightarrow$ Find a "good" $f : \mathcal{X} \mapsto \{-1, +1\}$.

What "good" means?

- Generalization error of f: $R(f) = \mathbb{P}(f(X) \neq Y)$.

Lemma

$f^*(x) := \text{sign}(2\eta(x) - 1)$ *verifies*

$$R(f^*) = \min_{f \text{ measurable}} R(f)$$

where $\eta(x) := \mathbb{P}(Y = 1|X = x)$ *conditional probability function.*

f^* is called the Bayes rule and is unknown !

Goal: estimate f^* from D_n with $\hat{f}_n$.

- excess risk of $\hat{f}_n$: $R(\hat{f}_n, f^*) = R(\hat{f}_n) - R(f^*)$.

Consistency and learning rates

Consider a classifier $\hat{f}_n$.

Definition

- $\hat{f}_n$ is **consistent** if

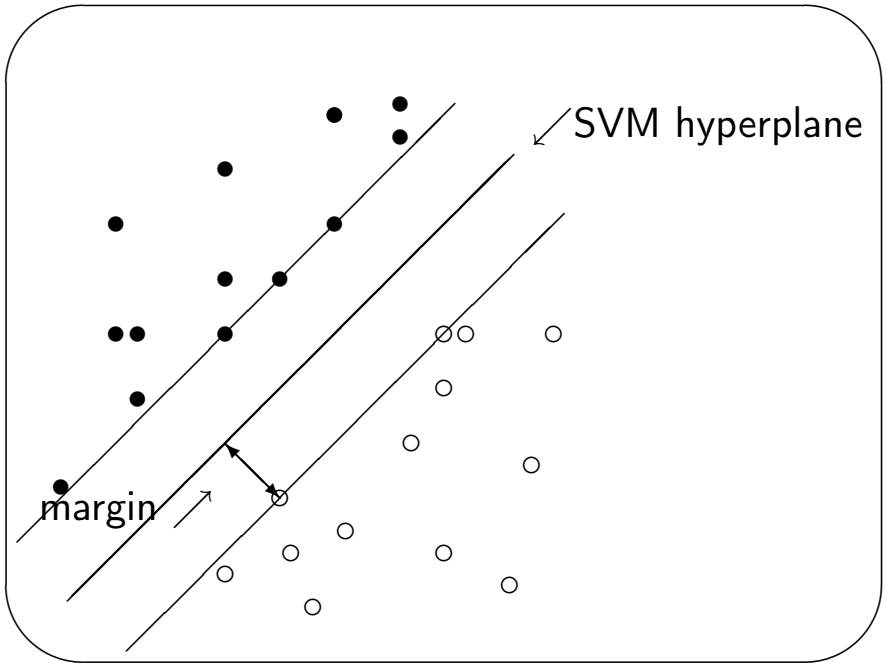
$$\mathbb{E}_{\pi^n} R(\hat{f}_n, f^*) \underset{n \to +\infty}{\longrightarrow} 0.$$

- We say that $\hat{f}_n$ **learns with rate** $(\psi_n)_{n \in \mathbb{N}}$ if there exists a constant $C > 0$ such that

$$\mathbb{E}_{\pi^n} R(\hat{f}_n, f^*) \leq C \psi_n.$$

Goal of Part I: Learning rates of SVM classifiers

Support Vector Machines: geometrical description



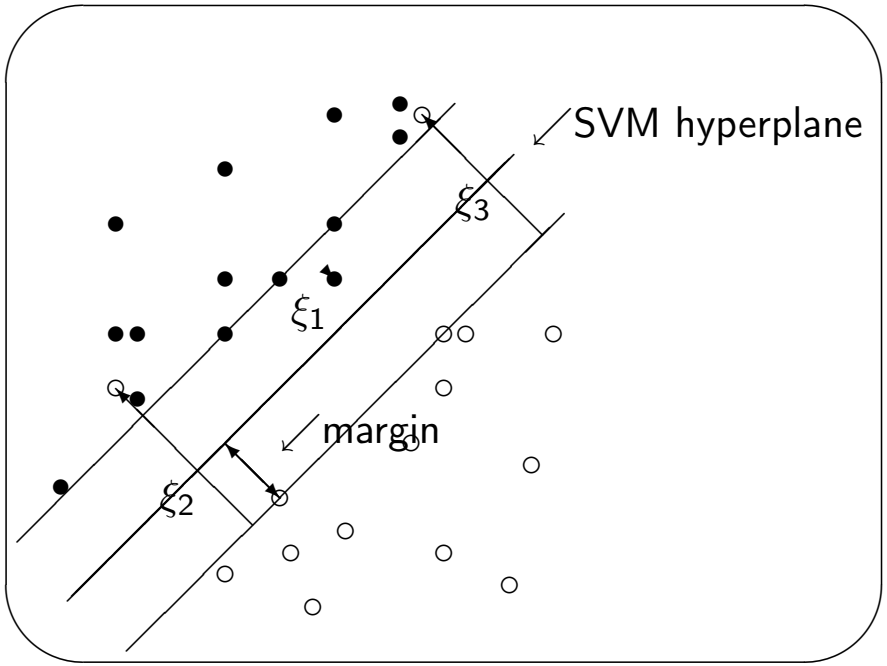
Linear case with no noise, $\mathcal{X} = \mathbb{R}^2$

Maximal margin hyperplane :

$$\begin{cases} \max\limits_{w,b} \ m \\ \\ \forall\, i = 1,..,n \ y_i f(x_i) \geq m, \end{cases}$$

where $f(x) =< w,x > +b$.

SVM : slack variables in the presence of noise



Linear case with noise, $\mathcal{X} = \mathbb{R}^2$

Wa add the slack variables ξ :

$$(*) \begin{cases} \max\limits_{w,b} \ (m - C\sum_{i=1}^{n} \xi_i) \\ \\ Y_i f_{w,b}(X_i) \geq 1 - \xi_i \, , \, \xi_i \geq 0, \end{cases}$$

where $f(x) =< w,x > +b$.

SVM Regularization

$(*)$ can be written:

$$\min_{f \in \mathcal{H}_K} \left(\frac{1}{n} \sum_{i=1}^{n} l(Y_i, f(X_i)) + \alpha_n \|f\|_K^2 \right),$$

where

- $l(y, f(x)) = (1 - yf(x))_+$ is the hinge loss (or SVM loss),
- $\mathcal{H}_K$ is a Reproducing Kernel Hilbert Space with norm $\|\cdot\|_K$,
- α_n smoothing parameter to determine explicitely.

Representer Theorem: $\hat{f}_n(x) = \sum_{i=1}^{n} Y_i v_i^* K(X_i, x)$, where v^* is sparse.

Reproducing Kernel Hilbert Space (RKHS)

Definition

- *A kernel is a symmetric and positive definite application* $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$.
- $\mathcal{H}_K$ *is a Hilbert space such that:*
  - $\mathcal{H}_K \subset \mathbb{R}^{\mathcal{X}}$.
  - $K(x,\cdot) \in \mathcal{H}_K, \forall x \in \mathcal{X}$.
  - $< f, K(x,\cdot) >_K = f(x), \forall f \in \mathcal{H}_K$.

  K *is called the reproducing kernel of* $\mathcal{H}_K$.

Given $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, we want to know :

$$f \in \mathcal{H}_K \Rightarrow \text{ What kind of regularity for } f \text{ ?}$$

Spectral representation of $\mathcal{H}_K$

Considering,

$$\begin{aligned} L_K : \quad L^2(\mathcal{X}) &\to \quad \mathbb{R}^{\mathcal{X}} \\ f &\longmapsto \quad \int_{\mathcal{X}} f(y)K(x,y)dy, \end{aligned}$$

we have $Im(L_K) = \mathcal{H}_K$.
If K continuous over $\mathcal{X}$ compact (Mercer kernel):

$$\mathcal{H}_K = \left\{ f \in L^2(\mathcal{X}) : f = \sum_{k\in\mathbb{N}} a_k\varphi_k \text{ et } \sum_{k\in\mathbb{N}} \frac{a_k^2}{\lambda_k} < \infty \right\},$$

where $L_K\varphi_k = \lambda_k\varphi_k$.

Convolution case

RBF or convolution kernel : $K(x,y) = \Phi(x-y)$, $x \in \mathbb{R}^d$.

Theorem

Consider a RBF kernel K with $\Phi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ and $0 < \mathcal{F}[\Phi] \in L^1(\mathbb{R}^d)$. Then :

$$\mathcal{H}_K = \{f \in L^2(\mathbb{R}^d) : \|f\|_K^2 = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{|\mathcal{F}[f](\omega)|^2}{\mathcal{F}[\Phi](\omega)} d\omega < \infty\}.$$

Regularity of $f \in \mathcal{H}_K$ in terms of the asymptotic behaviour of its Fourier transform.

Sobolev smooth kernel

Definition

A kernel K_s is called **Sobolev smooth kernel** with exponent $s > d$ if $\mathcal{H}_{K_s}$ is such that:

$$\mathcal{H}_{K_s} = \mathcal{W}_{\frac{s}{2}}^2 := \{f \in L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} |\mathcal{F}[f](\omega)|^2 (1 + \|\omega\|^2)^{s/2} d\omega < \infty\}.$$

Example :

- $K(x,y) = \exp(-\sigma\|x-y\|)$ (Laplacian kernels).
- $K(x,y) = \exp(-\sigma\|x-y\|^2)$ (Gaussian kernels) **are not** Sobolev smooth.

Assumptions for learning rates

π has **margin parameter** $q > 0$ if there exists a constant $c_0 > 0$ such that for all sufficiently small $t > 0$,

$$\mathbb{P}(|2\eta(X) - 1| \leq t) \leq c_0 t^q.$$

Related to the local slope of $\eta(x) = \mathbb{P}(Y = 1|X = x)$ at the level $\frac{1}{2}$.

Learning rates using Sobolev spaces

Theorem

Let π be a distribution over $\mathbb{R}^d \times \{-1,1\}$ such that:

- *π has margin assumption $q \in [0, +\infty]$,*
- *$f^* \in \mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$ for $s > 0$.*

Consider the SVM minimization with Sobolev smooth kernel K_r, with $r > 2s \vee d$.
If we choose α_n such that

$$\alpha_n = n^{-\frac{r(r-s)(q+1)}{s(r(q+2)-d)+d(r-s)(q+1)}},$$

then there exists a constant C such that

$$\mathbb{E}R(\hat{f}_{\alpha_n}, f^*) \leq Cn^{-\frac{rs(q+1)}{s(r(q+2)-d)+d(r-s)(q+1)}}.$$

Problem of adaptation

Construction of $\hat{f}_{\alpha_n}$ depends on

$$\alpha_n(\mathrm{r},\mathrm{d},\mathbf{q},\mathbf{s}) = n^{-\frac{r(r-s)(q+1)}{s(r(q+2)-d)+d(r-s)(q+1)}},$$

where q and s are **unknown** parameters ! $\Rightarrow \hat{f}_{\alpha_n}$ is non-adaptive.

Goal of Part II : Build a data-dependent (or adaptive) classifier.

Aggregation for adaptation

Principle of the method:

- Split the data $D_n = (D_{n_1}^1, D_{n_2}^2)$.
- Construct a family of SVM with $D_{n_1}^1$

$$\{\hat{f}_{\alpha_1}, \ldots, \hat{f}_{\alpha_M}\} \text{ where } \alpha_1, \ldots \alpha_M \in \Lambda \text{ is a grid}$$

- Calculate from $D_{n_2}^2$ a sequence of weights w_k, for $k \in \{1 \ldots M\}$.

$\Rightarrow$ Aggregate $\tilde{f}_n$ is defined by :

$$\tilde{f}_n = \sum_{k=1}^{n} w_k \hat{f}_{\alpha_k}.$$

Aggregation

Precisely :

- Choice of the weights : exponential weights

$$\omega_k^{(n)} = \frac{\exp\left(\sum_{i=n_1+1}^{n} Y_i \hat{f}_{\alpha_k}(X_i)\right)}{\sum_{k' \in \Lambda} \exp\left(\sum_{i=n_1+1}^{n} Y_i \hat{f}_{\alpha_{k'}}(X_i)\right)}.$$

- Choice of the grid

$$\Lambda = \left\{ \alpha_k = n_2^{-\phi_k} : \phi_k = \frac{1}{2} + k n_2^{-b}, k = 0, \ldots, \lfloor \frac{(2r-d)n_2^b}{2d} \rfloor \right\},$$

from previous theorem.

Performances of the aggregate

Theorem

If we take $n_2 = \lceil a \frac{n}{\log n} \rceil$, *under the assumptions of last theorem, there exists a constant* C *such that*

$$\mathbb{E}R(\tilde{f}_n, f^*) \leq Cn^{-\frac{rs(q+1)}{s(r(q+2)-d)+d(r-s)(q+1)}}.$$

Now $\tilde{f}_n$ is computable !

Implementation

We consider two cases:

- Sobolev case: $\tilde{f}_n$ comes from the previous description. Here $K_\sigma(x,y) = \exp\left(-\sigma\|x-y\|\right)$.
- Gaussian case : $\tilde{f}_n$ from Steinwart and Scovel (2007). Here $K_\sigma(x,y) = \exp\left(-\sigma\|x-y\|^2\right)$.

Classification Datasets

Dataset	d	n	p	realizations
Banana	2	400	4900	100
Titanic	3	150	2051	100
Thyroid	5	140	75	100
Diabetis	8	468	300	100
Breast-cancer	9	200	77	100
Flare-solar	9	666	400	100
Heart	13	170	100	100
Image	18	1300	1010	20
Waveform	21	400	4600	100

$$\text{"Dataset"} = \{(D_n^1, T_p^1), (D_n^2, T_p^2), \ldots, (D_n^{100}, T_p^{100})\}.$$

Experimental results

Dataset	Laplace Aggregate	Gaussian Aggregate
Banana	11.31± 0.57	11.43±0.84
Titanic	22.77±1.13	22.57±0.79
Thyroid	5.45±2.68	6.31±2.97
Diabetis	28.34±2.27	27.80±2.06
Breast-cancer	32.74±5.16	32.13±4.77
Flare-solar	35.69±1.93	34.87±1.82
Heart	22.12±3.98	22.62±3.77
Image	3.95±0.74	5.66±0.74
Waveform	14.12±0.72	15.04±0.79

Experimental results

Dataset	Laplace Aggregate	Gaussian Aggregate	Rästch et al. (2001)
Banana	11.31± 0.57	11.43±0.84	11.53±0.66
Titanic	22.77±1.13	22.57±0.79	22.42±1.02
Thyroid	5.45±2.68	6.31±2.97	4.80±2.19
Diabetis	28.34±2.27	27.80±2.06	23.53±1.76
Breast-cancer	32.74±5.16	32.13±4.77	26.04±4.74
Flare-solar	35.69±1.93	34.87±1.82	32.43±1.82
Heart	22.12±3.98	22.62±3.77	15.95±3.26
Image	3.95±0.74	5.66±0.74	2.96±0.6
Waveform	14.12±0.72	15.04±0.79	9.88±0.83

Conclusion and comments

We have presented a classifier with a mathematical background.
Advantages:

- no tuning parameter to adjust.
- reasonable computing time.

Possible improvements:

- large dimension
- average of aggregates, temperature parameter, ...

... Thanks for your attention !