



école —————
normale —————
supérieure —————
paris-saclay —————

université
PARIS-SACLAY

Introduction to Statistical Learning

Nicolas Vayatis

Lecture # 4 - Statistical analysis of mainstream ML algorithms

Part I - Margin bounds and application to SVM

Machine Learning Methods

Optimization is central

Some popular examples :

- Sparse linear models \rightarrow convex optimization (gradient methods)
- Kernel ridge regression \rightarrow convex optimization (quadratic optimization)
- Deep learning \rightarrow nonconvex optimization (stochastic gradient descent) + implicit regularization (tricks)

At the end of the day :

loss+training data+functional class+optimization \rightarrow random rule \hat{f}_n

Main theoretical objectives of the course

- Take a well-known ML algorithm which operates in \mathcal{F} : it produces a (random) sequence of decision rules $(\hat{f}_n)_{n \geq 1}$ in \mathcal{F} . Then show :
- **Convergence of estimation error :**

$$L(\hat{f}_n) \rightarrow \inf_{\mathcal{F}} L \text{ almost surely as } n \rightarrow \infty ,$$

- **Upper bounds :** with probability at least $1 - \delta$, there exists some constant c such that :

$$L(\hat{f}_n) - \inf_{\mathcal{F}} L \leq C(\mathcal{F}, n) + c \sqrt{\frac{\log(1/\delta)}{n}} ,$$

where $C(\mathcal{F}, n) = O(1/\sqrt{n})$ after processing some complexity/stability measure

Key principle to lower bias : Regularized optimization

- Objective : aim at consistency $L(\hat{f}_n) \rightarrow L^*$ almost surely as $n \rightarrow \infty$.
- Take \mathcal{F} a *very large* space and define a proper penalty term :

$$C_n(f) = \underbrace{\hat{L}_n(f)}_{\text{Training error}} + \lambda \underbrace{\text{pen}(f, n)}_{\text{Regularization}}$$

- Example : ridge regression where $f(x) = \theta^T x$:
 $\hat{L}_n(f) = \frac{1}{n} \sum_{i=1}^n (Y_i - \theta^T X_i)^2$ and $\text{pen}(f, n) = \frac{1}{n} \|\theta\|_2^2$
- The penalty grows with the complexity of f and vanishes when $n \rightarrow \infty$

Stability of ML algorithms (from last lecture)

Definition of (uniform) stability

- Consider an algorithm which provides an estimator \tilde{h}_n on a sample of size n and we denote \tilde{h}'_n the estimator resulting from the same sample where one observation was changed.
- For a given loss function ℓ , We say that the algorithm is (uniformly) γ -stable if there exists a constant γ for which we have : for any training sample, and for any pair (x, y) ,

$$|\ell(y, \tilde{h}_n(x)) - \ell(y, \tilde{h}'_n(x))| \leq \gamma$$

Error bound based on stability

- Consider a loss function ℓ which is uniformly bounded by $M > 0$ and \tilde{h}_n is the output of a γ -uniformly stable learning algorithm.
- We have, with probability at least $1 - \delta$:

$$L(\tilde{h}_n) \leq \hat{L}_n(\tilde{h}_n) + \gamma + (2n\gamma + M)\sqrt{\frac{\log(1/\delta)}{2n}}$$

- Proof left as an exercise

Stability of regularized methods

- Consider the case of linear models estimated by minimizing an error based on convex and ρ -Lipschitz loss, with an ℓ_2 -penalty : for any $\lambda > 0$

$$\hat{\beta} = \hat{\beta}_n^R(\lambda) = \arg \min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(Y_i, \beta^T X_i) + \lambda \|\beta\|_2^2 \right\}$$

- It can be shown that this algorithm is stable with parameter γ such that :

$$\gamma \leq \frac{2\rho^2}{n\lambda}$$

Oracle inequality

- Set $L(\beta) = \mathbb{E}(\ell(Y, \beta^T X))$ and

$$\widehat{L}_n(\beta) = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, \beta^T X_i)$$

- We have, assuming the loss is convex and ρ -Lipschitz : for any $\lambda > 0$,

$$\mathbb{E}(L(\widehat{\beta})) \leq \inf_{\beta \in \mathbb{R}^d} \{L(\beta) + \lambda \|\beta\|_2^2\} + \frac{2\rho^2}{n\lambda}$$

PAC bound

- Assume the loss is convex and ρ -Lipschitz, and all $\|\beta\|$ bounded by B
- Set $\lambda = \sqrt{\frac{2\rho^2}{nB^2}}$, we then have :

$$\mathbb{E}(L(\hat{\beta})) \leq \inf_{\beta \in \mathbb{R}^d} L(\beta) + \rho B \sqrt{\frac{8}{n}}$$

- and, for any $\epsilon > 0$ and $n \geq \frac{8\rho^2 B^2}{\epsilon^2}$, we also have :

$$\mathbb{E}(L(\hat{\beta})) \leq \inf_{\beta \in \mathbb{R}^d} L(\beta) + \epsilon$$

Margin maximization and Linear Support Vector Machines

Margin of a hyperplane

- Consider a hyperplane $\mathbb{H} = \mathbb{H}(\beta, b) = \{u \in \mathbb{R}^d : \beta^T u + b\}$
- Distance of a point x to \mathbb{H} :

$$d(x, \mathbb{H}) = \inf_{u \in \mathbb{H}} \|x - u\|$$

- If $\|\beta\| = 1$ then $d(x, \mathbb{H}) = |\beta^T x + b|$
- Definition of the margin M relatively to a set of points x_1, \dots, x_n and a hyperplane $\mathbb{H}(\beta, b)$ with $\|\beta\| = 1$:

$$M(\beta, b) = \min_{i \in \{1, \dots, n\}} |\beta^T x_i + b|$$

Linear classification problem

Constraints with slack variables

- Setup : consider binary classification problem with supervised classification data $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ with $X_i \in \mathbb{R}^d$ and $Y_i \in \{-1, +1\}$
- Strategy for linear classification : find a hyperplane \mathbb{H} which maximizes the margin (excluding "noisy" labels).
- Relaxed set of constraints assuming $\|\beta\| = 1 : \forall i,$

$$Y_i \cdot (\beta^T X_i + b) \geq M \cdot (1 - \xi_i)$$

where $\xi_i \geq 0$ are *slack* variables.

- We can release the norm of β and set $M = 1/\|\beta\|$

Linear SVM

- Margin maximization : for $\lambda > 0$, solve

$$\min_{\beta, b, \xi_1, \dots, \xi_n} \left(\lambda \|\beta\|^2 + \frac{1}{n} \sum_{i=1}^n \xi_i \right)$$

subject to : $\forall i, Y_i \cdot (\beta^T X_i + b) \geq 1 - \xi_i$, and $\xi_i \geq 0$.

- Equivalent formulation with hinge loss :

$$\min_{\beta, b} \left(\lambda \|\beta\|^2 + \frac{1}{n} \sum_{i=1}^n \ell((X_i, Y_i), (\beta, b)) \right)$$

where $\ell((x, y), (\beta, b)) = \max\{0, 1 - y(\beta^T x + b)\}$

- Oracle inequality holds on a bounded set because hinge loss is $\|x\|$ -Lipschitz.

Some comments

- SVM as an example of convex risk minimization
- Oracle inequality and PAC bound hold under mild assumptions on the data
- Stability bounds are far more accurate than VC bounds
- Using duality, one can show that solving regularized ERM with hinge loss over a linear class boils down to solving a *quadratic* optimization problem.
- Furthermore :
 - there exist coefficients $\alpha_1, \dots, \alpha_n$ such that $\beta = \sum_{i \in I} \alpha_i X_i$ where $I = \{i : |\beta^T X_i + b| = 1\}$
 - the dual formulation only implies the original data by means of the Gram matrix with coefficients $X_i^T X_j$, for any i, j .

From linear SVM to general SVM

RKHS theory in a nutshell

Theorem.

Let $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ a kernel that is symmetric and positive.

Then, there exists :

- a Hilbert space $(\mathcal{H}_k, \langle \cdot, \cdot \rangle)$, called the Reproducing Kernel Hilbert Space
- a mapping $\Phi : \mathbb{R}^d \rightarrow \mathcal{H}_k$ such that :

$$\forall u, v \in \mathbb{R}^d, \quad k(u, v) = \langle \Phi(u), \Phi(v) \rangle$$

Plus, we have the reproducing property :

$$\forall h \in \mathcal{H}_k, \quad \forall u \in \mathbb{R}^d, \quad h(u) = \langle h, k(u, \cdot) \rangle$$

and $\|h\|_k = \sqrt{\langle h, h \rangle}$

Principle of Support Vector Machines

- Consider $(\mathcal{H}_k, \langle \cdot, \cdot \rangle)$ the Reproducing Kernel Hilbert Space corresponding to the kernel $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ symmetric and positive
- Class of soft classifiers :

$$h \in \mathcal{H}(X) \doteq \left\{ h = \sum_{i=1}^n \alpha_i k(X_i, \cdot) : \alpha_1, \dots, \alpha_n \in \mathbb{R} \right\} \subset \mathcal{H}_k$$

- Optimization problem : set $\lambda > 0$

$$\hat{h}_\lambda = \arg \min_{\mathcal{H}_k} \left\{ \sum_{i=1}^n (1 - Y_i h(X_i))_+ + \lambda \|h\|_k^2 \right\}$$

Key property of SVM

- By the representer's theorem (admitted), it suffices to minimize over $\mathcal{H}(X)$ instead of \mathcal{H}_k
- Note that, if $h \in \mathcal{H}(X)$:

$$\|h\|_k^2 = \sum_{i,j} \alpha_i \alpha_j k(X_i, X_j)$$

Some tools :

Contraction Principle and Concentration Inequality

Contraction principle

Theorem. (Ledoux, Talagrand (1991))

Consider $\psi : \mathbb{R} \rightarrow \mathbb{R}$ a Lipschitz function with constant κ

Then, for any class \mathcal{F} of real-valued functions, we have :

$$\widehat{R}_n(\psi \circ \mathcal{F}) \leq \kappa \widehat{R}_n(\mathcal{F})$$

Uniform bound with Rademacher average

Reminder

Proposition.

Consider \mathcal{F} a class of functions from \mathcal{Z} to $[0, 1]$

Then, with probability at least $1 - \delta$:

$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}(f(Z_1)) - \frac{1}{n} \sum_{i=1}^n f(Z_i) \right) \leq 2R_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

and

$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}(f(Z_1)) - \frac{1}{n} \sum_{i=1}^n f(Z_i) \right) \leq 2\hat{R}_n(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

Consistency of global methods based on optimization

The case of kernel-based classification

Global methods (e.g. CRM)

- Based on empirical minimization of error functionals
- Example in the case of *soft* classifiers $h : \mathbb{R}^d \rightarrow \mathbb{R}$
- Convex risk minimization, with φ positive convex cost function :

$$\hat{A}(h) = \frac{1}{n} \sum_{i=1}^n \varphi(-Y_i h(X_i))$$

- Note that if $h \in \text{span}(\mathcal{H})$ with \mathcal{H} some class of classifiers, then the minimization problem is convex.
- Main issue : complexity of the class \mathcal{H} of candidate decision rules

Rademacher complexity of SVM

- Let X_1, \dots, X_n be an n -sample in \mathbb{R}^d , and denote by K the Gram matrix with coefficients $k(X_i, X_j)$, $1 \leq i, j \leq n$.
- Introduce the subspace of functions with bounded RKHS norm :

$$\mathcal{F}_\tau = \{h \in \mathcal{H}_k : \|h\|_k \leq \tau\}$$

- We then have :

$$\hat{R}_n(\mathcal{F}_\tau) \leq \frac{\tau \sqrt{\text{trace}(K)}}{n}$$

- In addition, if we have : $k(X_i, X_i) \leq R^2$ for $1 \leq i \leq n$, then

$$\hat{R}_n(\mathcal{F}_\tau) \leq \frac{\tau R}{\sqrt{n}}$$

Margin loss

- Fix $\rho > 0$
- The *margin loss* is defined, for any $u, v \in \mathbb{R}$, as :
 $\ell(u, v) = m_\rho(uv)$ where

$$m_\rho(t) = \begin{cases} 0 & \text{if } \rho \leq t \\ 1 - \frac{t}{\rho} & \text{if } 0 \leq t \leq \rho \\ 1 & \text{if } t \leq 0 \end{cases}$$

- Empirical margin error on a sample D_n :

$$\hat{L}_{n,\rho}(f) = \frac{1}{n} \sum_{i=1}^n m_\rho(Y_i f(X_i))$$

Margin bounds for SVM classification

Theorem. (Fixed margin)

Let \mathcal{H}_k the RKHS with bounded kernel $k \leq R^2$.

Fix $\rho \in (0, 1)$, and $\delta > 0$. Then with probability at least $1 - \delta$, we have, for any SVM classifier g :

$$L(g) \leq \hat{L}_{n,\rho}(g) + 2 \left(\frac{\tau R}{\rho \sqrt{n}} \right) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

and

$$L(g) \leq \hat{L}_{n,\rho}(g) + 2 \left(\frac{\tau \sqrt{\text{trace}(K)}}{\rho n} \right) + 3 \sqrt{\frac{\log(2/\delta)}{2n}}$$