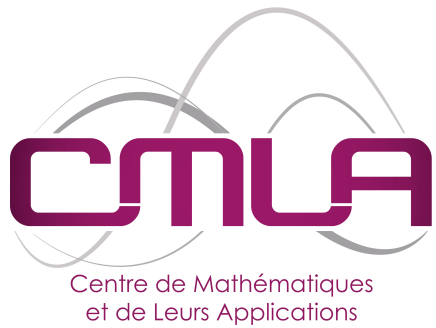
Introduction to Statistical Learning #4

Nicolas Vayatis (CMLA - ENS Cachan)



MVA 2016-2017

Course outline

- Chapter 1 - Modeling aspects:
  - Data, decision, risk, optimality
- Chapter 2 - Tools:
  - Concentration inequalities and complexity measures
- Chapter 3 - Theory:
  - Consistency and error bounds of mainstream learning algorithms
- Chapter 4 - Advanced topics:
  - Multiclass classification, ranking, link prediction, ...

Overview of Chapter 3 - Statistical analysis of learning algorithms

1. Consistency of local methods
2. Consistency of global methods
    a. Support Vector Machines
    b. Boosting
    c. Neural networks
3. Aggregation, resampling and randomization: Bagging, Random Forests
4. Complexity regularization

Statistical model: Classification data

- (X, Y) random pair with distribution P over $\mathbb{R}^d \times \{-1, +1\}$
- Joint distribution P described by (P_X, η)
- Posterior probability function:

$$\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\} \ , \quad \forall x \in \mathbb{R}^d$$

Learning problem: Classification

- Classifier: $g : \mathbb{R}^d \to \{0, +1\}$
- Classification error: $L(g) = \mathbb{P}\{g(X) \neq Y\}$

$$L(g) = \mathbb{E}\big(\eta(X) \cdot \mathbb{I}\{g(X) = 0\} + (1 - \eta(X)) \cdot \mathbb{I}\{g(X) = 1\}\big)$$

- Bayes rule: $g^*(x) = \mathbb{I}\{\eta(x) > 1/2\}\ , \quad \forall x \in \mathbb{R}^d$
- Bayes error: $L^* = \inf_g L(g)$

Main requirement: consistency

- Performance/error functional $L : \mathcal{F} \to \mathbb{R}_+$
- Bayes error: $L^* = \inf_f L(f)$
- Sequence of n-samples $(D_n)_{n \geq 1}$
- Sequence of empirical decision rules $(\widehat{f}_n)_{n \geq 1}$ with $\widehat{f}_n = \widehat{f}(\cdot, D_n)$
- True error of $\widehat{f}_n$ conditionally on D_n given by $L(\widehat{f}_n)$

  $$\textit{Example: } L(\widehat{f}_n) = \mathbb{P}(Y \cdot \widehat{f}_n(X) < 0 \mid D_n)$$

- The sequence of decision rules $(\widehat{f}_n)_{n \geq 1}$ is said to be **L-consistent** (respectively, strongly L-consistent) if

  $$L(\widehat{f}_n) \to L^* \text{ as } n \to \infty \ ,$$

  holds in probability (respectively, almost-surely).

Efficient classification algorithms for high dimensional data

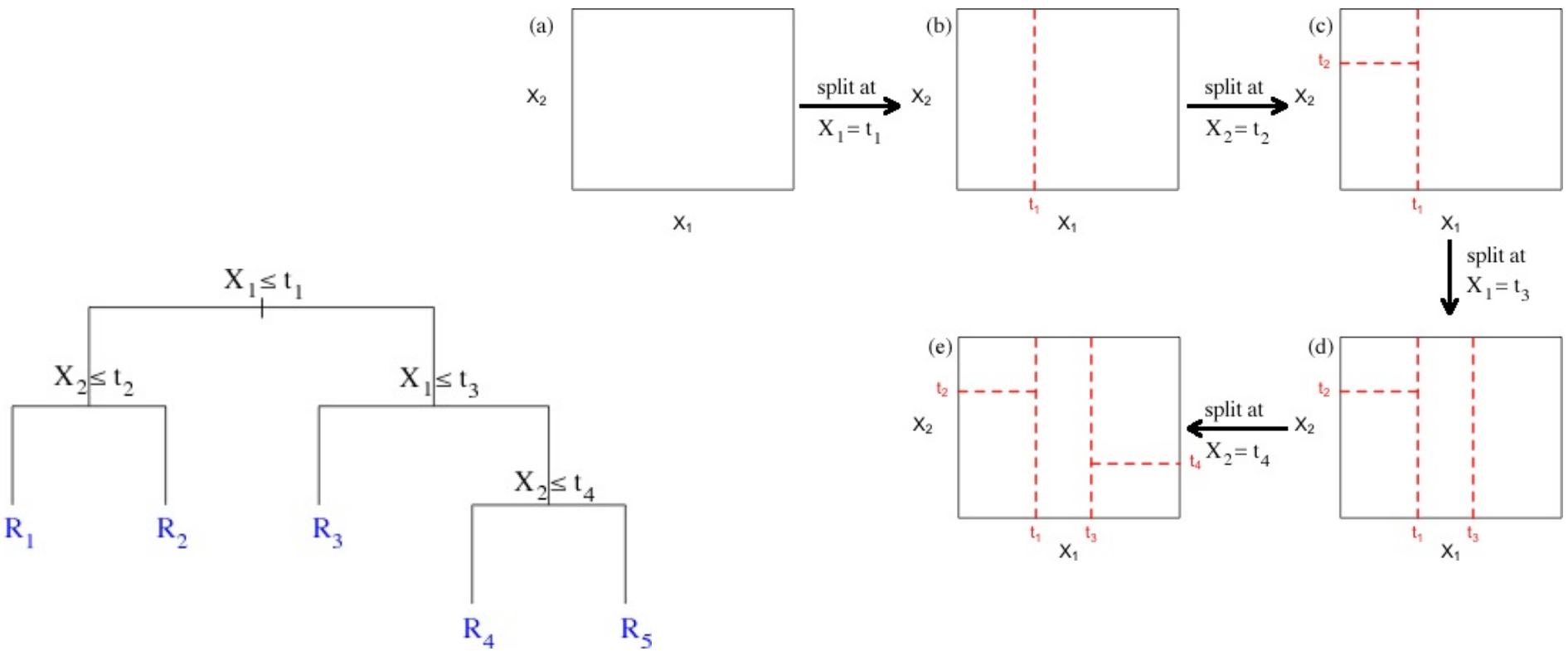
1. Local averaging
   - Histogram or Kernel rules
   - Nearest Neighbors
   - Partitioning methods: decision trees (CART, C4.5, ...)
2. Global methods
   - Neural Networks: minimize (smooth version of) classification error
   - Support Vector Machines, Boosting: minimize convex surrogate of classification error
3. Meta-algorithms: Aggregation and randomization
   - Bagging, Boosting, Random Forests: use aggregation, resampling and randomization
4. Statistical ingredient: complexity regularization with penalties

1. Consistency of local methods

Local methods

- Key ingredient 1: a concept of neighborhood of prediction point
- Key ingredient 2: decision through a majority vote
- Main examples:
  - Kernel rules (here 'kernels' refer to regularizing functions)
  - Nearest neighbor rule
  - Partitioning rules (including histograms)

Decision Trees - Principles of recursive partitioning



- Geometry of splits: perpendicular, linear, binary or more, ...
- Impurity criterion: entropy, Gini, classification error
- Algorithms: CART, C4.5, ...

2. Global methods consistency

a. Support Vector Machines

Global methods (e.g. CRM)

- Data in supervised learning: sample $D_n = \{(X_i, Y_i) \ : \ i = 1, \ldots, n\}$ with Y-labels in $\{-1, +1\}$
- Based on empirical minimization of error functionals
- Example in the case of classification with decision rule $f \ : \ \mathbb{R}^d \to \mathbb{R}$
- Convex risk minimization, with φ positive convex cost function:

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

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

RKHS theory in a nutshell

Theorem

Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ 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 $\Phi : \mathbb{R}^d \to \mathcal{H}_k$ 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

- Optimization problem: set $\lambda > 0$

$$\hat{f}_\lambda = \arg\min_{\mathcal{H}_k} \left\{ \sum_{i=1}^{n} (1 - y_i f(x_i))_+ + \lambda \|f\|_k \right\}$$

- Class of candidate decision rules: $g = \mathrm{sgn}(f)$ where

$$f \in \mathcal{F}(X) \doteq \left\{ h = \sum_{i=1} \alpha_i k(x_i, \cdot) \ : \ \alpha_1, \ldots, \alpha_n \in \mathbb{R} \right\} \subset \mathcal{H}_k$$

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

$$\|f\|_k^2 = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j)$$

Rademacher complexity of SVM

- Consider a sample $x_1, \ldots, x_n$
- 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 RKHS norm bounded by M:

$$\mathcal{F}_M = \{h \in \mathcal{H}_k \ : \ \|h\|_k \leq M\}$$

- We then have:

$$\widehat{R}_n(\mathcal{F}_M) \leq \frac{M\sqrt{\mathrm{Tr}\,(K)}}{n}$$

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

$$\widehat{R}_n(\mathcal{F}_M) \leq \frac{MR}{\sqrt{n}}$$

The concept of margin

- Assume that data are linearly separable: there exists $\theta \in \mathbb{R}^d$ and $\theta_0 \in \mathbb{R}$ such that

$$\forall i \in \{1, \dots, n\} \ , \quad y_i(\theta^T x_i + \theta_0) > 0$$

- Recall that the distance of any point $x_0 \in \mathbb{R}^d$ to the hyperplane $\theta^T x + \theta_0 = 0$ is given by:

$$\frac{|\theta^T x_0 + \theta_0|}{\|\theta\|}$$

- The margin is defined by:

$$\rho = \min_{i=1,\dots,n} \frac{|\theta^T x_i + \theta_0|}{\|\theta\|}$$

Beyond the linearly separable case - 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

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

- Empirical margin error on a sample D_n:

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

Contraction principle

- Consider $\psi : \mathbb{R} \to \mathbb{R}$ a Lipschitz function with constant κ
- Then, for any class $\mathcal{F}$ of real-valued functions, we have:

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

Reminder from Chapter 2 - Uniform bound

- Consider $\mathcal{F}$ a class of functions from $\mathcal{Z}$ to $[0,1]$
- Then, with probability at least $1-\delta$:

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

and

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

Margin bounds for SVM classification (for fixed ρ)

- Let $\mathcal{H}_k$ the RKHS related to kernel k
- Fix $\rho \in (0,1)$, and $\delta > 0$, then with probability at least $1-\delta$, we have, for any SVM classifier g:

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

and

$$L(g) \leq \widehat{L}_{n,\rho}(g) + 2\left(\frac{M\sqrt{\mathrm{Tr}\,(K)}}{\rho n}\right) + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

Margin bounds for SVM classification (uniform in ρ)

- Let $\mathcal{H}_k$ the RKHS related to kernel k
- Fix $\delta > 0$, then with probability at least $1 - \delta$, we have, for any SVM classifier g and any $\rho \in (0, 1)$:

$$L(g) \leq \widehat{L}_{n,\rho}(g) + 4\left(\frac{MR}{\rho\sqrt{n}}\right) + \sqrt{\frac{\log\log_2(2/\rho)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}$$

and

$$L(g) \leq \widehat{L}_{n,\rho}(g) + 4\left(\frac{M\sqrt{\mathrm{Tr}\,(K)}}{\rho n}\right) + \sqrt{\frac{\log\log_2(2/\rho)}{n}} + 3\sqrt{\frac{\log(4/\delta)}{2n}}$$

2. Global methods consistency

b. Boosting

Main ingredients for boosting aggregation

- Data sample $D_n = \{(X_i, Y_i) \ : \ i = 1, \ldots, n\}$
- Base class $\mathcal{H}$ of *weak* classifiers such as decision trees
- Cost function: here $\varphi(u) = \exp(-u)$
- Discrete distributions over D_n denoted by Π_t, $t \geq 1$
- Weighted empirical classification error:

$$\widehat{\varepsilon}_t(h) = \sum_{i=1}^{n} \Pi_t(i) \mathbb{I}\{h(X_i) \neq Y_i\}$$

- AdaBoost (Schapire 1990, Freund & Schapire 1995)

A 'deterministic' procedure: AdaBoost

1. **Initialization.** Π_1 is the uniform distribution on $\{1, \ldots, n\}$
2. **Boosting iterations.** For $t = 1, \ldots, T$, compute :

$$\widehat{h}_t = \underset{h \in \mathcal{H}}{\arg\min}\, \widehat{\varepsilon}_t(h) \quad \text{and set} \quad \widehat{\varepsilon}_t^* = \widehat{\varepsilon}_t(\widehat{h}_t)$$

3. **Boosting distribution and weight updates.** if $\widehat{\varepsilon}_t^* < 1/2$
   - the weight

$$w_t = \frac{1}{2} \log\left(\frac{1 - \widehat{\varepsilon}_t^*}{\widehat{\varepsilon}_t^*}\right)$$

   - the distribution

$$\Pi_{t+1}(i) \propto \Pi_t(i) \exp\left(-w_t Y_i \cdot \widehat{h}_t(X_i)\right)$$

4. **Output - Aggregation step.** Compute the boosting aggregate:

$$\widehat{f}_n^B(x) = \sum_{t=1}^{T} w_t \widehat{h}_t(x) \;, \quad \widehat{g}_n^B = \operatorname{sgn}(\widehat{f}_n^B)$$

Empirical error bound for AdaBoost

- Let $\widehat{f}_n^B = \sum_{t=1}^{T} w_t \widehat{h}_t$ the AdaBoost decision function after T rounds and assume that, for all t, we have $\widehat{\varepsilon}_t^* < 1/2$
- Then, we have:

$$\widehat{L}_n\left(\widehat{f}_n^B\right) \leq \exp\left(-2\sum_{t=1}^{T}\left(\frac{1}{2} - \widehat{\varepsilon}_t^*)^2\right)\right)$$

- Moreover, if $\gamma \leq \frac{1}{2} - \widehat{\varepsilon}_t^*$, for all $t \leq T$, then:

$$\widehat{L}_n\left(\widehat{f}_n^B\right) \leq \exp(-2\gamma^2 T)$$

Key complexity argument for consistency

- Let $\epsilon_1, \ldots, \epsilon_n$ be IID Rademacher random variables $\{-1, +1\}$
- Empirical Rademacher complexity

$$\widehat{R}_n(\mathcal{F}) = \mathbb{E}\left(\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \epsilon_i f(X_i) \mid X_1, \ldots, X_n \right)$$

- Magic!

$$\widehat{R}_n\big(\mathrm{conv}(\mathcal{H})\big) = \widehat{R}_n(\mathcal{H})$$

- And under VC dimension assumption ($V < +\infty$) on $\mathcal{H}$, we have (admitted):

$$R_n(\mathcal{H}) = \mathbb{E}(\widehat{R}_n(\mathcal{H})) = O\left(\sqrt{\frac{V}{n}}\right)$$

Margin bounds for boosting (convex) aggregates

- Let $\mathcal{H}$ denote a set of classifiers with finite VC dimension V
- Fix $\rho > 0$, and $\delta > 0$, then with probability at least $1 - \delta$, we have, for any function $f \in \mathrm{conv}(\mathcal{H})$:

$$L(f) \leq \widehat{L}_{n,\rho}(f) + \frac{2}{\rho}\sqrt{\frac{2V \log(en/V)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

and

$$L(f) \leq \widehat{L}_{n,\rho}(f) + \frac{2}{\rho}\widehat{R}_n(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

- Remark: for an AdaBoost decision function $\widehat{f}_n^B$, use $\widehat{f}_n^B / \|w\|_1$ above instead of f

Empirical margin error bound for boosting aggregate

- Let $\widehat{f}_n^B = \sum_{t=1}^{T} w_t \widehat{h}_t$ the AdaBoost decision function after T rounds and assume that, for all t, we have $\widehat{\varepsilon}_t^* < 1/2$
- Then, for any $\rho \in (0, 1)$, we have:

$$\widehat{L}_{n,\rho}\left(\widehat{f}_n^B / \|w\|_1\right) \leq 2^T \prod_{t=1}^{T} \sqrt{(\widehat{\varepsilon}_t^*)^{1-\rho}(1-\widehat{\varepsilon}_t^*)^{1+\rho}}$$

Interpretation as coordinate descent

- Let the empirical convex risk with exponential loss: for $\mathbf{w} \in \mathbb{R}^T$

$$\widehat{A}_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \exp\left(-Y_i \sum_{t=1}^{T} w_t h_t(X_i) \right)$$

- let $\mathbf{e_t}$ be the unit vector on the t-th coordinate and $\mathbf{w}_{t-1} = (w_1, \ldots, w_{t-1}, 0, \ldots, 0)^T$

- We have that: optimal element at each iteration is h_t

$$\mathbf{e_t} = \arg\min_t \left. \frac{d\widehat{A}_n(\mathbf{w}_{t-1} + \eta \mathbf{e_t})}{d\eta} \right|_{\eta=0}$$

- and also that: optimal weight is as in AdaBoost

$$\frac{d\widehat{A}_n(\mathbf{w}_{t-1} + \eta \mathbf{e_t})}{d\eta} = 0 \Leftrightarrow \eta = \frac{1}{2} \log\left(\frac{1 - \widehat{\varepsilon}_t^*}{\widehat{\varepsilon}_t^*} \right)$$

Statistical formulation and assumptions

- Let $\mathcal{F}_1 = \mathrm{conv}(\mathcal{H})$ and $\mathcal{F}_\lambda = \lambda \cdot \mathcal{F}_1$
- Boosting estimator:

$$\widehat{f}^{\lambda} = \operatorname*{arg\,min}_{f \in \mathcal{F}_\lambda} \widehat{A}(f)$$

- Denseness property: assume P and $\mathcal{H}$ such that

$$\lim_{\lambda \to \infty} \inf_{f \in \mathcal{F}_\lambda} A(f) = A^*$$

Consistency result

- Assume $\varphi \in \{\exp, \mathrm{logit}\}$
- Assume that $\mathcal{H}$ has finite VC dimension
- Consider $\lambda_1, \lambda_2, \ldots$ a positive sequence such that:

$$\lambda_n \to \infty \quad \text{and} \quad \lambda_n \varphi'(\lambda_n) \sqrt{\frac{\log n}{n}} \to 0$$

- Then:

$$L(\mathrm{sgn}(\widehat{f}^{\lambda_n})) \to L^* \ , \quad \text{almost surely}$$

- Fast rates result with model selection: Blanchard, Lugosi and V. (2003)