Introduction to Statistical Learning #3

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:

Overview of Chapter 2 - Tools

1. Motivations: empirical risk minimization
2. Tools from probability: inequalities
3. Tools from complexity theory

1 . Motivations

ERM and sample complexity

True error and empirical error

- Consider the binary classification prediction problem: $Y \in \{0, 1\}$
- Classifiers or predictors of the form: $g : \mathbb{R}^d \to \{0, 1\}$
- True error: $L(g) = \mathbb{P}\{Y \neq g(X)\}$
- Given a sample $D_n = \{(X_i, Y_i) : i = 1, \ldots, n\}$, the empirical error of g is defined as:

$$\widehat{L}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Y_i \neq g(X_i)\}$$

What can be learned? - achievable case

- Case where candidates are $g(x) = \mathbb{I}\{x \in R\}$, and R is a rectangle with axis-orthogonal edges
- Posterior probability : $\eta(x) = \mathbb{I}\{x \in R^*\}$, hence $Y = \eta(X)$
- A simple algorithm:

  given a sample D_n, choose a rectangle $\widehat{R}_n$ with axis-orthogonal edges such that:

  - $Y_i = \mathbb{I}\{X_i \in \widehat{R}_n\}$, *for all i*
  - $\widehat{R}_n$ *has minimum volume*

- Denote $\widehat{g}_n(x) = \mathbb{I}\{x \in \widehat{R}_n\}$

How many samples to learn?

- Goal: achieve a precision of ϵ with confidence of $1 - \delta$
- Formally: $\mathbb{P}\{L(\widehat{g}_n) > \epsilon\} \leq \delta$
- Sample complexity in the rectangle example:

$$n = n(\epsilon, \delta) \geq \frac{4}{\epsilon} \log\left(\frac{4}{\delta}\right)$$

First result for the finite case - achievable goal

- Assume there is a finite family $\mathcal{G}$ of candidates
- Assume also that $g^* \in \mathcal{G}$
- Consider an algorithm such that:

  for any IID sample D_n, it returns $\widehat{g}_n$ such that $\widehat{L}_n(\widehat{g}_n) = 0$

- Then, we have, for any ϵ, δ, that $\mathbb{P}\{L(\widehat{g}_n) > \epsilon\} \leq \delta$ if

$$n = n(\epsilon, \delta) \geq \frac{1}{\epsilon}\left(\log|\mathcal{G}| + \log\left(\frac{1}{\delta}\right)\right)$$

Second result for the finite case - unachievable goal

- Assume there is a finite family $\mathcal{G}$ of candidates
- Now we may as well have $g^* \notin \mathcal{G}$
- Then, we have, for any δ, with probability at least $1 - \delta$:

$$\forall g \in \mathcal{G} \ , \quad L(g) \leq \widehat{L}_n(g) + \sqrt{\frac{\log |\mathcal{G}| + \log\left(\frac{1}{\delta}\right)}{2n}}$$

- Technical result involved: Hoeffding's inequality

Empirical risk minimization (ERM)

- Set the ERM classifier as:

$$\widehat{g}_n = \arg\min_{g \in \mathcal{G}} \widehat{L}_n(g)$$

- Define the best classifier in the class as:

$$\overline{g} = \arg\min_{g \in \mathcal{G}} L(g)$$

- We have:

$$L(\widehat{g}_n) - L(\overline{g}) \leq 2 \sup_{g \in \mathcal{G}} |L(g) - \widehat{L}_n(g)|$$

- Need for uniform inequalities! What about the countable case?

2 . Probability inequalities

Hoeffding's lemma

- Consider Z a random variable such that:
  - $\mathbb{E}(Z) = 0$
  - $Z \in [a, b]$ almost surely
- Then, for any $s > 0$, we have:

$$\mathbb{E}\left(e^{sZ}\right) \leq \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

- Subgaussian behavior of the Laplace transform of bounded random variables

Hoeffding's inequality

- Consider $Z_1, \ldots, Z_n$ IID over $[0, 1]$ and $\overline{Z}_n = \frac{1}{n} \sum_{i=1}^{n} Z_i$
- We have, for any $t > 0$

$$\mathbb{P}\{\overline{Z}_n - \mathbb{E}(Z_1) > t\} \leq \exp(-2nt^2)$$

and

$$\mathbb{P}\{\overline{Z}_n - \mathbb{E}(Z_1) < -t\} \leq \exp(-2nt^2)$$

- Implies the strong law of large numbers (by Borel-Cantelli lemma)

Beyond IID: Martingale difference - definition

- Consider $V = (V_1, \ldots, V_n, ...)$ and $Z = (Z_1, \ldots, Z_n, ...)$ two sequences of random variables
- We call V_n a martingale difference sequence wrt Z_n if, for any n we have:
  - V_n is a function of $Z_1, \ldots, Z_n$
  - $\mathbb{E}(V_{n+1} \mid Z_1, \ldots, Z_n) = 0$

Azuma's inequality

- Consider V_n a a martingale difference sequence wrt Z_n
- Assume that, for any n, there exists $c_n \geq 0$ such that:

$$Z_n \leq V_n \leq Z_n + c_n$$

- We have, for any $t > 0$

$$\mathbb{P}\left(\sum_{i=1}^{n} V_i > t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

and

$$\mathbb{P}\left(\sum_{i=1}^{n} V_i < -t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

Bounded differences functions - definition

- Consider a metric space $\mathcal{Z}$
- Consider a function $h : \left(\mathcal{Z}\right)^n \to \mathbb{R}$ of n vector-valued variables $z_1, \ldots, z_n$
- We say that h is a function with bounded differences if there exist $c_1, \ldots, c_n > 0$ such that:

$$\sup_{z_1,\ldots,z_n,z_i' \in \mathcal{Z}} |h(z_1, \ldots, z_n) - h(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \leq c_i$$

- Concept of regularity based on componentwise bounded variations property

McDiarmid's inequality

- Assume h is a function with bounded differences and denote by $c_1, \ldots, c_n > 0$ the upper bounds on its componentwise variations
- We have, for any $t > 0$

$$\mathbb{P}\left(h(Z_1, \ldots, Z_n) - \mathbb{E}\big(h(Z_1, \ldots, Z_n)\big) > t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

and

$$\mathbb{P}\left(h(Z_1, \ldots, Z_n) - \mathbb{E}\big(h(Z_1, \ldots, Z_n)\big) < -t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

- Generalization of Hoeffding's inequality and law of large numbers!

3 . Complexity measures

a. Rademacher complexity

Rademacher random variables - definition

- A Rademacher random variable ε satisfies:
  - $\epsilon \in \{-1, +1\}$ almost surely
  - $\mathbb{P}(\varepsilon = -1) = \mathbb{P}(\varepsilon = +1) = \frac{1}{2}$
- Rademacher random variables are Bernoulli-type (uniform) sign random variables

Rademacher complexity - definitions

- Consider $\mathcal{F}$ a class of bounded functions over some metric space $\mathcal{Z}$
- Consider a sample of $D_n = (Z_1, \ldots, Z_n)$ of IID random variables and a sample $\varepsilon_1, \ldots, \varepsilon_n$ of IID Rademacher random variables
- The empirical Rademacher complexity of $\mathcal{F}$ wrt to the sample D_n is defined as:

$$\widehat{R}_n(\mathcal{F}) = \mathbb{E}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(Z_i) \middle| D_n \right)$$

- The Rademacher complexity of $\mathcal{F}$ (deterministic quantity) is then given by the sequence:

$$R_n(\mathcal{F}) = \mathbb{E}\big(\widehat{R}_n(\mathcal{F})\big)$$

- Quantification of the correlation of the class $\mathcal{F}$ with a random noise vector

Concentration of Rademacher complexity

- Define

$$h(Z_1, \ldots, Z_n) = \widehat{R}_n(\mathcal{F}) = \mathbb{E}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(Z_i) \middle| D_n \right)$$

- The function h satisfies the bounded differences condition with $c_i = \frac{1}{n}$, for any i
- Therefore, we have, by McDiarmid's inequality, with probability at least $1 - \delta$:

$$R_n(\mathcal{F}) \leq \widehat{R}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Application to a uniform bound - Proposition

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

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

and

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

- Easily applies to bounding the ERM performance for binary classification

Proof sketch

- Step 1 - Bounded differences function ($c_i = 1/n$)

$$\Phi(Z_1, \ldots, Z_n) = \sup_{f \in \mathcal{F}} \left(\mathbb{E}\big(f(Z_1)\big) - \frac{1}{n} \sum_{i=1}^{n} f(Z_i) \right)$$

- Step 2 - Upper bound on $\mathbb{E}\Phi(Z_1, \ldots, Z_n)$

$$\mathbb{E}\Phi(Z_1, \ldots, Z_n) \leq 2R_n(\mathcal{F})$$

  by a series of classical arguments: symmetrization #1, Jensen's inequality, symmetrization #2, sub-additivity of the sup

- Step 3 - McDiarmid's inequality

  (applied twice)

Application to ERM - Theorem

- Let $\mathcal{G}$ be a class of classifiers from $\mathbb{R}^d$ to $\{0, 1\}$
- Consider $\widehat{g}_n$ the ERM classifier:

$$\widehat{g}_n = \arg\min_{g \in \mathcal{G}} \widehat{L}_n(g)$$

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

$$L(\widehat{g}_n) \leq \inf_{g \in \mathcal{G}} L(g) + R_n(\mathcal{G}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

and

$$L(\widehat{g}_n) \leq \inf_{g \in \mathcal{G}} L(g) + \widehat{R}_n(\mathcal{G}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

3 . Complexity measures

b. Combinatorial concepts

Growth function - definition

- Consider $\mathcal{G}$ a class of bounded functions over $\mathbb{R}^d$
- The growth function is

$$\gamma(\mathcal{G}, n) = \max_{x_1,\ldots,x_n \in \mathbb{R}^d} |\{(g(x_1), \ldots, g(x_n)) \ : \ g \in \mathcal{G}\}|$$

- Note that, for any n:

$$\log_2 \gamma(\mathcal{G}, n) \leq n$$

Massart's lemma

- Consider C a finite subset of $\mathbb{R}^n$ and $R = \max_{z \in C} \|z\|_2$
- Consider a sample $\varepsilon_1, \ldots, \varepsilon_n$ of IID Rademacher random variables
- We then have:

$$\mathbb{E}\left(\sup_{z \in C} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i z_i \middle| D_n \right) \leq \frac{R\sqrt{2 \log |C|}}{n}$$

where $z = (z_1, \ldots, z_n)$

Application to upper bounding Rademacher average

- Consider $\mathcal{G}$ a class of functions over $\mathbb{R}^d$ with values in $\{-1, +1\}$
- Then, we have, by Massart's lemma:

$$R_n(\mathcal{F}) \leq \sqrt{\frac{2 \log \gamma(\mathcal{G}, n)}{n}}$$

VC dimension - definition

- Consider a class of sets $\mathcal{C}$ in $\mathbb{R}^d$
- Define the growth function $\gamma(\mathcal{C}, n)$ of $\mathcal{C}$ as the growth function of

$$\mathcal{G} = \{g = \mathbb{I}_C \ : \ C \in \mathcal{C}\}$$

- The VC dimension is the largest index such that there exist a set of n points in $\mathbb{R}^d$ for which all its subsets can be intercepted by elements of $\mathcal{C}$, *i.e.*

$$V(\mathcal{C}) = \max\{n \ : \ \gamma(\mathcal{C}, n) = 2^n\}$$

VC dimension - examples

- Hyperplanes in $\mathbb{R}^d$: $V = d + 1$
- Axis-aligned rectangles in $\mathbb{R}^2$: $V = 4$
- Just any rectangles in $\mathbb{R}^2$: $V = 7$
- Triangles in $\mathbb{R}^2$: $V = 7$
- Convex polygons in $\mathbb{R}^2$: $V = +\infty$

Sauer's lemma

- Consider a class of sets $\mathcal{C}$ in $\mathbb{R}^d$ with $V(\mathcal{C}) = V$
- Then, for any $n > 0$, we have:

$$\gamma(\mathcal{C}, n) \leq \sum_{i=1}^{V} \binom{n}{i}$$

- In particular, for $n \geq V$, we have:

$$\gamma(\mathcal{C}, n) \leq \left(\frac{en}{V}\right)^V$$

Summary of complexity measures

- From finest complexity measure to more explicit
- Rademacher bounded by growth function

$$R_n(\mathcal{G}) \leq \sqrt{\frac{2 \log \gamma(\mathcal{G}, n)}{n}}$$

- Rademacher bounded by VC dimension

$$R_n(\mathcal{G}) \leq \sqrt{\frac{2\big(1 + \log(n/V)\big)}{(n/V)}}$$

- Upper bound on empirical Rademacher average also available with high confidence with an extra order $n^{-1/2}$-term

Next class - Chapter 3

- Consistency and error bounds of mainstream learning algorithms
  - Algorithms: Support Vector Machines, Boosting, Neural networks
  - Optimization problems formulations, Rademacher complexities, consistency results
  - Complexity regularization
  - Aggregation principle, resampling and randomization: Bagging, Random Forests