Mathematics of Machine Learning - Summer School

Lecture 1
Introduction

June 28, 2021

Patrick Rebeschini
Department of Statistics, University of Oxford

Statistical/Computational Learning Theory

Problem formulation (out-of-sample prediction):

- Given n data $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. from $\mathbf{P}$ (**unknown**)
- Consider the *population risk* $r(a) = \mathbf{E}\,\phi(a(X), Y)$

Goal: **Compute** $A \in \sigma\{(X_i, Y_i)_{i=1}^n\}$ such that $\underbrace{r(A) - \inf_a r(a)}_{\text{excess risk}}$ is **small**

What does it mean to solve the problem **optimally**?

- **Statistics:** A is minimax-optimal w.r.t. the class of distrib. $\mathcal{P}$ if

$$\mathbf{E}\,r(A) - \inf_a r(a) \quad \sim \inf_{A \in \sigma\{Z_1, \ldots, Z_n\}} \sup_{\mathbf{P} \in \mathcal{P}} \left\{ \mathbf{E}\,r(A) - \inf_a r(a) \right\}$$

- **Runtime:** Computing A takes same time to read the data, i.e. $O(nd)$ cost
- **Memory:** Storing $O(1)$ data point at a time, i.e. $O(d)$ storage cost
- **Distributed computations:** Runtime $O(1/m)$ if we have m machines
- (communication, privacy, robustness...)

Offline statistical learning: prediction

1. Observe training data $Z_1, \ldots, Z_n$ i.i.d. from unknown distribution
2. Choose action $A \in \mathcal{A} \subseteq \mathcal{B}$
3. Suffer an expected/population loss/risk $r(A)$, where

$$\boxed{a \in \mathcal{B} \longrightarrow r(a) := \mathbf{E}\, \ell(a, Z)}$$

with ℓ is an prediction loss function and Z is a new test data point

Goal: Minimize the estimation error defined by the following decomposition

$$\boxed{\underbrace{r(A) - \inf_{a \in \mathcal{B}} r(a)}_{\text{excess risk}} = \underbrace{r(A) - \inf_{a \in \mathcal{A}} r(a)}_{\text{estimation error}} + \underbrace{\inf_{a \in \mathcal{A}} r(a) - \inf_{a \in \mathcal{B}} r(a)}_{\text{approximation error}}}$$

as a function of n and notions of "complexity" of the set $\mathcal{A}$ of the function ℓ

Note: Estimation/Approximation trade-off, a.k.a. complexity/bias

Goal - Applications

- The data distribution is unknown so also the risk r can not be computed
- Nevertheless, r is used as a way to assess the performance of the algorithm
- **Goal:** Derive upper bounds for the estimation error
- **Bounds in expectation:**

$$\mathbf{E}\, r(A) - r(a^\star) \leq \boxed{\texttt{Expectation}}$$

- **Bounds in probability:** For any $\varepsilon \geq 0$,

$$\mathbf{P}\Big(r(A) - r(a^\star) \geq \varepsilon\Big) \leq \boxed{\texttt{UpperTail}(\varepsilon)}$$

or, equivalently, for any $\delta \in [0, 1]$,

$$\mathbf{P}\Big(r(A) - r(a^\star) < \boxed{\texttt{UpperTail}^{-1}(\delta)}\Big) \geq 1 - \delta$$

ERM and Uniform Learning

- A natural framework is given by the empirical risk minimization (ERM)

$$a \in \mathcal{B} \longrightarrow R(a) := \frac{1}{n} \sum_{i=1}^{n} \ell(a, Z_i)$$

- A natural algorithm is given by the minimizer of the ERM

$$A^\star \in \operatorname*{argmin}_{a \in \mathcal{A}} R(a)$$

- **Uniform Learning:** The estimation error is bounded by

$$\underbrace{r(A^\star) - r(a^\star)}_{\text{estimation error for ERM}} \leq \underbrace{\sup_{a \in \mathcal{A}} \{r(a) - R(a)\} + \sup_{a \in \mathcal{A}} \{R(a) - r(a)\}}_{\texttt{Statistics}}$$

- Statistical Learning deals with bounding the `Statistics` term (Vapnik 1995)
- **Generalization Error:** $r(a) - R(a) \approx \frac{1}{n^{(\text{test})}} \sum_{i=1}^{n^{(\text{test})}} \ell(a, Z_i^{(\text{test})}) - \frac{1}{n} \sum_{i=1}^{n} \ell(a, Z_i)$

Goal - Theory

To analyse the ERM algorithm, we need to develop tools to:

- Control the suprema of random processes:

$$\mathbf{E} f(Z_1, \dots, Z_n) \leq \boxed{?}$$

with $f(Z_1, \dots, Z_n) = \sup_{a \in \mathcal{A}} \{R(a) - r(a)\}$

- Control the concentration of random processes:

$$\mathbf{P}\Big(f(Z_1, \dots, Z_n) - \mathbf{E}\, f(Z_1, \dots, Z_n) \geq \varepsilon \Big) \leq \boxed{\texttt{UpperTail}_f(\varepsilon)}$$

$$\mathbf{P}\Big(f(Z_1, \dots, Z_n) - \mathbf{E}\, f(Z_1, \dots, Z_n) < \boxed{\texttt{UpperTail}_f^{-1}(\delta)} \Big) \geq 1 - \delta$$

Q. Can the ERM rule/algorithm $A^\star$ be computed?
(we depart from classical learning theory and also consider computational issues)

Computational aspects

- The ERM is in general intractable
- We need to approximately compute it
- We will consider stochastic optimisation methods to minimize R.
- New error decomposition that highlight the statistical/computational parts

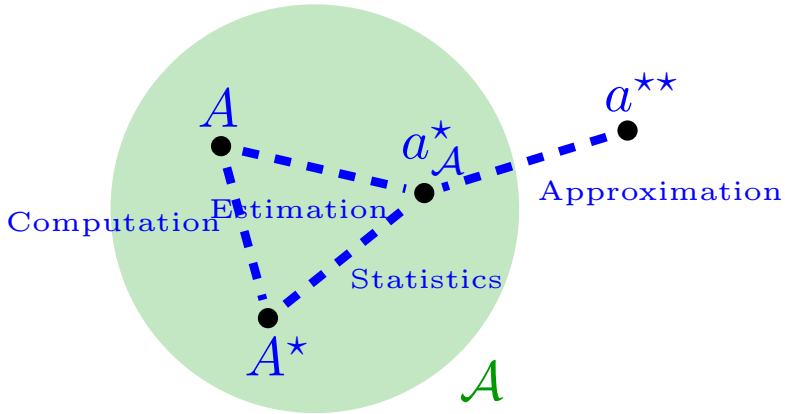
$$r(A) - r(a^\star) \leq \underbrace{R(A) - R(A^\star)}_{\texttt{Optimization}} + \underbrace{\sup_{a \in \mathcal{A}} \{r(a) - R(a)\} + \sup_{a \in \mathcal{A}} \{R(a) - r(a)\}}_{\texttt{Statistics}}$$

- Key insight (Bousquet and Bottou 2008)

$$\boxed{\text{Bound on } \texttt{Optimisation}} \sim \boxed{\text{Bound on } \texttt{Statistics}}$$

It is only necessary to run an optimization algorithm until we are guaranteed to find a estimator with an accuracy of the same order as the statistical fluctuations of the problem

Explicit regularization: uniform convergence



- Estimation/approximation: $r(A) - r(a^{\star\star}) = \underbrace{r(A) - r(a^\star)}_{\text{Estimation}} + \underbrace{r(a^\star) - r(a^{\star\star})}_{\text{Approximation}}$
- Classical error decomposition for estimation error:

$$\underbrace{r(A) - r(a^\star)}_{\text{Estimation}} = r(A) - R(A) + R(A) - R(A^\star) + \underbrace{R(A^\star) - R(a^\star)}_{\leq 0} + R(a^\star) - r(a^\star)$$

$$r(A) - r(a^{\star\star}) \leq \underbrace{2 \sup_{a \in \mathcal{A}} |r(a) - R(a)|}_{\text{Statistics}} + \underbrace{R(A) - R(A^\star)}_{\text{Computation}} + \underbrace{r(a^\star) - r(a^{\star\star})}_{\text{Approximation}}$$

Offline statistical learning: estimation

1. Observe training data $Z_1, \ldots, Z_n$ i.i.d. from distr. parametrized by $a^\star \in \mathcal{A}$
2. Choose a parameter $A \in \mathcal{A}$
3. Suffer a loss $\ell(A, a^\star)$ where ℓ is an estimation loss function

Goal: Minimize the estimation loss $\ell(A, a^\star)$ as a function of n and notions of "complexity" of the set $\mathcal{A}$ of the function ℓ

Online statistical learning

At every time step $t = 1, 2, \ldots, n$:

1. Choose an action $A_t \in \mathcal{A}$
2. A dynamic data point Z_t is sampled from an unknown distribution
3. Suffer an expected/population loss/risk $r(A_t)$, where

$$\boxed{a \in \mathcal{B} \longrightarrow r(a) := \mathbf{E}\, \ell(a, Z)}$$

with ℓ a prediction loss function and Z is a new data point

Goal: Minimize the (normalized) (pseudo-)regret defined as

$$\boxed{\frac{1}{n} \sum_{t=1}^{n} r(A_t) - \inf_{a \in \mathcal{A}} r(a)}$$

as a function of n and notions of "complexity" of the set $\mathcal{A}$ of the function ℓ

Probability Bounds: Concentration inequalities

Concentration phenomenon

If $X_1, \ldots, X_n$ are independent (or weakly dependent) random variables, then $f(X_1, \ldots, X_n)$ is "close" to its mean $\mathbf{E}[f(X_1, \ldots, X_n)]$ provided that $x_1, \ldots, x_n \to f(x_1, \ldots, x_n)$ is not too "sensitive" to any of the coordinates x_i.

- If $X_1, \ldots, X_n$ are i.i.d. mean μ **(Problem 1.1)**:

$$\left\{ \mathbf{E}\left[\left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right)^p \right] \right\}^{1/p} \leq \frac{c_p}{\sqrt{n}},$$

  E.g., **variance** ($p = 2$) captures how close random variable is to its mean

  These notions of "closeness" capture **size** of fluctuations

- We need notion of "closeness" that captures **distribution** of fluctuations:

$$\mathbf{P}\Big(f(Z_1, \ldots, Z_n) - \mathbf{E}\, f(Z_1, \ldots, Z_n) \geq \varepsilon \Big) \leq \boxed{\texttt{UpperTail}_f(\varepsilon)}$$

$$\mathbf{P}\Big(f(Z_1, \ldots, Z_n) - \mathbf{E}\, f(Z_1, \ldots, Z_n) < \boxed{\texttt{UpperTail}_f^{-1}(\delta)} \Big) \geq 1 - \delta$$

Markov's Inequality and Chernoff's bounds

Markov's inequality is the main result to prove tail inequalities

Markov's Inequality (Proposition 6.1)

For any non-negative random variable X we have, for any $\varepsilon \geq 0$,

$$\mathbf{P}(X \geq \varepsilon) \leq \frac{\mathbf{E}X}{\varepsilon}$$

Proof: $X = X1_{X \geq \varepsilon} + X1_{X < \varepsilon} \geq \varepsilon 1_{X \geq \varepsilon}$, where we used that $X \geq 0$

Chernoff's Bound (Proposition 6.2)

For any random variable X and any $\lambda \geq 0$ we have, for any $\varepsilon \in \mathbb{R}$,

$$\mathbf{P}(X \geq \varepsilon) \leq e^{-\lambda\varepsilon}\, \mathbf{E}\, e^{\lambda X}$$

Proof: Exponentiate and apply Markov's inequality: $\mathbf{P}(X \geq \varepsilon) = \mathbf{P}(e^{\lambda X} \geq e^{\lambda\varepsilon}) \leq \frac{\mathbf{E}\, e^{\lambda X}}{e^{\lambda\varepsilon}}$

On the Summer School

- Lecturers: Patrick Rebeschini (theory); Tomas Vaškevičious (practicals)
- Pointers to Propositions/Lemmas/Theorems/Problems in the slides refer to the lecture notes and materials available at
  http://www.stats.ox.ac.uk/~rebeschi/teaching/AFoL/20/index.html
- Hand-on practicals in Python are available on the School's GitHub page
  https://github.com/alan-turing-institute/mathematics-of-ml-course
- Logistics of the School available at
  https://hackmd.io/@VAuPdHDeQGer_2vJb_c2Ig/maths-ml-ss
- To send (anonymous) feedback:
  https://docs.google.com/forms/d/
  1kf3QihO23f54JdXCLkc9CBgdhVWLp74KlL-d_AqSCzE