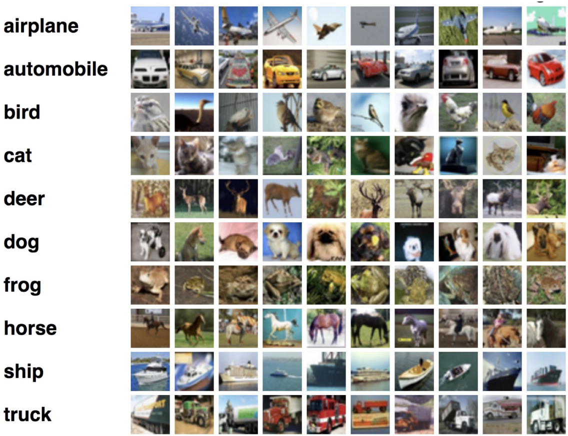
Algorithmic Foundations of Learning

Lecture 2
Maximal Inequalities and Rademacher complexity

Patrick Rebeschini

Department of Statistics
University of Oxford

Offline statistical learning: prediction



Offline learning: prediction
Given a batch of observations (images & labels) interested in predicting the label of a new image

Offline statistical learning: prediction

1. Observe training data $Z_1, \dots, Z_n$ 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

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: derive bounds in expectation

- Goal:

$$\mathbf{E}\, \underbrace{r(A^\star) - r(a^\star)}_{\text{estimation error for ERM}} \lesssim \frac{f(\text{dimension})}{n^\alpha}$$

- By uniform learning, it suffices to bound the suprema of random processes:

$$\mathbf{E} g(Z_1, \ldots, Z_n) \le \frac{f(\text{dimension}, \text{complexity of } \mathcal{A})}{n^\alpha}$$

with $g(Z_1, \ldots, Z_n) = \sup_{a \in \mathcal{A}} \{r(a) - R(a)\} = \sup_{a \in \mathcal{A}} \left\{ \mathbf{E}\ell(a, Z) - \frac{1}{n} \sum_{i=1}^{n} \ell(a, Z_i) \right\}$

- We aim to derive a uniform, non-asymptotic Law of Large Numbers
- In machine learning, dimension can be $\gg 10^6$, e.g., number of pixels
- Ideally, $f(\text{dimension}) \ll$ dimension, e.g., $f(\text{dimension}) \sim \log(\text{dimension})$
- Ideally, $\alpha = 1$ (fast rate)

Hoeffding's Lemma (Lemma 2.1)

Let X be a bounded random variable $a \le X - \mathbf{E}X \le b$. Then, for any $\lambda \in \mathbb{R}$,

$$\boxed{\mathbf{E}\, e^{\lambda(X - \mathbf{E}X)} \le e^{\lambda^2 (b-a)^2/8}}$$

Proof

- W.l.o.g., take $\mathbf{E}X = 0$. Let $\psi(\lambda) = \log \mathbf{E}\, e^{\lambda X}$

$$\psi'(\lambda) = \frac{\mathbf{E}[X e^{\lambda X}]}{\mathbf{E}\, e^{\lambda X}} \qquad \psi''(\lambda) = \frac{\mathbf{E}[X^2 e^{\lambda X}]}{\mathbf{E}\, e^{\lambda X}} - \left(\frac{\mathbf{E}[X e^{\lambda X}]}{\mathbf{E}\, e^{\lambda X}}\right)^2$$

- $\psi''(\lambda)$ is the variance of X under the distribution $\mathbf{Q}(\mathrm{d}x) = \frac{e^{\lambda x}}{\mathbf{E}\, e^{\lambda X}} \mathbf{P}(\mathrm{d}x)$
- $\psi''(\lambda) = \mathbf{Var}_{\mathbf{Q}}\left(X - \frac{a+b}{2}\right) \le \mathbf{E}_{\mathbf{Q}}\left[\left(X - \frac{a+b}{2}\right)^2\right] \le \frac{(b-a)^2}{4}$
- Fundamental Thm of Calculus: $\psi(\lambda) = \int_0^\lambda \int_0^\mu \psi''(\rho) \mathrm{d}\rho \mathrm{d}\mu \le \frac{\lambda^2 (b-a)^2}{8}$

□

Maximum of finitely many bounded random variables (Proposition 2.2)

Let $X_1, \dots, X_n$ be n centered random variables bounded in the interval $[a, b]$.

$$\mathbf{E} \max_{i \in [n]} X_i \le \frac{b-a}{\sqrt{2}} \sqrt{\log n}$$

Proof

- $X = \max_{i \in [n]} X_i$. Exponentiate. Jensen's ineq. as $x \to e^{\lambda x}$ ($\lambda > 0$) is convex:

$$\mathbf{E} X = \frac{1}{\lambda} \log e^{\lambda \mathbf{E} X} \le \frac{1}{\lambda} \log \mathbf{E}\, e^{\lambda X}$$

- Bound maximum of non-negative numbers by the sum:

$$\mathbf{E}\, e^{\lambda X} = \mathbf{E}\, e^{\lambda \max_{i \in [n]} X_i} = \mathbf{E} \max_{i \in [n]} e^{\lambda X_i} \le \mathbf{E} \sum_{i=1}^{n} e^{\lambda X_i} = \sum_{i=1}^{n} \mathbf{E}\, e^{\lambda X_i}$$

- Put everything together and use Hoeffding's lemma ($\mathbf{E}\, e^{\lambda X_i} \le e^{\lambda^2 (b-a)^2/8}$):

$$\mathbf{E} \max_{i \in [n]} X_i \le \frac{1}{\lambda} \log \sum_{i=1}^{n} e^{\lambda^2 (b-a)^2/8} = \frac{1}{\lambda} \log n + \frac{\lambda (b-a)^2}{8}$$

- Optimizing the bound $\alpha/\lambda + \lambda\beta$ over $\lambda > 0$ yields the minimum is at $\lambda = \sqrt{\alpha/\beta}$ and the optimal value $2\sqrt{\alpha\beta} = (b-a)\sqrt{\log n / 2}$ □

Bound in expectation for finitely-many actions

Bound in expectation (Proposition 2.3)

If the loss function ℓ is bounded by c, we have

$$\boxed{\mathbf{E} \max_{a \in \mathcal{A}} \{r(a) - R(a)\} \leq c \frac{\sqrt{2 \log |\mathcal{A}|}}{\sqrt{n}}}$$

Proof: Same as above, using the independence of the data $Z_1, \ldots, Z_n$
(note that for each $a \in \mathcal{A}$, $r(a) - R(a)$ is a centered random variable as $\mathbf{E} R(a) = r(a)$)

- Recall wish: $\boxed{\mathbf{E} \sup_{a \in \mathcal{A}} \{r(a) - R(a)\} \leq \frac{f(\text{dimension}, \text{complexity of } \mathcal{A})}{n^{\alpha}}}$
- The dimension of the data is superseded by the boundedness assumption
- $\alpha = 1/2$, slow rate
- When $|\mathcal{A}| < \infty$, $\log |\mathcal{A}|$ is a valid notion of complexity of the problem
- When $|\mathcal{A}| = \infty$, upper bound is trivial and we need another notion of complexity

Rademacher complexity

Rademacher complexity (Definition 2.5)

The Rademacher complexity of a set $\mathcal{T} \subseteq \mathbb{R}^n$ is defined as

$$\boxed{\mathtt{Rad}(\mathcal{T}) := \mathbf{E} \sup_{t \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^{n} \Omega_i t_i}$$

where $\Omega_1, \ldots, \Omega_n \in \{-1, 1\}$ are i.i.d. uniform random variables (Rademacher)

- Measures of complexity: describes how well elements in $\mathcal{T}$ can replicate the sign pattern of a uniform random signal in $\mathbb{R}^n$ (see **Problem 1.5**)
- Useful properties:
  - $\boxed{\mathtt{Rad}(c\mathcal{T} + v) = |c| \, \mathtt{Rad}(\mathcal{T})}$ (Proposition 2.6)
  - $\boxed{\mathtt{Rad}(\mathcal{T} + \mathcal{T}') = \mathtt{Rad}(\mathcal{T}) + \mathtt{Rad}(\mathcal{T}')}$ (Proposition 2.7)
  - $\boxed{\mathtt{Rad}(\mathrm{conv}(\mathcal{T})) = \mathtt{Rad}(\mathcal{T})}$ (Proposition 2.8)

    with $\mathrm{conv}(\mathcal{T}) = \{\sum_{j=1}^{m} w_j t_j : w \in \Delta_m, t_1, \ldots, t_m \in \mathcal{T}, m \in \mathbb{N}\}$

Rademacher complexity

Massart's Lemma (Lemma 2.9)

Let $\mathcal{T} \subseteq \mathbb{R}^n$ and let $v \in \mathbb{R}^n$ be any vector. We have

$$\texttt{Rad}(\mathcal{T}) \leq \max_{t \in \mathcal{T}} \|t - v\|_2 \frac{\sqrt{2 \log |\mathcal{T}|}}{n}$$

Proof: Similar to ones given above. **Problem 1.6**

Contraction property - Talagrand's Lemma (Lemma 2.10)

Let $\mathcal{T} \subseteq \mathbb{R}^n$. For each $i \in \{1, \dots, n\}$, let $f_i : \mathbb{R} \to \mathbb{R}$ be a γ-Lipschitz function. Then,

$$\texttt{Rad}((f_1, \dots, f_n) \circ \mathcal{T}) \leq \gamma \, \texttt{Rad}(\mathcal{T})$$

with $(f_1, \dots, f_n) \circ \mathcal{T} := \{(f_1(t_1), \dots, f_n(t_n)) \in \mathbb{R}^n : t \in \mathcal{T}\}$

Proof: **Problem 1.7**