Mathematics of Machine Learning - Summer School

Lecture 3
Bernstein's Concentration Inequalities. Fast Rates

June 29, 2021

Patrick Rebeschini
Department of Statistics, University of Oxford

Sub-Gaussian and Bernstein Random Variables

Sub-Guassian (Definition 6.5)

A random variable X is *sub-Gaussian* with *variance proxy* $\sigma^2 > 0$ if

$$\boxed{\mathbf{E}\, e^{\lambda(X-\mathbf{E}X)} \leq \exp(\sigma^2\lambda^2/2) \qquad \text{for any } \lambda \in \mathbb{R}}$$

- $\psi^\star(\varepsilon) = \varepsilon^2/(2\sigma^2)$
- **Bounded r.v.'s**: if $a \leq X - \mathbf{E}X \leq b$ then $\sigma^2 = \frac{(b-a)^2}{4}$ (Hoeffding's Lem. 2.1)

One-sided Bernstein's condition (Definition 7.1)

A random variable X satisfies the *one-sided Bernstein's condition* with $b > 0$ if

$$\boxed{\mathbf{E}\, e^{\lambda(X-\mathbf{E}X)} \leq \exp\left(\frac{(\mathbf{Var}\, X)\lambda^2/2}{1-b\lambda}\right) \qquad \text{for any } \lambda \in [0, 1/b)}$$

- $\psi^\star(\varepsilon) = \frac{\mathbf{Var}\, X}{b^2} h(\frac{b\varepsilon}{\mathbf{Var}\, X})$ with $h(u) = 1 + u - \sqrt{1+2u}$ for $u > 0$
- **Bounded above r.v.'s**: if $X - \mathbf{E}X \leq c$ then $b = c/3$ (Proposition 7.4)

Hoeffding's Inequality vs Bernstein's Inequality

Consider $X_1, \dots, X_n \sim X$ i.i.d. bounded in $[-c, c]$

- **Upper-tail bounds:**

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}X \geq \varepsilon\right) \leq e^{-n\varepsilon^2/(2c^2)} \qquad \text{(Hoeffding's)}$$

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}X \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2/2}{\mathbf{Var}X + c\varepsilon/3}\right) \qquad \text{(Bernstein's)}$$

- **Upper-confidence bounds:**

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}X < \sqrt{\frac{2c^2\log(1/\delta)}{n}}\right) \geq 1-\delta \qquad \text{(Hoeffding's)}$$

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}X < \frac{c}{3n}\log(1/\delta) + \sqrt{\frac{2(\mathbf{Var}X)\log(1/\delta)}{n}}\right) \geq 1-\delta \qquad \text{(Bernstein's)}$$

If $\mathbf{Var}X = 0$ then we get fast rate $\Rightarrow$ need to understand **noise** in learning

Back to Binary Classification

To understand main ideas to get fast rate, consider binary classification:

- $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \{-1, 1\}$
- Admissible action set $\mathcal{A} \subseteq \mathcal{B} := \{a : \mathbb{R}^d \to \{-1, 1\}\}$
- **True** loss function $\ell(a, (x, y)) = 1_{a(x) \neq y}$

$$r(a) = \mathbf{P}(a(X) \neq Y) \qquad a^\star \in \operatorname*{argmin}_{a \in \mathcal{A}} r(a) \qquad a^{\star\star} \in \operatorname*{argmin}_{a \in \mathcal{B}} r(a)$$

$$R(a) = \frac{1}{n} \sum_{i=1}^{n} 1_{a(X_i) \neq Y_i} \qquad A^\star \in \operatorname*{argmin}_{a \in \mathcal{A}} R(a)$$

The **Bayes decision rule** $a^{\star\star}$ reads

$$a^{\star\star}(x) \in \operatorname*{argmax}_{\hat{y} \in \mathcal{Y}} \mathbf{P}(Y = \hat{y} | X = x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ -1 & \text{if } \eta(x) \leq 1/2 \end{cases}$$

with the **unkown** **regression function** $\eta(x) := \mathbf{P}(Y = 1 | X = x)$
(η captures noise of unkown generative model)

Regression Function: Excess Risk and Bayes Risk

(Theorem 7.6)

For any $a \in \mathcal{B}$ $\boxed{r(a) - r(a^{\star\star}) = \mathbf{E}\big[|2\eta(X) - 1| 1_{a(X) \neq a^{\star\star}(X)}\big]}$

$$\boxed{r(a^{\star\star}) = \mathbf{E}\min\{\eta(X), 1 - \eta(X)\} \leq \frac{1}{2}}$$

$r(a^{\star\star}) = 1/2$ if and only if $\eta(X) = 1/2$ (Y contains no information on X)

- η **close to** $1/2$**:** large Bayes risk large; small excess risk
- η **away from** $1/2$**:** small Bayes risk large; large excess risk

Fast Rate: Massart's Condition

Massart's Noise Condition (Definition 7.7)

There exists $\gamma \in (0, 1/2]$ such that

$$\boxed{\mathbf{P}\left(\left|\eta(X) - \frac{1}{2}\right| \geq \gamma\right) = 1}$$

($\gamma = 0$ would mean condition is void)

Fast Rate in Binary Classification (Theorem 7.10)

Let $a^{\star\star} \in \mathcal{A}$ so that $a^\star = a^{\star\star}$. If Massart's condition holds with $\gamma \in (0, 1/2]$,

$$\boxed{\mathbf{P}\left(r(A^\star) - r(a^\star) \leq \frac{\log(|\mathcal{A}|/\delta)}{\gamma n}\right) \geq 1 - \delta}$$

Fast rate if $|\mathcal{A}| < \infty$

- Massart's condition is strong: η uniformly bounded away from $1/2$
- Weaker conditions: η arbitrarily close to $1/2$, but with small probability

Proof of Theorem 7.6 (Part I)

- Error decomposition: $r(A^\star) - r(a^\star) \leq R(a^\star) - R(A^\star) - (r(a^\star) - r(A^\star))$

$$G(a) := R(a^\star) - R(a) - (r(a^\star) - r(a)) = R(a^\star) - R(a) - \mathbf{E}[R(a^\star) - R(a)]$$
$$= \frac{1}{n}\sum_{i=1}^{n}(g(a, Z_i) - \mathbf{E}g(a, Z_i))$$

with $g(a, z) = 1_{a^\star(x)\neq y} - 1_{a(x)\neq y}$

- The above yields $r(A^\star) - r(a^\star) \leq G(A^\star)$
- Bernstein's inequality for bounded random variables yields, for any $a \in \mathcal{A}$,

$$\mathbf{P}(G(a) \geq \varepsilon) \leq \exp\left(-\frac{n\mathbf{Var}\,g(a,Z)}{b^2} h\left(\frac{b\varepsilon}{\mathbf{Var}\,g(a,Z)}\right)\right)$$

- Setting the right-hand side to $\delta/|\mathcal{A}|$, using that $h^{-1}(u) = u + \sqrt{2u}$ for $u > 0$

$$\mathbf{P}\left(G(A^\star) < \frac{b}{n}\log(|\mathcal{A}|/\delta) + \sqrt{\frac{2(\mathbf{Var}\,g(A^\star, Z))\log(|\mathcal{A}|/\delta)}{n}}\right)$$
$$\geq \mathbf{P}\left(\bigcap_{a\in\mathcal{A}}\left\{G(a) < \frac{b}{n}\log(|\mathcal{A}|/\delta) + \sqrt{\frac{2(\mathbf{Var}\,g(a, Z))\log(|\mathcal{A}|/\delta)}{n}}\right\}\right) \geq 1-\delta$$

Proof of Theorem 7.6 (Part II)

- As for any $a \in \mathcal{A}$ we have $|g(a,Z)| = 1_{a(X) \neq a^\star(X)}$, then

$$\mathbf{Var}\, g(a,Z) \leq \mathbf{E}[g(a,Z)^2] = \mathbf{P}(a(X) \neq a^\star(X))$$

  and from Theorem 7.6 and Massart's noise condition we have

$$r(a) - r(a^\star) = \mathbf{E}[|2\eta(X) - 1| 1_{a(X) \neq a^\star(X)}] \geq 2\gamma \mathbf{P}[a(X) \neq a^\star(X)],$$

  which yields $\mathbf{Var}\, g(a,Z) \leq \frac{1}{2\gamma}(r(a) - r(a^\star))$

- Using that $r(A^\star) - r(a^\star) \leq G(A^\star)$, we can conclude

$$\mathbf{P}\left(r(A^\star) - r(a^\star) < \frac{2}{3n} \log(|\mathcal{A}|/\delta) + \sqrt{\frac{(r(A^\star) - r(a^\star)) \log(|\mathcal{A}|/\delta)}{\gamma n}} \right) \geq 1 - \delta.$$

  The proof follows by solving the expression in the event with respect to the excess risk $r(A^\star) - r(a^\star)$, using that $x < 2\alpha/3 + \sqrt{x\alpha/\gamma}$ for $x \in [0,1]$, with $\alpha > 0$ and $\gamma \in (0, 1/2]$, implies $x < \alpha/\gamma$.

Interpolation Slow and Fast Rate: Tsybakov's Condition

Tsybakov's Noise Condition (Definition 7.11)

There exist $\alpha \in (0, 1)$, $\beta > 0$, and $\gamma \in (0, 1/2]$ such that, for all $t \in [0, \gamma]$,

$$\boxed{\mathbf{P}\left(\left|\eta(X) - \frac{1}{2}\right| \leq t\right) \leq \beta t^{\alpha/(1-\alpha)}}$$

Interpolation Slow and Fast Rate in Binary Classification (Theorem 7.13)

Let $a^{\star\star} \in \mathcal{A}$. If Tsybakov's condition holds for $\alpha \in (0, 1)$, $\beta > 0$, $\gamma \in (0, 1/2]$,

$$\boxed{\mathbf{P}\left(r(A^\star) - r(a^\star) \leq c\left(\frac{\log(|\mathcal{A}|/\delta)}{n}\right)^{\frac{1}{2-\alpha}}\right) \geq 1 - \delta}$$

for a given constant c that depends on α, β, γ.

- if $\alpha \to 0$ then we recover slow rate (condition becomes void)
- if $\alpha \to 1$ then we recover fast rate (condition recovers Massart's)

Note: $A^\star$ does **not** depend on α: <u>it automatically adjusts to the noise level!</u>