Mathematical Machine Learning Theory, M1407.002700
E. Ryu
Spring 2024

Homework 2
Due 5pm, Wednesday, March 27, 2024

Problem 1: *Variance of bounded RVs.* Let $X \in [a,b]$ with $a < b$ be a random variable. Show that

$$\mathrm{Var}(X) \le \frac{(b-a)^2}{4}.$$

Hint. Show that

$$\mathrm{Var}(X) \le \mathbb{E}\big[(X - \tfrac{b+a}{2})^2\big].$$

Problem 2: *Sample complexity with Hoeffding.* Let $X_1, \dots, X_N \in [a,b]$ be IID random variables with mean $\mu \in \mathbb{R}$. Let $\bar{X} = \frac{1}{N}\sum_{i=1}^N X_i$. Show that

$$N \ge \frac{(b-a)^2}{2\varepsilon^2}\log(2/\delta) \quad \Rightarrow \quad \mathbb{P}(|\bar{X} - \mu| < \varepsilon) \ge 1 - \delta,$$

for all $\varepsilon > 0$ and $\delta > 0$.

Problem 3: *Sample complexity with Bernstein.* Let $X_1, \dots, X_N \in [a,b]$ be IID random variables with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}$. Let $\bar{X} = \frac{1}{N}\sum_{i=1}^N X_i$. Show that

$$N \ge \left(\frac{2\sigma^2}{\varepsilon^2} + \frac{2(b-a)}{3\varepsilon}\right)\log(2/\delta) \quad \Rightarrow \quad \mathbb{P}(|\bar{X} - \mu| < \varepsilon) \ge 1 - \delta,$$

for all $\varepsilon > 0$ and $\delta > 0$.

Problem 4: *Strictly convex losses admit unique Bayes optimal predictors.* Let $C \subseteq \mathbb{R}^d$ be a nonempty convex set. We say a function $f\colon C \to \mathbb{R}$ is *strictly convex* if

$$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta)f(y), \qquad \forall\, x \ne y \in C,\ \theta \in (0,1).$$

Assume $\tilde{\mathcal{Y}}$ is nonempty convex and $\ell(y', y)$ is *strictly* convex in $y' \in \tilde{\mathcal{Y}}$ for all $y \in \mathcal{Y}$. Then,

$$f^\star(X) \in \underset{y' \in \tilde{\mathcal{Y}}}{\operatorname{argmin}} \underset{Y \sim P_{Y|X}}{\mathbb{E}} [\ell(y', Y) \,|\, X]$$

is unique (up to a P-measure 0 set), if it exists. In other words, show that the set $\operatorname{argmin}_{y' \in \tilde{\mathcal{Y}}}\{\cdots\}$ has exactly 0 or 1 elements.

Remark. Existence of the Bayes optimal predictor should not be taken for granted. Simple settings such as (unregularized) logistic regression with separable data may fail to have a Bayes optimal predictor. We will return to this in the future.

Problem 5: *Estimation error decomposition without minimizer.* Let $\mathcal{R}$ be the true risk, and assume $|\mathcal{R}[f]| < \infty$ for all $f \in \mathcal{F}$. Likewise, let $\hat{\mathcal{R}}$ be the empirical risk, and assume $|\hat{\mathcal{R}}[f]| < \infty$ for all $f \in \mathcal{F}$. Assume

$$\inf_{f' \in \mathcal{F}} \mathcal{R}[f'] > -\infty,$$

but do not assume $\operatorname{argmin}_{f' \in \mathcal{F}} \mathcal{R}[f']$ exists. Show the following bound on the estimation error:

$$\mathcal{R}[\hat{f}] - \inf_{f' \in \mathcal{F}} \mathcal{R}[f'] \leq \sup_{f \in \mathcal{F}} \{\mathcal{R}[f] - \hat{\mathcal{R}}[f]\} + \sup_{f \in \mathcal{F}} \{\hat{\mathcal{R}}[f] - \mathcal{R}[f]\} + (\hat{\mathcal{R}}[\hat{f}] - \inf_{f \in \mathcal{F}} \hat{\mathcal{R}}[f]).$$

Problem 6: *Computation and data complexity for PAC guarantee with covering number.* Assume $\ell(\cdot, Y)$ is G-Lipschitz for all $Y \sim P_Y$ and $0 \leq \ell(f(X), Y) \leq \ell_\infty$ for all $f \in \mathcal{F}$ and $(X, Y) \sim P$. Assume the function class $\mathcal{F}$ has an covering number $m(\varepsilon) \leq C_{\text{cov}}/\varepsilon^d$ for some $C_{\text{cov}} > 0$. Assume we have access to IID training data $\mathcal{D} = (X_1, Y_1), \ldots, (X_N, Y_N) \sim P$ with $N \geq 1$. Consider a machine learning algorithm that uses the N data points in $\mathcal{D}$ and K amount of computational cost (number of floating point operations) to compute $\hat{f} \in \mathcal{F}$ such that

$$\hat{\mathcal{R}}[\hat{f}] - \inf_{f \in \mathcal{F}} \hat{\mathcal{R}}[f] \leq C_{\text{opt}} \sqrt{\frac{N}{K}}.$$

for some $C_{\text{opt}} > 0$. Let $\eta \in (0, 1/2)$ and $\varepsilon > 0$.

(a) Show that if

$$N^{2\eta} \geq \frac{1}{4} + \frac{1}{d} \log C_{\text{cov}} + \frac{1}{2} \log N,$$

then

$$\mathcal{R}[\hat{f}] - \inf_{f' \in \mathcal{F}} \mathcal{R}[f'] \leq \frac{4G + \sqrt{8\ell_\infty^2}\left(\sqrt{d} + \sqrt{\log(2/\delta)}\right)}{N^{1/2-\eta}} + C_{\text{opt}} \sqrt{\frac{N}{K}}$$

with probability $> 1 - \delta$.

(b) Show that if

$$\left(\frac{8G + \sqrt{32\ell_\infty^2}\left(\sqrt{d} + \sqrt{\log(2/\delta)}\right)}{\epsilon}\right)^{\frac{2}{1-2\eta}} \leq N, \qquad \frac{4C_{\text{opt}}^2 N}{\epsilon^2} \leq K,$$

holds, then

$$\mathcal{R}[\hat{f}] - \inf_{f' \in \mathcal{F}} \mathcal{R}[f'] \leq \varepsilon \qquad \text{with probability} > 1 - \delta.$$

Problem 7: *Basic properties of Rademacher complexity.* Show the following.

(a) $\mathcal{H} \subset \mathcal{H}'$, then $\text{Rad}_N(\mathcal{H}) \leq \text{Rad}_N(\mathcal{H}')$

(b) $\text{Rad}_N(\mathcal{H} + \mathcal{H}') \leq \text{Rad}_N(\mathcal{H}) + \text{Rad}_N(\mathcal{H}')$

(c) $\text{Rad}_N(\alpha\mathcal{H}) \leq |\alpha| \text{Rad}_N(\mathcal{H})$

(d) $\text{Rad}_N(\mathcal{H}) = \text{Rad}_N(\text{conv}(\mathcal{H}))$

Problem 8: *Computation and data complexity for PAC guarantee with Rademacher complexity.* Assume $\ell(\cdot, Y)$ is G-Lipschitz for all $Y \sim P_Y$ and $0 \le \ell(f(X), Y) \le \ell_\infty$ for all $f \in \mathcal{F}$ and $(X, Y) \sim P$. Let $\phi\colon \mathcal{X} \to \mathbb{R}^d$ be a given feature function such that $\|\phi(X)\|_2 \le R$ (P-almost surely) for all X. Let

$$\mathcal{F} = \left\{ f_\theta(x) = \theta^\intercal \phi(X) \,\middle|\, \|\theta\|_2 \le D,\, \theta \in \mathbb{R}^d \right\}$$

for some D such that $0 < D < \infty$. Assume we have access to IID training data $\mathcal{D} = (X_1, Y_1), \ldots, (X_N, Y_N) \sim P$ with $N \ge 1$. Consider a machine learning algorithm that uses the N data points in $\mathcal{D}$ and K amount of computational cost (number of floating point operations) to compute $\hat{f} \in \mathcal{F}$ such that

$$\hat{\mathcal{R}}[\hat{f}] - \inf_{f \in \mathcal{F}} \hat{\mathcal{R}}[f] \le C_{\text{opt}} \sqrt{\frac{N}{K}}.$$

for some $C_{\text{opt}} > 0$. Let $\eta \in (0, 1/2)$ and $\varepsilon > 0$.

(a) Show that

$$\mathcal{R}[\hat{f}] - \inf_{f' \in \mathcal{F}} \mathcal{R}[f'] \le \frac{4DGR + \ell_\infty \sqrt{2\log(2/\delta)}}{\sqrt{N}} + C_{\text{opt}} \sqrt{\frac{N}{K}}$$

with probability $> 1 - \delta$.

(b) Show that if

$$K \ge \frac{C_{\text{opt}}^2 N^2}{(4DGR + \ell_\infty \sqrt{2\log(2/\delta)})^2}, \qquad N \ge \frac{(8DGR + \ell_\infty \sqrt{8\log(2/\delta)})^2}{\varepsilon^2}$$

furthermore holds, then

$$\mathcal{R}[\hat{f}] - \inf_{f' \in \mathcal{F}} \mathcal{R}[f'] \le \varepsilon \qquad \text{with probability } > 1 - \delta.$$

Problem 9: *Linear algebra review for pseudo-inverses.* Let $v_1, \ldots, v_r \in \mathbb{R}^d$ be an orthonormal set of vectors. Let

$$V = \begin{bmatrix} v_1 & \cdots & v_r \end{bmatrix} \in \mathbb{R}^{d \times r}.$$

Show the following.

(a) $V^\intercal V = I$.

(b) $VV^\intercal \theta = \theta$ if and only if $\theta \in \mathcal{R}(V)$.

Problem 10: *Pseudo-inverses for full-rank matrices.* Let $A \in \mathbb{R}^{N \times d}$, and let $A^\dagger$ denote the pseudo-inverse. Show the following.

- If A has full column rank (which requires that $N \ge d$), then $A^\dagger = (A^\intercal A)^{-1} A^\intercal$.
- If A has full row rank (which requires that $N \le d$), then $A^\dagger = A^\intercal (AA^\intercal)^{-1}$.