18.657. Fall 2105
Rigollet October 18, 2015

Problem set #2 (due Wed., October 21)

SHOULD BE TYPED IN LaTeX

Problem 1. Rademacher Complexities and beyond

Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$ and let $X_1, \ldots, X_n$ be iid copies of a random variable $X \in \mathcal{X}$. Moreover, let $\sigma_1, \ldots, \sigma_n$ be n i.i.d. $\mathsf{Rad}(1/2)$ random variables and let $g_1, \ldots, g_n$ be n i.i.d. $N(0,1)$. Assume that all these random variables are mutually independent.

1. Prove the *desymmetrization inequality*:

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i\big[f(X_i) - \mathbb{E}[f(X)]\big]\Big|\Big] \le 2\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \big[f(X_i) - \mathbb{E}[f(X)]\big]\Big|\Big]$$

2. Prove the Rademacher/Gaussian process comparison inequality

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \sigma_i f(X_i)\Big] \le \sqrt{\frac{\pi}{2}}\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\sum_{i=1}^n g_i f(X_i)\Big]$$

Define $R_n(\mathcal{F}) = \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\frac{1}{n}\Big|\sum_{i=1}^n \sigma_i f(X_i)\Big|\Big]$. Let $\mathcal{F}$ and $\mathcal{G}$ be two set of functions from $\mathcal{X}$ to $\mathbb{R}$ and recall that $\mathcal{F}+\mathcal{G} = \{f+g \,:\, f\in\mathcal{F}, g\in\mathcal{G}\}$.

3. Let $h \in \mathbb{R}^{\mathcal{X}}$ be a given function and define $\mathcal{F}+h = \{f+h \,:\, f\in\mathcal{F}\}$. Show that

$$R_n(\mathcal{F}+\{h\}) \le R_n(\mathcal{F}) + \frac{\|h\|_\infty}{\sqrt{n}}\,,$$

where $\|h\|_\infty = \sup_{x\in\mathcal{X}}|h(x)|$.

4. Let $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be k sets of functions from $\mathcal{X}$ to $\mathbb{R}$. Show that

$$R_n(\mathcal{F}_1 + \cdots, \mathcal{F}_k) \le \sum_{j=1}^k R_n(\mathcal{F}_j)\,.$$

5. Show that this inequality derived in 4. is in fact an equality when the $\mathcal{F}_j$s are the same.

Problem 2. Covering and packing

Definition: A set $P \subset T$ is called an *ε-packing* of the metric space (T, d) if $d(f, g) > \varepsilon$ for every $f, g \in P$, $f \neq g$. The largest cardinality of an ε-packing of (T, d) is called the *packing number* of (T, d):

$$D(T, d, \varepsilon) = \sup \big\{ \operatorname{card}(P) \,:\, P \text{ is an } \varepsilon \text{ packing of } (T, d) \big\}$$

Recall that $N(T, d, \varepsilon)$ denotes the ε-covering number of (T, d).

1. Show that
$$D(T, d, 2\varepsilon) \leq N(T, d, \varepsilon) \leq D(T, d, \varepsilon)$$

Let M be an $n \times m$ random matrix with entries that are i.i.d $\mathsf{Rad}(1/2)$ entries. We are interested in its operator norm

$$\|M\| = \sup_{\substack{u \in \mathbb{R}^n \,:\, |u|_2 \leq 1 \\ v \in \mathbb{R}^m \,:\, |v|_2 \leq 1}} u^\top M v \,.$$

2. Show that
$$\|M\| \leq 2 \max_{\substack{u \in N_n \\ v \in N_m}} u^\top M v \,,$$
where N_n and N_m are $\frac{1}{4}$-nets of the unit Euclidean balls of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively.

3. Conclude that
$$\mathbb{E}\|M\| \leq C\big(\sqrt{m} + \sqrt{n}\big) \,.$$

Problem 3. Chaining

Let $\mathcal{F}$ be the class of all *nondecreasing functions* from $[0, 1]$ to $[0, 1]$.

1. Show that for any $x = (x_1, \ldots, x_n) \in [0, 1]^n$, the covering number of $(\mathcal{F}, d_\infty^x)$ satisfy:
$$N(\mathcal{F}, d_\infty^x, \varepsilon) \leq n^{2/\varepsilon} \,.$$

2. Using the chaining bound, show that
$$\mathcal{R}_n(\mathcal{F}) \leq C\sqrt{\frac{\log n}{n}}$$

3. Show that there is indeed a strict improvement over the bound obtained using the theorem in section 5.2.1

Problem 4. Kernel ridge regression

Consider the regression model:

$$Y_i = f(x_i) + \xi_i, \quad , i = 1, \dots, n$$

where $x_1, \dots, x_n$ are fixed design points in $\mathbb{R}^d$, $\xi = (\xi_1, \dots, \xi_n) \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^n$ with known covariance matrix $\Sigma \succ 0$ and $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown regression function.

Let W be an RKHS on $\mathbb{R}^d$ with reproducing kernel k. Define $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ and $\mathtt{g} = [g(x_1), \dots, g(x_n)]^\top$ for any function g. Define the estimator $\hat{f}$ of f by

$$\hat{f} = \operatorname*{argmin}_{g \in W} \left\{ \psi(\mathbf{Y} - \mathtt{g}) + \mu \|g\|_W^2 \right\}$$

where $\|\cdot\|_W$ denotes the Hilbert norm on W, $\psi(\mathbf{x}) = \mathbf{x}^\top \Sigma^{-1/2} \mathbf{x}$ and $\mu > 0$ is a tuning parameter to be chosen later.

1. Prove the representer theorem, i.e., that there exists a vector $\theta \in \mathbb{R}^n$ such that

$$\hat{f}(x) = \sum_{i=1}^n \theta_i k(x_i, x), \qquad \text{for any } x \in \mathbb{R}^d$$

2. Prove that the vector $\hat{\mathtt{f}} = [\hat{f}(x_1), \dots, \hat{f}(x_n)]^\top$ satisfies

$$(K\Sigma^{-1/2} + \mu I_n)\hat{\mathtt{f}} = K\Sigma^{-1/2}\mathbf{Y},$$

where I_n is the identity matrix of $\mathbb{R}^n$ and K denotes the symmetric $n \times n$ matrix with elements $K_{i,j} = k(x_i, x_j)$.

3. Prove that the following inequality holds

$$\psi(\mathtt{f} - \hat{\mathtt{f}}) \le \inf_{g \in W} \left\{ \psi(\mathtt{f} - \mathtt{g}) + 2\mu \|g\|_W^2 \right\} + \frac{1}{\mu} \Big\| \sum_{i=1}^n Z_i k(x_i, \cdot) \Big\|_W^2,$$

where $Z_1, \dots, Z_n$ are iid $\mathcal{N}(0, 1)$.

4. Conclude that

$$\mathbb{E}\psi(\mathtt{f} - \hat{\mathtt{f}}) \le \inf_{g \in W} \left\{ \psi(\mathtt{f} - \mathtt{g}) + 2\mu \|g\|_W^2 \right\} + \frac{1}{\mu} \mathbf{Tr}(K),$$

where $\mathbf{Tr}(K)$ denotes the trace of K.

5. Assume now that k is the Gaussian kernel:

$$k(x, x') = e^{-|x-x'|_2^2}$$

Show that there exists a choice of μ for which

$$\mathbb{E}\psi(\mathtt{f} - \hat{\mathtt{f}}) \leq 2\|f\|_W \sqrt{2n}\,.$$

18.657 Mathematics of Machine Learning
Fall 2015