36-755, Fall 2017
Homework 6

Due Mon Dec 4 by 5:00pm in Jisu's mailbox

1. Let $\mathcal{F}$ be a collection of functions from $\mathbb{R}^d$ into $[0, b]$, for some $b > 0$. For each $\delta > 0$, let $N_\infty(\delta, \mathcal{F})$ denote the δ-covering number of $\mathcal{F}$ in the d_∞ distance given by

$$d_\infty(f, g) = \sup_{x \in \mathbb{R}^d} |f(x) - g(x)|, \quad f, g \in \mathcal{F}.$$

Let $(X_1, \ldots, X_n)$ be an i.i.d. sample from some distribution P on $\mathbb{R}^d$ and P_n be the associated empirical measure. Show that

$$\mathbb{P}\left(\|P_n - P\|_{\mathcal{F}} > \epsilon\right) \leq 2N_\infty(\epsilon/3, \mathcal{F}) e^{-\frac{2n\epsilon^2}{9b^2}} \quad \epsilon > 0.$$

Hint: for any $\epsilon > 0$, consider a minimal $\epsilon/3$ covering of $\mathcal{F}$. Then, for each $f \in \mathcal{F}$, there exists a function $\overline{f}$ in the cover (which one depends on f) such that $d_\infty(f, \overline{f}) \leq \epsilon/3$. Run with it...

2. **Reading Assignment.**

Reproduce the proof of Theorem 2.1 in the following paper, which provides dimension-free performance of k-means in Hilbert spaces.

Biau, G., Devroye, L. and Lugosi, G. (2008). On the Performance of Clustering in Hilbert Spaces, IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 54, NO. 2, 781–790.

You may assume that $\mathcal{H} = \mathbb{R}^d$

3. Recall the relative VC bounds: for a class $\mathcal{A}$ of sets in $\mathbb{R}^d$ and an i.i.d. sample $(X_1, \ldots, X_n)$ from a probability distribution P,

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} \frac{P(A) - P_n(A)}{\sqrt{P(A)}} > \epsilon\right) \leq 4S_{\mathcal{A}}(2n) e^{-n\epsilon^2/4}, \quad \epsilon > 0,$$

and

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} \frac{P_n(A) - P(A)}{\sqrt{P_n(A)}} > \epsilon\right) \leq 4S_{\mathcal{A}}(2n) e^{-n\epsilon^2/4}, \quad \epsilon > 0,$$

where $S_{\mathcal{A}}(n)$ is the n-shattering coefficient of $\mathcal{A}$, i.e.

$$\max_{x_1^n} |\mathcal{A}(x_1^n)| = \max_{x_1^n} |x_1^n \cap A, A \in \mathcal{A}|$$

where x_1^n denotes an n-tuple of points in $\mathbb{R}^d$. See, e.g.,

- Vapnik, V., Chervonenkis, A.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications 16 (1971) 264280.
- M. Anthony and J. Shawe-Taylor, "A result of Vapnik with applica- tions," Discrete Applied Mathematics, vol. 47, pp. 207-217, 1993.

(a) Show that

$$\mathbb{P}\left(\exists A \in \mathcal{A} \colon P(A) > \epsilon \text{ and } P_n(A) \leq (1-t)P(A)\right) \leq 4S_{\mathcal{A}}(2n) e^{-n\epsilon t^2/4},$$

for all $t \in (0, 1]$ and $\epsilon > 0$. What do you obtain when $t = 1$?

(b) Show that, uniformly over all the sets $A \in \mathcal{A}$,

$$P(A) \le P_n(A) + 2\sqrt{P_n(A)\frac{\log S_{\mathcal{A}}(2n) + \log\frac{4}{\delta}}{n}} + 4\frac{\log S_{\mathcal{A}}(2n) + \log\frac{4}{\delta}}{n},$$

with probability at least $1-\delta$.

(c) Let B be a closed ball in $\mathbb{R}^d$ (of arbitrary center and radius). Let k be a positive integer. Then $P_n(B) > \frac{k}{n}$ if and only if B contains more than k sample points. Show that, for any $\delta \in (0,1)$ and with $k \ge C' d \log n$ for some $C' > 0$,there exists a constant C_δ (depending on δ and C') such that, with probability at least $1-\delta$, every ball B satisfies the following conditions:

i. if $P(B) > C_\delta \frac{d\log n}{n}$, then $P_n(B) > 0$;
ii. if $P(B) \ge \frac{k}{n} + \frac{C_\delta}{n}\sqrt{kd\log n}$, then $P_n(B) \ge \frac{k}{n}$;
iii. if $P(B) \le \frac{k}{n} - \frac{C_\delta}{n}\sqrt{kd\log n}$, then $P_n(B) \le \frac{k}{n}$;

Hint: use the fact that the VC dimension of the class of all closed Euclidean balls in $\mathbb{R}^d$ is $d+1$.

Use the previous inequalities to reproduce the proof of Theorem 1 in *Kamalika Chaudhuri, Sanjoy Dasgupta, Samory Kpotufe, Ulrike von Luxburg: Consistent Procedures for Cluster Tree Estimation and Pruning. IEEE Trans. Information Theory 60(12): 7900-7912 (2014)*

4. **Another symmetrization inequality**

Let $\mathcal{A}$ be a countable class of sets in $\mathbb{R}^d$ and $X = (X_1,\dots,X_n) \overset{i.i.d.}{\sim} P$. Let P_n^X be the empirical measure corresponding to the sample X. Let $Y = (Y_1,\dots,Y_n) \overset{i.i.d.}{\sim} P$ be a ghost sample, independent of Y, and P_n^Y the corresponding empirical measure. Prove that

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n^X(A) - P(A)| > \epsilon\right) \le 2\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n^X(A) - P_n^Y(A)| > \epsilon/2\right),$$

for $n\epsilon^2 \ge 2$.

Proceed as follows:

(a) Show that, if $n\epsilon^2 \ge 2$, for all $A \in \mathcal{A}$,

$$\mathbb{P}\left(|P_n^X(A) - P(A)| > \epsilon/2\right) \le 1/2.$$

(b) Prove the following claim: let $(Z_k, k = 1,2,\dots)$ be a sequence (finite or infinite) of random variables and $(Z'_k, k = 1,2,\dots)$ be a ghost sequence with the same distribution and independent of it. Suppose that $\mathbb{P}(|Z_k| > \epsilon/2) = \mathbb{P}(|Z'_k| > \epsilon/2) \le 1/2$ for all k. Then

$$\mathbb{P}\left(\sup_k |Z_k| > \epsilon\right) \le 2\mathbb{P}\left(\sup_k |Z_k - Z'_k| > \epsilon/2\right).$$

Hint: Define the events: $A_1 = \{|Z_i| > \epsilon\}$ and for $k \ge 2$,

$$A_k = \{|Z_1| \le \epsilon, \dots, |Z_{k-1}| \le \epsilon, |Z_k| > \epsilon\}.$$

Then,

$$\frac{1}{2}\mathbb{P}\left(\max_k |Z_k| > \epsilon\right) = \frac{1}{2}\sum_k \mathbb{P}(A_k).$$

Proceed...

5. Exercise 4.10.

6. **When is the sample an ϵ cover of the support?** Suppose that $X = (X_1, \ldots, X_n)$ is an i.i.d. sample from a probability distribution supported on $\mathcal{S}$, assumed to be a compact subset of $\mathbb{R}^d$ with non-empty interior (this means that $\mathcal{S}$ is the smallest closed and bounded subset of $\mathbb{R}^d$ of dimension d such that $P(\mathcal{S}) = 1$). In many problems in geometric and topological data analysis, it is often desirable that X be an ϵ-cover of $\mathcal{S}$, which is equivalent to

$$\mathcal{S} \subset \bigcup_{i=1}^{n} B(X_i, \epsilon), \tag{1}$$

where $B(x, \epsilon)$ is the closed Euclidean ball centered at x and of radius ϵ. Assume that there exists a $a > 0$ such that

$$\inf_{x \in \mathcal{S}} P(B(x,r)) \geq \min\left\{1, \frac{r^d}{a}\right\}, \quad \forall r > 0.$$

The above requirement is known as the *standard condition* and amounts to assuming (i) that P has a Lebesgue density bounded away from 0 over its support and (ii) that $\mathcal{S}$ does not get arbitrarily narrow or exhibit cusp-like shapes protruding outwards.

   (a) For a given ϵ, find a lower bound on n such that, with high probability, X is an ϵ-cover of $\mathcal{S}$.

   (b) The union of balls of radius ϵ centered at the sample points is an estimator of $\mathcal{S}$, known as the Devroye-Wise estimator. The Devroy-Wise estimator of $\mathcal{S}$ is consistent when ϵ can be chosen as a function of n, written as ϵ_n, in such a way that $\epsilon_n \to 0$ and (1) holds with probability tending to 1 as $n \to \infty$,. Find a scaling for ϵ_n that satisfies both conditions.

   Hint: Take a look at this paper: Antonio Cuevas and Ricardo Fraiman. A plug-in approach to support estimation. Ann. Statist., 25(6):23002312, 1997.

7. Exercise 5.11.