36-755, Fall 2016
Homework 6

Due Wed Nov 9 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 $A \in \mathcal{A}$,

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

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 $k \geq C' d \log n$ 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 $P(B) \geq \frac{k}{n} + \frac{C_\delta}{n}\sqrt{kd\log n}$, then $P_n(B) \geq \frac{k}{n}$;

iii. if $P(B) \leq \frac{k}{n} - \frac{C_\delta}{n}\sqrt{kd\log n}$, then $P_n(B) \leq \frac{k}{n}$;

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

This result is used to prove consistency of density based clustering in the paper
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. **More on using relative deviations.** Let $X_1, \ldots, X_n$ be an i.i.d. sample from P, a probability distribution on $\mathbb{R}^d$. For a given $h > 0$ consider the following estimator for the Lebesgue density of P:

$$\hat{f}_h(x) = \frac{1}{n}\frac{1}{h^d V_d}\sum_{i=1}^{n} 1\left(\|x - X_i\| \leq h\right), \quad x \in \mathbb{R}^d,$$

where V_d is the volume of the Euclidean unit ball in $\mathbb{R}^d$. Consider the function

$$f_h(x) = \mathbb{E}\left[\hat{f}_h(x)\right], \quad x \in \mathbb{R}^d.$$

(a) Show that f_h is a density (i.e. it is non-negative and integrate to 1). *Hint: you may use the fact that the volume of any Euclidean ball of radius* h *in* $\mathbb{R}^d$ *is* $h^d V_d$.

(b) We are interested in finding out how much f_h and $\hat{f}_h$ differ in the L_∞ norm. Use the relative deviation bounds to compute an upper bound for the probability

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}^d} |f_h(x) - \hat{f}_h(x)| \geq \epsilon\right),$$

where $\epsilon > 0$. *Hint: use again the fact that the VC dimension of the class of all closed Euclidean balls in* $\mathbb{R}^d$ *is* $d+1$.

5. Exercise 5.11.