ECE 543 STATISTICAL LEARNING THEORY **SPRING 2019**
PROBLEM SET 2 **Due: Upload to Compass by beginning of class, Tuesday, February 19**

2. Expected Risk Minimization and Abstract Tools for Uniform Approximation

Assigned reading: Chapters 5-8.1 of the course notes.
Additional recommended reading: Shalev-Shwartz and Ben-David, *Understanding Machine Learning, from Theory to Algorithms,* Chapters 3-6, 26 and Appendix B.

(Problems and solutions)

1. **[Empirical cumulative distribution functions on the reals, revisited]**
Let F be a cumulative distribution function (CDF) on $\mathbb{R}$ and let $X_1, \ldots, X_n$ be n independent random variables with distribution F. Denote the empirical distribution function by $F_n(c) = \frac{1}{n}\sum_{i\in[n]} \mathbf{1}_{\{X_i \le c\}}$ for $c \in \mathbb{R}$ and let $\Delta_n = \sup_{c\in\mathbb{R}} |F_n(c) - F(c)|$.

   (a) Find two upper bounds on $E[\Delta_n]$ that follow from the general tools based on Rademacher averages and VC dimension in the course notes. Begin by explaining how the problem fits the framework presented in Section 6.1 of the notes.
   Solution: Let $\mathcal{F} = \{f_c : c \in \mathbb{R}\}$, where $f_c(x) = \mathbf{1}_{\{x\le c\}}$ for $c \in \mathbb{R}$. Then for any CDF F, if P is the probability distribution on $\mathbb{R}$ with CDF F (in other words, $P((-\infty, c]) = F(c)$ for all $c \in \mathbb{R}$, then $F_n(c) = \mathsf{P}_n(f_c)$ and $F(c) = P(f_c)$. Therefore, $\Delta_n = \sup_{f\in\mathcal{F}} |\mathsf{P}_n(f) - P(f)|$, so the abstract framework introduced in Section 6.1 applies. The VC dimension of $\mathcal{F}$ is $V(\mathcal{F}) = 1$, because any singleton set $\{c\}$ is shattered by $\{f_{c-1}, f_{c+1}\}$ and no two point set $\{c, c'\}$ is shattered because all the functions in $\mathcal{F}$ are nondecreasing. Therefore,

$$\mathbb{E}[\Delta_n] \overset{(a)}{\le} 2\mathbb{E}[R_n\mathcal{F}(X^n)] \overset{(b)}{\le} 4\sqrt{\frac{\log(n+1)}{n}}$$
$$\mathbb{E}[\Delta_n] \le 2\mathbb{E}[R_n(\mathcal{F}(X^n))] \overset{(c)}{\le} C'\sqrt{\frac{1}{n}}$$

   for some universal constant C', where (a) follows from Theorem 6.1 (based on the symmetrization technique), (b) follows from Theorem 7.1 and (c) follows from Theorem 7.2.

   (b) Using your answer to (a) and the McDiarmid inequality, find an upper bound on the upper tail of the distribution of Δ_n. For example, you could find an upper bound on probabilities of the form $\mathbb{P}\left\{\Delta_n \ge g(n) + \frac{t}{\sqrt{n}}\right\}$ for positive values of t, where g is some function of n of your choice. (You may apply results from Problem 1, Problem Set 1 for this part.)
   Solution: Combining with Problem 1 of Problem Set 1 yields

$$\mathbb{P}\left\{\Delta_n \ge 4\sqrt{\frac{\log(n+1)}{n}} + \frac{t}{\sqrt{n}}\right\} \le \mathrm{e}^{-2t^2}$$
$$\mathbb{P}\left\{\Delta_n \ge \frac{C'+t}{\sqrt{n}}\right\} \le \mathrm{e}^{-2t^2}$$

2. **[Is the Rademacher average a bound on the maximum deviation with probability one?]**
See the notes for the context of Theorem 6.1, which states: "Let $\mathcal{F}$ be a class of functions $f : \mathsf{Z} \to [0, 1]$. Then for any $P \in \mathcal{P}(\mathsf{Z})$, $\mathbb{E}[\Delta_n(Z^n)] \le 2\mathbb{E}[R_n(\mathcal{F}(Z^n))]$." This problem investigates whether there is a universal finite constant c such that $\Delta_n(z^n) \le cR_n(\mathcal{F}(z^n))$ for all $z^n \in \mathsf{Z}^n$.

   (a) What does $\Delta_n(z^n)$ depend on? What does $R_n(\mathcal{F}(z^n))$ depend on? (Be sure to include z^n!)
   Solution: $\Delta(z^n) = \sup_{f\in\mathcal{F}} |P(f) - \mathsf{P}_n(f)|$, where P_n is the empirical distribution of z^n. Hence, it depends on z^n, P, and $\mathcal{F}$. (It is a measure of how well P_n approximates P.)
   $R_n(\mathcal{F}(z^n)) = E_\sigma[\sup_{f\in\mathcal{F}} \frac{1}{n}|\sum_{i\in[n]} \sigma_i f(z_i)|]$, which only depends on z^n and $\mathcal{F}$.

(b) Consider the following example. $\mathsf{Z} = \{0, 1\}$, $\mathcal{F} = \{f\}$ where $f(z) = z$ for $z \in \mathsf{Z}$, and $P\{0\} = P\{1\} = 1/2$. Give simple expressions for $\Delta(z^n)$ and $R_n(\mathcal{F}(z^n))$ for an arbitrary $z^n \in \mathsf{Z}^n$.

Solution: Since there is only one function in $\mathcal{F}$, the supremum over f is unnecessary. Also, $P(f) = 1/2$, and $\mathsf{P}_n(f) = k/n$, where k is the number of ones in z^n. Therefore, $\Delta(z^n) = |1/2 - k/n|$ and $R_n(\mathcal{F}(z^n)) = E_\sigma[\frac{1}{n}|\sum_{i\in[n]} \sigma_i z_i|] = E_\sigma[\frac{1}{n}|\sum_{i\in[k]} \sigma_i|] = \frac{2}{n}\mathbb{E}\left[|Y - k/2|\right]$, where Y has the binomial$(k, 1/2)$ distribution. (Nice upper and lower bounds on $\mathbb{E}\left[|Y - k/2|\right]$ are given in https://www.cs.bgu.ac.il/~karyeh/sharp-mad.pdf. All we need for (c) is $\mathbb{E}\left[|Y - k/2|\right] = 0$ for $k = 0$.)

(c) Find $\max_{z^n} \Delta(z^n)/R_n(\mathcal{F}(z^n))$ for the example in (b).

Solution: The ratio takes its maximum value $+\infty$ for $k = 0$. This shows, as suggested by (a), that no universal constant as described at the beginning of the problem exists.

3. **[Illustration of the shifting algorithm used in a proof of the Sauer- Shelah lemma]**
Suppose the shifting algorithm used in the proof of the Sauer-Shelah lemma is applied to the set $\mathcal{U}$ represented by the rows of the matrix shown:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

(a) Which subsets of $\{1, 2, 3, 4, 5\}$ are shattered by $\mathcal{U}$? (Include $\emptyset$.) What is the number of these sets?

Solution: Original $\mathcal{U}$ shatters $\emptyset$, all singleton sets, and $\{1, 2\}, \{1, 3\}, \{2, 3\}, \{2, 4\}, \{2, 5\}, \{3, 4\}, \{3, 5\}$. So $\mathbb{S}_5(\mathcal{U}) = 13$.

(b) Find the new $\mathcal{U}$ obtained by running steps 1-5 of the shifting algorithm on the set $\mathcal{U}$ in the previous part to produce a new $\mathcal{U}$ (so each column is processed once; you don't need to display all intermediate steps), and answer the questions in (a) for the new $\mathcal{U}$.

Solution:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \mathbf{0} & 1 & 0 & 0 & 0 \\ \mathbf{0} & 1 & 1 & 0 & 0 \\ 0 & 1 & \mathbf{0} & 1 & 0 \\ 0 & \mathbf{0} & 1 & \mathbf{0} & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \mathbf{0} & 1 & 0 \\ 0 & 0 & 0 & \mathbf{0} & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Now $\mathcal{U}$ shatters $\emptyset$, all singleton sets, and $\{2, 3\}, \{2, 4\}, \{3, 5\}$. So $\mathbb{S}_5(\mathcal{U}) = 9$.

(c) Repeat (b) as long as changes occur.

Solution: No more changes occur. The set $\mathcal{U}$ found in part (b) is downward closed.

4. **[The VC dimension of elliptical regions]**
Find an upper bound on the VC dimension of the set of (non-degenerate, closed) elliptical regions in $\mathbb{R}^d$ for $d \geq 1$. Such a region has the form $\{x : x^T Ax + b^T x \leq 1\}$ for some positive definite (thus symmetric) $d \times d$ matrix A and d dimensional vector b.

Solution: This is a subset of a Dudley class with linear dimension $\frac{d(d+1)}{2} + d = \frac{d^2+3d}{2.}$. Basically we relax the constraint that A be positive definite to A being symmetric. Therefore, $V(\mathcal{C}) \leq \frac{d^2+3d}{2}$. A quick search on the internet shows this bound is tight.

5. **[Classification among a finite number of concepts is a PAC learnable problem]**
Consider the concept classification problem $(\mathsf{X}, \mathcal{P}, \mathcal{C})$ described in Section 8.1 of the notes. *Suppose* $\mathcal{C}$ *has finite cardinality* M. The goal of this problem is to *directly* find an expression for $n(\epsilon, \delta, M)$ such that the following PAC guarantee holds for the ERM estimator $\widehat{C}_n$: For any $\epsilon, \delta > 0$ and any $P \in \mathcal{P}$, if $n \geq n(\epsilon, \delta, M)$, then $P^n\{L_P(\widehat{C}_n) > L_P^*(\mathcal{C}) + \epsilon\} \leq \delta$. We will also see what $n(\epsilon, \delta, M)$ is produced by the Vapnik-Chernovenkis theory.

   (a) Write $\mathcal{C} = \{C_i : i \in [M]\}$ and $\widehat{C}_n = C_{\widehat{i}_n}$. Fix $P \in \mathcal{P}$ and $\epsilon > 0$. Let $B_P = \{j \in [M] : L_P(C_j) > L_P^*(\mathcal{C}) + \epsilon\}$. The set B_P is the set of bad concepts for P. Then $P^n\{L_P(\widehat{C}_n) > L_P^*(\mathcal{C}) + \epsilon\} = P^n\{\widehat{i}_n \in B_P\}$. Let i^* denote a concept index such that $L_P(C_{i^*}) = L_P^*(\mathcal{C})$. Of course, $i^* \notin B_P$. The event $\{\widehat{i}_n \in B_P\}$ can happen only if:

$$L_{\mathsf{P}_n}(C_{i^*}) \geq L_P^*(\mathcal{C}) + \frac{\epsilon}{2} \qquad \text{or} \qquad (L_{\mathsf{P}_n}(C_j) \leq L_P^*(\mathcal{C}) + \frac{\epsilon}{2} \text{ for some } j \in B_P.)$$

   Apply the Hoeffding inequality and union bound to find a suitable upper bound on $P^n\{\widehat{i}_n \in B_P\}$.
   Solution: For each j fixed, $L_{\mathsf{P}_n}(C_j)$ is a sum of n independent random variables, each with values in the interval $[0, 1/n]$. Also, $E_{P^n}[L_{\mathsf{P}_n}(C)] = L_P(C_j)$. Therefore, by the Hoeffding inequality, for any j,

$$P^n\left\{L_{\mathsf{P}_n}(C_j) - L_P(C_j) \geq \quad \frac{\epsilon}{2}\right\} \leq \mathrm{e}^{-n\epsilon^2/2}$$
$$P^n\left\{L_{\mathsf{P}_n}(C_j) - L_P(C_j) \leq -\frac{\epsilon}{2}\right\} \leq \mathrm{e}^{-n\epsilon^2/2}$$

   Thus, the event $L_{\mathsf{P}_n}(C_{i^*}) \geq L_P^* + \frac{\epsilon}{2}$ and each of the events $L(\mathsf{P}_n, C_j) \leq L_P^* + \frac{\epsilon}{2}$ for $j \in B_P$, has probability less than or equal to $\mathrm{e}^{-n\epsilon^2/2}$. Since there are at most M such events, by the union bound, $P^n\{\widehat{i}_n \in B_P\} \leq M\mathrm{e}^{-n\epsilon^2/2}$.

   (b) Identify $n(\epsilon, \delta, M)$ such that the PAC property holds based on (a).
   Solution: By part (a), the PAC property holds if n is so large that $M\mathrm{e}^{-n\epsilon^2/2} \leq \delta$, or equivalently, if $n \geq n(\epsilon, \delta, M)$, where $n(\epsilon, \delta, M) = \frac{2}{\epsilon^2} \log \frac{M}{\delta}$.

   (c) Apply the second bound given in Theorem 8.1 of the notes to identify another value of $n(\epsilon, \delta, M)$ such that the PAC property holds.
   Solution: The VC dimension is less than or equal to $\log_2 M$, so by Theorem 8.1, the PAC property is satisfied if $\epsilon \leq C\sqrt{\frac{\log_2 M}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}$, so $n(\epsilon, \delta, M) = \frac{1}{\epsilon^2}\left\{C\sqrt{\log_2 M} + \sqrt{2\log(1/\delta)}\right\}^2$ works. (Using the inequality $(a+b)^2 \leq 2(a^2+b^2)$, we see this value of n is larger than the answer found in (b) by at most a fixed constant.)

6. **[No free lunch theorem in terms of VC dimension]**
The goal is to leverage Problem 3 of Problem Set 1 to prove the converse portion of the fundamental theorem of concept learning. Consider an agnostic concept learning problem $(\mathsf{X}, \mathcal{P}, \mathcal{C})$ such that $\mathcal{P}$ is the set of all probability distributions on $\mathsf{Z} = \mathsf{X} \times \{0, 1\}$. Show that if $V(\mathcal{C}) = \infty$ then the problem is not PAC learnable.

   Solution: We argue by contradiction. We can restrict attention to realizable probability distributions, because that makes the learning problem easier in the sense of PAC learnability. So, instead of considering probability distributions over Z, we consider pairs (P, C^*) such that $P \in \mathcal{P}(\mathsf{X})$ and $C^* \in \mathcal{P}$, with the interpretation that a fresh data sample has the form $Z = (X, \mathbf{1}_{\{X \in C^*\}})$, where X has distribution P. Suppose $V(\mathcal{C}) = \infty$ and $\mathcal{A}$ is an arbitrary learning algorithm. It suffices to show that $\mathcal{A}$ is not PAC. For that it suffices to show that for any n there exits a choice of $P \in \mathcal{P}$ and a choice of $C_n^* \in \mathcal{C}$ such that, for $\widehat{C}_n = \mathcal{A}(Z^n)$, $P^n\{P(C_n^* \triangle \widehat{C}_n) > 0.1\} \geq 1/6$. So fix $n \geq 1$. Since $V(\mathcal{C}) = +\infty > 2n$, there exists a set of the form $\{x_1, \ldots, x_{2m}\} \subset \mathsf{X}$ that is shattered by $\mathcal{C}$. So for each binary length $2n$ vector b, there exists a corresponding set $C_b \in \mathcal{C}$ such that $b = (\mathbf{1}_{\{x_1 \in C_b\}}, \ldots, \mathbf{1}_{\{x_{2n} \in C_b\}})$. Let P denote the uniform probability distribution over $x_1, \ldots, x_{2n}$. The features $X_1, \ldots, X_n$ are thus independently and uniformly distributed over $\{x_1, \ldots, x_{2n}\}$, which is informationally equivalent to the selector variables in

Problem 3 of Problem Set 1. By part (c) of that problem, there exists a deterministic binary $2n$-vector b (depending on the algorithm $\mathcal{A}$) such that with probability at least 1/6 (where the randomness is with respect to the selection of the $2n$ X's), the accuracy of $\widehat{C}_n$ when C_b is the true concept is less than 90% with probability at least 1/6. In other words, we can take $C_n^* = C_b$.

7. **[Some variations on the Iris flower data]**
The python programming problem for this problem set is explained within the .ipynb file. You can see a static version at http://nbviewer.jupyter.org/urls/courses.engr.illinois.edu/ece543/sp2019/ece543_PythonProblem2.ipynb?flush_cache=true and download the ipynb file the static version or directly from https://courses.engr.illinois.edu/ece543/sp2019/ece543_PythonProblem2.ipynb. The problem inolves a bit of use of the kernel method and support vector machines/classifiers, which will be examined later in the course in more detail.