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 to be handed in:

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 \(F_n(c) = \frac{1}{n} \sum_{i \in [n]} \mathbf{1}_{\{X_i \leq c\}}\) 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.

   (b) Using your answer to (a) and the McDiarmid inequality, find an upper bound on the upper tail of the distribution of \(\Delta_n\). For example, you could find an upper bound on probabilities of the form \(P\left\{\Delta_n \geq g(n) + \sqrt{\frac{t}{n}}\right\}\) for positive values of \(t\), where \(g\) is some function of \(n\) of your choice. (You may apply results from Problem 3, Problem Set 1 for this part.)

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 \(F\) be a class of functions \(f: Z \rightarrow [0, 1]\). Then for any \(P \in \mathcal{P}(Z)\), \(\mathbb{E}[\Delta_n(Z^n)] \leq 2\mathbb{E}[R_n(F(Z^n))]\)” This problem investigates whether there is a universal finite constant \(c\) such that \(\Delta_n(z^n) \leq cR_n(F(z^n))\) for all \(z^n \in Z^n\).

   (a) What does \(\Delta_n(z^n)\) depend on? What does \(R_n(F(z^n))\) depend on? (Be sure to include \(z^n\)!)  

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

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

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 \(U\) represented by the rows of the matrix shown:

   \[
   \begin{pmatrix}
   1 & 0 & 0 & 0 & 0 \\
   1 & 1 & 0 & 0 & 0 \\
   1 & 1 & 1 & 0 & 0 \\
   0 & 1 & 1 & 1 & 1 \\
   0 & 0 & 1 & 0 & 0 \\
   0 & 0 & 1 & 1 & 0 \\
   0 & 0 & 0 & 1 & 1 \\
   0 & 0 & 0 & 0 & 0 \\
   \end{pmatrix}
   \]

   (a) Which subsets of \(\{1, 2, 3, 4, 5\}\) are shattered by \(U\)? (Include 0.) What is the number of these sets? In other words, what is the shatter coefficient \(S_2(U)\)?

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

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

5. **[Classification among a finite number of concepts is a PAC learnable problem]**
Consider the concept classification problem \((X, P, C)\) described in Section 8.1 of the notes. Suppose \( C \) 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 \( \hat{C}_n \):

\[
\text{For any } \epsilon, \delta > 0 \text{ and any } P \in \mathcal{P}, \text{ if } n \geq n(\epsilon, \delta, M), \text{ then } P^n\{L_P(\hat{C}_n) > L_P(C) + \epsilon\} \leq \delta.
\]
We will also see what \( n(\epsilon, \delta, M) \) is produced by the Vapnik-Chernovenkis theory.

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

The event \( \{i_n \in B_P\} \) can happen only if:

\[
L_{P_n}(C_{i^*}) \geq L_P^*(C) + \frac{\epsilon}{2} \quad \text{or} \quad (L_{P_n}(C_j) \leq L_P^*(C) + \frac{\epsilon}{2} \text{ for some } j \in B_P.)
\]

Apply the Hoeffding inequality and union bound to find a suitable upper bound on \( P^n\{i_n \in B_P\} \).

(b) Identify \( n(\epsilon, \delta, M) \) such that the PAC property holds based on (a).

(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.

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 \((X, P, C)\) such that \( P \) is the set of all probability distributions on \( Z = X \times \{0, 1\} \). Show that if \( V(C) = \infty \) then the problem is not PAC learnable.

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 involves a bit of use of the kernel method and support vector machines/classifiers, which will be examined later in the course in more detail.