CSC 665: Homework 2

Chicheng Zhang

October 14, 2019

Please complete the following set of exercises. You must write down your solutions **on your own**. If you have discussed with your classmates on any of the questions, please indicate so in your solutions. The homework is due **on Oct 15, 12:30pm, on Gradescope**. You are free to cite existing theorems from the textbook and course notes.

Problem 1

Do Exercise 2.3 in (Shalev-Shwartz and Ben-David, 2014). For item 2, you can assume that the joint distribution of (X_1, X_2) is continuous over $\mathbb{R}^2$.

Solution

First we set up some notation. Denote by D_X the marginal distribution of D over $\mathcal{X} = \mathbb{R}^2$. Suppose $h^\star = h_{(a_1^\star, b_1^\star, a_2^\star, b_2^\star)}$ is the underlying classifier that separates the data. Denote by training set $S = \{x_1, \ldots, x_n\}$; for every example x, we use (x^1, x^2) to denote its feature representation.

1. Denote by $\hat{h} = h_{(\hat{a}_1, \hat{b}_1, \hat{a}_2, \hat{b}_2)}$ the classifier produced by A. Note that for $j = 1, 2$, $\hat{a}_j = \min \left\{ x_i^j : i \in \{1, \ldots, n\}, y_i = 1 \right\}$, $\hat{b}_j = \max \left\{ x_i^j : i \in \{1, \ldots, n\}, y_i = 1 \right\}$. Therefore, for $j = 1, 2$, $[\hat{a}_j, \hat{b}_j] \subset [a_j^\star, b_j^\star]$. Hence,
$$[\hat{a}_1, \hat{b}_1] \times [\hat{a}_2, \hat{b}_2] \subset [a_1^\star, b_1^\star] \times [a_2^\star, b_2^\star].$$
Therefore, the negative region of $\hat{h}$ is a superset of the negative region of $h^\star$, implying that $\hat{h}$ classifies all negative training examples correctly. In addition, by the definition of $\hat{a}_j$ and $\hat{b}_j$'s, all positive training examples x_i lies in rectangle $[\hat{a}_1, \hat{b}_1] \times [\hat{a}_2, \hat{b}_2]$, hence classified as positive by $\hat{h}$. Therefore A is an ERM.

2. (a) See item 1.

   (b) We use the definition of a_1, b_1, a_2, b_2 in the hint. If S contains positive examples in all rectangles R_1,R_2,R_3,R_4, then $\hat{a}_1 \le a_1$, $\hat{b}_1 \ge b_1$, $\hat{a}_2 \le a_2$, $\hat{b}_2 \ge b_2$ hold simultaneously, that is, $[a_1, b_1] \times [a_2, b_2] \subset [\hat{a}_1, \hat{b}_1] \times [\hat{a}_2, \hat{b}_2]$. Now,
$$\begin{aligned}
\mathrm{err}(\hat{h}, D) &= \mathbb{P}_{x \sim D_X}(x \notin [\hat{a}_1, \hat{b}_1] \times [\hat{a}_2, \hat{b}_2] \wedge x \in [a_1^\star, b_1^\star] \times [a_2^\star, b_2^\star]) \\
&= \mathbb{P}_{x \sim D_X}(x \notin [a_1, b_1] \times [a_2, b_2] \wedge x \in [a_1^\star, b_1^\star] \times [a_2^\star, b_2^\star]) \\
&= \mathbb{P}_{x \sim D_X}((x_1 < a_1 \vee x > b_1 \vee x_2 < a_2 \vee x_2 > b_2) \wedge x \in [a_1^\star, b_1^\star] \times [a_2^\star, b_2^\star]) \\
&= \mathbb{P}_{x \sim D_X}(x \in (R_1 \cup R_2 \cup R_3 \cup R_4)) \\
&\le \mathbb{P}(R_1) + \mathbb{P}(R_2) + \mathbb{P}(R_3) + \mathbb{P}(R_4) \le 4 \times \frac{\epsilon}{4} \le \epsilon.
\end{aligned}$$

(c) For each $j \in \{1,2,3,4\}$, as $\mathbb{P}(R_j) = \frac{\epsilon}{4}$, and $n \geq \frac{4 \ln \frac{4}{\delta}}{\epsilon}$,

$$\mathbb{P}(S \text{ does not contain examples in } R_j) = \mathbb{P}(\forall i \in \{1,\ldots,n\}, x_i \notin R_j) = (1 - \frac{\epsilon}{4})^n \leq e^{-\frac{n\epsilon}{4}} \leq \frac{\delta}{4}.$$

(d) By union bound and the last item,

$$\mathbb{P}(\exists j, S \text{ does not contain examples in } R_j) \leq \sum_{j=1}^{4} \mathbb{P}(S \text{ does not contain examples in } R_j) \leq \delta.$$

Therefore, with probability $1-\delta$, S contains examples in all R_j's, in which case $\hat{h}$ has error at most ϵ by item (b).

3. Denote by $R(f_1, g_1, \ldots, f_d, g_d) = \times_{i=1}^{d} [f_i, g_i]$ as the hyper-rectangle determined by its $2d$ sides. For every j, denote by a_j (resp. b_j) be such that $R_{2j-1} \triangleq R(a_1^\star, b_1^\star, \ldots, a_j^\star, a_j, \ldots, a_1^\star, b_1^\star)$ (resp. $R_{2j} \triangleq R(a_1^\star, b_1^\star, \ldots, b_j, b_j^\star, \ldots, a_1^\star, b_1^\star)$) has probability $\frac{\epsilon}{2d}$. Now suppose $n \geq \frac{2d \ln \frac{2d}{\delta}}{\epsilon}$, we have

$$\mathbb{P}(\exists j, S \text{ does not contain examples in } R_j) \leq \sum_{j=1}^{2d} \mathbb{P}(S \text{ does not contain examples in } R_j) \leq 2d(1 - \frac{\epsilon}{2d})^n \leq \delta.$$

Therefore, with probability $1-\delta$, we have at least one training example in all the regions R_j's. For the rest of the proof suppose this fact happens. consider $\hat{h} = h_{\hat{a}_1, \hat{b}_1, \ldots, \hat{a}_d, \hat{b}_d}$. Therefore, for all j, $\hat{a}_j \leq a_j$ and $\hat{b}_j \geq b_j$. Hence,

$$\begin{aligned} \mathrm{err}(\hat{h}, D) &\leq \mathbb{P}_{x \sim D_X}(x \notin \times_{j=1}^{d} [\hat{a}_i, \hat{b}_i] \wedge x \in \times_{i=1}^{d} [a_i^\star, b_i^\star]) \\ &= \mathbb{P}_{x \sim D_X}(x \in \cup_{j=1}^{2d} R_j) \\ &\leq \sum_{j=1}^{2d} \mathbb{P}(R_j) \leq 2d \cdot \frac{\epsilon}{2d} \leq \epsilon. \end{aligned}$$

Therefore, A has a (ϵ, δ)-PAC sample complexity of $\frac{2d \ln \frac{2d}{\delta}}{\epsilon}$.

4. The algorithm A has a time complexity of $O(nd)$, as it need to scan through all training examples and compute the $\hat{a}_j$ and $\hat{b}_j$'s, minimum abd maximum values of j-th coordinate over all positive training examples. As we can $n = O(\frac{d}{\epsilon} \ln \frac{d}{\delta})$, the time complexity is $O(\frac{d^2}{\epsilon} \ln \frac{d}{\delta})$, which is polynomial in d, $\frac{1}{\epsilon}$, $\ln \frac{1}{\delta}$.

Problem 2

1. Show the following inequality: for positive a, b and x, if $x > 2a \ln(2a) + 2b$, then $x > a \ln x + b$.

2. Show the following basic inequality: for n, d such that $n \geq 2$ and $n \geq d$, $\binom{n}{\leq d} \leq n^{d+1}$.

3. Consider l hypothesis classes $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_l$, where $\mathrm{VC}(\mathcal{H}_i) = v \geq 1$. Define $\mathcal{H} \triangleq \cup_{i=1}^{l} \mathcal{H}_i$. Show that there exists some constant $c > 0$ such that

$$\mathrm{VC}(\mathcal{H}) \leq c \cdot \big(v \ln(v) + \ln(l)\big).$$

4. Let $\mathcal{H} = \left\{\text{sign}(\langle w, x\rangle) : w \in \mathbb{R}^d, |\{i : w_i \neq 0\}| = k\right\}$ be the set of k-sparse homogenenous linear classifiers in $\mathbb{R}^d$, where $k \leq d$. Show that there exists some constant $c > 0$ such that

$$\text{VC}(\mathcal{H}) \leq c \cdot (k \ln d).$$

5. Consider l hypothesis classes $\mathcal{H}_1, \ldots, \mathcal{H}_l$, where $\text{VC}(\mathcal{H}_i) = d_i \geq 1$. Suppose f is a function from $\{\pm 1\}^l$ to $\{\pm 1\}$ (for example, the majority function $f(z_1, \ldots, z_l) = \text{sign}(\sum_{i=1}^l z_i)$ or the parity function $f(z_1, \ldots, z_l) = \prod_{i=1}^l z_i$). Define $\mathcal{H} \triangleq \left\{f(h_1(x), \ldots, h_l(x)) : h_1 \in \mathcal{H}_1, \ldots, h_l \in \mathcal{H}_l\right\}$. Show that there exists some constant $c > 0$ such that

$$\text{VC}(\mathcal{H}) \leq c \left(\sum_{i=1}^l d_i\right) \ln \left(\sum_{i=1}^l d_i\right).$$

Solution

1. We use proof by contradiction. Suppose the conclusion does not hold, i.e.

$$x \leq a \ln x + b. \tag{1}$$

By hint, we know that $\ln \frac{x}{2a} \leq \frac{x}{2a} - 1$. This implies that $a \ln x + b \leq \frac{x}{2} - a + a \ln(2a) + b$. Combining this inequality with Equation (1), we have

$$x \leq \frac{x}{2} - a + a \ln(2a) + b$$

Therefore, $x \leq 2a \ln(2a) + 2b$, contradicting with the assumption that $x > 2a \ln(2a) + 2b$.

2. Using the fact that $\binom{n}{i} \leq n^i$, we have

$$\binom{n}{\leq d} = \sum_{i=0}^d \binom{n}{i} \leq \sum_{i=0}^d n^i = \frac{n^{d+1} - 1}{n - 1} \leq n^{d+1}.$$

where the last inequality uses the fact that $n \geq 2$.

3. Consider a set of n points $S = \{x_1, \ldots, x_n\}$ shattered by $\mathcal{H}$. It suffices to show that there exists a cosntant c such that $n \leq c(v \ln v + \ln l)$.

First, note that by definition of shattering, $|\Pi_{\mathcal{H}}(S)| = 2^n$. Second, note that

$$\begin{aligned}
\Pi_{\mathcal{H}}(S) &= \left\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\right\} \\
&= \left\{(h(x_1), \ldots, h(x_n)) : h \in \cup_{i=1}^l \mathcal{H}_i\right\} \\
&= \cup_{i=1}^l \left\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}_i\right\} \\
&= \cup_{i=1}^l \Pi_{\mathcal{H}_i}(S).
\end{aligned}$$

Therefore, $|\Pi_{\mathcal{H}}(S)| \leq \sum_{i=1}^l |\Pi_{\mathcal{H}_i}(S)|$. Now, by Sauer's Lemma, each $|\Pi_{\mathcal{H}_i}(S)| \leq \binom{n}{\leq v}$, implying that $|\Pi_{\mathcal{H}}(S)| \leq l \cdot \binom{n}{\leq v}$.

To summarize, we know that $2^n \leq l \cdot \binom{n}{\leq v}$. Now by the inequality in item 2, we know that

$$n \leq \log l + (v + 1) \log n,$$

therefore, there exists a constant c_1 such that

$$n \leq c_1(v \ln n + \ln l).$$

Now, applying the contrapositive of item 1, we have that there exists a constant c such that

$$n \leq c(v \ln v + \ln l).$$

4. For every $S \subset \{1, \ldots, d\}$, denote by $\mathcal{H}_S = \left\{\text{sign}(\langle w, x \rangle) : w \text{ takes zeros on } \bar{S}\right\}$, in other words, the set of homogeneous linear classifiers whose normal vectors are supported on S.

   Note that $\mathcal{H}_S$ can be alternatively written as $\left\{\text{sign}(\sum_{i \in S} w_i \cdot x_i) : w \in \mathbb{R}^d\right\}$, specifically, it ignores the feature outside S. As the VC dimension of the set of l-dimensonal homogeneous linear classifiers equals m, $\text{VC}(\mathcal{H}_S) = |S|$.

   Now, $\mathcal{H} = \cup_{S:|S|=k} \mathcal{H}_S$, where each $\mathcal{H}_S$ has VC dimension k and there are $\binom{d}{k}$ sets in the union. Therefore, by item 3.

$$\text{VC}(\mathcal{H}) \le c(k \ln k + \ln \binom{d}{k}) \le c(k \ln k + k \ln d).$$

   where the last inequality uses the fact that $\binom{d}{k} \le d^k$.

5. The proof is similiar to that of item 3. Consider a set of n points $S = \{x_1, \ldots, x_n\}$ shattered by $\mathcal{H}$. It suffices to show that there exists a constant c such that $n \le c((\sum_{i=1}^{l} d_i) \ln\left(\sum_{i=1}^{l} d_i\right))$.

   First, note that by definition of shattering, $|\Pi_{\mathcal{H}}(S)| = 2^n$. Second, note that

$$\begin{aligned}
|\Pi_{\mathcal{H}}(S)| &= |\left\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\right\}| \\
&= |\left\{(f(h_1(x_1), \ldots, h_l(x_1)), \ldots, f(h_1(x_n), \ldots, h_l(x_n))) : h_1 \in \mathcal{H}_1, \ldots, h_l \in \mathcal{H}_l\right\}| \\
&\le |\left\{((h_1(x_1), \ldots, h_l(x_1)), \ldots, (h_1(x_n), \ldots, h_l(x_n))) : h_1 \in \mathcal{H}_1, \ldots, h_l \in \mathcal{H}_l\right\}| \\
&= \prod_{i=1}^{l} |\Pi_{\mathcal{H}_i}(S)|.
\end{aligned}$$

   In words, suppose instead of assigning a binary labeling ($\{\pm 1\}$) on each example, we assign a $\pm 1^l$ labeling on each example, using a l-tuple of classifier $(h_1, \ldots, h_l)$. There are $\prod_{i=1}^{l} |\Pi_{\mathcal{H}_i}(S)|$ many such "composite labelings", and each composite labeling induces one labeling of S by $\mathcal{H}$.

   Now, by Sauer's Lemma, $\prod_{i=1}^{l} |\Pi_{\mathcal{H}_i}(S)| \le \prod_{i=1}^{l} \binom{n}{d_i} \le \prod_{i=1}^{l} n^{d_i+1} \le n^{2(\sum_{i=1}^{l} d_i)}$. Combining this fact with $\Pi_{\mathcal{H}}(S) = 2^n$, and using the contrapositive of item 1, we get the conclusion.

Problem 3

In this exercise, we will show that, under the *realizable setting*, with hypothesis class $\mathcal{H}$ having VC dimension d, ERM (in fact, the consistency algorithm) will have a PAC sample complexity of $O\left(\frac{1}{\epsilon}(d \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})\right)$. Suppose $S = \{Z_1, \ldots, Z_m\}$ a set of m training examples drawn iid from distribution D, where each $Z_i = (X_i, Y_i)$ is a labeled example. In addition, $\mathcal{F} = \left\{\mathbf{1}(h(x) \neq y) : h \in \mathcal{H}\right\}$ is the zero-one loss function class. Our proof will mostly follow the steps for showing agnostic PAC sample complexity given in the lecture.

1. **Double Sampling Trick.** Fix a training set S. Suppose $\mathbb{E}_S f(Z) = 0$ and $\mathbb{E}_D f(Z) \ge \epsilon$. Show that for a fresh set of random examples S' of size m ($m \ge \frac{16}{\epsilon}$) sampled iid from D:

$$\mathbb{P}_{S' \sim D^m}\left(\mathbb{E}_{S'} f(Z) \ge \frac{\epsilon}{2}\right) \ge \frac{1}{2}.$$

2. **Conditioning.** Denote by events

$$E' \triangleq \left\{\text{there exists } f \in \mathcal{F}, \mathbb{E}_S f(Z) = 0, \mathbb{E}_{S'} f(Z) \ge \frac{\epsilon}{2}\right\},$$

$$E \triangleq \left\{\text{there exists } f \in \mathcal{F}, \mathbb{E}_S f(Z) = 0, \mathbb{E}_D f(Z) \ge \epsilon\right\}.$$

   Show $\mathbb{P}_{S,S' \sim D^m}(E'|E) \ge \frac{1}{2}$, and conclude that $\mathbb{P}_{S \sim D^m}(E) \le 2\mathbb{P}_{S,S' \sim D^m}(E')$.

3. **Symmetrization.** Introduce $\sigma = (\sigma_1, \ldots, \sigma_m)$ where each $\sigma_i \in \{\pm 1\}$. Denote by

$$(W_i, W_i') = \begin{cases} (Z_i, Z_i') & \sigma_i = +1, \\ (Z_i', Z_i) & \sigma_i = -1. \end{cases}$$

Show that

$$\mathbb{P}_{S,S' \sim D^m}(E') = \mathbb{P}_{S,S' \sim D^m, \sigma \sim \mathrm{R}^m}\left(\text{exists } f \in \mathcal{F}, \sum_{i=1}^m f(W_i) = 0, \sum_{i=1}^m f(W_i') \geq \frac{m\epsilon}{2}\right),$$

where R is the Rademacher distribution, i.e. uniform in $\{\pm 1\}$.

4. **The randomness in Rademacher random variables.** Fix two size m training sets S and S'. Show that for a fixed classifier f in $\mathcal{F}$,

$$\mathbb{P}_{\sigma \sim \mathrm{R}^m}\left(\sum_{i=1}^m f(W_i) = 0, \sum_{i=1}^m f(W_i') \geq \frac{m\epsilon}{2}\right) \leq \exp\left(-\frac{m\epsilon}{4}\right).$$

5. Use the above items to conclude that for $m \geq \frac{16}{\epsilon}$,

$$\mathbb{P}_{S \sim D^m}(\text{there exists } f \in \mathcal{F}, \mathbb{E}_S f(Z) = 0, \mathbb{E}_D f(Z) \geq \epsilon) \leq 2\mathcal{S}(\mathcal{F}, 2m) \exp\left\{-\frac{m\epsilon}{4}\right\}. \tag{2}$$

In addition, show that ERM has a PAC sample complexity of $O\left(\frac{1}{\epsilon}(d \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})\right)$.

Solution

1. Let $p \triangleq \mathbb{E}_D f(Z)$. Suppose $S' = \{Z_1, \ldots, Z_m\}$. Then denote by $Y_i = f(Z_i)$, we have $\mathbb{E}_{S'} f(Z) = \frac{1}{m}\sum_{i=1}^m Y_i$. Note that $Y_i \sim \text{Bernoulli}(p)$ iid.
In the above notation, $\mathbb{P}_{S' \sim D^m}\left(\mathbb{E}_{S'} f(Z) \geq \frac{\epsilon}{2}\right) \geq \frac{1}{2}$ is equivalent to $\mathbb{P}(\sum_{i=1}^m Y_i \geq \frac{mp}{2}) \geq \frac{1}{2}$, which is equivalent to

$$\mathbb{P}(\sum_{i=1}^m Y_i < \frac{mp}{2}) \leq \frac{1}{2}.$$

This can be easily seen by Chernoff bound for Bernoulli random variables and the fact that $m \geq \frac{16}{\epsilon}$:

$$\mathbb{P}(\sum_{i=1}^m Y_i < \frac{mp}{2}) \leq e^{-\frac{mp}{16}} \leq e^{-1} \leq \frac{1}{2}$$

2. We first show a basic probability fact: suppose random variables (U, V) (taking values in $\mathcal{U}$ and $\mathcal{V}$) have a joint probability density p. Suppose A is a subset of $\mathcal{U}$ and B is a subset of $\mathcal{U} \times \mathcal{V}$; in addition, for all u in A, $\mathbb{P}((U,V) \in B | U = u) \geq \frac{1}{2}$. Then $\mathbb{P}((U,V) \in B | U \in A) \geq \frac{1}{2}$. To see this, note that

$$\begin{aligned} \mathbb{P}((U,V) \in B, U \in A) &= \int\int p(u,v)\mathbf{1}((u,v) \in B)\mathbf{1}(u \in A)dudv \\ &= \int(\int p(u|v)\mathbf{1}((u,v) \in B)dv)p(v)\mathbf{1}(u \in A)du \\ &\geq \int \frac{1}{2}\mathbf{1}(u \in A)du = \frac{1}{2}\mathbb{P}(U \in A). \end{aligned}$$

Now, apply the fact with $U = S$, $V = S'$, $A = \big\{S : \text{there exists } f \in \mathcal{F}, \mathbb{E}_S f(Z) = 0, \mathbb{E}_D f(Z) \geq \epsilon\big\}$, $B = \big\{(S, S') : \text{there exists } f \in \mathcal{F}, \mathbb{E}_S f(Z) = 0, \mathbb{E}_{S'} f(Z) \geq \epsilon/2\big\}$, along with item 1, we immediate get $\mathbb{P}(E'|E) \geq \frac{1}{2}$.

As $\frac{1}{2} \leq \mathbb{P}(E'|E) = \frac{\mathbb{P}(E' \cup E)}{\mathbb{P}(E)} \leq \frac{\mathbb{P}(E')}{\mathbb{P}(E)}$, we have $\mathbb{P}(E) \leq 2\mathbb{P}(E')$.

3. We first show that for a fixed $\sigma \in \{\pm 1\}^m$,

$$
\begin{aligned}
& \mathbb{P}_{S,S' \sim D^m}\left(\text{exists } f \in \mathcal{F}, \sum_{i=1}^m f(Z_i) = 0, \sum_{i=1}^m f(Z_i') \geq \frac{m\epsilon}{2}\right) \\
= \;& \mathbb{P}_{S,S' \sim D^m}\left(\text{exists } f \in \mathcal{F}, \sum_{i=1}^m f(W_i) = 0, \sum_{i=1}^m f(W_i') \geq \frac{m\epsilon}{2}\right).
\end{aligned}
$$

The reason is that, $(Z_i, Z_i')_{i=1}^m$ has the same distribution as $(W_i, W_i')_{i=1}^m$. To see this, denote by $p_Z(z)$ the probability density function of Z. We see that both random vectors has the probability density of $\prod_{i=1}^{2m} p_Z(z_i)$. This means that, for any $2m$-ary function h,

$$\mathbb{E}h(Z_1, Z_1', \ldots, Z_m, Z_m') = \mathbb{E}h(W_1, W_1', \ldots, W_m, W_m').$$

Specifically, taking function h to be

$$h(z_1, z_1', \ldots, z_m, z_m') = \mathbf{1}\left(\text{exists } f \in \mathcal{F}, \sum_{i=1}^m f(Z_i) = 0, \sum_{i=1}^m f(Z_i') \geq \frac{m\epsilon}{2}\right),$$

we proved Equation 3. As Equation 3 holds for any fixed $\sigma \in \{\pm 1\}^m$, taking expectation on both sides (wrt the randomness in $\sigma \sim \mathrm{R}^m$), we have

$$
\begin{aligned}
& \mathbb{P}_{S,S' \sim D^m}\left(\text{exists } f \in \mathcal{F}, \sum_{i=1}^m f(Z_i) = 0, \sum_{i=1}^m f(Z_i') \geq \frac{m\epsilon}{2}\right) \\
= \;& \mathbb{E}_{\sigma \sim \mathrm{R}^m} \mathbb{P}_{S,S' \sim D^m}\left(\text{exists } f \in \mathcal{F}, \sum_{i=1}^m f(W_i) = 0, \sum_{i=1}^m f(W_i') \geq \frac{m\epsilon}{2}\right).
\end{aligned}
$$

which yields the desired result.

4. Consider three cases: (1) there exists some i, $(f(Z_i), f(Z_i')) = (1, 1)$; (2) $\sum_{i=1}^m f(Z_i) + f(Z_i') < \frac{m\epsilon}{2}$; (3) both (1) and (2) are not satisfied. In case (1), it is impossible that $\sum_{i=1}^m f(W_i) = 0$ regardless of the choice of σ; in case (2), as $\sum_{i=1}^m f(W_i') \leq \sum_{i=1}^m f(W_i) + f(W_i') = \sum_{i=1}^m f(Z_i) + f(Z_i') < \frac{m\epsilon}{2}$, the probability of interest is also zero.

In case (3), we note that there are at least $\frac{m\epsilon}{2}$ i's such that $(f(Z_i), f(Z_i'))$ is $(1, 0)$ or $(0, 1)$ - let us call these locations $\mathcal{I}$. Note that a random σ will make $f(W_i)$ taking value 0 with probability $\frac{1}{2}$ at locations i in $\mathcal{I}$. Therefore,

$$
\begin{aligned}
\mathbb{P}_{\sigma \sim \mathrm{R}^m}\left(\sum_{i=1}^m f(W_i) = 0, \sum_{i=1}^m f(W_i') \geq \frac{m\epsilon}{2}\right) \;&\leq\; \mathbb{P}(f(W_i) = 0, \text{ for all } i \in \{1, \ldots, m\}) \\
&=\; \mathbb{P}(f(W_i) = 0, \text{ for all } i \in \mathcal{I}) \\
&\leq\; 2^{-|\mathcal{I}|} \leq 2^{-\frac{m\epsilon}{2}} \leq \exp\left\{-\frac{m\epsilon}{4}\right\}.
\end{aligned}
$$

5. Fix a training set S and S'; observe that there are at most $\mathcal{S}(\mathcal{F}, 2m)$ different configurations of $(f(Z_1), f(Z_1'), \ldots, f(Z_m), f(Z_m'))$ that functions f in $\mathcal{F}$ can induce. item 4 implies that for each such configuration,
$$\mathbb{P}_{\sigma\sim \mathrm{R}^m}\left(\sum_{i=1}^m f(W_i) = 0, \sum_{i=1}^m f(W_i') \geq \frac{m\epsilon}{2}\right) \leq \exp\left\{-\frac{m\epsilon}{4}\right\}.$$
By union bound (over all possible configurations),
$$\mathbb{P}_{\sigma\sim \mathrm{R}^m}\left(\text{exists } f \in \mathcal{F}, \sum_{i=1}^m f(W_i) = 0, \sum_{i=1}^m f(W_i') \geq \frac{m\epsilon}{2}\right) \leq S(\mathcal{F}, 2m)\exp\left\{-\frac{m\epsilon}{4}\right\}.$$
Taking average over the random choices of S and S', and apply items 2 and 3, we have
$$\begin{aligned}\mathbb{P}(E) \leq 2\mathbb{P}(E') &\leq 2\mathbb{P}_{S\sim D^m, S'\sim D^m, \sigma\sim \mathrm{R}^m}\left(\text{exists } f \in \mathcal{F}, \sum_{i=1}^m f(W_i) = 0, \sum_{i=1}^m f(W_i') \geq \frac{m\epsilon}{2}\right)\\ &\leq 2S(\mathcal{F}, 2m)\exp\left\{-\frac{m\epsilon}{4}\right\}.\end{aligned}$$
Lastly, by Sauer's Lemma, $S(\mathcal{F}, 2m) \leq (\frac{2em}{d})^d$. Therefore, to make the right hand side of Equation (2) at most δ, it suffices to let $m \geq \frac{c}{\epsilon}(d \ln \frac{m}{d} + \ln \frac{1}{\delta})$ for large enough constant c. Observe that
$$\begin{aligned} & m \geq \frac{c}{\epsilon}(d\ln\frac{m}{d} + \ln\frac{1}{\delta}) \\ \Leftarrow \quad & m \geq c\ln\frac{1}{\delta} \text{ and } m \geq \frac{c}{\epsilon}\cdot d\ln\frac{m}{d} \\ \Leftarrow \quad & m \geq c\ln\frac{1}{\delta} \text{ and } \frac{m}{d} \geq \frac{c}{\epsilon}\ln\frac{m}{d} \\ \Leftarrow \quad & m \geq c\ln\frac{1}{\delta} \text{ and } \frac{m}{d} \geq \frac{2c}{\epsilon}\ln\frac{2c}{\epsilon} \\ \Leftarrow \quad & m \geq c\ln\frac{1}{\delta} + \frac{2cd}{\epsilon}\ln\frac{2c}{\epsilon}. \end{aligned}$$
where the second to last line uses the inequality in item 1 of problem 2, and the last line uses the fact that $m \geq a + b$ implies that $m \geq a$ and $m \geq b$. This concludes the PAC sample complexity upper bound.

Problem 4

In this exercise, we develop sample complexity lower bounds for *realizable* PAC learning using Le Cam's method and Assouad's method. Suppose hypothesis class $\mathcal{H}$ has VC dimension $d \geq 2$, and it shatters examples $z_0, z_1, \ldots, z_{d-1}$. In additon, suppose $\epsilon, \delta \in (0, \frac{1}{8})$ are target error and target failure probability. A learning algorithm $\mathcal{A}$ is a mapping from training set S to $\{\pm 1\}$. In the following, you can use the elementary fact that for $x \in (0, \frac{1}{2})$, $e^{-x} \geq 1 - x \geq e^{-2x}$.

1. Consider D_{-1} and D_{+1} as follows: for every i in $\{\pm 1\}$,
$$D_i(x, y) = \begin{cases} 1 - 2\epsilon, & (x, y) = (z_0, -1), \\ 2\epsilon, & (x, y) = (z_1, i), \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\min_{h \in \mathcal{H}} \mathrm{err}(h', D_i) = 0$ for both $i \in \{\pm 1\}$. For every j in $\{\pm 1\}$, denote by $P_j((x_i, y_i)_{i=1}^m) = \prod_{i=1}^m D_j(x_i, y_i)$ the distribution over training sets (observations). Use Le Cam's method to show that for any hypothesis tester f, there exists an i in $\{\pm 1\}$, such that

$$\mathbb{P}_i(f(S) \neq i) > \frac{1}{2}(1 - 4\epsilon)^m.$$

2. Conclude that for any learning algorithm $\mathcal{A}$, if sample size $m \leq \frac{1}{8\epsilon} \ln \frac{1}{4\delta}$, then there exists an i in $\{\pm 1\}$,

$$\mathbb{P}_i(\mathrm{err}(\hat{h}, D_i) > \epsilon) > \delta.$$

3. For every $\tau \in \{\pm 1\}^{d-1}$, consider D_τ as follows:

$$D_\tau(x, y) = \begin{cases} 1 - 4\epsilon, & (x, y) = (z_0, -1), \\ \frac{4\epsilon}{d-1}, & (x, y) = (z_i, \tau_i) \text{ for some } i \in \{1, \ldots, d-1\}, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\min_{h \in \mathcal{H}} \mathrm{err}(h', D_\tau) = 0$ for all $\tau \in \{\pm 1\}^{d-1}$. For every τ, denote by $P_\tau((x_i, y_i)_{i=1}^m) = \prod_{i=1}^m D_\tau(x_i, y_i)$ the distribution over training sets (observations).

Use Assouad's method to show that for any hypothesis tester $f_1, \ldots, f_{d-1}$, there exists $\tau \in \{\pm 1\}^{d-1}$,

$$\mathbb{E}_\tau \left[\sum_{j=1}^{d-1} \mathbf{1}(f_j(S) \neq \tau_j) \right] > \frac{d-1}{2} \left(1 - \frac{4\epsilon}{d-1} \right)^m.$$

4. Conclude that for any learning algorithm $\mathcal{A}$, suppose that sample size $m \leq \frac{d-1}{128\epsilon}$, then there exists a $\tau \in \{\pm 1\}^d$, such that

$$\mathbb{P}_\tau(\mathrm{err}(\hat{h}, D_\tau) > \epsilon) > \frac{1}{4}.$$

Solution

1. Consider observation $S_0 = ((z_0, -1), \ldots, (z_0, -1))$. Note that

$$\mathbb{P}_{-1}(S_0) = \prod_{i=1}^m D_{-1}(z_0, -1) = (1 - 2\epsilon)^m = \prod_{i=1}^m D_{+1}(z_0, -1) = \mathbb{P}_{+1}(S_0).$$

In addition, all expression in the above equation is at least $(1 - 4\epsilon)^m$. By Le Cam's Lemma, this implies that for any hypothesis tester f, there exists i in $\{\pm 1\}$, such that

$$\mathbb{P}_i(f(S) \neq i) \geq \frac{1}{2} \sum_{S \in (\{z_0, z_1\} \times \{pm1\})^m} \min(P_{-1}(S), P_{+1}(S)) \geq \frac{1}{2} \min(P_{-1}(S_0), P_{+1}(S_0)) = \frac{1}{2}(1 - 4\epsilon)^m.$$

2. Suppose learning algorithm $\mathcal{A}$ returns a classifier $\hat{h}$ that depends on S. Define hypothesis tester f such that $f(S) = \hat{h}(z_1)$ that also depends on S. It can be seen that

$$\mathrm{err}(h, D_i) = 2\epsilon \mathbf{1}(h(z_1) \neq i).$$

Therefore, if $f(S) = \hat{h}(z_1) \neq i$, then $\mathrm{err}(\hat{h}, D_i) \geq 2\epsilon$.

By item 1, we know that there exists i such that

$$\mathbb{P}_i(\mathrm{err}(h, D_i) \geq 2\epsilon) \geq \mathbb{P}_i(f(S) \neq i) \geq \frac{1}{2}(1 - 4\epsilon)^m.$$

As $m \leq \frac{1}{8\epsilon} \ln \frac{1}{4\delta}$, $\frac{1}{2}(1 - 4\epsilon)^m \geq \frac{1}{2} e^{-8\epsilon m} \geq \frac{1}{2} e^{-\ln \frac{1}{4\delta}} \geq 2\delta > \delta$. This conclude the proof of the item.

3. For every j in $\{1,\dots,d-1\}$, denote by $A_j = \{S = (x_i, y_i)_{i=1}^m : x_i \neq z_j \text{ for all } i\}$. Now, for every $\tau \overset{j}{\sim} \tau'$, for all S in A_j, we have:

$$\mathbb{P}_\tau(S) = \prod_{i=1}^n D_\tau(x_i, y_i) = \prod_{i=1}^n D_{\tau'}(x_i, y_i) = \mathbb{P}_{\tau'}(S).$$

where the second equality is from the fact that for all i, $D_\tau(x_i, y_i) = D_\tau(x_i, y_i)$, as $x_i \neq z_j$ for all i. In addition,

$$\mathbb{P}_{S \sim P_\tau}(A_j) = \prod_{i=1}^m \mathbb{P}((x_i, y_i) \sim D_\tau)(x_i \neq z_j) = (1 - \frac{4\epsilon}{d-1})^m.$$

Therefore,

$$\begin{aligned}
\|P_\tau \wedge P'_\tau\| &= \sum_{S \in (\{z_0,\dots,z_{d-1}\} \times \{\pm 1\})^m} \min(P_\tau(S), P_\tau(S)) \\
&\geq \sum_{S \in A} \min(P_\tau(S), P_{\tau'}(S)) \\
&= \sum_{S \in A} \min(P_\tau(S), P_{\tau'}(S)) = P_\tau(A_j) \geq (1 - \frac{4\epsilon}{d-1})^m.
\end{aligned}$$

As the above holds for any j, by Assouad's Lemma, for any hypothesis tester $f_1, \dots, f_{d-1}$, there exists $\tau \in \{\pm 1\}^{d-1}$,

$$\begin{aligned}
\mathbb{E}_\tau \left[\sum_{j=1}^{d-1} \mathbf{1}(f_j(S) \neq \tau_j) \right] &\geq \frac{d-1}{2} \min_{\tau, \tau' : \tau \sim \tau'} \|P_\tau \wedge P'_\tau\| \\
&\geq \frac{d-1}{2} (1 - \frac{4\epsilon}{d-1})^m.
\end{aligned}$$

4. Suppose learning algorithm $\mathcal{A}$ returns a classifier $\hat{h}$ that depends on S. Define hypothesis tester $f = (f_1, \dots, f_{d-1})$ such that $f_i(S) = \hat{h}(z_i)$ that also depends on S. It can be seen that

$$\text{err}(h, D_\tau) = \frac{4\epsilon}{d-1} \sum_{j=1}^{d-1} \mathbf{1}(h(z_j) \neq \tau_j).$$

Therefore, if $\sum_{j=1}^{d-1} \mathbf{1}(h(z_j) \neq \tau_j) > \frac{d-1}{4}$, then $\text{err}(h, D_\tau) > \epsilon$.

By item 3 and the choice of $m \leq \frac{d-1}{128\epsilon}$, we know that there exists τ such that

$$\mathbb{E}_\tau \sum_{j=1}^{d-1} \mathbf{1}(h(z_j) \neq \tau_j) \geq \frac{d-1}{2}(1 - \frac{4\epsilon}{d-1})^m > \frac{d-1}{2} e^{-\frac{8\epsilon m}{d-1}} > \frac{d-1}{2}(1 - \frac{16m}{d-1}) \geq \frac{d-1}{2} \cdot \frac{7}{8}.$$

Let random variable $Y = \sum_{j=1}^{d-1} \mathbf{1}(h(z_j) \neq \tau_j)$. Note that $Y \leq d-1$. Denote by $a = \mathbb{P}_\tau(Y > \frac{d-1}{4})$. We have

$$\mathbb{E}_\tau Y = \mathbb{E}_\tau[Y \mathbf{1}(Y > \frac{d-1}{4})] + \mathbb{E}_\tau[Y \mathbf{1}(Y \leq \frac{d-1}{4})] \leq (d-1)a + \frac{d-1}{4}(1-a) = \frac{(d-1)}{4} + \frac{3(d-1)}{4} a.$$

In conjunction with $\mathbb{E}_\tau Y > \frac{d-1}{2} \cdot \frac{7}{8}$, this implies that $a = \mathbb{P}_\tau(Y > \frac{d-1}{4}) > \frac{1}{4}$. Therefore,

$$\mathbb{P}_\tau(\text{err}(h, D_\tau) \geq 2\epsilon) > \frac{1}{4}.$$

Problem 5 (No need to submit)

In this problem, we develop an alternative proof of Sauer's Lemma: any hypotheis class $\mathcal{H}$ with VC dimension d can have at most $\binom{n}{\leq d}$ labelings on any dataset $S = \{z_1, \ldots, z_n\}$. Throughout, we will be using the notation that

$$\binom{\{1, \ldots, n\}}{d+1} \triangleq \left\{(i_1, \ldots, i_{d+1}) : 1 \leq i_1 < \ldots < i_{d+1} \leq n\right\}$$

to denote the set of $(d+1)$-tuples whose entries are distinct. Note that $\left|\binom{\{1,\ldots,n\}}{d+1}\right| = \binom{n}{d+1}$.

1. Show that for any indices $I = (i_1, \ldots, i_{d+1}) \in \binom{\{1,\ldots,n\}}{d+1}$, there exists a string $s_I \in \{\pm 1\}^{d+1}$, such that none of the labelings in
$$L_I = \left\{b \in \{\pm 1\}^n : (b_{i_1}, \ldots, b_{i_{d+1}}) = s_I\right\}$$
are achievable by classifiers in $\mathcal{H}$.

2. Show the following basic facts:
   (a) For a finite set A and a function f, denote by $f(A) = \left\{f(a) : a \in A\right\}$. Then $|f(A)| \leq |A|$, where $|B|$ denotes the cardinality of set B.
   (b) Suppose $\mathcal{I}$ is a set of indices. Given a collection of sets $\{L_I\}_{I\in\mathcal{I}}$ and a function f,
$$\left|\bigcup_{I\in\mathcal{I}} f(L_I)\right| \leq \left|\bigcup_{I\in\mathcal{I}} L_I\right|. \tag{3}$$

3. Use the above two facts to conclude that
$$\left|\bigcup_{I\in\binom{\{1,\ldots,n\}}{d+1}} L_I\right| \geq \sum_{i=d+1}^{n} \binom{n}{i}.$$
(Hint: consider functions $f_1, \ldots, f_n$, where $f_i(s_1, \ldots, s_n) = (s_1, \ldots, s_{i-1}, -1, s_{i+1}, \ldots, s_n)$ is the function that sets a length n string's i-th entry to -1. Iteratively applying Equation (3) for $f_1, \ldots, f_n$, what do you get?)

4. Use item 3 to conclude that $|\Pi_{\mathcal{H}}(S)| \leq \binom{n}{\leq d}$.

Solution

1. For any set of $d+1$ indices $I = i_1, \ldots, i_{d+1}$, observe that $(z_{i_1}, \ldots, z_{i_d})$ are not shatterable by $\mathcal{H}$. Therefore, there exists a string s_I such that for all h in $\mathcal{H}$,
$$(h(z_{i_1}), \ldots, h(z_{i_{d+1}})) \neq s_I.$$
Hence none of the labelings in
$$L_I = \left\{b \in \{\pm 1\}^n : (b_{i_1}, \ldots, b_{i_{d+1}}) = s_I\right\}$$
are achievable by classifiers in $\mathcal{H}$ on $\{z_1, \ldots, z_n\}$.

2. (a) follows from the simple observation that each element in A can induce at most one element in $f(A)$ (when two elements in A are mapped to the same element strict inequality is achieved). (b) follows from the observation that $\cup_{I\in\mathcal{I}} f(L_I) = f(\cup_{I\in\mathcal{I}} L_I)$ and (a).

3. We apply functions $f_1, \dots, f_n$ sequentially to L_I's, with the caveat that for the application of f_i, we only apply to L_I's such that $i \in I$. For example, if $n = 4$, $I = \{2, 3\}$, then we only apply f_2 and f_3 on L_I. Formally, define L_I^0 as L_I for all I, and for all i in $\{1, \dots, n\}$, $L_I^i = \begin{cases} f_i(L_I^{i-1}), & i \in I, \\ L_I^{i-1}, & i \notin I. \end{cases}$

   At step 1 (where we convert L_I^0's to L_I^1's), define $A_1 = \cup_{I \in \mathcal{I}: 1 \in I} L_I^0$, and $B_1 = \cup_{I \in \mathcal{I}: 1 \notin I} L_I^0$. Note that B_1 is closed under negation on the first coordinate: if $a \in B$, then a^1 (the string that negates the first coordinate on a) will also be in a.

   Therefore, for a string s, $s \notin B_1$ if and only if $f_1(s) \notin B_1$. This implies that $f_1(A_1) - B_1 \subset f_1(A_1 - B_1)$. Therefore,
$$|f_1(A_1) - B_1| \leq |f_1(A_1 - B_1)| \leq |A_1 - B_1|.$$

   Furthermore,
$$|f_1(A_1) \cup B_1| = |B_1| + |f_1(A_1) - B_1| \leq |B_1| + |A_1 - B_1| = |A_1 \cup B_1|.$$

   we have $| \cup_{I \in \mathcal{I}} L_I^1| = |f_1(A_1) \cup B_1| \leq |A_1 \cup B_1| = | \cup_{i \in \mathcal{I}} L_I^0|$.

   Similarly, we have that $| \cup_{I \in \mathcal{I}} L_I^i| \leq | \cup_{I \in \mathcal{I}} L_I^{i-1}|$ for all i in $\{1, \dots, n\}$. In addition, observe that $\cup_{I \in \mathcal{I}} L_I^n = \left\{(b_1, \dots, b_n) : \# \left\{i : b_i = -1\right\} \geq d + 1\right\}$, which has size $\binom{n}{\geq d+1}$. This concludes that
$$\binom{n}{\geq d+1} \leq | \cup_{I \in \mathcal{I}} L_I^n| \leq \dots \leq | \cup_{I \in \mathcal{I}} L_I^1| \leq | \cup_{i \in \mathcal{I}} L_I|.$$

4. The number "forbidden" patterns, i.e. patterns unachievable by $\mathcal{H}$ is $|\cup_{i \in \mathcal{I}} L_I| \geq \binom{n}{\geq d+1}$. This implies that the number of allowed patterns by $\mathcal{H}$ is at most $2^n - \binom{n}{\geq d+1} = \binom{n}{\leq d}$.