Homework Assignment 5

Due in class, Monday April 23rd

SDS 384-11 Theoretical Statistics

1. (VC dimension) Compute the VC dimension of the following function classes

   (a) Circles in $\mathcal{R}^2$

   For any three points which are not collinear , we can easily draw a circle that includes all three of them, any two of them, any one of them, or none of them, so VC dimension is at least 3.

   However, for any set of four points, they are not shattered. We show this by constructing a counterexample in several cases:

      i. Collinear: the labeling +-+- (going along the line) is impossible, among numerous others.

      ii. Convex hull is a triangle: then the labeling with +(the three points of the triangle) and -(the interior point) is not possible.

      iii. Convex hull is a quadrilateral: let (x_1, x_2) be the points separated by the long diagonal and (y_1, y_2) be the points separated by the short diagonal. At least one of the labelings $\{+x_1, +x_2, -y_1, -y_2\}$, $\{-x_1, -x_2, +y_1, +y_2\}$ will not be achieved. If they were both possible, then this would mean that the circles satisfying the two labelings can have four non-overlapping regions, which is not possible. (is it possible with ellipses?)Since some set of 3 points is shattered by the class of circles, and no set of 4 points is, the VC dimension of the class of circles is 3. Note that

   (b) Axis aligned rectangles in $\mathcal{R}^2$ It is easy to see that one can shatter four points. Consider 5 points and the following cases.

      i. They are all collinear. In which case, a trivial alternative labels cannot be shattered by axis aligned rectangles.

      ii. If they are not all collinear, then draw the largest rectangle through the largest and smallest x and y coordinates. Either all five points are on this rectangle, or one is inside. In the first case, do an alternative labeling, this cannot be shattered by a axis aligned rectangle. In the second case, label everyone on the rectangle as one label, and the one inside with the opposite label.

   (c) Axis aligned squares in $\mathcal{R}^2$ It is easy to construct 3 points which are shattered. Let us take 4 points. Let the leftmost point be L, rightmost R, top one T and bottom one B. Draw a rectangle like last question through these points. If there are ties, then it is easy to label them so that they cannot be shattered. So let us think about the case where there are no ties. In this case, if the rectangle is a square, i.e. the distance between x coordinates of L and R (d_x) is equal to the distance between y coordinates of T and B (d_y) are such that $d_x \geq d_y$, then (L+, R+, T-, B-) cannot be shattered. If $d_x < d_y$ then (L-, R-, T+, B+) cannot be shattered. So VC dimension is 4.

2. In class, you upper bounded the Rademacher complexity of a function class. Now you will derive a lower bound.

   (a) For function classes $\mathcal{F}$ with function values in $[0,1]$, prove that $E\|\hat{P}_n - P\|_{\mathcal{F}} \geq \frac{\mathcal{R}_{\mathcal{F}}}{2} - \sqrt{\frac{\log 2}{2n}}$. *Hint: may be it is easier to start from* $\mathcal{R}_{\mathcal{F}}$ *and show that* $\mathcal{R}_F \leq 2E\|\hat{P}_n - P\|_{\mathcal{F}} + \sqrt{\frac{2\log 2}{n}}$. *In order to do this, you would need to add and subtract* $E[f(X)]$ *and then use triangle inequality.*

$$
\begin{aligned}
\mathcal{R}_{\mathcal{F}} &= \frac{1}{n} E \sup_{f \in \mathcal{F}} \sum_i \epsilon_i f(X_i) \\
&= \frac{1}{n} E \sup_{f \in \mathcal{F}} \left(\sum_i \epsilon_i \left(f(X_i) - E[f(X_i)]\right) + \sum_i \epsilon_i E[f(X_1)] \right) \\
&\leq \underbrace{E \sup_{f \in \mathcal{F}} \sum_i \epsilon_i \left(f(X_i) - E[f(X_i)]\right)}_{A} + \underbrace{\frac{|\sum_i \epsilon_i|}{n}}_{B}
\end{aligned}
$$

   The first part can be bounded as:

$$
\begin{aligned}
(A) &= E \sup_{f \in \mathcal{F}} \sum_i \epsilon_i \left(f(X_i) - E[f(X_i')]\right) \\
&= E \sup_{f \in \mathcal{F}} \sum_i \epsilon_i E_{X'} \left(f(X_i) - f(X_i')\right) \\
&\leq E_{X,X',\epsilon} \sup_{f \in \mathcal{F}} \sum_i \epsilon_i \left(f(X_i) - f(X_i')\right) \\
&= E_{X,X'} \sup_{f \in \mathcal{F}} \sum_i \left(f(X_i) - Ef(X_i) + Ef(X_i') - f(X_i')\right) \\
&\leq 2E\|\hat{P}_n - P_n\|
\end{aligned}
$$

   As for the part (B) we use the finite class lemma.

   (b) Now prove that $\|P - \hat{P}_n\|_{\mathcal{F}} \geq E\|P - \hat{P}_n\|_{\mathcal{F}} - \epsilon$ with probability at least $1 - \exp(-cn\epsilon^2)$ for some constant c. This can be done easily shown using McDiarmid's inequality and the fact that the bounded differences are upper bounded by $2/n$.

   (c) Recall the class of all subsets with finite size in $[0,1]$? Prove that then Rademacher complexity of this class is at least $1/2$. What does this imply?
   Let $\mathcal{S} := \{S \subset [0,1] | S is finite\}$ and $\mathcal{F}_{\mathcal{S}} = \{1_S : S \in \mathcal{S}\}$. Furthermore, $\mathcal{R}_{\mathcal{F}} = 1/nE\sup_{S \in \mathcal{S}} |\sum_i \epsilon_i 1_S(X_i)|$. Now we either have $\sum_i 1(\epsilon_i = 1) \geq n/2$ or we have $\sum_i 1(\epsilon_i = -1) \geq n/2$. Then the rademacher complexity is maximized at $S_+ = \{i : 1(\epsilon_i = 1)\}$ or $S_- = \{i : 1(\epsilon_i = -1)\}$. The maximum is larger than $1/2$. This combined with the previous lower bound suggests that for this function class the ULLN does not hold and so $\mathcal{F}_{\mathcal{S}}$ is not a GC class.

3. In this exercise, we explore the connection between VC dimension and metric entropy. Given a set class $\mathcal{S}$ with finite VC dimension ν, we show that the function class $\mathcal{F}_S := 1_S, S \in \mathcal{S}$ of indicator functions has metric entropy at most

$$N(\delta; \mathcal{F}_{\mathcal{S}}, L^1(P)) \leq \left(\frac{K \log(3e/\delta)}{\delta}\right)^{\nu} \qquad \text{For a constant } K \tag{1}$$

Let $\{1_{S^1}, \ldots, 1_{S^N}\}$ be a maximal delta packing in the $L^1(P)$ norm, so that:

$$\|1_{S_i} - 1_{S_j}\|_1 = E[|1_{S_i}(X) - 1_{S_j}(X)|] > \delta \qquad \text{for all } i \neq j$$

This is an upper bound on the δ covering number. **This theorem is due to Dudley, Haussler and the proof is inspired by "Empirical processes: theory and applications" by Jon Wellner.

(a) Suppose that we generate n samples $X_i, i = 1, \ldots, n$ drawn i.i.d. from P. Show that the probability that every set S_i picks out a different subset of $\{X_1, \ldots, X_n\}$ is at least $1 - \binom{N}{2}(1-\delta)^n$.

$$\begin{aligned}
P(S_i \text{ pick different subsets}) &= P(\forall i \neq j S_i \neq S_j) \\
&= 1 - P(\exists i \neq j S_i = S_j) \\
&= 1 - \binom{N}{2} P(\forall k \in [n]\ 1(k \in S_i) = 1(k \in S_j)) \\
&= 1 - \binom{N}{2}(1 - P(|1_{S_i}(X_1) - 1_{S_j}(X_1)| = 1))^n \\
&= 1 - \binom{N}{2}(1 - E|1_{S_i}(X_1) - 1_{S_j}(X_1)|) \\
&\leq 1 - \binom{N}{2}(1-\delta)^n
\end{aligned}$$

(b) Using part (a), show that for $N \geq 2$ and $n = \lceil 2 \log N/\delta \rceil$, there exists a set of n points from which $\mathcal{S}$ picks out at least N subsets, and conclude that $N \leq \left(\frac{3e \log N}{\nu\delta}\right)^{\nu}$.
In order to have $1 - \binom{N}{2}(1-\delta)^n > 1 - N^2 e^{-n\delta} > 0$ we need $n \geq 2 \log N/\delta$. Since the probability that P(S_i pick different subsets) is bigger than zero, there exists a set of n points so that $\mathcal{F}$ picks up N different subsets. Clearly, $N \leq \Pi_{\mathcal{F}}(n) \leq (en/\nu)^{\nu}$. Plugging in the value of n gives the answer.

(c) Use part (b) to show that Eq (1) holds with $K := 3e^2/(e-1)$. *Hint: Note that you have* $\frac{N^{1/\nu}}{\log N} \leq \frac{3e}{\nu\delta}$. *Let* $g(x) = x/\log x$. *We are solving for* $g(m^{1/\nu}) \leq 3e/\delta$. *Prove that* $g(x) \leq y$ *implies* $x \leq \frac{e}{e-1} y \log y$.
We have:

$$g(N^{1/\nu}) \leq \frac{3e}{\delta}$$

Since $g(x)$ is minimized by $x = e$ and is increasing, $y \geq g(x)$ for $x \geq e$ imply,

$$\begin{aligned}
\log y \geq \log x - \log\log x = \log x\,(1 - \log\log x/\log x) &\geq \log x(1 - 1/e) \\
x \leq y \log x < y \log y(1 - 1/e)^{-1}
\end{aligned}$$

Hence, $N^{1/\nu} \leq \frac{3e^2}{\delta(e-1)} \log(3e\delta)$, i.e.

$$N(\delta; \mathcal{F}_{\mathcal{S}}, L^1(P)) \leq \left(\frac{K \log(3e/\delta)}{\delta}\right)^{\nu} \quad \text{For a constant } K \tag{2}$$

where $K = 3e^2/(e-1)$