Idan Drori - 205983406

Theory Questions

1. PAC learnability of ℓ_2-balls around the origin.

Given a real number $R \geq 0$ define the hypothesis $h_R : \mathbb{R}^d \to \{0,1\}$ by,

$$h_R(x) = \begin{cases} 1 & \|x\|_2 \leq R \\ 0 & \text{otherwise} \end{cases}$$

Consider the hypothesis class $\mathcal{H}_{\text{ball}} = \{h_R | R \geq 0\}$. Prove directly that $\mathcal{H}_{\text{ball}}$ is PAC learnable in the realizable case. How does the sample complexity depend on the dimension d? Explain.

Solution: Given a sample $S = \{(x_1, h_R(x_1)), ..., (x_n, h_R(x_n))\}$, we'll define an ERM algorithm A as follows: $A(S) := h_s$, where $h_s = \max_{x \in [x_i]_{i=1}^n} \|x\|_2 \, | h_R(x) = 1$.

We'll show that $\mathcal{H}_{ball}$ is PAC learnable by A with a sample complexity of $N(\varepsilon, \delta) = -\frac{\ln \delta}{\varepsilon}$.

First, it follows that $e_P(A(S)) = P[B_R \, B_s]$ and so $e_P(A(S)) > \varepsilon \Leftrightarrow P[B_R \, B_s] > \varepsilon$. Because P is continuous, there exists some radius $R_\varepsilon \in \mathbb{R}$ such that $P[R_\varepsilon \leq \|X\| \leq R] = P[B_R \, B_{R_\varepsilon}] = \varepsilon$. So, if there's at least one point $x_t \in S$ such that $R_\varepsilon \leq \|x_t\| \leq R$ then $h_R(x_t) = 1$, and then by the definition of A, $s \geq R_\varepsilon$ and $B_{R_\varepsilon} \subseteq B_s$, and so $P[B_R \, B_s] \leq P[B_R \, B_{R_\varepsilon}]$. This condition is equivalent to:

$$P[e_P(A(S)) > \varepsilon] = P[\|x_1\|, ..., \|x_n\| \leq R_\varepsilon] = \prod_{i=1}^{n} (1 - \varepsilon) = (1-\varepsilon)^n \leq e^{-\varepsilon n}$$

And so

$$e^{-\varepsilon n} \leq \delta \Leftrightarrow n > -\frac{\ln \delta}{\varepsilon}$$

And from here we get:

$$n > -\frac{\ln \delta}{\varepsilon} \Rightarrow e^{-\varepsilon n} \leq \delta \Rightarrow P[e_P(A(S)) > \varepsilon] < \delta$$

As required. And as we can see, the sample complexity is not affected by d. $\square$

2. PAC in expectation.

Consider learning in the realizable case. We say a hypothesis class $\mathcal{H}$ is ***PAC learnable in expectation*** using algorithm A if there exists a function $N(a) : (0,1) \to \mathbb{N}$ such that $\forall a \in (0,1)$ and for any distribution P (realizable by $\mathcal{H}$), given a sample set S such that $|S| \geq N(a)$, it holds that,

$$\mathbb{E}[e_p(A(S))] \leq a$$

Show that $\mathcal{H}$ is PAC learnable *if and only if* $\mathcal{H}$ is PAC learnable in expectation.

Solution: ($\Leftarrow$) We assume that $\mathcal{H}$ is PAC learnable in expectation. Then there exists a function $N(a)$ such that for all ε, δ if the sample size is larger than $N(\varepsilon \cdot \delta)$ then $\mathbb{E}[e_P(A(S))] \leq \varepsilon \cdot \delta$. We define $N'(\varepsilon, \delta) = N(\varepsilon \cdot \delta)$, and from Markov's inequality we get that for $n \geq N'(\varepsilon, \delta)$:

$$P[e_P(A(S)) > \varepsilon] \leq P[e_P(A(S)) \geq \varepsilon] \leq \frac{\mathbb{E}[e_P(A(S))]}{\varepsilon} \leq \frac{\varepsilon \cdot \delta}{\varepsilon} = \delta$$

($\Rightarrow$) We assume that $\mathcal{H}$ is PAC learnable with A. Let N_p be a complexity function of $\mathcal{H}$, and we'll define $N(a) = N_p\left(\frac{a}{2}, \frac{a}{2}\right)$. Then for $|S| > N(a)$ we get:

$$\begin{aligned}\mathbb{E}[e_P(A(S))] = \mathbb{E}&\left[e_P(A(S)) | e_P(A(S)) < \frac{a}{2}\right] \cdot P\left[e_P(A(S)) < \frac{a}{2}\right] \\ &+ \mathbb{E}\left[e_P(A(S)) | e_P(A(S)) \geq \frac{a}{2}\right] \cdot P\left[e_P(A(S)) \geq \frac{a}{2}\right]\end{aligned}$$

Because $|S| \geq N\left(\frac{a}{2}, \frac{a}{2}\right)$ we get that:

$$\begin{aligned}&P\left[e_P(A(S)) \geq \frac{a}{2}\right] \leq \frac{a}{2} \\ &\mathbb{E}\left[e_P(A(S)) | e_P(A(S)) \geq \frac{a}{2}\right], P\left[e_P(A(S)) < \frac{a}{2}\right] \leq 1 \\ &\mathbb{E}\left[e_P(A(S)) | e_P(A(S)) < \frac{a}{2}\right] \leq \frac{a}{2}\end{aligned}$$

And finally we get that:

$$\mathbb{E}[e_P(A(S))] \leq \frac{a}{2} + \frac{a}{2} = a$$

So $\mathcal{H}$ is PAC learnable in expectation.

$\square$

3. Union of intervals.

Determine the VC-dimension of $\mathcal{H}_k$ - the subsets of the real line formed by the union of k intervals. Prove your answer.

Solution: We'll show that $VCdim(\mathcal{H}_k) = 2k$.

We first show that $VCdim(\mathcal{H}_k) \geq 2k$:

Let $C_n = \left\{\frac{i}{2k}\right\}_{i=1}^{2k}$. Let $(s_i)_{i=1}^{2k}$ be a dichotomy such that $s_i \in \{0, 1\}$ for all $i \in [n]$. Let $\varepsilon = \frac{1}{2k} \cdot \frac{1}{2}$:

$$\overline{I} = \bigcup_{i \in [k] \wedge s_0 = 0} \left[\frac{i}{n} - \varepsilon, \frac{1}{n} + \varepsilon\right]$$

Let $I = \overline{\overline{I}}$, we'll show that for $h_{I \cup \partial(I)}$ for all $i \in [n]$ it follows that $h_{I \cup \partial(I)}(C_i) = s_i$ (where $\partial(I)$ is the boundary of I). For $i \in [2k]$, if $s_i = 0$ then $C_i \in \left(\frac{i}{n} - \varepsilon, \frac{i}{n} + \varepsilon\right) \subseteq \overline{I}$, and so $C_i \notin I \cup \partial(I)$ and $h_{I \cup \partial(I)}(C_i) = 0 = s_i$. If $s_i = 1$ then $C_i \in \left(\frac{i}{n} - \varepsilon, \frac{i}{n} + \varepsilon\right) \not\subseteq \overline{I}$, this is because the intersection between the intervals that define $\overline{I}$ are only on the boundaries, and so $h_{I \cup \partial(I)}(C_i) = 1 = s_i$. If we assume the dichotomy has m zeros, the the number of intervals of I is at most $2m$, because every interval added to $\overline{I}$ can either split an existing interval into two intervals, or it'll join with another interval in $\overline{I}$, so the total number of interval will stay the same. Meaning, #Intervals in I $\leq 2m \leq 2k$.

So we get that $\mathcal{H}$ shatters the set C, then $VCdim(\mathcal{H}) \geq |C| = 2k$.
We'll now show that $VCdim(\mathcal{H}) \leq 2k$:

Let $C = \{c_1, ..., c_{2k+1}\}$, and let $s = (s_1, ..., s_{2k+1})$ be the dichotomy such that $s_i = \begin{cases} 0 & i \text{ is even} \\ 1 & i \text{ is odd} \end{cases}$.

We assume for the sake of contradiction that there exists h_I such that $s = (h_I(c_1), ..., h_I(c_{2k+1}))$. I has k intervals, but no single interval can have two points in C because they're separated by another non-empty interval. But C has $k+1$ points that are in I, and so there must be two points that belong to the same interval, in contradiction. $\square$

4. Inhomogeneous linear classifiers.

Prove that the VC-dimension of $\mathcal{H}_d$, the class of inhomogeneous linear classifiers in $\mathbb{R}^d$, is $d+1$. $\mathcal{H}_d$ is the class of hypotheses of the form

$$h_{w,b}(x) = sign(w \cdot x - b),$$

where $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$.

Solution: We'll show that $VCdim(\mathcal{H}_d) \geq d+1$:

Let $C = (e_1, ..., e_d, 0)$, let $S = (s_1, ..., s_{d+1})$ be some dichotomy, Let $b = -s_{d+1} + \frac{1}{2}, w = (2s_1 - 1, ..., 2s_d - 1)$. For all $i \in [d]$:

$$h_{w,b}(e_i) = sign(w \cdot e_i + s_{d+1}) = sign(2s_i - 1 + b) = s_i$$

This because if $s_i = 1$ then $2s_i - 1 - b = 1 - b \geq \frac{1}{2}$, and if $s_i = 0$ then $2s_i - 1 = -1 - b \leq -\frac{1}{2}$. And $h_{w,b}(0) = sign(0 - b) = sign(-b) = sign\left(s_{d+1} - \frac{1}{2}\right) = sign(s_{d+1})$. We get that $\mathcal{H}_d$ shatters C, and so $VCdim(\mathcal{H}_d) \geq d+1$. We'll show that $VCdim(\mathcal{H}_d) \leq d+1$:

Let $C = (x_1, ..., x_{d+2})$, and we'll show that C is not shattered. We assume for the sake of contradiction that C is shattered. Then for every $S = (s_1, ..., s_{d+2})$ there exists $b \in \mathbb{R}, w \in \mathbb{R}^d$ such that $h_{w,b}(x_i) = s_i$. Let $C' = (y_1, ..., y_{d+2})$ where $y_i = \begin{pmatrix} x_i \\ -1 \end{pmatrix} \in \mathbb{R}^{d+1}$. Then we get that $C' \subset \mathbb{R}^{d+1}$ while it has $d+2$ elements, and so C' is linearly dependant, so for some $a_i \in \mathbb{R}$:

$$a_{d+2} y_{d+2} = \sum_{i \in [d+1]} a_i y_i$$

We'll assume without loss of generality that $a_{d+2} = 1$. Then we'll define $S' = (s_1, ..., s_{d+2})$ where for each $i \in [d+1] : s_i = 1 \Leftrightarrow a_i \geq 0$ and $s_{d+2} = 0$. We've assumed that there exists some $h_{w,b}$ such that:

$$(s_1, ..., s_{d+1}, 0) = (h(x_1), ..., h(x_{d+2}))$$

But,

$$(w, b) \cdot y_{d+2} = (w, b) \cdot \sum_{i \in [d+1]} a_i \cdot \begin{pmatrix} x_i \\ -1 \end{pmatrix} = \sum_{i \in [d+1]} a_i \cdot (w, b) \cdot \begin{pmatrix} x_i \\ -1 \end{pmatrix} = \sum_{i \in [d+1]} a_i \cdot (wx_i - b)$$

And because $s_i = 1 \Leftrightarrow a_i \geq 0$ we get that:

$$\sum_{i \in [d+1]} a_i \cdot (wx_i - b) \geq 0)$$

And so $h(x_{d+2}) = 1$, in contradiction. $\square$

5. Prediction by polynomials.

Given a polynomial $p : \mathbb{R} \to \mathbb{R}$ define the hypothesis $h_p : \mathbb{R}^2 \to \{0,1\}$ by,

$$h_p(x_1, x_2) = \begin{cases} 1 & p(x_1) \geq x_2 \\ 0 & \text{otherwise} \end{cases}$$

Determine the VC-dimension of $\mathcal{H}_{\text{poly}} = \{h_p | p \text{ is a polynomial}\}$. You can use the fact that given n distinct values $x_1, ..., x_n \in \mathbb{R}$ and $z_1, ..., z_n \in \mathbb{R}$ there exists a polynomial p of degree $n-1$ such that $p(x_i) = z_i$ for every $1 \leq i \leq n$.
Solution: We'll show that $VCdim(\mathcal{H}_{poly}) = \infty$. Let $n \in \mathbb{N}$, let $C = \{(1,1), ..., (n,n)\}$, well denote $x_i = (i,i)$. We'll show that $\mathcal{H}_{poly}$ shatters C.
Let $S = (s_1, ..., s_n)$ by some dichotomy. Then there exists some polynomial P such that $P(i) = i + |1 - s_i|$ for all $i \in [n]$. And so for every $i \in [n]$ it follows that $P(i) = \begin{cases} i & s_i = 1 \\ i-1 & s_i = 0 \end{cases}$, and so we get:

$$h_P(x_i) = \begin{cases} 1 & P(i) \geq i \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 1 & s_i = 1 \\ 0 & s_i = 0 \end{cases}$$

And so C is shattered by $\mathcal{H}_{poly}$. So for every $n \in \mathbb{N}$ there exists some C of size n that's shattered by $\mathcal{H}_{poly}$, then we conclude $VCdim(\mathcal{H}_{poly}) = \infty$ $\square$

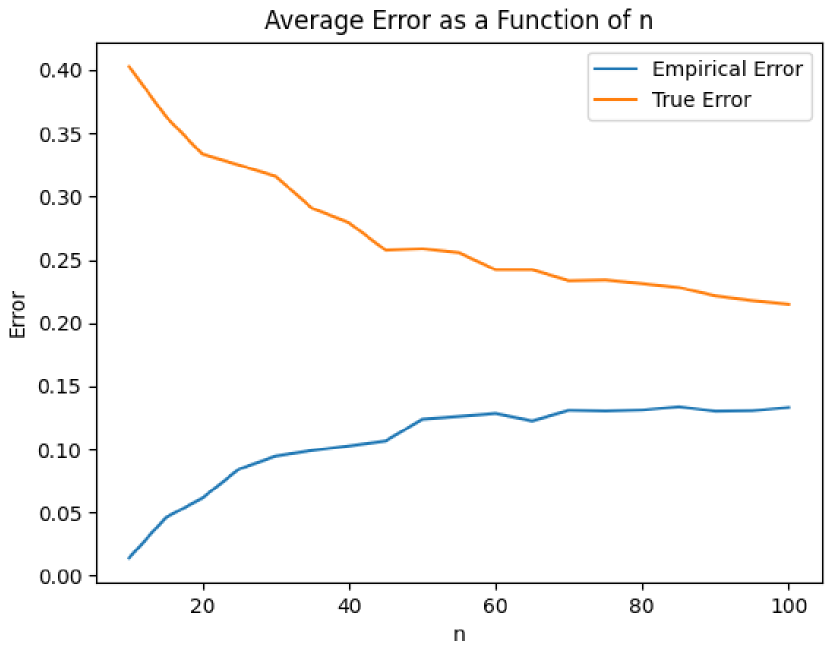
Programming Assignment

(a) We've seen that for a binary Y with zero-one loss, the optimal h which minimizes $e_p(h)$ is a Maximum-A-Posteriori classifier. With the given probability, the maximal $P[Y = 1 \mid X = x]$ is given when $x \in [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1]$, and so we get:

$$h(x) = \underset{h \in \mathcal{H}_{10}}{\arg\min}\, e_P(h) = \underset{y \in \{0,1\}}{\arg\max}\, P[Y = y \mid X = x] = \begin{cases} 1 & , x \in [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1] \\ 0 & , \text{Otherwise} \end{cases}$$

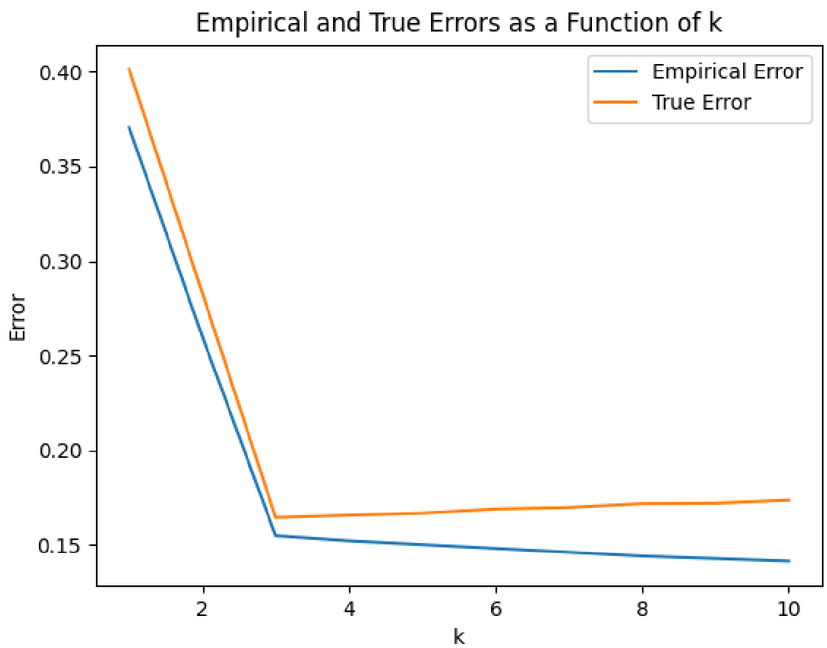
There are less than 10 intervals, so $h \in \mathcal{H}_{10}$.

(b) Plot:



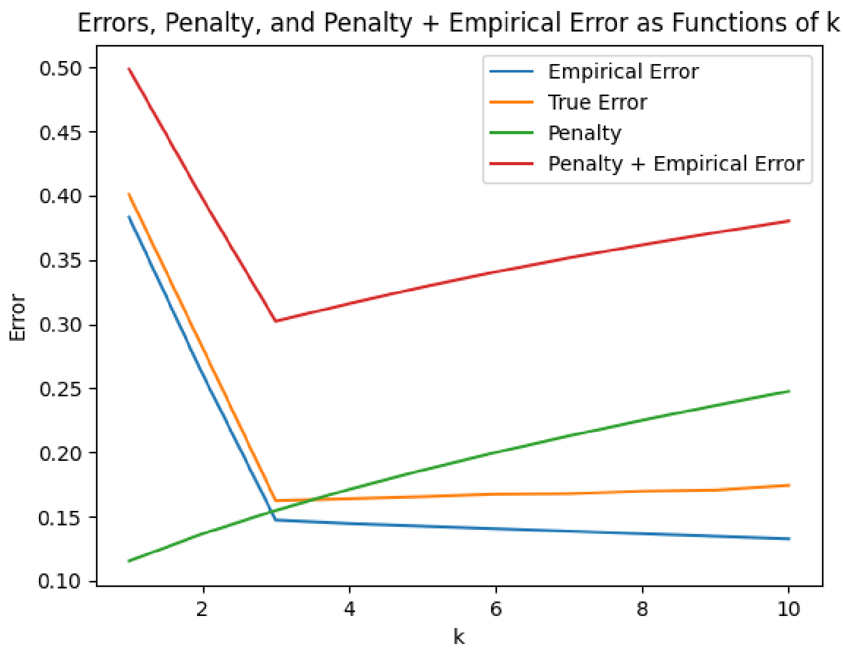
We can see that the empirical error grows as n grows, and at the same time the true error diminishes. The empirical error grows as n grows, because the chance to get a low probability label from P increases with more samples. On the other hand, the true error diminishes because with more samples, we become "more representative" of the actual distribution P.

(c) Plot:



As we can see, both the empirical and true error drop sharply until $k = 3$, and from that point the true error climbs slightly, and the empirical error keeps going down at a gradual pace. The MAP is in $\mathcal{H}_{10}$ and is made out of 3 intervals, so it makes sense that the best true error is at $k = 3$. When $k \leq 3$ we probably have some underfitting, because the actual distribution has 3 intervals. And when $k \geq 3$ the empirical error going down is probably a case of overfitting, again, because the actual distribution has 3 intervals, which also causes the true error to rise.

(d) Plot:



The empirical and true errors behave in the same way as the previous question. The penalty grows with k because $VCdim(\mathcal{H}_k) = 2k$, so $2 \cdot \sqrt{\frac{VCdim(\mathcal{H}_k)+\ln\frac{2}{0.1}}{n}} = 2 \cdot \sqrt{\frac{2k+\ln\frac{2}{0.1}}{n}}$, meaning it grows similarly to $\sqrt{k}$. And finally we can that the minimum of Penalty+Empirical Error does happen at $k = 3$, as expected considering the best hypothesis from (a).

(e) Using holdout validation we get the best hypothesis at $k = 3$. We showed the in previous question that the hypothesis with the lowest true error is one where $k = 3$, so this is the result we expect. The best hypothesis we got was:

$$(0.001772200820015557, 0.20104033369187824),$$
$$(0.40203239381634387, 0.6008933816014819),$$
$$(0.8000421753784073, 0.9988574437000303)$$

Which is quite close to the MAP we showed in (a) to be optimal.