Introduction to Machine Learning-Exercise 2

Odai Agbaria

Theory Questions

1. **(15 points) 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,

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

Consider the hypothesis class $\mathcal{H}_{ball} = \{h_R \mid R \geq 0\}$. Prove directly (without using the Fundamental Theorem of PAC Learning) that $\mathcal{H}_{ball}$ is PAC learnable in the realizable case (assume for simplicity that the marginal distribution of X is continuous). How does the sample complexity depend on the dimension d? Explain.

Proof:

To prove that that H_{ball} is PAC learnable by an algorithm A(which we will define in the next few lines), we are going to prove that for every ε, δ there exists $N(\varepsilon, \delta)$ such that for every training set S consist of $n \geq N(\varepsilon, \delta)$ it holds that: $P[e_p(A(S)) > \varepsilon] < \delta$.

The algorithm A:

Given a set of data points $S = (X_i, Y_i)$ of size n, choose h_{RA} such that $R_A = max_{Xi\ such\ that\ Yi=1}||Xi||_2$, intuitively we return the smallest ball that is consistent with the data(such a ball's radius is as said above is the maximum distance of the origin among the data points with $Yi = 1$) .

***H_{ball}* is PAC learnable by A:**

Let it be ε, δ>0.

H_{ball} is realizable, hence there exists R' such that $e_p(h_{R'})$=0. Notice that $R_A \leq R'$, that's almost immediately by our algorithm, cause if $e_p(h_{R'})$=0 then for every Xi with Yi=1 then it holds that $||Xi||_2 \leq R'$, in particular $max_{Xi\ such\ that\ Yi=1}||Xi||_2 = R_A \leq R'$.

then, lets notice that by our algorithm A if Y=0 then our chosen hypothesis certainly predicts Y=0, and it can't be wrong cause if Y=0 then $||X||_2 > R' > R_A$, thus the only way to predict some label wrongly is that for those X with Y=1 that have distances of the origin more than R_A but less than R', formally:

$$ep(A(S)) = P\left[R_A < ||X||_2 \leq R'\right] =_{X\ is\ continuous} P\left[||X||_2 \leq R'\right] - P[||X||_2 \leq R_A]$$

If $P\left[||X||_2 \leq R'\right] \leq \varepsilon$ then:

$$ep(A(S)) = P\left[R_A < ||X||_2 \leq R'\right] = P\left[||X||_2 \leq R'\right] - P\left[||X||_2 \leq R_A\right] \leq \varepsilon - (+) \leq \varepsilon$$

Meaning that in this case $P[e_p(A(S)) > \varepsilon] = 0 < \delta$. this holds for any $N(\varepsilon, \delta)$.

Else, since the marginal distribution of X is continuous then there exists R_ε such that $P\left[R_\varepsilon \le \left||X|\right|_2 \le R'\right] \le \varepsilon$, meaning that for every epsilon there is a radius R_ε which holds that the probability of x to be inside the big ball defined by the radius R' but not inside the small ball defined with R_ε is at most ε.

So now we need to prove that the space between the balls defined with R_A and R' is contained in the space between the balls defined with R_ε and R' with probability at least 1- δ.

We can consider two cases:

- there exists x in S such that: $R_\varepsilon \le \left||X|\right|_2 \le R_A$: in this case it holds immediately that:
  $R_\varepsilon \le R_A \le R' \rightarrow ep(A(S)) = P\left[R_A < \left||X|\right|_2 \le R'\right] \le P\left[R_\varepsilon < \left||X|\right|_2 \le R'\right] = \varepsilon$

  Meaning that in this case $P\left[e_p(A(S)) > \varepsilon\right] = 0 < \delta$. this holds for any $N(\varepsilon, \delta)$.

- there is no such x:
  this is the only case where $e_p(A(S)) > \varepsilon$, so here we want to choose $N(\varepsilon, \delta)$ to guarantee that this $P\left[e_p(A(S)) > \varepsilon\right]$ *is at most* δ.

$$P\left[e_p(A(S)) > \varepsilon\right] = P[\forall i: X_i < R_\varepsilon \text{ or } X_i > R'] = (1-\varepsilon)^n \le e^{-\varepsilon n}$$

We choose R_ε such that $P\left[R_\varepsilon \le \left||X|\right|_2 \le R'\right] \le \varepsilon$

For every a>0: $1 - a \le e^{-a}$

$P\left[e_p(A(S)) > \varepsilon\right] \le e^{-\varepsilon n} < \delta \rightarrow$ *we should choose* $N(\varepsilon, \delta)$ *to be any n greater than:* $\frac{\ln\left(\frac{1}{\delta}\right)}{\varepsilon}$

because: $e^{-\varepsilon n} < \delta \rightarrow \frac{1}{e^{\varepsilon n}} < \delta \rightarrow \frac{1}{\delta} < e^{\varepsilon n} \rightarrow \ln\left(\frac{1}{\delta}\right) < \varepsilon n \rightarrow \frac{\ln\left(\frac{1}{\delta}\right)}{\varepsilon} < n.$

And thus H_{ball} is PAC learnable by A. and we can see that the sample complexity does not depend on d, I think the reason of that because this is a Cartesian look at the hypothesis class, but we can look at it in another way-the polar look- and then we can say that the shape and the size of the ball is not affected by the number of dimensions, it is affected only by the radius R.s ■

2. **(15 points) 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 $|S| \ge N(a)$, it holds that,

$$\mathbb{E}\left[e_P(A(S))\right] \le a.$$

Show that $\mathcal{H}$ is PAC learnable *if and only if* $\mathcal{H}$ is PAC learnable in expectation (Hint: For one direction, use the law of total expectation. For the other direction, use Markov's inequality).

Proof:

First Direction: *H is PAC learnable → H is PAC learnable in expectation*:

Assume H is PAC learnable by an algorithm A then for every $\varepsilon, \delta \in (0,1)$ there exists $N(\varepsilon, \delta)$ such that for every training set S which contains at least $N(\varepsilon, \delta)$ samples then:

$$P\left[ep\left(A(S)\right) > \varepsilon\right] < \delta$$

We will prove that H is PAC learnable in expectation by the same algorithm A.

Let $a \in (0,1)$.

Consider $N(a) = N(\frac{a}{2}, \frac{a}{2})$. then by the law of total expectation:

$$E\left[ep\left(A(S)\right)\right] = E\left[ep\left(A(S)\right)\middle|ep\left(A(S)\right) > \frac{a}{2}\right] * P\left(p\left(A(S)\right) > \frac{a}{2}\right) + E\left[ep\left(A(S)\right)\middle|ep\left(A(S)\right) \leq \frac{a}{2}\right] * P\left(p\left(A(S)\right) \leq \frac{a}{2}\right) \leq P\left(p\left(A(S)\right) > \frac{a}{2}\right) + E\left[ep\left(A(S)\right)\middle|ep\left(A(S)\right) \leq \frac{a}{2}\right]$$

The last transition is due to the fact that $E\left[ep\left(A(S)\right)\middle|ep\left(A(S)\right) \leq \frac{a}{2}\right]$ was multiplied by $P\left(p\left(A(S)\right) \leq \frac{a}{2}\right) \leq 1$, and to that $P\left(p\left(A(S)\right) > \frac{a}{2}\right)$ was multiplied by $E\left[ep\left(A(S)\right)\middle|ep\left(A(S)\right) > \frac{a}{2}\right] \leq 1$ because the true error is at most 1 since we are using the zero-one loss.

Now, notice that since H is PAC learnable via A then $P\left(p\left(A(S)\right) > \varepsilon = \frac{a}{2}\right) < \delta = \frac{a}{2}$, that's by our choice of $N(\varepsilon, \delta)$, and that $E\left[ep\left(A(S)\right)\middle|ep\left(A(S)\right) \leq \frac{a}{2}\right] \leq \frac{a}{2}$ derived immediately by the condition that $ep\left(A(S)\right) \leq \frac{a}{2}$, so we get:

$$E\left[ep\left(A(S)\right)\right] \leq P\left(p\left(A(S)\right) > \frac{a}{2}\right) + E\left[ep\left(A(S)\right)\middle|ep\left(A(S)\right) \leq \frac{a}{2}\right] \leq \frac{a}{2} + \frac{a}{2} = a$$

Then H is PAC learnable in expectation by A.

Second Direction: *H is PAC learnable in expectation → H is PAC learnable*:

Assume H is PAC learnable in expectation by an algorithm A, then for every $a \in (0,1)$ there exists a $N(a)$ such that for every training set S which contains at least $N(a)$ samples then:

$$E\left[ep\left(A(S)\right)\right] \leq a$$

We will prove that H is PAC learnable by the same algorithm A.

Let $\varepsilon, \delta \in (0,1)$.

Consider $N(\varepsilon, \delta) = N\left(\frac{\varepsilon * \delta}{2}\right)$.

Then using Markov's inequality (we can use it since ep(A(S)) is non-negative-zero one loss -random variable and with finite E):

$$P(ep(A(S) > \varepsilon) \leq \frac{E\left(ep\left(A(S)\right)\right)}{\varepsilon} \leq \frac{\varepsilon * \delta}{2\varepsilon} = \frac{\delta}{2} < \delta$$

Markov's inequality:

Since we chose $a = \frac{\varepsilon * \delta}{2}$

Thus, H is PAC learnable via A.

To sum up, H is PAC learnable **if and only** if H is PAC learnable in expectation. ■

3. **(10 points) Union of intervals.** Determine the VC-dimension of $\mathcal{H}_k$ - the subsets of the real line formed by the union of k intervals (see the programming assignment for a formal definition of $\mathcal{H}$). Prove your answer.

Answer:

VC-dimension of $Hk\ is\ 2k$.

First, let's prove that $\boldsymbol{Hk \geq 2k}$: for this we will show that there exists a set of size 2k which is shattered by Hk.

Consider the set $C = \{x_1, x_2, \dots, x_{2k}\}\ where\ x_i = \frac{i}{2k}$. Let $S = \{s_1, s_2, \dots, s_{2k}\}$ be some dichotomy.

Define $I = \{[l_1, u_1], [l_2, u_2], \dots, [l_k, u_k]\}$ where:

$$\forall\ 1 \leq i \leq k:\ \ l_i = \begin{cases} \frac{2i-1}{2k}\ if\ s_{2i-1} = 1 \\ \frac{2i-0.75}{2k}\ if\ s_{2i-1} = 0 \end{cases}, u_i = \begin{cases} \frac{2i}{2k}\ if\ s_{2i} = 1 \\ \frac{2i-0.25}{2k}\ if\ s_{2i} = 0 \end{cases}$$

C is shattered by Hk, for this we need to prove that $h_I(x_i) = s_i$

$$x_{2i-1} \in [l_i, u_i]\ \boldsymbol{iff}\ l_i = \frac{2i-1}{2k}\ \boldsymbol{iff}\ s_{2i-1} = 1$$

And by the way we defined I, x_{2i-1} cannot be in any other interval except $[l_i, u_i]$ so in this case $h_I(x_{2i-1}) = 1\ \boldsymbol{iff} s_{2i-1} = 1$, meaning that $h_I(x_{2i-1}) = s_{2i-1}$.

$$x_{2i} \in [l_i, u_i]\ \boldsymbol{iff}\ u_i = \frac{2i}{2k}\ \boldsymbol{iff}\ s_{2i} = 1$$

And by the way we defined I, x_{2i} cannot be in any other interval except $[l_i, u_i]$ so in this case $h_I(x_{2i}) = 1\ \boldsymbol{iff} s_{2i} = 1$, meaning that $h_I(x_{2i}) = s_{2i}$.

Now, let's prove that $\boldsymbol{Hk \leq 2k}$: for this we will show that for any set of size > 2k it cannot be shattered by Hk.

Let $C = \{x_1, x_2, \dots, x_{2k+1}\}$, Assume without loss of generality that $x_1 < x_2 < \cdots < x_{2k+1}$.

Consider the next dichotomy: $S = \{s_1, s_2, \dots, s_{2k+1}\}\ where\ s_i = i\ mod\ 2$ meaning $S = \{1,0,1, \dots,0,1\}$

We will prove that this dichotomy cannot be generated by Hk then C is not shattered by Hk.

Assume by contradiction that there exists I such that $h_I(x_i) = s_i$. Notice that then $h_I(x_{2i+1}) = 1$ and $h_I(x_{2i}) = 0$. thus, there are k+1(with the odd indexes) points with the label 1, and since we have k disjoint intervals there must be at least 2 points (with odd indexes) which are in the same interval. Denote them x_{2i-1}, x_{2j-1} but then x_{2i} is also in the same interval and thus $h_I(x_{2i}) = 1$, but $s_{2i} = 0$.

We got a contradiction. Hence, we conclude that for every set C of size > 2k it cannot be shattered by Hk because there exists at least one dichotomy that cannot be generated by Hk.

To sum up: VC-dimension of $\boldsymbol{Hk\ is\ 2k}$. ■

4. **(10 points) 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 all hypotheses of the form

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

where $\boldsymbol{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ (Hint: Proceed along the lines of the proof for homogeneous linear classifiers from recitation 3. For the upper bound, given a sample $\mathbf{x}_1, \ldots, \mathbf{x}_{d+2} \in \mathbb{R}^d$, construct a new set of points $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{d+2}$ by appending a constant entry of 1 to each of the $\mathbf{x}_i$'s. What can you say about the new set of points as a subset of $\mathbb{R}^{d+1}$?)

Proof:

We will prove that the VC-dimension of H_d is d+1.

First, let's prove that $H_d \geq d+1$:

Consider the sample $C = \{x_1, x_2, \ldots, x_d, x_{d+1}\} = \{e_1, e_2, \ldots, e_d, 0^d\}$.

Let $S = \{s_1, s_2, \ldots, s_{d+1}\}$ be some dichotomy.

Then define w, b as follows:

- if $s_{d+1} = 0$:
  define $w = \{2s_1, 2s_2, \ldots, 2s_d\}$ and $b = 1$. Then:
  $for\ every\ 1 \leq i \leq d$: $sign(w * x_i - b) = sign(w * e_i - b) = sign(2s_i - 1) = s_i$ because if $s_i = 0$ then $sign(2s_i - 1) = sign(-1) = 0 = s_i$ and if $s_i = 1$ then $sign(2s_i - 1) = sign(1) = s_i$
  $for\ d+1$: $sign(w * x_i - b) = sign(w * 0 - 1) = sign(-1) = 0 = s_{d+1}$
- if $s_{d+1} = 1$:
  define $w = \{2s_1 - 2, 2s_2 - 2, \ldots, 2s_d - 2\}$ and $b = -1$. Then:
  $for\ every\ 1 \leq i \leq d$: $sign(w * x_i - b) = sign(w * e_i - b) = sign(2s_i - 2 + 1) = sign(2s_i - 1) = s_i$ because if $s_i = 0$ then $sign(2s_i - 1) = sign(-1) = 0 = s_i$ and if $s_i = 1$ then $sign(2s_i - 1) = sign(1) = s_i$
  $for\ d+1$: $sign(w * x_i - b) = sign(w * 0 + 1) = sign(+1) = 1 = s_{d+1}$

  Hence H_d shatters C.

Now, let's prove that $H_d \leq d+1$:

for this we will show that for any set of size > d+1 it cannot be shattered by H_d.

Let $C = \{x_1, x_2, \ldots, x_{d+2}\}$.

Define $v_i = [x_i, -1] = (x_{i_1}, x_{i_2}, \ldots, x_{i_n}, -1) \in R^{d+1}$. $\{v_1, \ldots, v_{d+2}\}$ is a set of d+2 vectors in R^{d+1}, thus they are linearly dependent, meaning there exists $a_1, \ldots, a_{d+2}$ such that:

$$a_{d+2} v_{d+2} = \sum_{i=1}^{d+1} a_i v_i$$

W.L.O.G $a_{d+2} = 1$.

Assume by contradiction that $\{x_1, x_2, \ldots, x_{d+2}\}$ can be shattered.

Consider $S = \{s_1, s_2, \dots, s_{d+2}\}$, $where\ s_i = 1\ if\ a_i \geq 0, and\ s_i = 0\ if\ a_i < 0\ for\ every\ 1 \leq i \leq d+1, and\ where\ s_{d+2} = 0$.

Let h be a hypothesis which realizes it, with parameters w,b. notice that get:

$$[w,b]v_{d+2} = \sum_{i=1}^{d+1} a_i[w,b]v_i = \sum_{i=1}^{d+1} a_i[w,b][x_i,-1] = \sum_{i=1}^{d+1} a_i(wx_i - b) \geq 0$$

And that is since:

$for\ every\ 1 \leq i \leq d+1: if\ a_i \geq 0\ then\ s_i = 1\ \ meaning\ that\ sign(wx_i - b) = 1\ and\ thus\ (wx_i - b) > 0 \rightarrow a_i(wx_i - b) \geq 0. and\ if\ a_i < 0\ then\ s_i = 0\ \ meaning\ that\ sign(wx_i - b) = 0\ and\ thus\ (wx_i - b) < 0 \rightarrow a_i(wx_i - b) \geq 0.$

But then it follows that $[w,b]v_{d+2} \geq 0 \rightarrow (wx_{d+2} - b) \geq 0$, and that is a contradiction since $s_{d+2} = 0$.

■

5. **(10 points) Prediction by polynomials.** Given a polynomial $P : \mathbb{R} \to \mathbb{R}$ define the hypothesis $h_P : \mathbb{R}^2 \to \{0, 1\}$ by,

$$h_P(x_1, x_2) = \begin{cases} 1 & P(x_1) \geq x_2 \\ 0 & otherwise. \end{cases}$$

Determine the VC-dimension of $\mathcal{H}_{poly} = \{h_P \mid P \text{ is a polynomial}\}$. You can use the fact that given n distinct values $x_1, \dots, x_n \in \mathbb{R}$ and $z_1, \dots, z_n \in \mathbb{R}$ there exists a polynomial P of degree $n-1$ such that $P(x_i) = z_i$ for every $1 \leq i \leq n$.

Answer:

$\text{VC} - \text{dimension of } \text{H}_{\text{poly}} \text{ is } \infty$.

To prove that we need to show that for every n>0, there exists a set $C\ such\ that\ |C| = n$, where $|H_c| = 2^n$, in other words that for this sample C we can get to every dichotomy $\{s_1, \dots, s_n\}\ using\ some\ h_p$.

Fix n>0.

Consider $C = \{(x_1, 0), \dots, (x_n, 0)\}$, where $x_i = i\ for\ every\ i$ Let $\{s_1, \dots, s_n\} \in (0,1)^n$ be some dichotomy. Notice that x_i are distinct by the way defined C, then by the hint provided in the question there exists a polynomial P of degree n-1 such that $P(x_i) = 1\ if\ s_i = 1\ and\ P(x_i) = -1\ if\ s_i = 0$.

Meaning that we defined $z_i = 1\ if\ s_i = 1, z_i = -1\ if\ s_i = 0$.

Consider $h_p\ defined\ by\ P\ above$, Thus we get that:

$$h_p(x_i, 0) = 1\ \boldsymbol{iff}\ P(x_i) \geq 0\ \boldsymbol{iff}\ P(x_i) = 1\ \boldsymbol{iff}\ s_i = 1$$

$$h_p(x_i, 0) = 0\ \boldsymbol{iff}\ P(x_i) < 0\ \boldsymbol{iff}\ P(x_i) = -1\ \boldsymbol{iff}\ s_i = 0$$

So, we got that $h_p(x_i, 0) = s_i$. Then H_{poly} shatters C.

■

Programming Assignment:

(a) **(8 points)** Assume that the true distribution $P[x, y] = P[y|x] \cdot P[x]$ is as follows: x is distributed uniformly on the interval $[0, 1]$, and

$$P[y = 1|x] = \begin{cases} 0.8 & \text{if } x \in [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1] \\ 0.1 & \text{if } x \in (0.2, 0.4) \cup (0.6, 0.8) \end{cases}$$

---

and $P[y = 0|x] = 1 - P[y = 1|x]$. Since we know the true distribution P, we can calculate $e_P(h)$ precisely for any hypothesis $h \in \mathcal{H}_k$. What is the hypothesis in $\mathcal{H}_{10}$ with the smallest error (i.e., $\arg\min_{h \in \mathcal{H}_{10}} e_P(h)$)?

Answer:

Since we know the true distribution of P, we can use the optimal predictor we saw in lecture 1: MAP(maximum a-posteriori) which is:

$$if\ P(Y = 1|X = x)\ is\ bigger\ then\ we\ predict\ h(x) = 1$$

$$if\ P(Y = 0|X = x)\ is\ bigger\ then\ we\ predict\ h(x) = 0$$

We proved that this is the optimal predictor in case of using binary classification, and zero one loss, which we are in this question.

So the hypothesis in H_{10} which has the smallest error is as follows:

$$h(x) = \begin{cases} 1\ if\ x \in [0,0.2]\ U\ [0.4,0.6]\ U\ [0.8,1] \\ 0\ \ if\ x \in [0.2,0.4]\ U\ [0.6,0.8] \end{cases}$$

that is because when $x \in [0,0.2]\ U\ [0.4,0.6]\ U\ [0.8,1]$ then $P(Y = 1|X = x)\ is\ bigger$, and if $x \in [0.2,0.4]\ U\ [0.6,0.8]\ P(Y = 0|X = x)\ is\ bigger$.

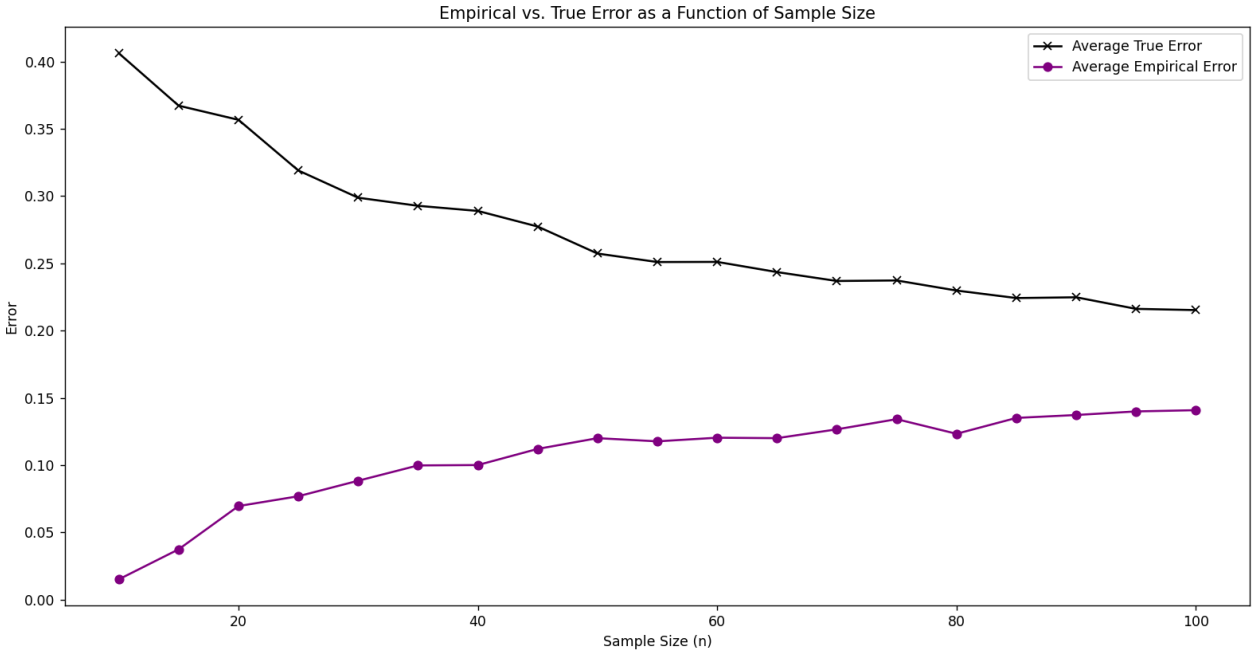
h provided above consists of less than 10 intervals thus it is in H_{10}.

(b) **(8 points)** Write a function that, given a list of intervals I, calculates the true error $e_P(h_I)$. Then, for $k = 3$, $n = 10, 15, 20, \ldots, 100$, perform the following experiment $T = 100$ times: (i) Draw a sample of size n and run the ERM algorithm on it; (ii) Calculate the empirical error for the returned hypothesis; (iii) Calculate the true error for the returned hypothesis. Plot the empirical and true errors, averaged across the T runs, as a function of n. Discuss the results. Do the empirical and true errors decrease or increase with n? Why?

Answer:

Below we can see the empirical and true errors as a function of n. we can see that the average true error decreases with n, but the average empirical error increases with n.

The average true error decreases with n since as long as we get more samples then we are getting closer to P, meaning we are more accurate thus we have less mistakes. On the other side ,I think that the average empirical error increases with n since now the probability of predicting labels with low probabilities gets higher, something like overfitting, the model tries to fit itself to the "noise" which gives us higher empirical error.



(c) **(8 points)** Draw a sample of size $n = 1500$. Find the best ERM hypothesis for $k = 1, 2, \ldots, 10$, and plot the empirical and true errors as a function of k. How does the error behave? Define k^* to be the k with the smallest empirical error for ERM. Does this mean the hypothesis with k^* intervals is a good choice?

Answer:

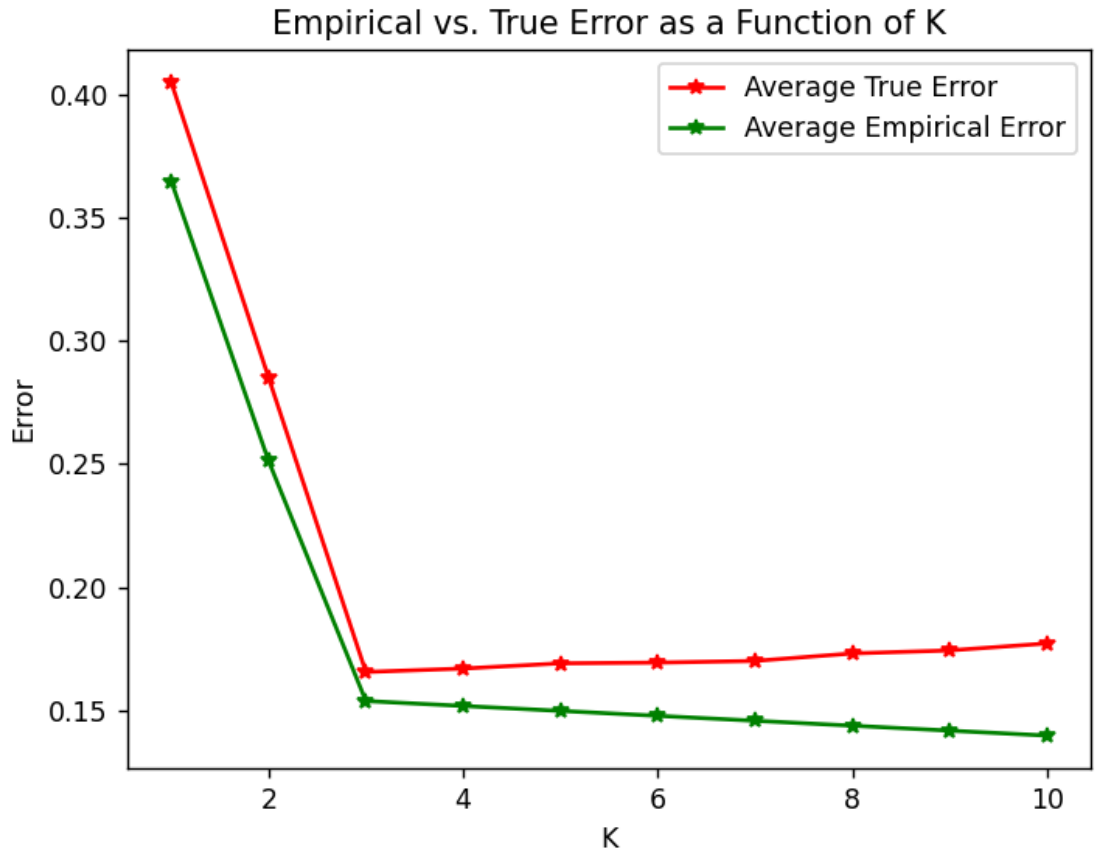
We can see the true error and empirical error as a function of k below in the plot.

We can notice that when k=3 that is the best true error (the minimal) we get, which makes sense since we know that the real P consists of three intervals.

Regarding to the empirical error, we can see a dramatic decrease in the empirical till k=3, but after that the empirical decreases but not that dramatically. But notice that if k<3 that is underfitting cause our model is too "simple" to get the real P since it consists of 3 intervals not less, but when k>3 there is overfitting since our model is more complex than our real P, but we can see that the empirical error decrease despite that, and that is cause our model fit the data better, but not the real P as we can see the true error increases after k=3.

Note: in the provided code we were asked to return the best k, and there was no mention of the best k with minimal true error or minimal empirical error, in the forum I saw that it doesn't matter, so I returned the k with the minimal empirical error which will be 10 as we can see from the plot, I could

simply change it to return the best true error and then it will return 3, but I think you intent was to the empirical error so we can compare with the next question.



(d) **(8 points)** Now we will use the principle of structural risk minimization (SRM), to search for a k that gives a good test error. Let $\delta = 0.1$:

- Use to following penalty function:

$$2\sqrt{\frac{\text{VCdim}(\mathcal{H}_k) + \ln\frac{2}{\delta}}{n}}$$

- Draw a data set of $n = 1500$ samples, run the experiment in (c) again, but now plot two additional lines as a function of k: 1) the penalty for the best ERM hypothesis and 2) the sum of penalty and empirical error.
- What is the best value for k in each case? is it better than the one you chose in (c)?

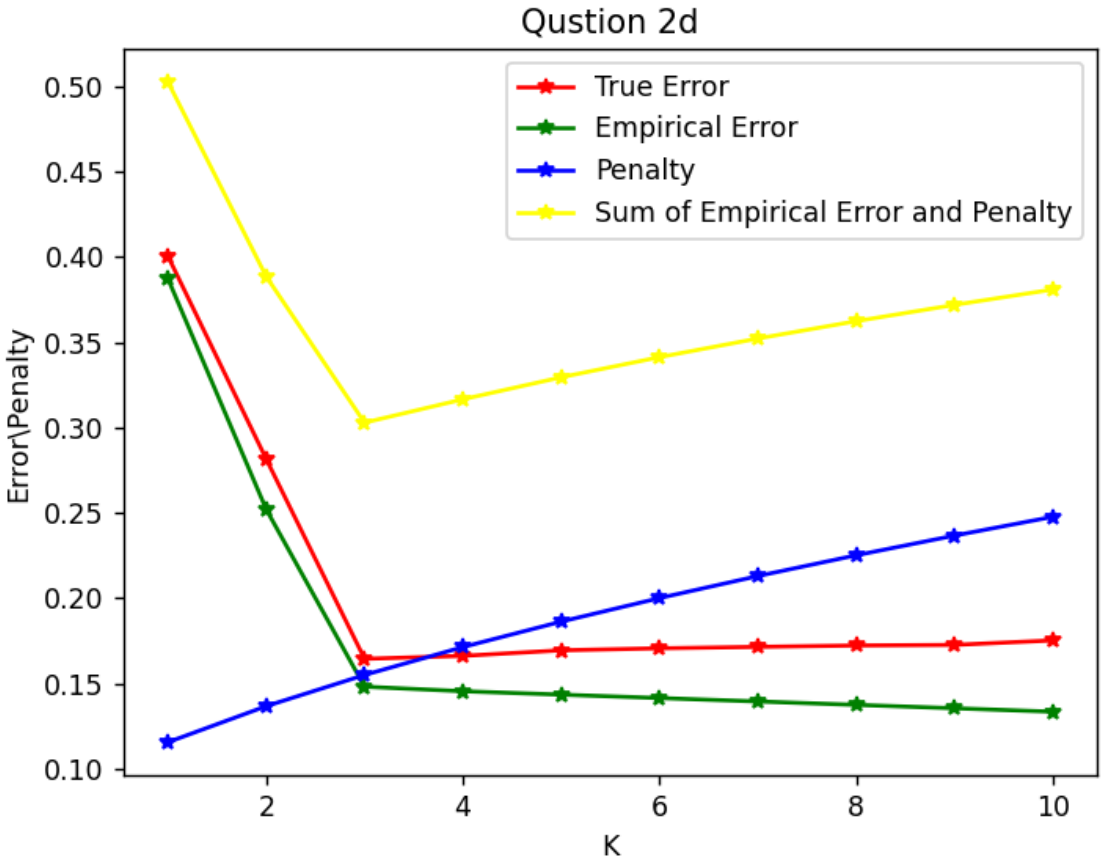
Answer:

We can see the desired plot below.

The true error and empirical is the same as before which we explained above.

We can see that the penalty increases with K since VC dimension of H_k is 2k(increases with k) and since the penalty increases when VC dimension of H_k increases, and that is the goal of the penalty: to make us pay more for "more complex" hypotheses.

The sum of the empirical error and the penalty is at its best when k=3(minimal), and that is what we want since the real P is with 3 intervals. We can say that now we got a better result from the previous question because as I said previously in the note, relying on the empirical error got us the wrong k(and that is what the function for "c" returned: k=10), but when we consider the penalty also we got the desired k(now this function in the provided code return k=3 as a best k) which is the best k since we know P and since it has the minimal true error.



(e) **(8 points)** Here we will use holdout-validation to search for a $k \in \{1, ..., 10\}$ that gives good test error. Draw a data set of $n = 1500$ samples and use 20% for a holdout-validation. Choose the best hypothesis and discuss how close this gets you to finding the hypothesis with optimal true error.

Answer:

We can see below the plot of the empirical error on the holdout set as a function of k.

Also here we can see that we have got a minimal error in k=3, after k=3 we can see that it's the same and I checked it by adding prints in the code, its really the same, and what I noticed in the intervals is

neither there is an “empty” interval which is added when we increase the k, or an interval which was in h_3 is divided to two intervals.

The code returns k=3 as the best choice, and that is what we hoped for, the best hypothesis we got is:

[(0.0005456276615157241, 0.20091079213206137),

(0.4000281320868144, 0.5989857001770413),

(0.7997716300535274, 0.9884491971490744)]

if we compare to the real one:

[(0, 0.2),

(0.4, 0.6),

(0.8, 1)]

We can see that the hypothesis returned is really close to the real one.