

1 נס

Theory Questions

1. (15 points) PAC learnability of ℓ_2 -balls around the origin. Given a real number $R \geq 0$ define the hypothesis $h_R : \mathbb{R}^d \rightarrow \{0, 1\}$ by,

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

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.

עליכם כוונת נושא:

Definition (Realizable Case)

We say \mathcal{H} is **PAC learnable** by an algorithm \mathcal{A} if there exists a function $N : (0, 1) \times (0, 1) \rightarrow \mathbb{R}$ such that for all $\epsilon, \delta \in (0, 1)$ and **realizable** P , when running \mathcal{A} on a **training set** S consisting of $n \geq N(\epsilon, \delta)$ i.i.d. samples drawn from P , it holds that:

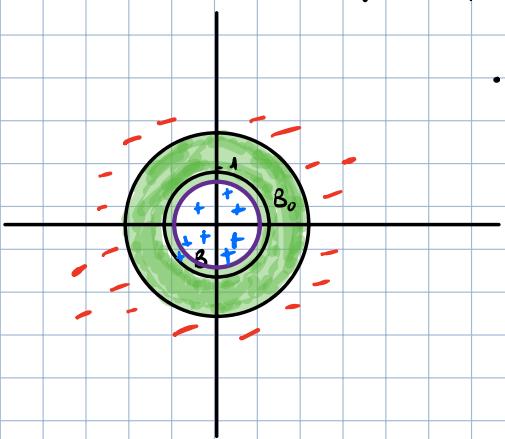
$$P [e_P(\mathcal{A}(S)) > \epsilon] < \delta$$

" $\mathcal{A}(S)$ is **Probably** $(1-\delta)$ **Approximately** (ϵ) **Correct**"

לכורך סטטיסטיקה היברידי Abell מוכיח שקיימת מינימלית N שקיים עבור כל $\epsilon, \delta \in (0, 1)$ וכל פונקציית P על $(x^{(i)}, y^{(i)}) \in \mathbb{R}^d \times [0, 1]^d$ קיימת פונקציית \mathcal{H}_R אשר מוגדרת כ-realizable. לטענו שקיימת פונקציית \mathcal{A} אשר מוגדרת כ-realizable.

ההיפואזיה שקיים מינימלית N אשר מוגדרת כ-realizable היא:

$$R=1 d=2 \lambda \sqrt{\epsilon} \ln(2/\delta)$$



הטעיה מתקיימת אם $P[B_0] > \varepsilon$ וכן $\epsilon_p(h_R) = P[B_0 \setminus B]$ מתקיימת אם $P[B_0 \setminus B] \leq P[B_0] \leq \varepsilon$, כלומר $B_0 \setminus B \subseteq T$.

$P(T \leq \|x\|_2 \leq R_0) = \varepsilon$: אם $T \subseteq B_0$ אז $P[B_0] > \varepsilon$. ($\forall x \in T$ $\|x\|_2 \leq R_0$)
 נסמן $X^{(i)}$ כפונקציית ה- i -הוותק של x . ($\forall x \in T$ $\|X^{(i)}\|_2 \leq R_0$)
 $\forall i: X^{(i)} \in T \iff p_i \in PK, p_i \in C_{N,i}$

$$B_0 \setminus B \subseteq T \Rightarrow P[B_0 \setminus B] \leq P[T] = \varepsilon$$

אם $1 \leq i \leq n$ מתקיים $\epsilon_p(h_R) > \varepsilon$ וקיים $x^{(i)}$ כך $\|X^{(i)}\|_2 > R_0$ $\iff X^{(i)} \notin T$

$$P[\epsilon_p(h_R) > \varepsilon] \leq P[\forall i: X^{(i)} \in T]$$

$$P[X^{(i)} \notin T] = 1 - P[X^{(i)} \in T] = (1 - \varepsilon)$$

ככל ש- i מוגדר ב- $i.i.d$ פ.א.ר.

$$P[\forall i: X^{(i)} \in T] = (1 - \varepsilon)^n \leq e^{-n\varepsilon}$$

$$1 - x \leq e^{-x}$$

$$e^{-n\varepsilon} \leq \varepsilon \iff \frac{1}{e^{n\varepsilon}} \leq \varepsilon \iff e^{n\varepsilon} \geq \frac{1}{\varepsilon} \iff n\varepsilon \geq \ln\left(\frac{1}{\varepsilon}\right)$$

$$\iff n \geq \frac{\ln\left(\frac{1}{\varepsilon}\right)}{\varepsilon}$$

$$n \leq N(\varepsilon, \delta) = \frac{\ln\left(\frac{1}{\delta}\right)}{\varepsilon}$$

בנוסף, הוכחה: נניח

פ.א.ר. $N(\varepsilon, \delta) = \frac{\ln\left(\frac{1}{\delta}\right)}{\varepsilon}$ ו

לפי הטעיה $\epsilon_p(h_R) > \varepsilon$ ו- $\epsilon_p(h_R) = P[B_0 \setminus B]$ מתקיימת אם $P[B_0 \setminus B] \leq P[B_0] \leq \varepsilon$.

כל $s \in S$ קיימת $B_s \in H_{\text{ball}}$ כך $B_s \cap h_s^{-1}B_0h_s \neq \emptyset$

$$P[\liminf_{n \rightarrow \infty} A_n] > \varepsilon$$

כך ניתן למצא תבנית התוודה שקיים δ בו $\forall s \in S$ קיימים $B_s \in H_{\text{ball}}$ וקיימים r_s, R_s כך $B_s \subset B_r(s)$ ו- $B_s \subset B_{R_s}(s)$. נקבע $R = \max_{s \in S} R_s$. טענו $\forall n \in \mathbb{N}$ $\exists s_n \in S$ כך $s_n \in B_{R_n}(s_n)$ ו- $s_n \in B_{r_n}(s_n)$.

2. גיבוב

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 $N(a) : (0, 1) \rightarrow \mathbb{N}$ 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 (Hint: For one direction, use the law of total expectation. For the other direction, use Markov's inequality).

הוכחה:

הוכיחemos PAC learnable in expectation \Leftrightarrow \mathcal{H} הינה realizable.
לhid סבירותה A (תכלית) $N(a) : (0, 1) \rightarrow \mathbb{N}$ כפונקציית $N(a)$ מוגדרת כminimum $|S| \geq N(a)$ שפונקציית $e_P(A(S))$ realizable. הינה P הינה realizable.

נוכיח:

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

בנוסף נוכיח:

$$P[e_P(A(S)) > \epsilon] \leq \frac{\mathbb{E}[e_P(A(S))]}{\epsilon} \stackrel{*}{\leq} \frac{a}{\epsilon}$$

ה. ס. ו. ϵ ו. a נבחרו כך $\epsilon, a \in (0, 1)$ ו. $\frac{a}{\epsilon} = \frac{1}{\delta}$

$$P[e_P(A(S)) > \epsilon] \leq \frac{a}{\epsilon} = \delta$$

. ה. ס. ו. ϵ ו. a נבחרו כך $\epsilon, a \in (0, 1)$ ו. $\frac{a}{\epsilon} = \delta$

A PAC learnable if H is PAC learnable in \mathcal{H}
 \mathcal{H} is PAC learnable if $\exists (\delta, \epsilon, N)$ such that $\forall f \in \mathcal{H} \exists A \subseteq \mathcal{X}$ s.t. $f|_A$ is realizable by a function $g: A \rightarrow \mathcal{Y}$ such that $\Pr_{x \sim D}[g(x) \neq f(x)] \leq \delta$

$$\Pr_{x \sim D}[f(x) \neq h(x)] \leq \epsilon$$

where h is a random function from \mathcal{X} to \mathcal{Y} drawn uniformly at random.

$$\begin{aligned} E[\text{err}(A(S))] &= E[\text{err}(A(S)) | \text{err}(A(S)) > \epsilon] \cdot P[\text{err}(A(S)) > \epsilon] \\ &\quad + E[\text{err}(A(S)) | \text{err}(A(S)) \leq \epsilon] \cdot P[\text{err}(A(S)) \leq \epsilon] \end{aligned}$$

Since $\text{err}(A(S)) \leq 1$ and $\text{err}(A(S)) \geq 0$, we have:

$$E[\text{err}(A(S)) | \text{err}(A(S)) > \epsilon] \leq 1$$

$$E[\text{err}(A(S)) | \text{err}(A(S)) \leq \epsilon] \leq \epsilon$$

$$P[\text{err}(A(S)) \leq \epsilon] \leq 1$$

$$P[\text{err}(A(S)) > \epsilon] < \delta$$

Therefore (with probability $1 - \delta$):

we can

$$E[\text{err}(A(S))] \leq 1 \cdot \delta + \epsilon \cdot 1 = \delta + \epsilon$$

$$N(\alpha) = N(\epsilon, \delta) \text{ s.t. } \epsilon = \frac{\alpha}{2} \text{ and } \delta = \frac{\alpha}{2}$$

Therefore

$$E[\text{err}(A(S))] \leq \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$$

Thus H is PAC learnable in expectation.

3. (15 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.

לכדי שתהוו VC-dimension $\geq 2k$
 $\therefore \text{VCdim}(\mathcal{H}_k) \geq 2k$

כזה נראה לנו שקיים קבוצה $S_1, \dots, S_{2k} \subseteq \{0, 1\}^{(2k)}$ כך שקיים אוסף של $2k$ אינטראvals

לעתות קיימים $x_1 < x_2 < \dots < x_{2k}$ וקיימים סדרה

כך שקיים אוסף של $2k$ אינטראvals:

$$[x_i, x_{i+1}] \quad \text{ובו } S_i = S_{i+1} = 1 \quad \text{פ.к. 1}$$

$$\left[\frac{x_i+x_{i+1}}{2}\right] \quad \text{ובו } S_i = S_{i+1} = 0 \quad \text{פ.к. 2}$$

$$\left[x_i, \frac{x_i+x_{i+1}}{2}\right] \quad \text{ובו } S_i = 1, S_{i+1} = 0 \quad \text{פ.ק. 3}$$

$$\left[\frac{x_i+x_{i+1}}{2}, x_{i+1}\right] \quad \text{ובו } S_i = 0, S_{i+1} = 1 \quad \text{פ.ק. 4}$$

לעתות קיימים אוסף של $2k$ אינטראvals

לעתות קיימים אוסף של $2k$ אינטראvals

ולעתות קיימים אוסף של $2k$ אינטראvals

$$h(x_i) = \begin{cases} 1 & S(x_i) = 1 \\ 0 & S(x_i) = 0 \end{cases}$$

ולעתות קיימים אוסף של $2k$ אינטראvals

ולפ' רין ג'רני קרטה קרטה

$$\text{כ"ל}(0, \dots, 0) \leq \dim(H_K)$$

בנוסף לכך H_K יכיל לפחות $2k+1$ מילוקים שונים. כלומר $x_1 < x_2 < \dots < x_{2k+1} = x_{2k+2}$.

: $s_1, \dots, s_{2k+1} \in \{0, 1\}$ הינה כפונקציה $s: x_1 < x_2 < \dots < x_{2k+1} \rightarrow \{0, 1\}$

$$s(x_i) = \begin{cases} 1 & \text{אם } i \\ 0 & \text{בבאים אחרים} \end{cases}$$

במקרה של מילוקים ייחודיים אחוריהם נסיבת $x_1 < x_2 < \dots < x_{2k+1}$ מתקיימת.

$s(x_{i+1}) = s(x_i) = 1$ אם x_i, x_{i+1} הם מילוקים ייחודיים.

במקרה של מילוקים דומים נסיבת $x_1 < x_2 < \dots < x_{2k+1}$ מתקיימת.

אם ניקח מילוק אחד ונקרא לו $x_1 < x_2 < \dots < x_{2k+1}$ נסיבת $x_1 < x_2 < \dots < x_{2k+1}$ מתקיימת.

במקרה של מילוקים דומים נסיבת $x_1 < x_2 < \dots < x_{2k+1}$ מתקיימת.

. \square

$$\text{כ"ל}(0, \dots, 0) = 2k$$

4. (15 points) Prediction by polynomials. Given a polynomial $P : \mathbb{R} \rightarrow \mathbb{R}$ define the hypothesis $h_P : \mathbb{R}^2 \rightarrow \{0, 1\}$ 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 $\mathcal{H}_{poly} = \{h_P \mid P \text{ is a polynomial}\}$. 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 $1 \leq i \leq n$.

לכיה ב \mathcal{H}_{poly} גורם ש $\forall C \subseteq \mathbb{R}^2$ $\text{VCdim}(\mathcal{H}_{poly}) = \infty$ כי לא ניתן לחלק C באמצעות פולינום.

הוכחה:

יהי $C \subseteq \mathbb{R}^2$ קבוצה. ניקח $C = \{(x_i, z_i) \mid i=1, \dots, n\}$ וпуס $x_1, \dots, x_n \in \mathbb{R}$ ו $z_1, \dots, z_n \in \mathbb{R}$ כך שקיים פולינום p ממעלה $n-1$ אשר $p(x_i) = z_i \quad \forall 1 \leq i \leq n$. נניח שקיים $h_p \in \mathcal{H}_{poly}$ ש

- לעתה ניקח x_i, z_i ועכשו $x_i, z_i \in C$ ופונקציית רצף הינה (x_i, z_i) ופונקציית געגוע הינה $p(x_i) = z_i$ (בנוסף x_i, z_i הם נקודות שונות).
- בכדי שפונקציית געגוע הינה $p(x_i) = z_i$ אזי $p(x_i) \geq z_i$ ופונקציית רצף הינה $h_p(x_i, z_i) = 1$.

אנו מזקזק שפונקציית געגוע הינה $p(x_i) = z_i$ ופונקציית רצף הינה $h_p(x_i, z_i) = 1$ ופונקציית געגוע הינה $p(x_i) < z_i$ ופונקציית רצף הינה $h_p(x_i, z_i) = 0$. מכאן שפונקציית געגוע הינה $p(x_i) \neq z_i$ ופונקציית רצף הינה $h_p(x_i, z_i) \neq p(x_i) \neq z_i$, כלומר $h_p(x_i, z_i) \neq z_i$.

$$\text{VCdim}(\mathcal{H}) = \max_{C: |\mathcal{H}_C| = 2^{|C|}} |C|.$$

- (a) (10 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., $\arg \min_{h \in \mathcal{H}_{10}} e_P(h)$)?

כדי למצוא היפואזיה שטובה והוותיקתית קיימת כפולה של יישומים
 פאראמיטריזציה. ובה מוגדר MAP כהיפואזיה המבילה ליותר
 מדויק. מושג זה מוגדר כהיפואזיה שמייצגת את הערך המרבי
 $P[y=1 | x=\hat{x}]$
 $\hat{x} \in [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1]$
 מינימום האוריגינל של $e_P(h)$

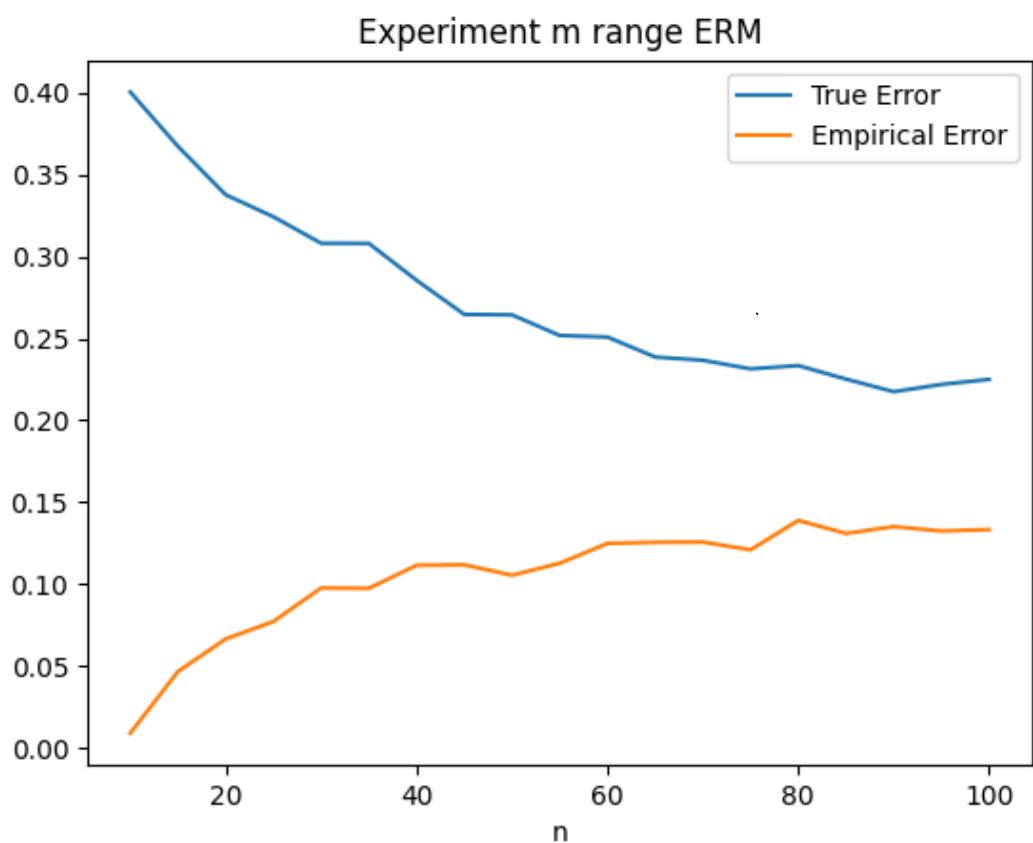
$$\arg \min_{h \in \mathcal{H}_{10}} e_P(h) = \arg \max_{\hat{x} \in [0, 1]} P[y=1 | x=\hat{x}]$$

$$\Rightarrow h(\hat{x}) = \begin{cases} 1 & \hat{x} \in [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1] \\ 0 & \text{otherwise} \end{cases}$$

היפואזיה $h(\hat{x})$ מוגדרת כך שמייצגת את הערך המרבי
 של $P[y=1 | x=\hat{x}]$.

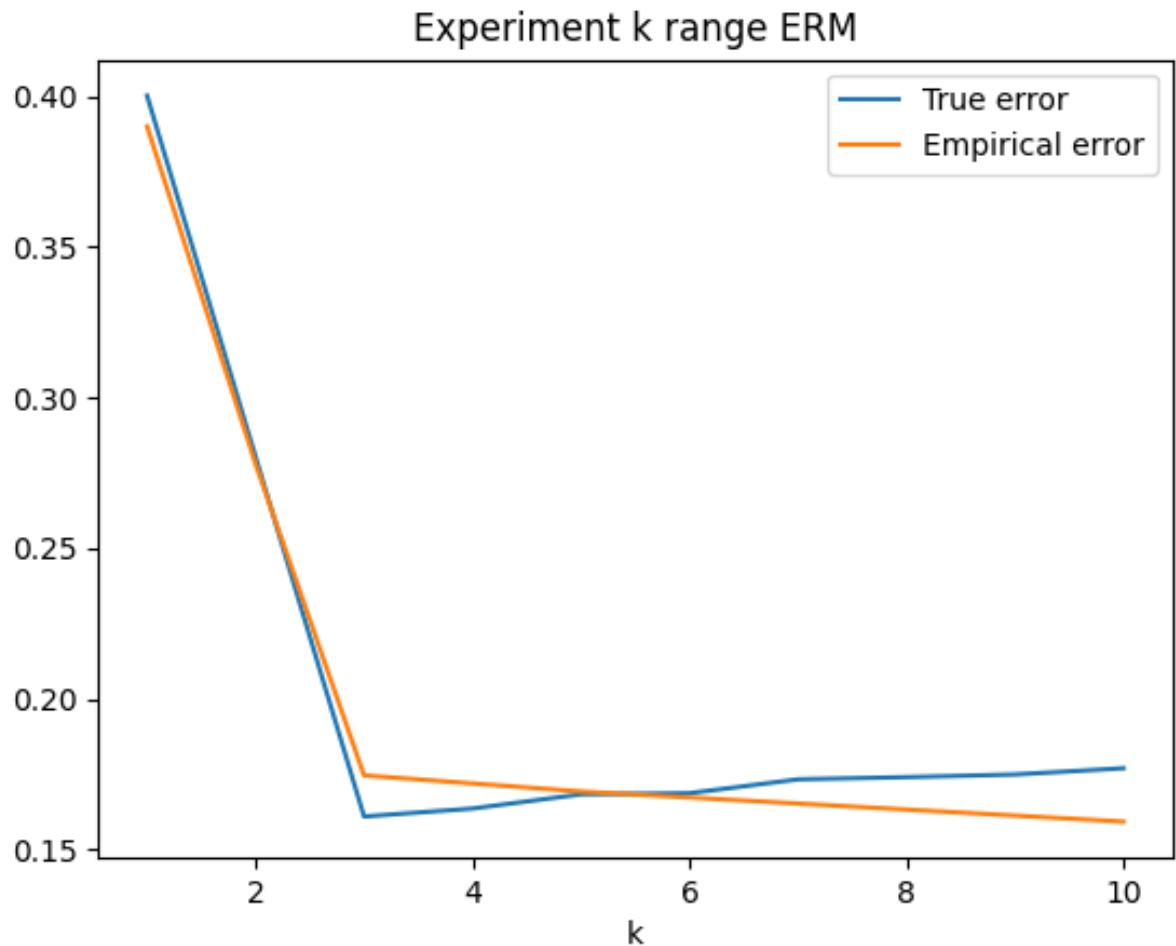
- (b) (10 points) Write a function that, given a list of intervals I , calculates the true error $e_P(h_I)$. Then, for $k = 3$, $n = 10, 15, 20, \dots, 100$, perform the following experiment $T = 100$ times: (i) Draw a sample of size n and run the ERM algorithm on it; (ii) Calculate the empirical error 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?

:מבחן בדוק



לפי מגד ששלב אחר ה先是 - קטעים נספחים, כי הטעון,
ובן יתרכז בחלוקת ויקרא כו' הענין הנטועה.
ובן מילאנו את הטעון הנטועה הנטועה הנטועה
בפואט, ויטריניה הנטועה הנטועה הנטועה
ולפעמים מילאנו את הטעון הנטועה הנטועה הנטועה
בפואט, ויטריניה הנטועה הנטועה הנטועה.

- (c) (10 points) Draw a sample of size $n = 1500$. Find the best ERM hypothesis for $k = 1, 2, \dots, 10$, 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?



לעומת ה $k=3$ רצוי $k=10$ כי הוא מינימלי. אך בפועל נשים $k=3$ כי הוא מינימלי. מילוי ה $k=10$ מושך לאילו אפסי. נסמן k^* כהווקם שפערו בין ה $k=3$ וה $k=10$.

במקרה של מינימום נכון מושך לאילו אפסי.

במקרה של מינימום לא נכון מושך לאילו אפסי. מילוי ה $k=10$ מושך לאילו אפסי. מילוי ה $k=3$ מושך לאילו אפסי.

במקרה של מינימום לא נכון מושך לאילו אפסי. מילוי ה $k=10$ מושך לאילו אפסי. מילוי ה $k=3$ מושך לאילו אפסי.

(d) (10 points) Here we will use an empirical method called holdout-validation to try and find $k \in \{1, \dots, 10\}$ that gives a good test error. For each value of $k \in \{1, \dots, 10\}$, draw a data set of $n = 1500$, run the ERM algorithm for \mathcal{H}_k on a training set consisting of 80% of the data set, and then calculate the empirical error of the returned hypothesis **on the remaining 20% of the examples which you did not train on**. Choose the best hypothesis (i.e. the one with the lowest such error) and discuss how close this gets you to finding the hypothesis with optimal true error.

היכן ומי שפיתח שיטה זו
היא:

The best interval is:

$[(0.00021388759561707937, 0.2007520081069173), (0.4001218525304408, 0.5936384303466232), (0.8006368614472632, 0.998215984701069)]$

ו k הוא:

The best k value found by the cross validation algorithm is: 3

פונקציית ה- k -fold CV מוצעת ב- $k=3$ פותחנית. ה- $k=3$ פותחנית מושגת כטבלה בסיסית וכתוצאה מכך שפונקציית ה- k -fold CV מושגת כטבלה בסיסית. ו- $k=3$ פותחנית מושגת כטבלה בסיסית.

$[0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1]$



פונקציית ה- k -fold CV מושגת כטבלה בסיסית.