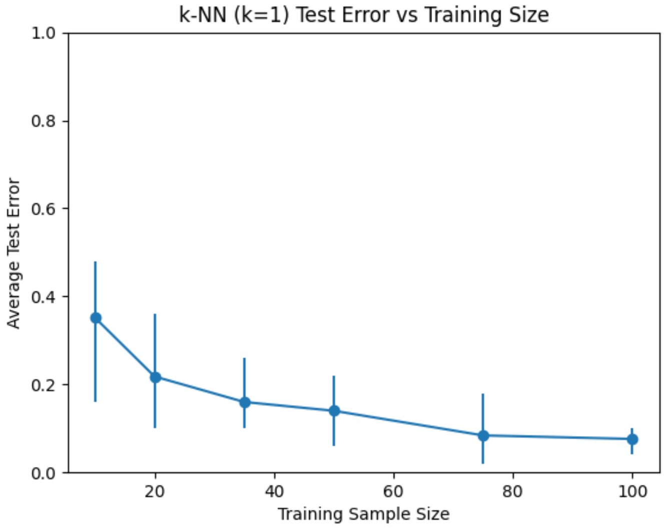
ML task 1 – 211314570 208300244

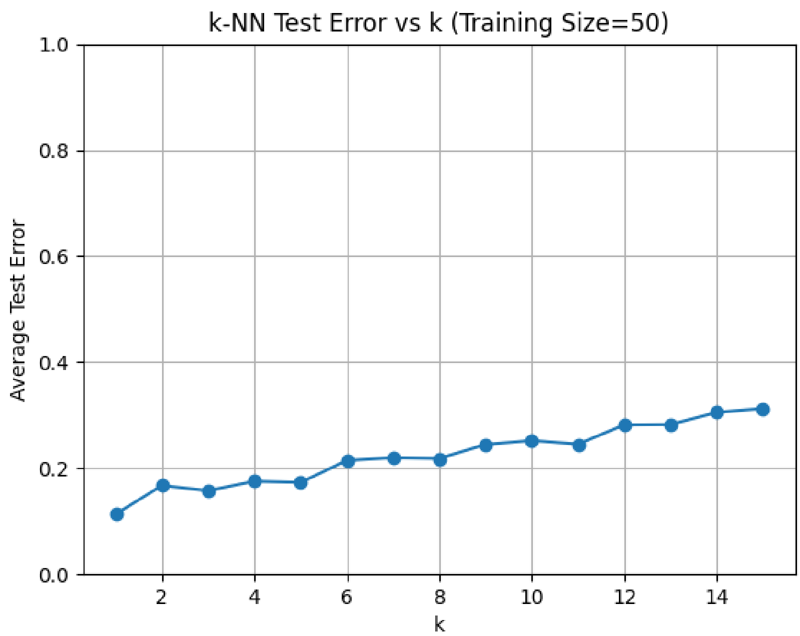
a)

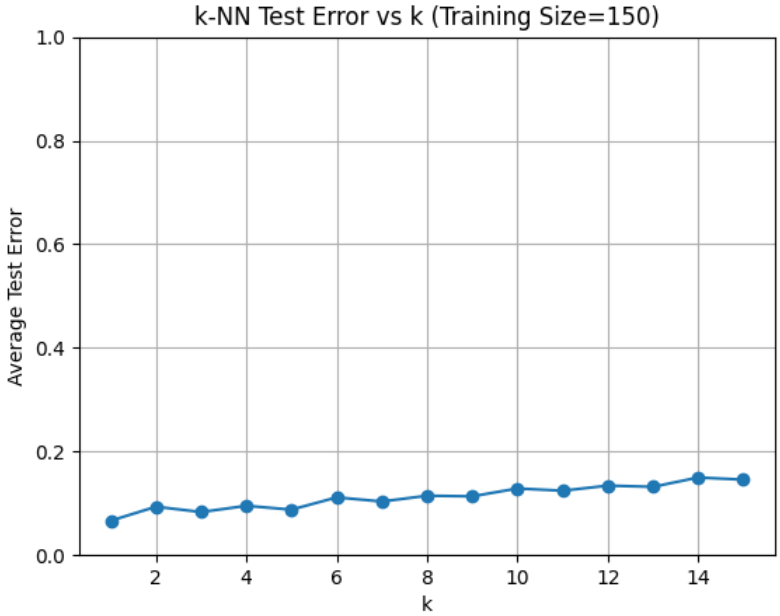


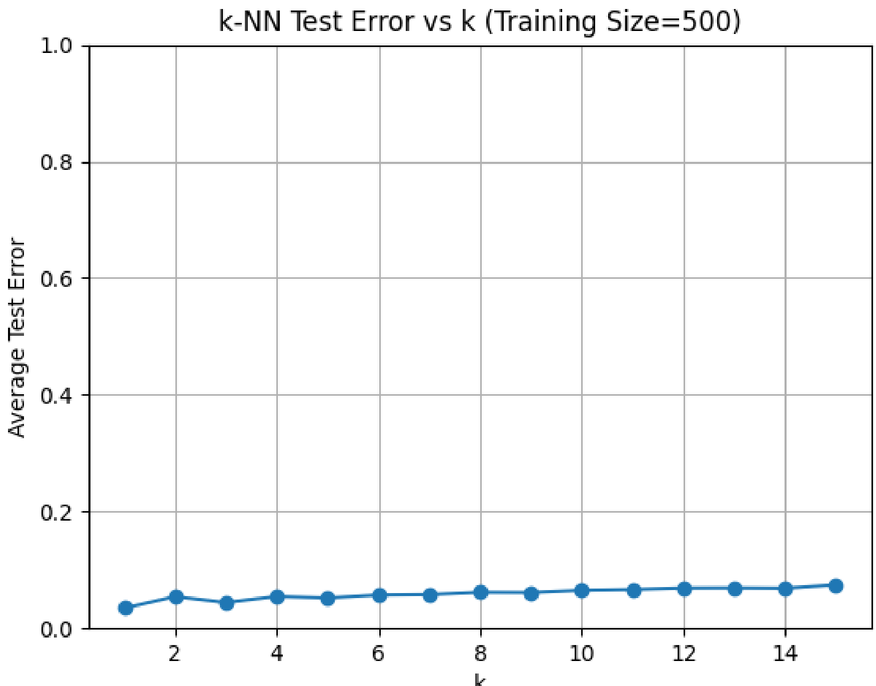
b) For k-NN, the graph shows that increasing the training sample size lowers the average test error. This happens because with more training data, the nearest neighbors for each test point are closer and more representative of the true class, so k-NN makes fewer mistakes. With a small training set, the nearest neighbor may be far away or mislabeled, leading to higher error.

c) Yes, the error bars get smaller as the training sample size increases.
This means the variability in test error decreases with larger training sets.
The reason is that with more training data, k-NN predictions become more stable and less sensitive to which random samples are chosen, so the error is more consistent across runs.

d)

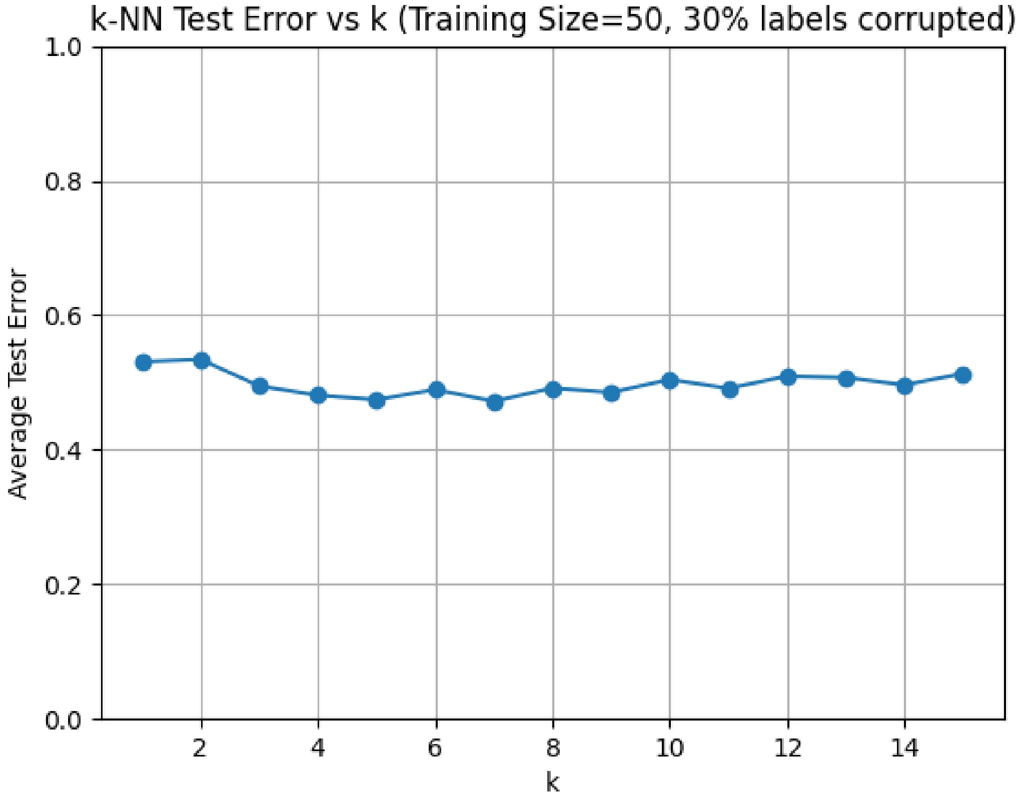




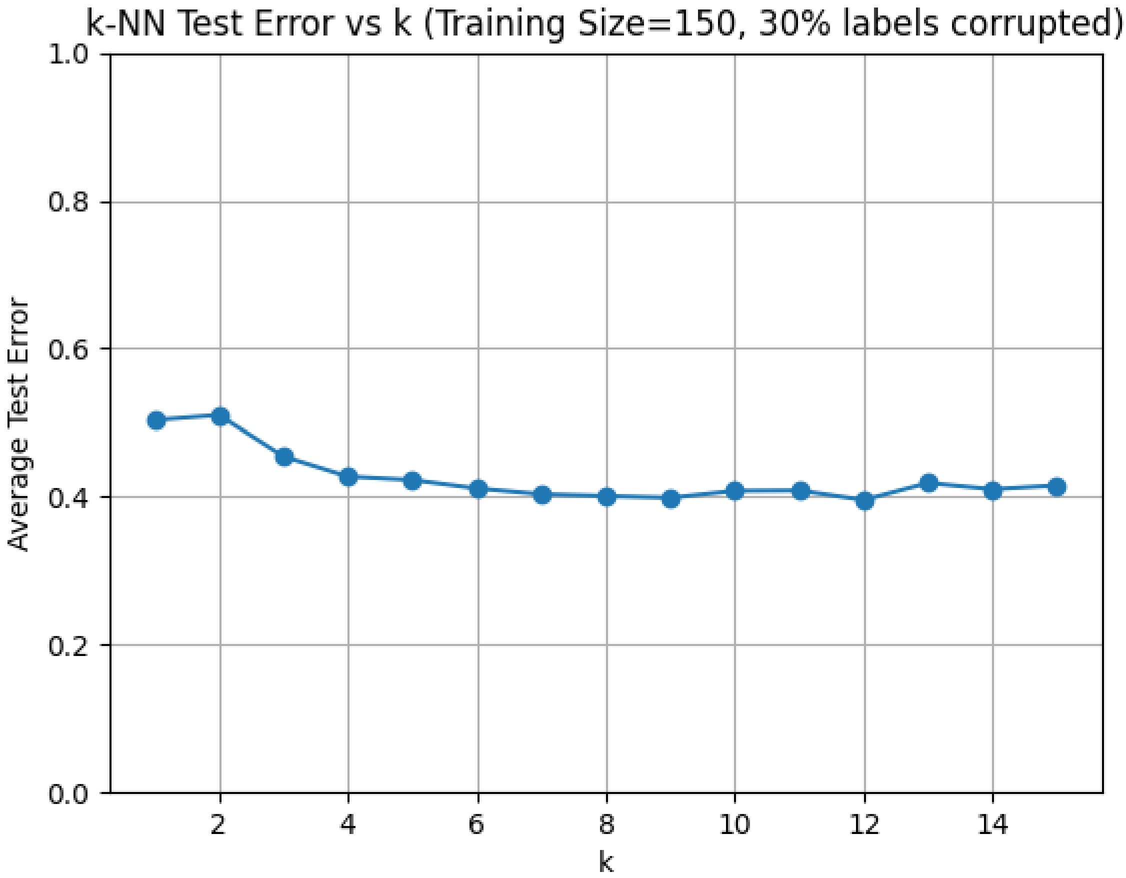



```
Optimal k for m=50: 1 (error=0.113)
Optimal k for m=150: 1 (error=0.067)
Optimal k for m=500: 1 (error=0.035)
```

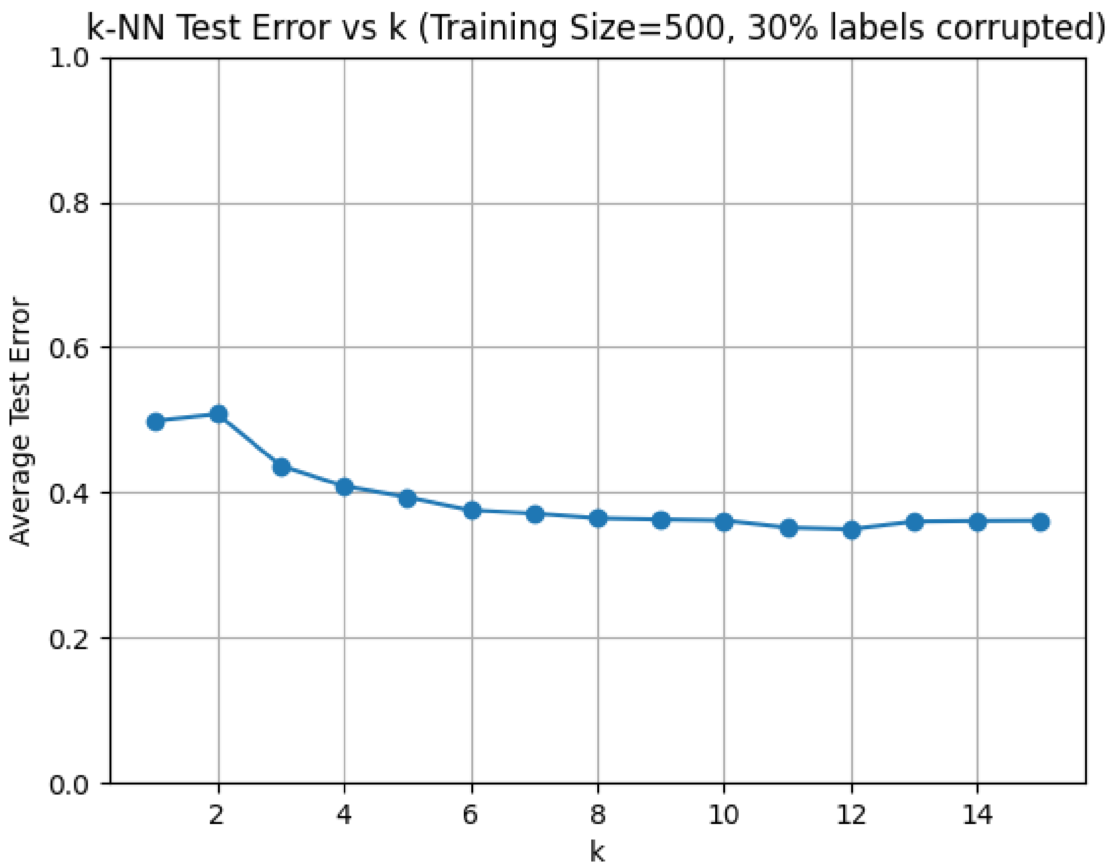
e)



k-NN Test Error vs k (Training Size=50, 30% labels corrupted)
Average Test Error
k
1.0
0.8
0.6
0.4
0.2
0.0
2
4
6
8
10
12
14



k-NN Test Error vs k (Training Size=150, 30% labels corrupted)
Average Test Error
k
1.0
0.8
0.6
0.4
0.2
0.0
2
4
6
8
10
12
14




```
Optimal k for m=50 (corrupted labels): 7 (error=0.472)
Optimal k for m=150 (corrupted labels): 12 (error=0.396)
Optimal k for m=500 (corrupted labels): 12 (error=0.350)
```

f) With corrupted labels, using a larger k helps average out the noise from incorrect labels, making the classifier more robust. In the clean data experiment, a small k is best because the labels are reliable.

In summary, label corruption increases the optimal k and the overall error, because higher k reduces the impact of noisy labels by majority voting.

Q3:

a) X should be the age and weight [age, weight] of the rabbit and Y should be the preferred food
b) $h_{bayes}(7,1)$ = lettuce
   $h_{bayes}(7,2)$ = lettuce
   $h_{bayes}(13,1)$ = carrot
   $h_{bayes}(13,2)$ = lettuce
   because the distribution is deterministic the err(h_{bayes}) is 0. err(h_{bayes}) =0
c)

Age	Preferred food	probability
7	Lettuce	60%
7	carrot	0%
13	Lettuce	25%
13	Carrot	15%

d)

$h_{bayes}(7)$ = lettuce
$h_{bayes}(13)$ = lettuce

err(h_{bayes}) = P[X = 7] * P[h(x)!=lettuce|X=7] + P[X = 13] * P[h(x)!=lettuce|X=13]
err(h_{bayes}) = 0.6 * 0 + 0.4 * (15/40) = 0.15

e)

$$\mathbb{E}_{S\sim\mathcal{D}^m}[\mathrm{err}(\hat{h}_S,\mathcal{D})] = \frac{k-1}{k}\sum_{x\in\mathcal{X}} p_x(1-p_x)^m.$$

k=2
m=2
px = the probability to draw x from the distribution

D = the distribution given in the table

Es~Dm[err(hs,D)] = 0.10*(1 - 0.10)^2 + 0.50*(1 - 0.50)^2 + 0.15*(1 - 0.15)^2 + 0.25*(1 - 0.25)^2 = 0.455

Q4:

(a) (4 points) Show that for a fixed S, a single number that depends on S (call it β_S) controls the probability $\Pr\left[\widehat{h}_S(a) = 1\right]$. Give a formula for β_S as a function of

$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\}.$$

Note that here the probability is over the randomness of $\widehat{h}_S$, since S is fixed.

נסמן: β_s - אחוז המופעים של התווית 1 ב-S.

$$\beta_s : S \to \mathbb{N}$$

$$\beta_s(S) = \frac{1}{|S|} \cdot \sum_{(x,y) \in S} \mathbb{1}\{y = 1\} = \frac{1}{m} \cdot \sum_{i=1}^{m} \mathbb{1}\{y_i = 1\}$$

$$\Rightarrow \Pr\left[\hat{h}_s(a) = 1\right] = \beta_s$$

זאת מכיוון שכל קלט x במבחן הוא שונה קרוב

ביותר של קלט מבחן, לכן "יבחר באופן אחיד שונה

לשכפו.

(b) (4 points) Calculate $\mathrm{err}\left(\widehat{h}_S, \mathcal{D}\right)$ as a function of β_S and ψ.

$$\mathrm{err}(\hat{h}_s, D) = P_{(x,y) \sim D}(\hat{h}_s(x) \neq y)$$

$$= P_{(x,y) \sim D}(\hat{h}_s(x) = 1, y = 0) + P_{(x,y) \sim D}(\hat{h}_s(x) = 0, y = 1)$$

$= P(\hat{h}_S(a)=1) \cdot P(Y=0 \mid X=a)$

$+ P(\hat{h}_S(a)=0) \cdot P(Y=1 \mid X=a)$

$= P(\hat{h}_S(a)=1) \cdot P(Y=0) + P(\hat{h}_S(a)=0) \cdot P(Y=1)$

$= \beta_S \cdot (1-\psi) + (1-\beta_S) \cdot \psi = \beta_S - \beta_S \cdot \psi + \psi - \beta_S \cdot \psi$

$= \beta_S + \psi - 2\beta_S \cdot \psi$

(c) (4 points) Calculate the expectation of β_S over samples, denoted $\mathbb{E}_{S\sim\mathcal{D}^m}[\beta_S]$, as a function of ψ.

$E_{S\sim D^m}[\beta_S] = E\left[\frac{1}{m}\sum_{i=1}^{m} \mathbb{I}[y_i=1]\right]$ $\quad y_i \in S \quad 1 \le i \le m$

Expected value of an indicator

$= \frac{1}{m}\sum_{i=1}^{m} E[\mathbb{I}[y_i=1]] = \frac{1}{m}\sum_{i=1}^{m} P(y_i=1)$

$x_i = a$ for all i $\qquad$ S is statistically i.i.d

$= \frac{1}{m}\sum_{i=1}^{m} P(y_i=1 \mid x_i=a) = \frac{1}{m}\sum_{i=1}^{m} P(Y=1 \mid X=a)$

$= \frac{1}{m} \cdot \sum_{i=1}^{m} \eta(a) = \frac{1}{m} \cdot m \cdot \psi = \psi$

(d) (3 points) Calculate the expected error of the 1-NN algorithm for $\mathcal{D}$ over random samples, denoted by

$$\text{err} := \mathbb{E}_{S\sim\mathcal{D}^m}\left[\text{err}\left(\hat{h}_S, \mathcal{D}\right)\right],$$

as a function of ψ.

$$E_{S\sim D^m}\left[err\left(\hat{h}_S, D\right)\right] \overset{(b)}{=} E\left[\beta_S + \psi - 2\beta_S \cdot \psi\right]$$

$$= E[\beta_S] + E[\psi] - 2E[\beta_S \cdot \psi]$$

$$\psi + \psi - 2\psi \cdot \psi = 2\psi - 2\psi^2$$

(e) (3 points) Calculate the error of the Bayes-optimal rule for $\mathcal{D}$, denoted by $\text{err}_{\text{bayes}}$, as a function of ψ.

The Bayes-optimal predictor will choose the most likely label given X. In our case, $P(X=a)=1$, so:

$$P(Y=1 \mid X=a) = P(Y=1) = \psi$$

$$P(Y=0 \mid X=a) = P(Y=0) = 1-\psi$$

Hence,
$$h_{Bayes} = \begin{cases} 1, & \psi > \frac{1}{2} \Rightarrow err(h_{Bayes}, D) = 1-\psi \\ 0, & \psi < \frac{1}{2} \Rightarrow err(h_{Bayes}, D) = \psi \end{cases}$$

$$\Rightarrow err(h_{Bayes}, D) = \min\{\psi, 1-\psi\}$$

(f) **(Bonus subsection: 3 points)**
Show that for any $\epsilon > 0$, there exists a value of ψ such that the ratio

$$\frac{\mathrm{err}}{\mathrm{err}_{\mathrm{bayes}}}$$

is larger than $2 - \epsilon$. Conclude that without knowing anything about the distribution, we cannot guarantee a factor lower than two (relative to the Bayes-optimal error) when using the 1-NN algorithm.

יהי $\varepsilon > 0$.

$$\frac{err}{err_{Bayes}} = \frac{2\psi - 2\psi^2}{\min\{\psi, 1-\psi\}} = \frac{2\psi(1-\psi)}{\min\{\psi, 1-\psi\}}$$

<u>מקרה א</u>: $\min\{\psi, 1-\psi\} = \psi$

$$\frac{2\psi(1-\psi)}{\psi} = 2\cdot(1-\psi)$$

ניקח $\psi = \frac{\varepsilon}{2}$, ונקבל:

$$\frac{err}{err_{Bayes}} = 2\left(1 - \frac{\varepsilon}{2}\right) > 2(1-\varepsilon)$$

<u>מקרה ב</u>: $\min\{\psi, 1-\psi\} = 1-\psi$

$$\frac{2\psi(1-\psi)}{1-\psi} = 2\cdot\psi$$

ניקח $\psi = 1$ ונקבל:

$$\frac{err}{err_{Bayes}} = 2 > 2 - \varepsilon$$

$\blacksquare$

Q5:

Question 5

(a) Need to prove:

$$\mathbb{P}_{\{S_j \sim \mathcal{D}^{m_j}\}_{\forall j \in \{1,\dots,K\}}}\left[\forall j \in \{1,\dots,K\},\ \widehat{\mathrm{err}}\left(h_{\mathrm{bad}}, S_j\right) = 0\right] \le (1-\epsilon)^b$$

Proof:

Starting from the L.H.S.:

$$\mathbb{P}_{\{S_j \sim \mathcal{D}^{m_j}\}_{\forall j \in \{1,\dots,K\}}}\left[\forall j \in \{1,\dots,K\},\ \widehat{\mathrm{err}}\left(h_{\mathrm{bad}}, S_j\right) = 0\right]$$

$$= \prod_{j=1}^{k} P_{S_j \sim D^{m_j}}\left[\hat{err}(h_{bad}, S_j) = 0\right]$$ Since $S_1, \dots, S_k$ are statistically independent

$$= \prod_{j=1}^{k} P_{S_j \sim D^{m_j}}\left[\forall (x,y) \in S_j,\ h_{bad}(x) = y\right]$$ By definition

$$= \prod_{j=1}^{k} \left(P_{(X,Y)\sim D}\left[h_{bad}(X) = Y\right]\right)^{m_j}$$ $S_j \sim D^{m_j}$

$$= \prod_{j=1}^{k} \left(1 - P_{(X,Y)\sim D}\left[h_{bad}(X) \neq Y\right]\right)^{m_j}$$

$$= \prod_{j=1}^{k} \left(1 - err(h_{bad}, D)\right)^{m_j}$$ By the definition of $err(h_{bad}, D)$

$$\le \prod_{j=1}^{k} (1-\varepsilon)^{m_j} = (1-\varepsilon)^{\sum_{j=1}^{k} m_j} = (1-\varepsilon)^b$$

(b) (9 points) **Formulate** a mathematical condition on $\{m_j\}_{j=1}^{K}$ that guarantees the prediction performance of

$$\mathbb{P}_{\{S_j\sim\mathcal{D}^{m_j}\}_{\forall j\in\{1,\ldots,K\}}}\left[\mathrm{err}\left(\widehat{h}_{SK},\mathcal{D}\right)\le\epsilon\right]\ge 1-\delta$$

where $\epsilon,\delta\in(0,1)$.

מכיוון ש- D ריאליזבילי ע״י H, $err(\hat{h}_{sk}, S_j)=0$

לכל $1\le j\le k$. לכן, "אימון" של $\hat{h}_{sk}$ על k הדגימות שקול

ל"אימון" על מדגם אחד גדול שגודלו $\{m_j\}_{j=1}^{k}$. שהפרדיקציה עליו

היא אפס - $err(\hat{h}_{sk}, \bigcup_{j=1}^{k} S_j) = \sum_{j=1}^{k} err(\hat{h}_{sk}, S_j) = 0$

לכן:

$$\{m_j\}_{j=1}^{k} \ge \frac{\log(|H|)+\log(\frac{1}{\delta})}{\epsilon}$$

$$\Vert$$

$$\sum_{j=1}^{k} m_j$$

Q6:

Question 6. Consider a binary classification problem with input space $\mathcal{X} = \mathbb{R}$ and label space $\mathcal{Y} = \{0, 1\}$.

Recall: $\mathbb{I}$[condition] is an indicator function that returns 1 if the condition is satisfied, and return 0 otherwise.

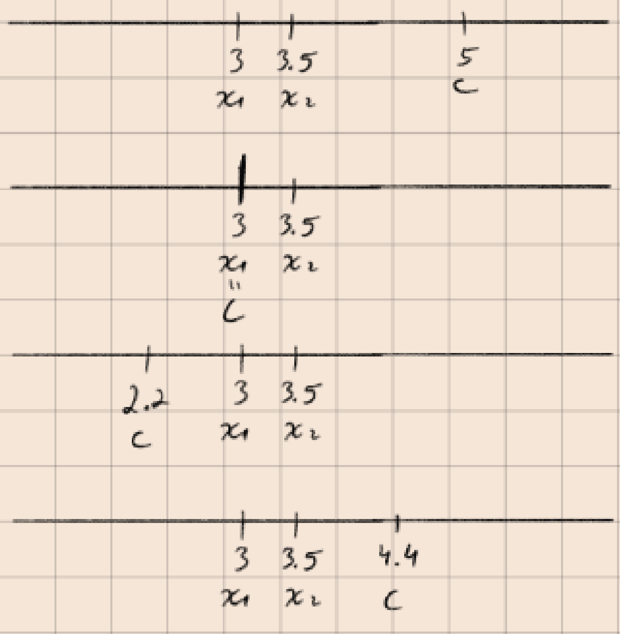
(a) (10 points) Consider the following function for a real input $x \in \mathbb{R}$:

$$f_c(x) = \mathbb{I}\left[|x - c| < 1\right]$$

where $c \in \mathbb{R}$ is a parameter of the function.
We define the hypothesis class $\mathcal{H} = \{f_c \mid c \in \mathbb{R}\}$.
What is the VC dimension of $\mathcal{H}$?
Prove your answer in detail.

הוכחה: ① קיימת קבוצה $X^{(2)} \subseteq \mathcal{X}$ שמנותצת (!) ע"י מחלקת ההשערות H. נבחר:

$X^{(2)} = \{x_1 = 3, x_2 = 3.5\}$



$y_1 = 0, y_2 = 0 \Rightarrow c = 5$

$y_1 = 1, y_2 = 1 \Rightarrow c = 3$

$y_1 = 1, y_2 = 0 \Rightarrow c = 2.2$

$y_1 = 0, y_2 = 1 \Rightarrow c = 4.4$

② נראה שכל קבוצה בגודל 3 אינה מנותצת ע"י H.

תהי $X \subseteq \mathcal{X}$ כאשר $|X| = 3$. נסמן: $X = \{x_1, x_2, x_3\}$

בה"כ $x_1<x_2<x_3$. נראה שלא קיים $f_c \in H$ המקיים

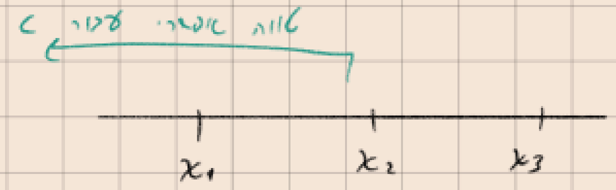
$y_1=1$, $y_2=0$, $y_3=1$.

נניח בשלילה שקיים f_c כזו. מכיוון ש- $f_c(x_1)=1$, $f_c(x_2)=0$

וגם $x_1<x_2$, בהכרח $c<x_2$. מכיוון ש- $x_3>x_2$,

$|c-x_3|>|c-x_2|\ge 1$, לכן $f_c(x_3)=0$ בסתירה.

$f_c(y_2)=0$



(2) נראה שכל קבוצה בגודל 3 אינה מנותצת ע"י H.

יהי $X\subseteq \mathcal{X}$ כאשר $|X|=3$. נסמן: $X=\{x_1,x_2,x_3\}$

בה"כ $x_1<x_2<x_3$. נראה שלא קיים $f_c\in H$ המקיים

$y_1=1$, $y_2=0$, $y_3=1$. נראה זאת ע"י הוכחה

טענה: $f_c(x_1)=1 \wedge f_c(x_3)=1 \to f_c(x_2)=1$

הוכחה: נניח $f_c(x_1)=f_c(x_3)=1$.

מהגדרת f_c, עבור $i\in\{1,3\}$ מתקיים

$|x_i-c|<1$

לכן

$$-1 < x_i - c < 1$$

$$\Downarrow$$

$$c-1 < x_i < c+1$$

כלומר בפרט:

$$c-1 < x_1 \quad \wedge \quad x_3 < c+1$$

כי $x_2 > x_1$ $\Downarrow$ $\qquad$ $\Downarrow$ כי $x_2 < x_3$

$$c-1 < x_2 \quad \wedge \quad x_2 < c+1$$

$$\Downarrow$$

$$c-1 < x_2 < c+1$$

$$\Downarrow$$

$$|x_2 - c| < 1$$

$$\Downarrow$$

$$f_c(x_2) = 1$$

$\blacksquare$