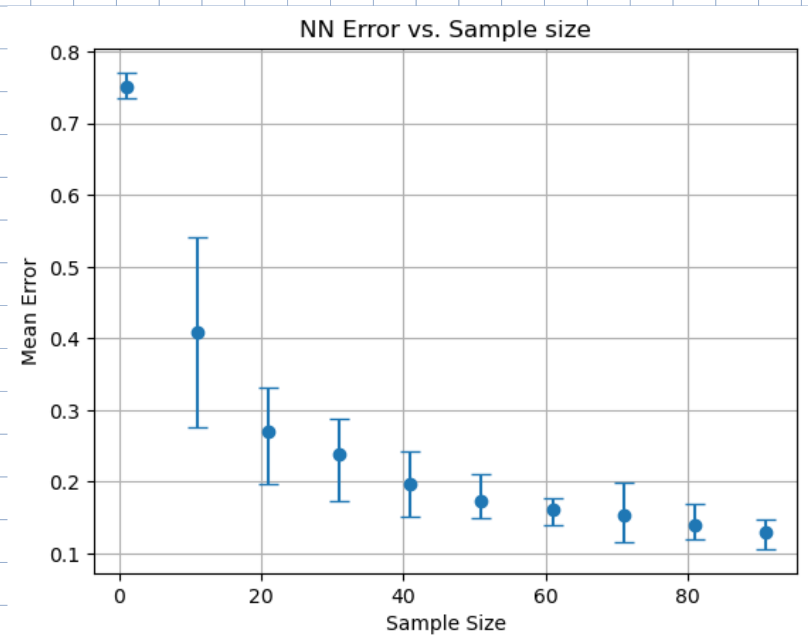
שאלה 2

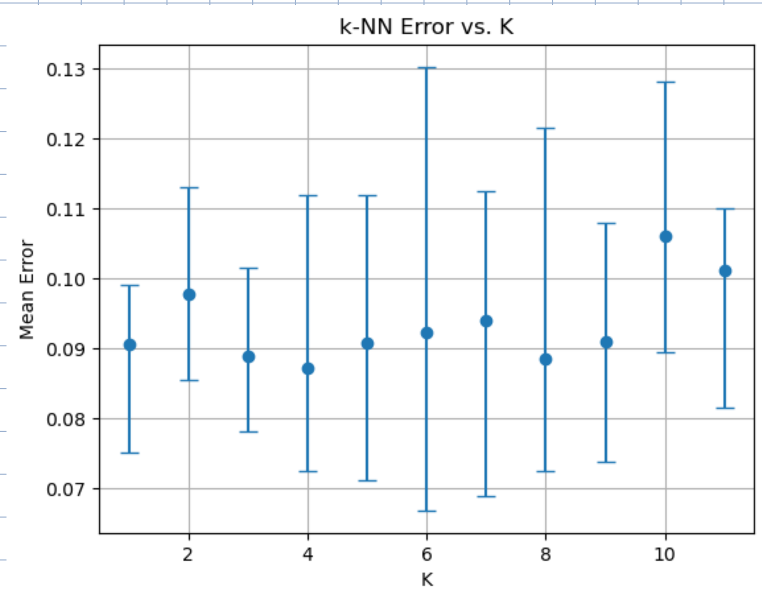
א) קיבלנו את הגרף:



ב) ניתן להבחין במגמת ירידה בשגיאה הממוצעת כתלות בגודל המדגם. המגמה צפויה מכיוון שאלגוריתם NN-1 משתמש במדגם כדי לסווג את הקלטים, ככל שהמדגם גדל יותר, הוא מייצג בצורה טובה יותר את ההתפלגות ולכן נצפה לראות ירידה בשגיאה. בנוסף ראינו בהרצאות כי עבור תוויות דטרמיניסטיות (מתקיים במקרה שלנו) אז כאשר גודל המדגם גדל אז האלגוריתם יתקרב ל bayes-error.

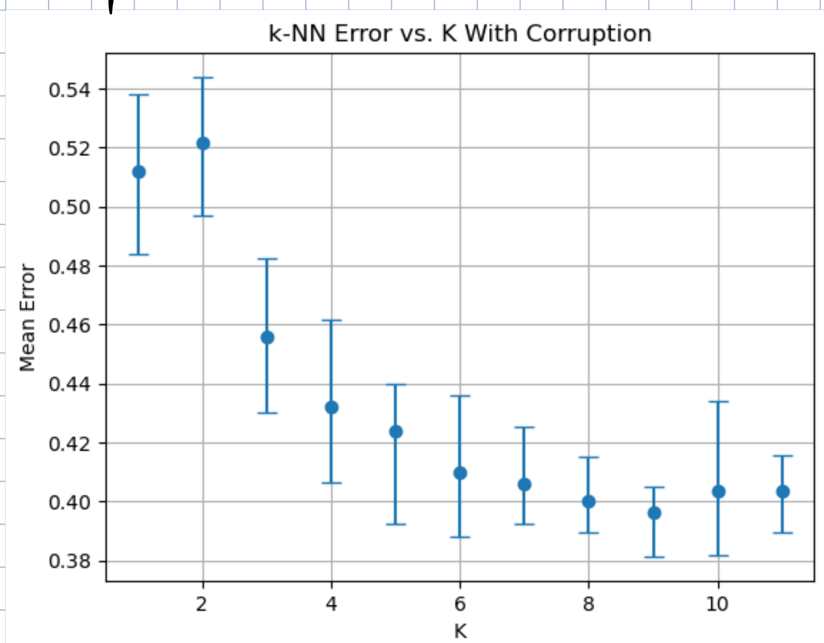
ג) ניתן לראות מגמת ירידה בגודל ה-error bars ככל שגודל המדגם גדל. זה נובע מכך שעבור מדגם קטן האלגוריתם שלנו מקבל מעט מאוד נק' ולכן קשה יותר להכליל. זה גורם לכך שהביצועים מאוד רגישים לנק' הספציפיות שהמדגם מכיל. אם המדגם הקטן מכיל "גיוון" בנק', כלומר מייצג יחסית טוב את ההתפלגות אז נקבל שגיאה קטנה יותר. אך ייתכן שנקבל מדגם קטן לא מייצג ולכן השגיאה תהיה גדולה. ככל שהמדגם גדל, הוא מייצג טוב יותר של ההתפלגות ולכן השונות קטנה יותר והשגיאה קטנה.

ד) הגרף שקיבלנו עבור השגיאה כתלות ב- k:



ה-k האופטימלי ע"פ הגרף הוא $k=4$. ניתן לראות שעבור $k>4$, מתקבלת עלייה בשגיאה. ראינו בהרצאה שאלגוריתם k-NN מתקרב לשגיאת בייס כאשר k וגם m שואפים אך m גדל בסדר גודל מעל. במקרה שלנו $m = 200$ ואינו משתנה ולכן עבור k הולכים וגדלים האלגוריתם נוטה לכיוון התיוג הנפוץ ביותר במדגם האימון, כלומר אנו מקבלים bias גבוה.

ה) הגרף שקיבלנו עבור השגיאה כתלות ב- k עם דאטה "מלוכלך":



1) עבור הדגימה הנקי, ראינו כבר בסעיף ד שב-k האופטימלי הוא 4 בעוד עבור הדגימה המלוכלך ה-k האופטימלי הוא 9. בנוסף בגרף מסעיף ג' המגמה היא שהשגיאה עולה עם k בעוד בסעיף הקודם אנו רואים שהשגיאה יורדת ככל ש-k גדל. ההבדל במגמה נובע מכך שללא "לכלוך" של הדגימה, האלגוריתם חוזה ע"פ מעט נקודות קרובות ומכיוון שאין רעש יוצא אנו מקבלים שגיאה יחסית נמוכה. בניסוי המלוכלך עבור ערכי k קטנים אנחנו חשופים יותר להשפעה של נקודות "מלוכלכות" ולכן השגיאה גבוהה מאוד עבור ערכי k נמוכים וככל ש-k גדל אנו נותנים פחות משקל לנק' מלוכלכות ולכן השגיאה יורדת. שווה לציין כי ערכי השגיאה בניסוי המלוכלך גדולים משמעותית מאלו המתקבלים בניסוי הנקי בגלל הלכלוך של הדגימה.

שאלה 3

א) נתחיל מתיאור התמך של X:

$$supp(X)=\{(20,5),(50,10),(50,15),(80,20)\}$$

ואת ההגדרה של bayes optimal predictor:

$$h_{bayes}(x)=\underset{y\in Y}{\arg\max}\ \eta_y(x) \quad \text{where: } \eta_y(x):=\underset{(x,y)}{\mathbb{P}}(Y=y|X=x)$$

כעת נחשב את $h_{bayes}(x)$ עבור כל x בתמך, נסמן לנוחות $\eta_{Apple}=\eta_A,\ \eta_{Orange}=\eta_O$.

$x=(80,20)$:

$$\eta_O((80,20))=\frac{0.2}{0.2}=1 \quad , \quad \eta_A((80,20))=0$$

$$\boxed{h_{bayes}((80,20))=\text{Orange}}$$

$x=(50,15)$:

$$\eta_O((50,15))=\frac{0.1}{0.1}=1 \quad , \quad \eta_A((50,15))=0$$

$$\boxed{h_{bayes}((50,15))=\text{Orange}}$$

$x=(50,10)$:

$$\eta_O((50,10))=\frac{0}{0.2}=0 \quad , \quad \eta_A((50,10))=\frac{0.2}{0.2}=1$$

$$\boxed{h_{bayes}((50,10))=\text{Apple}}$$

$x=(20,5)$:

$$\eta_O((20,5))=\frac{0.1}{0.5}=0.2 \quad , \quad \eta_A((20,5))=\frac{0.4}{0.5}=0.8$$

$$\eta_A((20,5))>\eta_O((20,5)) \Rightarrow \boxed{h_{bayes}((20,5))=\text{Apple}}$$

ב) כדי לחשב את ה error bayes optimal של D נשתמש בנוסחה שראינו בהרצאה:

$$err(h,D)=\sum \mathbb{P}[X=x]\left(1-\eta_{h(x)}(x)\right)$$

נציב בנוסחה:

$$\mathbb{P}[(20,5)]\cdot(1-0.8)+\mathbb{P}[(50,10)](1-1)+\mathbb{P}[(50,15)](1-1)$$

$$+\mathbb{P}[(80,20)](1-1)=\mathbb{P}[(20,5)]\cdot 0.2=0.5\cdot 0.2=\boxed{0.1}$$

ג) שגיאת האופרוקסימציה של מחלקת היפותזות H מוגדרת ע"י:

$$err_{app}:\min_{h\in H} err(h,D)$$

ובמקרה שלנו מחלקת ההיפותזות היא:

$$H=\{h(x)=Orang,\ h(x)=Apple\}$$

נחשב את השגיאה עבור כל אחת מהפונקציות (נסמן):

$$h_1: h(x)=Apple\quad,\ h_2: h(x)=Orange$$

$$err_{h_1}=\mathbb{P}[X=x\mid y=Orange]=0.1+0.1+0.2=0.4$$

$$err_{h_2}=\mathbb{P}[X=x\mid y=Apple]=0.4+0.2=0.6$$

ולכן מוגדרת:

$$err_{app}(H,D)=0.4$$

3) נשתמש בנוסחה שראינו בהרצאה:

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

במקרה שלנו $k=m=2$ ונקבל:

$$\mathbb{E}_{S\sim D^m}[err(\hat{h}_S,D)]=\frac{1}{2}\sum_{x\in X}p_x(1-p_x)^2=\frac{1}{2}\left[0.4\cdot 0.6^2+0.2\cdot 0.75^2+0.15\cdot 0.85^2\right.$$

$$\left.+0.2\cdot 0.8^2\right]=\frac{1}{2}\cdot 1.285=0.642$$

עבור התפלגות D'' ניתן להשתמש בנוסחה בסוד עבור התפלגות D לא מכיוון שפיתחנו את הנוסחה עם הנחת תנויות דטרמיניסטיות. כלומר יותחן אלו"ם רק כאשר קיבלנו x שורק ראינו וגם אינו בניתוח. בסוד בתנויות לא דטרמיניסטיות ניתן לטעות גם על x שראינו בתהליך הלמידה.

שאלה 4

נתון לנו כי $x\in[0,1]$ ו-$y\in\{0,1\}$ ועבור $k\geq 2$ זוגי נגדיר את מחלקת ההיפותזות:

$$\mathcal{H}=\{f_{a_1\ldots a_k} \mid \forall i\in\{1,\ldots,k\},\ a_i\in[0,1],\ \text{and } a_1\leq a_2\leq\ldots\leq a_k\}$$

$$\forall x\in[0,1],\ f_{a_1\ldots a_k}(x)=\mathbb{I}[\exists \text{ odd integer } i\in\{1,\ldots,k-1\} \text{ s.t } x\in[a_i,a_{i+1}]]$$

נתונה התפלגות D מעל $X\times Y$ כך שההסתברות $\mathbb{P}$ עבור $X=x$ $D\sim(X,Y)$ מתפלגת אחיד על $[0,1]$.

(א) עבור המחלקות - $\mathcal{H}_2,\mathcal{H}_4,\mathcal{H}_6$ נטען כי $\mathcal{H}_6$ מקבלת את שגיאת האפרוקסימציה המינימלית על D. נזכיר כי שגיאת האפרוקסימציה מוגדרת כ:

$$err_{app}=\min_{h\in\mathcal{H}} err(h,D)$$

מחלקות ההיפותזות הנתונות מכילות אחת את השניה, מתקיים:

$$\mathcal{H}_2\subseteq\mathcal{H}_4\subseteq\mathcal{H}_6$$

מכיוון שכל $\mathcal{H}_2$ מאפשרת לנו לחלק את המקטע $[0,1]$ לשני מקטעים לכל היותר, אז גם כ- $\mathcal{H}_4$ ו-$\mathcal{H}_6$ ניתן לשלב את החלוקה הזאת ע"י הגדרת מקטעים בגודל 0. וכנ"ל לגבי $\mathcal{H}_4$ ו- $\mathcal{H}_6$. לכן אם קיימת היפותזה h ב-$\mathcal{H}_2$ או $\mathcal{H}_4$ עבורה מתקבלת שגיאת האפרוקסימציה אז $h\in\mathcal{H}_6$. בהכרח קיימת $h^*\in\mathcal{H}_6$ כך ש-$h^*\notin\mathcal{H}_4,\mathcal{H}_2$ אז ייתכן כי h^* כבעלת תלגז שגיאה קטנה יותר מאשר h ולכן מתקיים:

$$\min_{h\in\mathcal{H}_2} err(h,D)\geq\min_{h\in\mathcal{H}_4} err(h,D)\geq\min_{h\in\mathcal{H}_6} err(h,D)$$

ב) נוכיח כי עבור $k\geq 2$ ו $m\leq k$ קיימת היפותזה $h\in H_k$ שמשחזרת את כל ההחלטות של אלגוריתם 1-NN מקיימת את כל ה-ERM.

אלגוריתם 1-NN על $S=\{(x_1,y_1)\ldots(x_k,y_k)\}$ יחלק את המקטע $[0,1]$ לכל היותר ל-m מקטעים המוגדרים ע"י המרחקים בין הנק' במדגם. אלגוריתם 1-NN ישיג שגיאה 0 עבור המדגם מכיוון שעבור כל x_i במדגם נחזיר את y_i.

כעת נרצה להראות שניתן לשחזר את התיוג של 1-NN בעזרת פונקציה מ-H_k.
נרצה לבחור את הנק' $a_1\leq\ldots\leq a_k$ כך שהתיוג עבור x_i יהיה y_i.
עבור k נק' אנחנו יכולים ליצור לכל היותר $k+1$ אינטרוולים, ונקבל $\frac{k}{2}$ אינטרוולים בעלי תיוג 1 ו $\frac{k}{2}+1$ אינטרוולים עם תיוג 0.

נבנה את הפונקציה $h\in H_k$ איטרטיבית:

- עבור הנק' (x_1,y_1) אם $y_1=0$ אז נרצה מקטע אשר מתויג כ-1. נבחר $a_1>x_1$ והמקטע $[0,a_1)$ מתויג ב-0 ולכן התיוג נכון.
אם $y_1=1$ אז נבחר $a_1<x_1$ ו $a_2>x$ ונקבל $x_1\in[a_1,a_2)$ והתיוג נכון ל-x_1.
- נמשיך כך עבור כל $i\in\{2,\ldots k\}$:
  - ← אם $y_i=y_{i-1}$, כלומר התיוג החדש זהה לזה של המקטע הקודם אז ניתן להרחיב את a_i כך ש $a_i>x_i$ ולכן $x_i\in[a_{i-1},a_i)$ ויום התיוג זהה אז הוא נכון.
  - ← אם $y_i\neq y_{i-1}$ אז נרצה להחליף אינדקס כדי ליצור מקטע חדש עם תיוג הפוך אז רק נדרוש ש $a_i<x_i$ ולכן נחליף תיוג מהמקטע הקודם ונקבל תיוג נכון.

מכיוון ש-$m\leq k$ אז בהכרח ש לנו מספיק נק' לבחור כדי ליצור מקטעים עם תיוגים מתאימים.

אם $m<k$ אז נוכל לבחור פשוט מקטעים באורך 0 כך שהתיוג בכל מקטע נשאר זהה לזה של 1-NN.

בסה"כ הראנו כי עבור כל S וכל תיוג ש 1-NN, נוכל למצוא פונקציה ב H_k אשר משחזרת את התיוג עם שגיאה אמפירית 0. מכיוון ש-0 זאת השגיאה המינימלית

ולכן הפונקציה מיישמת את כל ה-ERM.

ג) נרצה להוכיח כי עבור $m=k+1$ כאשר $k\geq 2$ קיים מדגם $S=\{(x_1,y_1),\dots,(x_m,y_m)\}$
ללא דוגמאות סותרות, שעבורו התוצאה המוחזרת ע"י 1-NN לא מקיימת את כלל
ERM עבור המחלקה H_k. נבחר את המדגם להיות:

$$S=\{(x_1=0,y_1=1),(x_2,y_2=0),\dots(x_m=1,y_m=1)\}$$

המדגם הוא בעל תוויות מתחלפות כאשר $x_1=0$, $x_m=1$ והתוויות של שניהם הן 1.
האלגוריתם 1-NN מסוג m נקודות מושלם מכיוון שהוא יחזיר את y_i עבור x_i
מכיוון שכל נקודה היא הקרובה ביותר לעצמה. 1-NN מקבל שגיאה אמפירית 0.
נראה כי אין $f\in H_k$ אשר עבורה ניתן לתייג עם שגיאה 0 את S.
בכדי לתייג את המדגם נכון f חייבת ש x_m יכלל במקטע בעל אינדקס אי זוגי.
אבל מכיוון ש $x_1=0$ ו $y_1=1$ אז בהכרח $a_1=0$ אחרת נטעה כבר שם. ומאחר
שהדוגמאות מתויגות בתוויות מתחלפות אז כל דוגמה דורשת הגדרה של a_i נוסף.
אז $y_k=0$ ולכן $a_k<x_k$ כדי לא לטעות בתיוג והמקטע האחרון הוא $[a_k,1]$
אך k זוגי ולכן x_m יתויג כ-0. כלומר בהכרח נטעה ונקבל שגיאה שונה מ-0.
כלומר למרות ש 1-NN מסוג נכון הוא לא מקיים את תנאי ה-ERM ב-H_k עבור
מדגם בגודל $m=k+1$.

שאלה 5

נתונה מחלקת ההיפותזות:

$$\mathcal{H}=\{f_{a,b,c,d} \mid a,b,c,d\in\mathbb{R},\ a<b<c<d\}$$

כאשר:

$$f_{a,b,c,d}(x)=\mathbb{I}[a\le x\le b \text{ or } c\le x\le d]$$

כלומר פונקציה המחזירה כאשר x נמצא באיחוד בין שני קטעים על הישר הממשי ו-0 אחרת. כדי למצוא את מימד VC של $\mathcal{H}$ נרצה להראות שני דברים:

1) קיימת קבוצה בגודל k אשר מנתצת את $\mathcal{H}$ כלומר ניתן לתייג אותה בכל דרך עם היפותזות ב-$\mathcal{H}$.

2) כל סט בגודל $k+1$ לא יכול להיות מתויג בכל דרך ע"י היפותזות ב-$\mathcal{H}$.

נטען כי $VC(\mathcal{H})=4$.

נראה תחילה כי קיימת קבוצה בגודל 4 שניתן לתייג בכל דרך ע"י היפותזות מ-$\mathcal{H}$:

נקח 4 נקודות $\{x_1,x_2,x_3,x_4\}$ כאשר $x_1<x_2<x_3<x_4$. אם התיוג שלהן הוא זהה, כלומר:

$$\{1,1,1,1\},\ \{0,0,0,0\}$$

אז ניתן לבחור $a=x_1, b=x_4$ עבור האפשרות הראשונה ו-$a>x_4$ עבור השניה.

אם התיוג דורש רק מקטע רציף אחד כדי לתייג אותו נכון, למשל תיוג מהצורה:

$$\{0,0,1,1\},\ \{1,1,0,0\},\ \{0,1,1,0\}$$

אז אנחנו יכולים לבחור מקטע יחיד שמכיל את הנק' המתויגות 1 ואת המקטע השני כך שהוא לא מכיל אף אחת מהנקודות האחרות ונקבל תיוג נכון.

עבור מקטעים הדורשים שני מקטעים רציפים כדי לתייג אותם, למשל מהצורה:

$$\{0,1,0,1\},\ \{1,0,0,1\},\ \{1,0,1,0\}$$

נוכל לבחור a,b,c,d כך שנקבל שני מקטעים זרים אשר מתייגים נכון.

כעת ניתן להראות כי עבור כל 5 נק' לא ניתן להשיג עבורן את כל התיוגים האפשריים עם פונקציות מ-$\mathcal{H}$? נסתכל על 5 נק' $x_1 \dots x_5 \in \mathbb{R}$. קיים תיוג עבור 5 הנק' הדורש שלושה מקטעים זרים, כלומר שאין 2 נק' עוקבות בעלות תיוג "1". תיוג לדוגמה: $\{1,0,1,0,1\}$ ומכיוון שמחלקת ההיפותזות שלנו מוגבלת לשני מקטעים זרים עבור כל קבוצה של 5 לא נוכל להשיג את התיוג הזה.

שאלה 6

נרצה למצוא תנאי על m (גודל) המדגם S כך שהאלגוריתם מחזיר $h_{S,S'}$ המקיים:

$$\mathbb{P}_{S\sim D^m, S'\sim D^n}\left[err(h_{S,S'}, D)\le \epsilon\right] \ge 1-\delta \quad \text{for } \epsilon,\delta\in(0,1)$$

נחסם את ההסתברות שהאלגוריתם החזיר לנו פרדיקטור בעל שגיאה גדולה מ-ϵ:

$$\mathbb{P}_{S\sim D^m, S'\sim D^n}\left[err(h_{S,S'}, D) > \epsilon\right] \overset{(1)}{=}$$

$$\mathbb{P}_{D^m, S'\sim D^n}\left[err(h_{S,S'}, D) > \epsilon \text{ and } \hat{err}(h_{S,S'}, S)=0 \text{ and } \hat{err}(h_{S,S'}, S')=0\right] \overset{(2)}{\le}$$

$$\mathbb{P}\left[\exists h\in \mathcal{H} \text{ s.t } err(h,D)>\epsilon \text{ and } \hat{err}(h,S)=0 \text{ and } err(h,S')=0\right] =$$

$$\mathbb{P}\left[\text{samples } S,S' \text{ are bad}\right]$$

(1) נובע מהגדרת ה-ERM כאשר D ריאליזבילית עבור $\mathcal{H}$.

(2) מכיוון שעברנו למקרה כללי יותר.

כעת נרצה למצוא גודל מדגם m שעבורו מתקיים:

$$\mathbb{P}_{S\sim D^m, S'\sim D^n}\left[S,S' \text{ are good samples}\right] \ge 1-\delta$$

נגדיר $h^{bad}\in\mathcal{H}$ להיות המסווג עבורה:

$$err(h^{bad}, D) > \epsilon$$

ומתקיים:

$$\mathbb{P}_{(X,Y)\sim D}\left[h^{bad}(X)\ne Y\right] = err(h^{bad}, D) > \epsilon$$

נחסם את ההסתברות ש-h^{bad} נראית "טובה" ל-$ERM_{\mathcal{H}}$:

$$\mathbb{P}\left[\hat{err}(h^{bad}, S)=0 \text{ and } err(h^{bad}, S')=0\right] =$$

$$\mathop{\mathbb{P}}_{S\sim D^m,\, S'\sim D^n}\left[\forall (x,y)\in S,\ (x',y')\in S',\ h^{bad}(x)=y \text{ and } h^{bad}(x')=y'\right] \overset{(3)}{=}$$

$$\mathop{\mathbb{P}}_{S\sim D^m}\left[\forall (x,y)\in S,\ h^{bad}(x)=y\right]\cdot \mathbb{P}\left[\forall (x',y')\in S'\ h^{bad}(x')=y'\right] \overset{(4)}{=}$$

$$\left(\mathop{\mathbb{P}}_{(X,Y)\in D}\left[h^{bad}(X)=Y\right]\right)^m \cdot \left(\mathop{\mathbb{P}}_{(X,Y)\in D}\left[h^{bad}(X)=Y\right]\right)^n = \left(\mathop{\mathbb{P}}_{(X,Y)\in D}\left[h^{bad}(X)=Y\right]\right)^{m+n} =$$

$$= \left(1-\mathop{\mathbb{P}}_{(X,Y)\sim D}\left[h^{bad}(X)\neq Y\right]\right)^{m+n} \overset{(5)}{\le} (1-\epsilon)^{m+n} \le e^{-\epsilon(m+n)}$$

(3) מכיוון ש-S' ו-S נדגמים ללא תלות

(4) מכיוון של (x,y) ו-(x',y') נדגמים סגמ"ים ניתן לעבור לכפל הסתברויות

(5) $\mathbb{P}\left[h^{bad}(X)\neq Y\right] = err(h^{bad}, D) > \epsilon$

כמו ונדרש שלא יתקיים עבור כל היפותזה "רעה" $h\in\mathcal{H}$. נגדיר: $I_{bad}\subseteq\{1,\dots,k\}$ להיות קבוצת האינדקסים עבור היפותזות "רעות" ב-$\mathcal{H}$.

חסמנו קודם את ההסתברות שהיפותזה נראית טובה ב המדגמים S, S' עבור $i\in I_{bad}$:

$$\mathop{\mathbb{P}}_{S\sim D^m,\, S'\sim D^n}\left[\hat{err}(h_i,S)=0 \text{ and } \hat{err}(h_i,S')=0\right] \le e^{-\epsilon(m+n)}$$

המדגמים S,S' הם לא טובים אם עבור $i\in I_{bad}$ כלשהו h_i נראית טובה עליהם.

$$\mathbb{P}[S,S' \text{ are not good}] = \mathop{\mathbb{P}}_{S\sim D^m,\, S'\sim D^n}\left[\exists i\in I_{bad} \text{ s.t } \hat{err}(h_i,S)=0 \text{ and } \hat{err}(h_i,S')=0\right] =$$

$$\mathop{\mathbb{P}}_{S\sim D^m,\, S'\sim D^n}\Big[\left(\hat{err}(h_1,S)=0 \text{ and } \hat{err}(h_1,S')=0\right) \text{OR} \dots$$

$$\left(\hat{err}(h_k,S)=0 \text{ and } \hat{err}(h_k,S')=0\right)\Big] \overset{(6)}{\le}$$

$$\sum_{i\in I_{bad}} \mathop{\mathbb{P}}_{S\sim D^m,\, S'\sim D^n}\left[\hat{err}(h_i,S)=0 \text{ and } \hat{err}(h_i,S')=0\right] \le |I_{bad}|\, e^{-\epsilon(m+n)} \le$$

$$|\mathcal{H}|\cdot e^{-\epsilon(m+n)}$$

(6) חסם האיחוד.

ולכן קיבלנו:

$$\underset{S\sim D^m, S'\sim D^n}{\mathbb{P}}\left[err(\hat{h}_{S,S'}, D) > \epsilon\right] \leq \mathbb{P}[S, S' \text{ are not good}] \leq |\mathcal{H}|e^{-\epsilon(m+n)}$$

נדרוש:

$$\underset{S\sim D^m, S'\sim D^n}{\mathbb{P}}\left[err(\hat{h}_{S,S'}, D) > \epsilon\right] \leq |\mathcal{H}|e^{-\epsilon(m+n)} \leq \delta$$

נסתכל על אי השוויון הימני:

$$|\mathcal{H}|e^{-\epsilon(m+n)} \leq \delta \Rightarrow e^{-\epsilon(m+n)} \leq \frac{\delta}{|\mathcal{H}|} \Rightarrow -\epsilon(m+n) \leq \log(\delta) - \log(|\mathcal{H}|)$$

$$\Rightarrow m+n \geq \frac{\log(|\mathcal{H}|) + \log(\frac{1}{\delta})}{\epsilon} = m \geq \frac{\log(|\mathcal{H}|) + \log(\frac{1}{\delta})}{\epsilon} - n$$

קיבלנו שעבור:

$$m \geq \frac{\log(|\mathcal{H}|) + \log(\frac{1}{\delta})}{\epsilon} - n$$

מתקיים:

$$\circledast \quad \underset{S\sim D^m, S'\sim D^n}{\mathbb{P}}\left[err(\hat{h}_{S,S'}, D) > \epsilon\right] \leq \delta$$

$$\underset{S\sim D^m, S'\sim D^n}{\mathbb{P}}\left[err(\hat{h}_{S,S'}, D) > \epsilon\right] + \underset{S\sim D^m, S'\sim D^n}{\mathbb{P}}\left[err(\hat{h}_{S,S'}, D) \leq \epsilon\right] = 1$$

$$\underset{S\sim D^m, S'\sim D^n}{\mathbb{P}}\left[err(\hat{h}_{S,S'}, D) \leq \epsilon\right] = 1 - \underset{S\sim D^m, S'\sim D^n}{\mathbb{P}}\left[err(\hat{h}_{S,S'}, D) > \epsilon\right]$$

נציב את $\circledast$:

$$\underset{S\sim D^m, S'\sim D^n}{\mathbb{P}}\left[err(\hat{h}_{S,S'}, D) \leq \epsilon\right] \geq 1 - \delta$$

עבור:

$$\boxed{m \geq \frac{\log(|\mathcal{H}|) + \log(\frac{1}{\delta})}{\epsilon} - n}$$