Machine learning- Ex1, Osher Elhadad 318969748

2. (15 pts) Polynomial regression

We showed that for a zero-order polynomial (namely, a constant $h_{\mathbf{w}}(x) = w_0$), the value that minimizes the mean squared error $Err(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(h_{\mathbf{w}}(x_i) - y_i)^2$ is the empirical *mean* of samples: $h_{\mathbf{w}}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$. Prove that for the case of zero-order polynomial with an absolute-value error

$$Err(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}|h_{\mathbf{w}}(x_i) - y_i| \quad ,$$

the optimal solution is the *median* of samples.

.2

בהוכחה זו נראה שהפתרון האופטימלי הוא $h_w(x) = w_0 = median\ of\ \{y_1, \dots, y_n\}$.

נשים לב כי $h_w(x) = w_0$ לכל x (פונקציה קבועה). לכן:

$$Err(w) = \frac{1}{n}\sum_{i=1}^{n}|h_w(x_i) - y_i| = \frac{1}{n}\sum_{i=1}^{n}|w_0 - y_i|$$

נסדר את ערכי ה-y כך שיהיו מסודרים בסדר עולה ($y_i < y_j$ לכל $1 \le i < j \le n$) ויהי $k \in \mathbb{N}$ $1 \le k \le n$. נשים לב כי הפונקציה $Err(w)$ היא בעצם מתחילה כפונקציה יורדת עבור w שואפים למינוס אינסוף (מכיוון שככל שערכו של ה-w קטן כך הוא מתרחק מכל ערכי ה-y_i ולכן בערך מוחלט השגיאה גדלה), עד לאיזו נקודה בה ככל שאנחנו מגדילים את w_0 יותר ויותר אנחנו בעצם מגדילים את ההפרש מכל ערכי ה-y_i ולכן מגדילים בערך מוחלט את השגיאה) ולכן אנו נמצא את ערך ה-w_0 עבורו השגיאה תהיה מינימלית ע״י גזירה והשוואה ל-0 כמו בהרצאה. כעת נחלק למקרים-

עבור n אי זוגי:

- **אם $y_1, \dots, y_k < w_0 < y_{k+1}, \dots, y_n$:**

$$Err(w) = \frac{1}{n}\sum_{i=1}^{n}|w_0 - y_i| = \frac{1}{n}\cdot\left(\sum_{i=1}^{k}(w_0 - y_i) + \sum_{i=k+1}^{n}(y_i - w_0)\right)$$

נגזור לפי w_0 ונשווה ל-0:

$$\frac{1}{n}\left(\left(\sum_{i=1}^{k} w_0 - y_i\right)' + \left(\sum_{i=k+1}^{n} y_i - w_0\right)'\right) = \frac{1}{n}\cdot(k + k - n) = \frac{1}{n}\cdot(2k - n)$$

$$\frac{1}{n}\cdot(2k - n) = 0$$

$$2k - n = 0$$

$$k=\frac{n}{2}$$

וקיבלנו סתירה לכך שk טבעי כפי שהגדרנו מכיוון שn הוא אי זוגי ולכן $\frac{n}{2}$ אינו טבעי, כלומר לא קיבלנו מינימום במקרה זה.

- **אם $y_1,\dots,y_{k-1}<(y_k=w_0)<y_{k+1},\dots,y_n$ (במידה וקיימים גם עוד y_j עבורם $y_j=w_0$ אזי נוריד כפילויות אלה ונשאיר רק את $y_k=w_0$ ונעבור למקרה האי זוגי או הזוגי עם n חדש לאחר הורדת הכפילויות):**

$$Err(w)=\frac{1}{n}\sum_{i=1}^{n}|w_0-y_i|=\frac{1}{n}\cdot\left(\left(\sum_{i=1}^{k-1}(w_0-y_i)\right)+(w_0-y_k)+\sum_{i=k+1}^{n}(y_i-w_0)\right)$$

$$=\frac{1}{n}\cdot\left(\sum_{i=1}^{k-1}(w_0-y_i)+\sum_{i=k+1}^{n}(y_i-w_0)\right)$$

$(w_0=y_k)$

נגזור לפי w_0 ונשווה ל-0:

$$\frac{1}{n}\left(\left(\sum_{i=1}^{k-1}w_0-y_i\right)'+\left(\sum_{i=k+1}^{n}y_i-w_0\right)'\right)=\frac{1}{n}\cdot(k-1+k-n)=\frac{1}{n}\cdot(2k-n-1)$$

$$\frac{1}{n}\cdot(2k-n-1)=0$$

$$2k-n-1=0$$

$$2k=n+1$$

$$k=\frac{n+1}{2}$$

ובמקרה זה קיבלנו כי עבור $k=\frac{n+1}{2}$ אזי הנגזרת שווה ל0 וכמו שהסברנו למעלה עבור $w_0=y_k=y_{\frac{n+1}{2}}$ (החציון עבור n אי זוגי במקרה שלנו) אזי ערך השגיאה הוא המינימלי.

עבור n זוגי:

- **אם $y_1,\dots,y_{k-1}<(y_k=w_0)<y_{k+1},\dots,y_n$ (במידה וקיימים גם עוד y_j עבורם $y_j=w_0$ אזי נוריד כפילויות אלה ונשאיר רק את $y_k=w_0$ ונעבור למקרה האי זוגי או הזוגי עם n חדש לאחר הורדת הכפילויות):**

$$Err(w)=\frac{1}{n}\sum_{i=1}^{n}|w_0-y_i|=\frac{1}{n}\cdot\left(\left(\sum_{i=1}^{k-1}(w_0-y_i)\right)+(w_0-y_k)+\sum_{i=k+1}^{n}(y_i-w_0)\right)$$

$$= \frac{1}{n} \cdot \left(\sum_{i=1}^{k-1} (w_0 - y_i) + \sum_{i=k+1}^{n} (y_i - w_0) \right)$$

$(w_0 = y_k)$

נגזור לפי w_0 ונשווה ל-0 :

$$\frac{1}{n}\left(\left(\sum_{i=1}^{k-1} w_0 - y_i\right)' + \left(\sum_{i=k+1}^{n} y_i - w_0\right)'\right) = \frac{1}{n} \cdot (k - 1 + k - n) = \frac{1}{n} \cdot (2k - n - 1)$$

$$\frac{1}{n} \cdot (2k - n - 1) = 0$$

$$2k - n - 1 = 0$$

$$2k = n + 1$$

$$k = \frac{n+1}{2}$$

וקיבלנו סתירה לכך שk טבעי כפי שהגדרנו מכיוון ש$n + 1$ הוא אי זוגי ולכן $\frac{n+1}{2}$ אינו טבעי, כלומר לא קיבלנו מינימום במקרה זה.

- **אם $y_1, \dots, y_k < w_0 < y_{k+1}, \dots, y_n$:**

$$Err(w) = \frac{1}{n} \sum_{i=1}^{n} |w_0 - y_i| = \frac{1}{n} \cdot \left(\sum_{i=1}^{k} (w_0 - y_i) + \sum_{i=k+1}^{n} (y_i - w_0) \right)$$

נגזור לפי w_0 ונשווה ל-0 :

$$\frac{1}{n}\left(\left(\sum_{i=1}^{k} w_0 - y_i\right)' + \left(\sum_{i=k+1}^{n} y_i - w_0\right)'\right) = \frac{1}{n} \cdot (k + k - n) = \frac{1}{n} \cdot (2k - n)$$

$$\frac{1}{n} \cdot (2k - n) = 0$$

$$2k - n = 0$$

$$k = \frac{n}{2}$$

ובמקרה זה קיבלנו כי עבור $k = \frac{n}{2}$ אזי הנגזרת שווה ל0 וכמו שהסברנו למעלה עבור $y_k = y_{\frac{n}{2}} < w_0 < y_{k+1} = y_{\frac{n}{2}+1}$ (החציון עבור n זוגי במקרה שלנו, נבחר את ערכו להיות ממש החציון- $\frac{y_{\frac{n}{2}} + y_{\frac{n}{2}+1}}{2}$, אך כל ערך בין $y_{\frac{n}{2}}$ ל $y_{\frac{n}{2}+1}$ יהיה אופטימלי) אזי ערך השגיאה הוא המינימלי.

וסה״כ גם למקרה שn זוגי וגם למקרה שהוא אי זוגי קיבלנו כי החציון הוא אכן הפתרון האופטימלי $h_w(x) = w_0 = median\ of\ \{y_1, \dots, y_n\}$ להביא את השגיאה למינימום.

3. (10 pts) Computational complexity of k-NN

You are given a dataset of n labeled samples, where each input sample is a vector in a d-dimensional Euclidean space $x_1, \ldots, x_n \in R^d$. You wish to apply a k-NN algorithm using the Euclidean distance as the distance measure.

(a) What is the runtime complexity and memory complexity,in terms of d and n, for training the classifier?

(b) What is the runtime complexity and the memory complexity, in terms of d and n, for inferring the label of a new sample x?

3. a) בעצם שלב האימון ל$k - NN$ מוגדר ע״י קבלת ווקטורי האימון והתוויות המתאימות לכל ווקטור ושמירתם בזיכרון. מלבד לפעולה זו של שמירתם בזיכרון אין באלגוריתם הנ״ל פעולות נוספות מקדימות לחיזוי מכיוון שבהמשך בשלב החיזוי אנו נקבל ווקטור חדש ונבדוק את מרחקו מכל שאר הווקטורים ששמרנו בשלב האימון ואז נסתכל על k הווקטורים הקרובים ביותר (לפי מרחק אוקלידי לפי מה שהוגדר בשאלה) והתווית השכיחה ביותר עבור k ווקטורים אלו תהיה התווית שאנו נחזיר כחיזוי (ה$class$ שחזינו). לכן בשלב האימון ל$k - NN$ אנו בעצם רק שומרים את הווקטורים (שאנו מקבלים לאימון) ובנוסף את התווית המתאימה לכל ווקטור (ה$class$ המתאים לווקטור), על מנת לבצע שלב זה בעצם אין חישוב שעלינו לבצע (אנו רק מעתיקים שני מצביעים לשני מבנה הנתונים אחד של ווקטורי האימון ואחד של התווית של כל ווקטור, נציין שאם מחשיבים את ההעתקה זו כהעתקת כל הזיכרון של כל ווקטורי הדוגמאות אזי מכיוון שישנם n ווקטורים וכל ווקטור הוא ממימד d לכן נקבל $O(nd)$ זמן) ולכן מבחינת **סיבוכיות זמן** אנו משתמשים ב**$O(1)$ זמן**. ומבחינת **סיבוכיות מקום** אנו רק מעתיקים שני מצביעים לשני מבנה הנתונים אחד של ווקטורי האימון ואחד של התווית של כל ווקטור ולכן גם מבחינת סיבוכיות המקום אנו משתמשים בסה״כ **$O(1)$ מקום**. (נציין שאם אנו מחשיבים את הזיכרון של הקלט כחלק מהזיכרון של השלב הזה או אף מעתיקים אלינו את כל הזיכרון ולא רק מצביע אזי אנו משתמשים בn דוגמאות של ווקטורים כך שכל ווקטור הוא ממימד d ובנוסף שומרים את התווית (ה$class$ המתאים לווקטור) עבור כל ווקטור במערך נפרד בגודל n (מספר הדוגמאות שאנו מקבלים), וסה״כ נקבל $O(n \cdot d + n) = O(n \cdot d)$ מקום).

b) בשלב החיזוי של דוגמה x (חיזוי ה$class$ המתאים לו) אנו נקבל ווקטור חדש x ונבדוק את מרחקו מכל שאר הווקטורים ששמרנו בשלב האימון ואז נסתכל על k הווקטורים הקרובים ביותר (לפי מרחק אוקלידי לפי מה שהוגדר בשאלה) והתווית השכיחה ביותר עבור k ווקטורים אלו תהיה התווית שאנו נחזיר כחיזוי (ה$class$ שחזינו). לכן למעשה בשלב הראשוני אנו נחשב מרחק בין הווקטור x לבין כל שאר n ווקטורי האימון (מרחק אוקלידי) שעבור חישוב זה נדרש לנו מבחינת סיבוכיות מקום $O(n)$ מקום מכיוון שאנו בעצם מחשבים כל פעם מרחק בין x לאחד מn ווקטורי האימון ושומרים כל ערך כזה במערך שגודלו n מרחקים (כאשר ליד כל מרחק נשמור את האינדקס של הווקטור שמתאים למרחק זה מבין n-1 ווקטורי האימון). (פה ישנה הנחה שהמקום שהשתמשתי בו לשלב a אינו נכלל במקום בשלב b , אחרת סיבוכיות המקום היא $O(nd)$). ועבור חישוב המרחק האוקלידי n פעמים עבור ווקטורים ממימד d נדרש לסיבוכיות של $O(nd)$ זמן. בשלב השני של החיזוי אנו בעצם מוצאים מיהם k הווקטורים שמרחקם מx (מתוך כל הn ווקטורים שקיבלנו באימון) הוא המינימלי. עבור מציאת k ערכים אלו ניתן לבצע מציאת k פעמים ערך מינימלי (של ערכי המרחקים) והוצאתו מהמערך וכך נדע מיהם הk ווקטורים הקרובים ביותר לx ולאחר מכן נמצא את התווית השכיחה ביותר עבור k ווקטורים אלו תהיה התווית שאנו נחזיר כחיזוי (ה$class$ שחזינו), זאת באמצעות מערך של k משתני מנייה לכל היותר לחישוב התווית השכיחה ביותר- מעבר על כל k הווקטורים הקרובים ביותר ועבור כל ערך תווית ייחודי מבין כל התוויות של k ווקטורים אלו ישנו מונה מייצג. לאחר מכן במעבר שוב על מערך המונים נצמא ערך מקסימלי בעזרת משנה נוסף ומעבר על מערך זה וכך נמצא את ערך התווית

השכיחה ביותר,כאשר סיבוכיות זמן הריצה של אלגוריתם זה הוא $O(kn + 2k) = O(kn)$ זמן ומבחינת סיבוכיות מקום אנו משתמשים בעוד k משתנים שבהם אנו שומרים את k האינדקסים של k הווקטורים הקרובים ביותר לx ועוד k משתני מונים לכל היותר עבור התוויות הייחודיות של K הווקטורים הללו. ואם נרצה לשפר אלגוריתם זה ניתן לעשות זאת באמצעות אלגוריתם $select$ שמבצע $select$ על האיבר הk-י במערך המרחקים (על ערכי המרחקים) ומכיוון שאלגוריתם זה מבצע גם $partition$ למערך המרחקים כך שכל $k-1$ המרחקים המינימליים נמצאים משמאל (באינדקסים נמוכים יותר) במערך המרחקים לאיבר הk-י שנמצא במקומו (הk), וע"י האינדקס של הווקטורים ששמור ליד כל מרחק ניתן לדעת מי אלו k הווקטורים שמרחקם מx הוא המינימלי, ולאחר מכן נמצא את התווית השכיחה ביותר עבור k ווקטורים אלו תהיה התווית שאנו נחזיר כחיזוי (ה$class$ שחזינו), זאת באמצעות מערך של k משתני מנייה לכל היותר לחישוב התווית השכיחה ביותר- מעבר על כל k הווקטורים הקרובים ביותר ועבור כל ערך תווית ייחודי מבין כל התוויות של k ווקטורים אלו ישנו מונה מייצג. לאחר מכן במעבר שוב על מערך המונים נצמא ערך מקסימלי בעזרת משנה נוסף ומעבר על מערך זה וכך נמצא את ערך התווית השכיחה ביותר. ונקבל כי אלגוריתם זה פועל בזמן ובמקום ליניארי לאורך המערך של המרחקים שהינו $O(n + k) = O(n)$ זמן ומקום. כלומר סיבוכיות הזמן שנקבל לכל שלב החיזוי היא $\boldsymbol{O(nd)}$ **זמן** ו**סיבוכיות המקום** היא $\boldsymbol{O(n)}$ **מקום** (בהנחה שלא מחשיבים את המקום מסעיף a במידה והיא $O(nd)$ אחרת נקבל סיבוכיות מקום $O(nd)$).

4. (15 pts) Regularized polynomial regression

We derived in class the solution for a zero-degree polynomial regression. Consider the problem of regularized polynomial regression.

$$Err(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(h_w(x_i) - y_i)^2 + \lambda||\mathbf{w}||^2 \quad .$$

(a) Derive the solution for a polynomial of degree 0: $h_w(x) = w_0$. Analyze the solution in the limit of $\lambda \to \infty$ and $\lambda \to 0$.

(b) Derive the solution for a polynomial of degree 1, $h_{\mathbf{w}}(x) = w_0 + w_1 x$, by computing the derivatives w.r.t. w_0 and w_1 and writing a system of two linear equations in w_0 and w_1. No need to solve the system. Analyze the solution in the limit of $\lambda \to \infty$ and $\lambda \to 0$.

4.a) עבור פתרון פולינומאלי עם דרגה 0: $h_w(x) = w = w_0$ ננתח את הפתרון לw_0 (נשים לב כי $||w||^2 = w_0^2$ מכיוון ש$w = w_0$), כמו כן נשים לב שהפונקציה היא פונקציה ממעלה שנייה של w_0 כאשר המקדם של w_0^2 הינו חיובי (גם בתוך הסכום וגם λ) ולכן נקבל שלפונקציה זו ישנה נקודת מינימום אחת עבור השוואת הנגזרת ל0.

$$Err(w) = \frac{1}{n}\sum_{i=1}^{n}(h_w(x_i) - y_i)^2 + \lambda||w||^2 = \frac{1}{n}\sum_{i=1}^{n}(w_0 - y_i)^2 + \lambda w_0^2$$

כעת נגזור לפי w_0 ונשווה ל0 :

$$\left(\frac{1}{n}\sum_{i=1}^{n}(w_0^2 - 2w_0y_i + y_i^2) + \lambda w_0^2\right)' = \frac{1}{n}\sum_{i=1}^{n}(2w_0 - 2y_i) + 2\lambda w_0$$

$$= \frac{1}{n}\sum_{i=1}^{n}(2w_0 - 2y_i) + 2\lambda w_0$$

$$\frac{1}{n}\sum_{i=1}^{n}(2w_0 - 2y_i) + 2\lambda w_0 = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}(w_0 - y_i) + \lambda w_0 = 0$$

$$w_0 + \frac{1}{n}\sum_{i=1}^{n}(-y_i) + \lambda w_0 = 0$$

$$w_0(1 + \lambda) = \frac{1}{n}\sum_{i=1}^{n}y_i$$

$$w_0 = \frac{1}{n(1 + \lambda)}\sum_{i=1}^{n}y_i$$

כאשר אנו משאיפים את $\lambda \to \infty$ אזי נקבל כי $Err'(w) = \frac{1}{n}\sum_{i=1}^{n}(2w_0 - 2y_i) + 2\lambda w_0$ ישאף ל∞ $Err'(w) \to \infty$ כאשר $w_0 > 0$ כלומר הוא יתן משקל הרבה יותר גדול ל$2\lambda w_0$ (כלומר לערך המקדם w_0) מאשר ל$\frac{1}{n}\sum_{i=1}^{n}(2w_0 - 2y_i)$ (נגזרת ממוצע השגיאות של המדגם) שהוא זניח במקרה זה.

כאשר אנו משאיפים את $\lambda \to \infty$ אזי נקבל גם כי $w_0 \to 0$ מכיוון שλ במכנה תשאיף את כל הביטוי ל0. ועבור $w_0 \to 0$ נקבל את הפתרון שיביא לטעות מינימלית במקרה זה.

כאשר אנו משאיפים את $\lambda \to 0$ אזי נקבל כי $Err'(w) = \frac{1}{n}\sum_{i=1}^{n}(2w_0 - 2y_i) + 2\lambda w_0$ ישאף ל$\frac{1}{n}\sum_{i=1}^{n}(2w_0 - 2y_i)$ $Err'(w) \to \frac{1}{n}\sum_{i=1}^{n}(2w_0 - 2y_i)$ כלומר הוא יתן משקל הרבה יותר גדול ל- $\frac{1}{n}\sum_{i=1}^{n}(2w_0 - 2y_i)$ (נגזרת ממוצע השגיאות של המדגם) מאשר לערך של $2\lambda w_0$ שהוא זניח במקרה זה.

כאשר אנו משאיפים את $\lambda \to 0$ אזי נקבל גם כי $w_0 \to \frac{1}{n}\sum_{i=1}^{n} y_i$ שזה בדיוק הw_0 עבורו קיבלנו ערך מינימלי של טעות כמו שמתואר בתיאור שאלה 2 וביצענו בכיתה, כלומר אנו מתעלמים מהנורמה שהוספנו ומתייחסים רק לסכום ההפרשים של w_0 מכל y_i, כאשר כל הסכום מחולק בn כפי שניתן לראות $Err(w) = \frac{1}{n}\sum_{i=1}^{n}(h_w(x_i) - y_i)^2 + \lambda||w||^2$ כאשר מזניחים את המחובר $\lambda||w||^2$ אזי נקבל בדיוק את אותה הגדרה של טעות כמו בתיאור שאלה 2- $Err(w) =$ $\frac{1}{n}\sum_{i=1}^{n}(h_w(x_i) - y_i)^2$ שעבורו אנו יודעים את הפתרון שיביא לטעות מינימלית.

b) עבור פתרון פולינומאלי עם דרגה 1: $h_w(x) = w = w_0 + w_1 x$ ננתח את הפתרון ל$h_w(x)$ (נשים לב כי $||w||^2 = w_0^2 + w_1^2$ מכיוון ש$w = w_0 + w_1 x$). כמו כן נשים לב שהפונקציה היא פונקציה ממעלה שנייה של w_0 ושל w_1 כאשר המקדם של w_0^2 הינו חיובי (גם בתוך הסכום וגם λ), והמקדם של w_1^2 הינו חיובי (גם בתוך הסכום- x_i^2 וגם λ) ולכן נקבל שלפונקציה זו ישנה נקודת מינימום עבור השוואת הנגזרת ל0 גם לw_0 וגם לw_1.

$$Err(w) = \frac{1}{n}\sum_{i=1}^{n}(h_w(x_i) - y_i)^2 + \lambda||w||^2 = \frac{1}{n}\sum_{i=1}^{n}(w_0 + w_1 x_i - y_i)^2 + \lambda(w_0^2 + w_1^2)$$

כעת נגזור לפי w_0 ונשווה ל0:

$$\left(\frac{1}{n}\sum_{i=1}^{n}(w_0^2 + 2w_0(w_1 x_i - y_i) + (w_1 x_i - y_i)^2) + \lambda w_0^2 + \lambda w_1^2\right)'$$
$$= \frac{1}{n}\sum_{i=1}^{n}(2w_0 + 2(w_1 x_i - y_i)) + 2\lambda w_0$$
$$= \frac{1}{n}\sum_{i=1}^{n}(2w_0 + 2w_1 x_i - 2y_i) + 2\lambda w_0$$

$$\frac{1}{n}\sum_{i=1}^{n}(2w_0 + 2w_1 x_i - 2y_i) + 2\lambda w_0 = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}(w_0 + w_1 x_i - y_i) + \lambda w_0 = 0$$

$$w_0 + \frac{1}{n}\sum_{i=1}^{n}(w_1x_i - y_i) + \lambda w_0 = 0$$

$$w_0(1+\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - w_1x_i)$$

$$w_0 = \frac{1}{n(1+\lambda)}\sum_{i=1}^{n}(y_i - w_1x_i)$$

כעת נגזור לפי w_1 ונשווה ל0 :

$$\left(\frac{1}{n}\sum_{i=1}^{n}(w_1^2x_i^2 + 2w_1x_i(w_0 - y_i) + (w_0 - y_i)^2) + \lambda w_0^2 + \lambda w_1^2\right)'$$
$$= \frac{1}{n}\sum_{i=1}^{n}(2x_i^2w_1 + 2x_i(w_0 - y_i)) + 2\lambda w_1$$
$$= \frac{1}{n}\sum_{i=1}^{n}(2x_i^2w_1 + 2x_iw_0 - 2x_iy_i) + 2\lambda w_1$$

$$\frac{1}{n}\sum_{i=1}^{n}(2x_i^2w_1 + 2x_iw_0 - 2x_iy_i) + 2\lambda w_1 = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}(x_i^2w_1 + x_iw_0 - x_iy_i) + \lambda w_1 = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}x_i^2w_1 + \frac{1}{n}\sum_{i=1}^{n}x_iw_0 + \frac{1}{n}\sum_{i=1}^{n}-x_iy_i + \lambda w_1 = 0$$

$$w_1\sum_{i=1}^{n}x_i^2 + n\lambda w_1 = \sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}x_iw_0$$

$$w_1\left(\sum_{i=1}^{n}x_i^2 + n\lambda\right) = \sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}x_iw_0$$

$$w_1 = \frac{\sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}x_iw_0}{\sum_{i=1}^{n}x_i^2 + n\lambda}$$

כאשר אנו משאיפים את $\lambda \to \infty$ אזי נקבל כי $\frac{\partial Err(w)}{\partial w_0} = \frac{1}{n}\sum_{i=1}^{n}(2w_0 + 2w_1x_i - 2y_i) +$ $2\lambda w_0$ ישאף ל$\infty \leftarrow \frac{\partial Err(w)}{\partial w_0}$ כאשר $w_0 > 0$ כלומר הוא יתן משקל הרבה יותר גדול ל$2\lambda w_0$ (הערך של w_0) מאשר ל$\frac{1}{n}\sum_{i=1}^{n}(2w_0 + 2w_1x_i - 2y_i)$ (נגזרת ממוצע השגיאות של המדגם) שהוא זניח במקרה זה.

בנוסף נקבל כי $2\lambda w_1 + \frac{1}{n}\sum_{i=1}^{n}(2x_i^2 w_1 + 2x_i w_0 - 2x_i y_i) = \frac{\partial Err(w)}{\partial w_1}$ ישאף ל∞ $\frac{\partial Err(w)}{\partial w_1} \to$ כאשר $w_1 > 0$ כלומר הוא יתן משקל הרבה יותר גדול ל$2\lambda w_1$ (הערך של w_1) מאשר ל$\frac{1}{n}\sum_{i=1}^{n}(2x_i^2 w_1 + 2x_i w_0 - 2x_i y_i)$ (נגזרת ממוצע השגיאות של המדגם) שהוא זניח במקרה זה.

כאשר אנו משאיפים את $\lambda \to \infty$ אזי נקבל גם כי $w_0 \to 0$ מכיוון שλ במכנה תשאיף את כל הביטוי ל0. ובנוסף גם נקבל כי $w_1 \to 0$ מכיוון שλ במכנה תשאיף את כל הביטוי ל0. ועבור $w_1 \to 0, w_0 \to 0$ נקבל את הפתרון שיביא לטעות מינימלית במקרה זה.

כאשר אנו משאיפים את $\lambda \to 0$ אזי נקבל כי $\frac{\partial Err(w)}{\partial w_0} = \frac{1}{n}\sum_{i=1}^{n}(2w_0 + 2w_1 x_i - 2y_i) +$ $2\lambda w_0$ ישאף ל$\frac{\partial Err(w)}{\partial w_0} \to \frac{1}{n}\sum_{i=1}^{n}(2w_0 + 2w_1 x_i - 2y_i)$ כלומר הוא יתן משקל הרבה יותר גדול ל$\frac{1}{n}\sum_{i=1}^{n}(2w_0 + 2w_1 x_i - 2y_i)$ (נגזרת ממוצע השגיאות של המדגם) מאשר ל $2\lambda w_0$ (ערך של w_0) שהוא זניח במקרה זה.

בנוסף נקבל כי $2\lambda w_1 + \frac{1}{n}\sum_{i=1}^{n}(2x_i^2 w_1 + 2x_i w_0 - 2x_i y_i) = \frac{\partial Err(w)}{\partial w_1}$ ישאף ל0 $\frac{\partial Err(w)}{\partial w_1} \to$ כלומר הוא יתן משקל הרבה יותר גדול ל$\frac{1}{n}\sum_{i=1}^{n}(2x_i^2 w_1 + 2x_i w_0 - 2x_i y_i)$ (נגזרת ממוצע השגיאות של המדגם) מאשר ל$2\lambda w_1$ (ערך של w_1) שהוא זניח במקרה זה.

כאשר אנו משאיפים את $\lambda \to 0$ אזי נקבל גם כי $w_0 \to \frac{1}{n}\sum_{i=1}^{n}(y_i - w_1 x_i)$. ובנוסף גם נקבל כי $w_1 \to \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i w_0}{\sum_{i=1}^{n} x_i^2}$. ועבור ערכים אלו נקבל את הפתרון שיביא לטעות מינימלית במקרה זה.

5. (10 pts) PAC learning: Sample-Complexity Monotonicity

Let $\mathcal{H}$ be a hypothesis class for a binary classification task. Suppose that $\mathcal{H}$ is PAC learnable and its sample complexity is given by $N(\epsilon, \delta)$. Show that N is monotonically non-increasing in each of its parameters. That is, show that given $\delta \in (0,1)$, and given $0 < \epsilon_1 \le \epsilon_2 < 1$, we have that $N(\epsilon_1, \delta) \ge N(\epsilon_2, \delta)$. Similarly, show that given $\epsilon \in (0,1)$, and given $0 < \delta_1 \le \delta_2 < 1$, we have that $N(\epsilon, \delta_1) \ge N(\epsilon, \delta_2)$.

5. תהי H מחלקת היפותזות עבור מטלת קלסיפיקציה בינארית. נניח שH היא $PAC\ learnable$ ולכן קיים אלגוריתם A שמוצא היפותזה $h^* = A(S)$ כך שלכל $\varepsilon, \delta > 0$ קיימת $N(\varepsilon, \delta)$ כך שלכל $n > N(\varepsilon, \delta)$ כך שלכל פיזור D מתקיים- $\Pr\left[Err_D\left(A(S_n)\right) < \varepsilon\right] > 1 - \delta$. כעת נוכיח ש$N$ הינה מונוטונית לא עולה לכל אחד מהפרמטרים ε, δ שלה. נחלק למקרים-

- נתון $\delta \in (0,1)$ ובהינתן $0 < \varepsilon_1 \le \varepsilon_2 < 1$ נוכיח כי $N(\varepsilon_1, \delta) \ge N(\varepsilon_2, \delta)$.
לפי הגדרת H כ$PAC\ learnable$ ומכיוון ש$\varepsilon_1, \varepsilon_2, \delta > 0$ אזי נקבל כי עבור ε_1, δ קיימת $N(\varepsilon_1, \delta)$ כך שלכל $n > N(\varepsilon_1, \delta)$ מתקיים $\Pr\left[Err_D\left(A(S_n)\right) < \varepsilon_1\right] > 1 - \delta$. בנוסף עבור $\varepsilon_2, , \delta$ קיימת $N(\varepsilon_2, \delta)$ (נבחר את ה$N(\varepsilon_2, \delta)$ המינימלית שמקיימת תנאי זה) כך שלכל $n > N(\varepsilon_2, \delta)$ מתקיים $\Pr\left[Err_D\left(A(S_n)\right) < \varepsilon_2\right] > 1 - \delta$. מכיוון ש$\varepsilon_1 \le \varepsilon_2$ לכן עבור $N(\varepsilon_1, \delta)$ לכל $n > N(\varepsilon_1, \delta)$ מתקיים $\Pr\left[Err_D\left(A(S_n)\right) < \varepsilon_1 \le \varepsilon_2\right] > 1 - \delta$ כלומר $\Pr\left[Err_D\left(A(S_n)\right) < \varepsilon_2\right] > 1 - \delta$ ולכן קיבלנו כי $N(\varepsilon_1, \delta)$ מקיימת את התנאי הרצוי לPAC עבור ε_2, δ וכי מתקיימת הסתברות זו לכל $n > N(\varepsilon_2, \delta)$ לכן סה"כ קיבלנו כי $N(\varepsilon_1, \delta) \ge N(\varepsilon_2, \delta)$ (מכיוון שבחרנו את ה$N(\varepsilon_2, \delta)$ המינימלית שמקיימת תנאי זה ו$N(\varepsilon_1, \delta)$ גם מקיימת תנאי זה).
- נתון $\varepsilon \in (0,1)$ ובהינתן $0 < \delta_1 \le \delta_2 < 1$ נוכיח כי $N(\varepsilon, \delta_1) \ge N(\varepsilon, \delta_2)$.
לפי הגדרת H כ$PAC\ learnable$ ומכיוון ש$\varepsilon, \delta_1, \delta_2 > 0$ אזי נקבל כי עבור ε, δ_1 קיימת $N(\varepsilon, \delta_1)$ כך שלכל $n > N(\varepsilon, \delta_1)$ מתקיים $\Pr\left[Err_D\left(A(S_n)\right) < \varepsilon\right] > 1 - \delta_1$. בנוסף עבור ε, δ_2 קיימת $N(\varepsilon, \delta_2)$ (נבחר את ה$N(\varepsilon, \delta_2)$ המינימלית שמקיימת תנאי זה) כך שלכל $n > N(\varepsilon, \delta_2)$ מתקיים $\Pr\left[Err_D\left(A(S_n)\right) < \varepsilon\right] > 1 - \delta_2$. מכיוון ש$\delta_1 \le \delta_2$ לכן $1 - \delta_1 \ge 1 - \delta_2$ לכן עבור $N(\varepsilon, \delta_1)$ לכל $n > N(\varepsilon, \delta_1)$ מתקיים $\Pr\left[Err_D\left(A(S_n)\right) < \varepsilon\right] > 1 - \delta_1 \ge 1 - \delta_2$ לכן $\Pr\left[Err_D\left(A(S_n)\right) < \varepsilon\right] > 1 - \delta_2$ ולכן קיבלנו כי $N(\varepsilon, \delta_1)$ מקיימת את התנאי הרצוי לPAC עבור ε, δ_2 וכי מתקיימת הסתברות זו לכל $n > N(\varepsilon, \delta_2)$ לכן סה"כ קיבלנו כי $N(\varepsilon, \delta_1) \ge N(\varepsilon, \delta_2)$ (מכיוון שבחרנו את ה$N(\varepsilon, \delta_2)$ המינימלית שמקיימת תנאי זה ו$N(\varepsilon, \delta_1)$ גם מקיימת תנאי זה).

וסה"כ הוכחנו עי N הינה מונוטונית לא עולה לכל אחד מהפרמטרים ε, δ שלה כדרוש.

6. (20 pts) PAC learnability of L2-balls around the origin

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

$$h_r = \begin{cases} 1 & ||x||_2 < r \\ 0 & otherwise \end{cases} \qquad (1)$$

Consider the hypothesis class $\mathcal{H} = \{h_r | r > 0\}$. Prove directly (without just using the fundamental theorem of PAC learning) that it 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.

6. נוכיח כי $H = \{h_r | r > 0\}$ היא $PAC\ learnable$. לפי ההגדרה של $PAC\ learnable$ נוכיח שקיים אלגוריתם A שמוצא היפותזה $h_{r^*} = A(S)$ כך שלכל $\varepsilon, \delta > 0$ קיימת $N(\varepsilon, \delta)$ כך שלכל $n > N(\varepsilon, \delta)$ כך שלכל פיזור D מתקיים- $\Pr[Err_D(A(S_n)) < \varepsilon] > 1 - \delta$.

נתאר אלגוריתם A אשר בהינתן קבוצת הדגימות S_n (n דגימות של ווקטורים- x_i השייכים ל $\mathbb{R}^d$ ולכל ווקטור מוצמדת תווית y_i השייכת ל $\{0,1\}$) בוחרת את ההיפותזה $h_{r_0} \in H$ הצמודה ביותר לדגימות, כלומר נבחר את r_0 להיות שווה לערך המקסימלי של $\max_{i \in \{j | j \in [n] \wedge y_j = 1\}} ||x_i||_2$ מבין כל הווקטורים שאנו מקבלים בS_n כך שהתיוג שלהם הוא 1, ובכך בעצם קיבלנו את הr_0 המינימלי שעבורו מתקיים לכל $i \in \{1, \dots, n\}$ כך ש$y_i = 1$ כי $||x_i||_2 < r_0$ ולכן בעצם עבור ההיפותזה שבחרנו לכל $i \in \{1, \dots, n\}$ כך ש $y_i = 1$ מתקיים כי $h_{r_0}(x_i) = 1$. מאחר ונתון כי זה $realizable\ case$ נקבל כי ההיפותזה האמיתית (זו שאנו מחפשים ונכונה עבור כל הפיזור D) $h_r \in H$. ונקבל מכך שהטעות היחידה שיכולה להיות לנו בהיפותזה h_{r_0} שבחרנו בעזרת האלגוריתם A היא טעות חד צדדית- כלומר עבור $x \in \mathbb{R}^d$ כלשהו אנו נסווג אותו בh_{r_0} כ0 כאשר זו טעות והסיווג האמיתי שלו (בh_r האמיתית) הוא 1, וזאת מכיוון ש$r_0 \leq r$ ולכן בהכרח כל x_i מקבוצת הדגימות שלנו עבורו $h_{r_0} = 1$ נקבל כי לפי הגדרת h_{r_0} אזי גם $y_i = 1$ ולכן בעצם נקבל כי $h_r(x_i) = y_i = 1$, כמו כן לכל $x \in \mathbb{R}^d$ בפיזור D נקבל כי אם $||x||_2 < r_0$ אזי מכיוון ש $r_0 \leq r$ לכן גם $||x||_2 < r$ ולכן מתקיים כי $h_r(x) = h_{r_0}(x) = 1$ כאשר תנאי זה מתקיים ($||x||_2 < r_0$), ולא תיתכן טעות מהסוג הזה (כלומר שאנו מסווגים ב h_{r_0} ווקטור שאנו מקבלים כ1, אך עבור h_r הוא מסווג כ0).

נרצה להראות כי מתקיים תנאי ה$PAC\ learnable$. יהיו $\varepsilon, \delta > 0$ נוכיח כי קיים $N(\varepsilon, \delta)$ כך שלכל $n > N(\varepsilon, \delta)$ ולכל פיזור D מתקיים $\Pr[Err_D(A(S_n)) < \varepsilon] > 1 - \delta$.

נסמן את המרחב (השטח הרב מימדי) בין המרחב שמוגדר ע״י r (כל השטח הרב מימדי שבתוכו נגדיר כי לכל $x \in \mathbb{R}^d$ מתקיים $h_r\ (x) = 1$ כלומר $||x||_2 < r$) לבין המרחב שמוגדר ע״י r_0 (כל השטח הרב מימדי שבתוכו נגדיר כי לכל $x \in \mathbb{R}^d$ מתקיים $h_{r_0}\ (x) = 1$ כלומר $||x||_2 < r$) ע״י M (כלומר חיסור בין השטחים הרב מימדיים שמגדיר h_r לבין h_{r_0}). ובעצם נצטרך שההסתברות של ווקטור במרחב D ליפול בתוך המרחב M (מרחב הטעות שלנו כי שם יכולה להיות טעות חד צדדית על נקודה בפילוג D כמו שתיארנו למעלה) תהיה קטנה מε. כלומר נרצה שההסתברות של דגימה כלשהי x_i בקבוצת הדגימות S_n ליפול מחוץ למרחב M תהיה גדולה מ$(1 - \varepsilon)$. ונקבל שעבור n דגימות בלתי תלויות אזי נרצה שההסתברות שכולן יפלו מחוץ למרחב M תהיה גדולה מ$(1 - \varepsilon)^n$. ונשים לב כי ההסתברות לכך שכל n הדגימות בקבוצת הדגימות S_n יהיו מחוץ

למרחב M זו בעצם ההסתברות לבחור h_{r_0} שהוא אינו טוב ויביא לטעות גדולה או שווה בהסתברותו מε , כלומר זו ההסתברות הבאה- $1-\Pr\left[Err_D\left(A(S_n)\right)<\varepsilon\right]$ עקב $A(S_n)=h_{r_0}$ וההסתברות לכך שהטעות של h_{r_0} על פני כל הנקודות במרחב תהיה קטנה מε היא השטח הרב מימדי של המרחב M ונרצה שההסתברות לבחור סט של n דגימות תביא לε > $Err_D\left(A(S_n)\right)$ תהיה קטנה מ $1-\delta$, כלומר נרצה ש$1-\Pr\left[Err_D\left(A(S_n)\right)<\varepsilon\right]=(1-\varepsilon)^n<\delta$. כעת נזכיר את אי השוויון הבא עבור y חיובי- $1-y\le e^{-y}$. ולכן

$$(1-\varepsilon)^n\le e^{-\varepsilon n}<\delta$$

$$-\varepsilon n<\ln(\delta)$$

$$n>-\frac{1}{\varepsilon}\ln(\delta)=\frac{1}{\varepsilon}\ln\left(\frac{1}{\delta}\right)$$

לסיכום קיבלנו כי לכל $n>N(\varepsilon,\delta)=\frac{1}{\varepsilon}\ln\left(\frac{1}{\delta}\right)$ מתקיים $1-\Pr\left[Err_D\left(A(S_n)\right)<\varepsilon\right]<\delta$ כלומר $\Pr\left[Err_D\left(A(S_n)\right)<\varepsilon\right]>1-\delta$.

ולכן **הוכחנו כי $\boldsymbol{H=\{h_r|r>0\}}$ היא אכן *PAC learnable*.**

נשים לב כי ה*sample complexity* הוא $N(\varepsilon,\delta)$ ונשים לב שלאורך כל התשובה שלנו אין התייחסות כלל למימד d ולכן למעשה לא משנה מה הוא d נקבל את אותה $N(\varepsilon,\delta)$ ולכן **ה*sample complexity* אינו תלוי במימד *d*.**