פיתרון מועד א' – טור 1

פיתרון שאלה 1 [10 נק']

ב. [9 נק'] נפתח את הפונקציה שיש להביא למינימום תוך השמטת גורמים שאינם תלויים בסיווג ושימוש בתכונות נוספות של argmax.

$$\hat{y} = \mathrm{argmax}_{\mathrm{y}} \underbrace{\Pr[y]}_{=1/3} \cdot \Pr[X_1 = x_1|y] \cdot \Pr[X_2 = x_2|y]$$

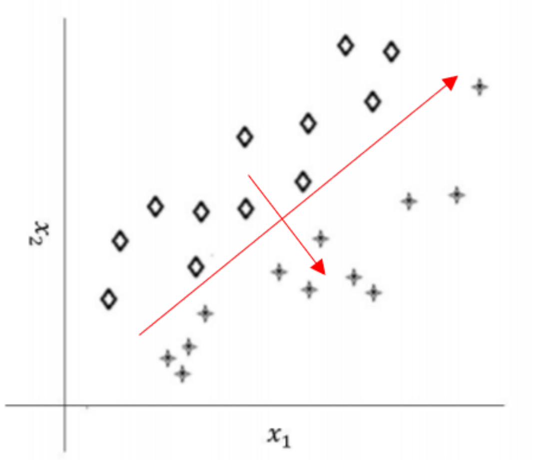
$$= \mathrm{argmax}_{\mathrm{y}} \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left\{-\frac{\left(x_1 - \mu_{y1}\right)^2}{2\sigma_1^2}\right\} \cdot \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left\{-\frac{\left(x_2 - \mu_{y2}\right)^2}{2\sigma_2^2}\right\}$$

$$= \mathrm{argmax}_{\mathrm{y}} \exp\left\{-\frac{\left(x_1 - \mu_{y1}\right)^2}{8} - \frac{\left(x_2 - \mu_{y2}\right)^2}{2}\right\} = \mathrm{argmin}_{\mathrm{y}} \left[\left(x_1 - \mu_{y1}\right)^2 + 4\left(x_2 - \mu_{y2}\right)^2\right]$$

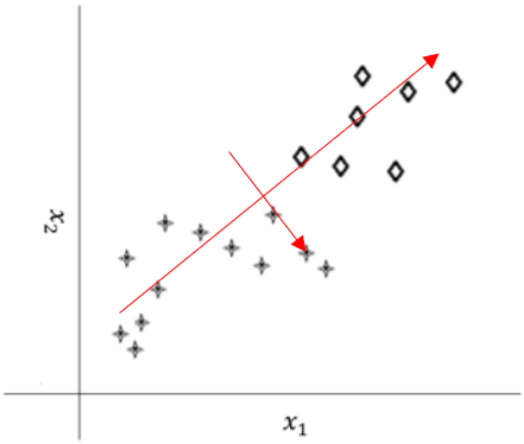
נותר להציב את הערכים המתאימים לפי סעיף א' לכל סיווג $y = 1,2,3$ ולבחור את הסיווג שמחזיר ערך מיטבי.

פיתרון שאלה 2 [20 נק']

בגרפים הבאים מובאים לפניכם שני מערכי נתונים ממימד 2. כל דוגמה מסווגת "◊" או "+".



מערך A



מערך B

א. [4 נק'] ציירו במחברת את כל ה-principal components (PC's) עבור כל אחד מהמערכים (אין צורך להעתיק את מערכי הנתונים, אבל יש לצייר מערכת צירים ברורה ואת הווקטורים המבוקשים באופן שהכיוון שלהם ברור, וכך שיהיה ברור מי ה-PC הראשי מבין אלה שציירתם). ראו ציור.

ב. [8 נק'] האם ניתן יהיה לסווג נכונה את התצפיות שבמערכים בעזרת מפריד לינארי הפועל על הנקודות כשהן מוטלות על ה-PC הראשון בלבד? הסבירו בקצרה.

מערך A: לא ניתן להפריד בין הנק' המוטלות על ה PC הראשון בעזרת מפריד לינארי. ניתן לראות שאילו היו מוטלות הנקודות על ה PC המשני אזי כן היה ניתן להפריד אותן בעזרת מפריד לינארי.

מערך B: כן, הנקודות המוטלות על ה PC הראשון פרידות לינארית.

ג. [8 נק'] לכל אחת מהטענות הבאות, כתבו במחברת התשובות האם היא <u>נכונה</u> או <u>לא נכונה</u> (אין צורך להסביר).

a. המטרה של PCA היא לפרש את המבנה הבסיסי של הנתונים במונחים של הרכיבים העיקריים הטובים ביותר לחיזוי משתנה הפלט (label). לא נכון, PCA מקטין את שגיאת ה-Reconstruction והוא בלתי תלוי ב-Labeling.

b. PC's ששונים מ-0 תמיד יהיו מאונכים זה לזה. נכון.

c. למערך נתונים d-מימדי (dataset המיוצג כמטריצה) תמיד יהיו בדיוק d רכיבים עיקריים (PC's) שונים מ-0.

לא נכון, דוגמא נגדית: מערך נתונים עם מאפיינים קורלטיביים.

d. ניתן לחשב התמרת k-PCA (הטלה ל-k הרכיבים הרכיבים העיקריים הראשונים) באופן שקול על-ידי אימון רשת עמוקה מסוג autoencoder עם שיכבה אחת נסתרת ברוחב k ופונקציית אקטיבציה טריוויאלית (פונק' הזהות).

נכון, הרשת בעצם טרנספורמציה לינארית עם שמקטינה את ה-Reconstruction error בדיוק כמו ה-PCA.

פיתרון שאלה 3 [25 נק']

היזכרו בבעיית ה-Least squares (LS): $\hat{\boldsymbol{w}} = \operatorname{argmin}_{\boldsymbol{w}\in\mathbb{R}^d}\|\mathbf{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 = \operatorname{argmin}_{\boldsymbol{w}\in\mathbb{R}^d}\sum_{i=1}^{m}(\boldsymbol{w}^\top\boldsymbol{x}_i - y_i)^2$

א. [4 נק'] יש כמה דרכים לראות את הפיתרון. נדגים אחת.

ראינו בקורס (ומהנוסחה לעיל) שהנורמה מתפרקת על שורות המטריצה $\mathbf{X}$ ועל שורות $\boldsymbol{y}$.

לכן, על ידי הוספת שורות, נבנה $\boldsymbol{y}' = \begin{bmatrix}\boldsymbol{y}\\ \mathbf{0}_{d\times 1}\end{bmatrix}$, $\mathbf{X}' = \begin{bmatrix}\mathbf{X}\\ \sqrt{\lambda}\mathbf{I}_{d\times d}\end{bmatrix}$.

מתקבל:

$$\left\|\begin{bmatrix}\mathbf{X}\\ \sqrt{\lambda}\mathbf{I}_{d\times d}\end{bmatrix}\boldsymbol{w} - \begin{bmatrix}\boldsymbol{y}\\ \mathbf{0}_{d\times 1}\end{bmatrix}\right\|_2^2 = \|\mathbf{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 + \|\sqrt{\lambda}\mathbf{I}_{d\times d}\boldsymbol{w} - \mathbf{0}_{d\times 1}\|_2^2 = \|\mathbf{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2$$

$\begin{bmatrix} x_1^1 & \cdots & x_2^1 \\ \vdots & & \vdots \\ x_1^d & \cdots & x_2^d \\ \lambda & & \\ & \lambda & \\ & & \ddots \\ & & & \lambda \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$

ב. [8 נק'] נציג פיתרון אחד (התקבלו גם פיתרונות נוספים, למשל שעמודות $\mathbf{X}$ לא מנורמלות וה-*features* בסדרי גודל שונים).

נניח ש-100 העמודות הראשונות של $\mathbf{X}$ כמעט זהות, למשל עמודות רעש או עמודות זהות ממש לצורך הפשטות.

נניח שאחריהן יש עוד 2 עמודות שמסבירות את וקטור התיוגים $\boldsymbol{y}$ באופן מושלם: $y(x) = 5x_{101} + 2x_{102}$.

שאר העמודות לא משנות.

הקופסה השחורה עשויה להחזיר את וקטור המשקלים $\boldsymbol{w} = \Big[\underbrace{10^8, 10^8, \dots}_{50\ weights}, \underbrace{-10^8, -10^8, \dots}_{50\ weights}, 5, 2, \underbrace{0, 0, \quad \dots \quad, 0}_{10^5-102\ weights}\Big]$,

שעבורו שגיאת האימון היא 0.

השיטה שהציע הצוות תחזיר את וקטור המשקלים ה"דליל" $\boldsymbol{w} = \Big[\underbrace{10^8, 10^8, \dots}_{50\ weights}, \underbrace{-10^8, -10^8, \dots}_{50\ weights}, \underbrace{0, 0, \quad \dots \quad, 0}_{10^5-100\ weights}\Big]$

שעבורו שגיאת האימון גבוהה.

עם זאת, קיים פיתרון דליל $\boldsymbol{w} = \Big[\underbrace{0, 0, \quad \dots \quad, 0}_{100\ weights}, 5, 2, \underbrace{0, 0, \quad \dots \quad, 0}_{10^5-102\ weights}\Big]$ שעבורו שגיאת האימון היא 0.

ג. [5 נק'] בכלליות רגולריזציה יכולה לעזור כי היא לא תאפשר משקלים כמו 10^8 עבור מאפיינים שלא תורמים להפרדה.

ובפרט, למדנו שרגרסיה ליניארית עם רגולריזציית $\mathcal{L}^1$ (LASSO) נוטה להחזיר פתרונות דלילים.

ד. [8 נק'] מטילים את הנתונים בעזרת PCA על ידי $\widetilde{\mathbf{X}} = \underbrace{\mathbf{X}}_{m\times 10^5} \cdot \underbrace{\mathbf{V}^{(100)}}_{10^5\times 100}$ ואז פותרים LS ומקבלים $\hat{\boldsymbol{w}}^{(100)} = LS(\widetilde{\mathbf{X}}, \boldsymbol{y})$.

ד.1. שיטה זו יכולה למנוע overfitting כי: (1) PCA נוטה להחליק רעשים ולעיתים התאמת-היתר נובעת מרעשים,

ובנוסף (2) המעבר לרגרסיה במימד נמוך יותר, מקטין את מרחב ההיפותזות ומקשה על הגעה להתאמת יתר.

ד.2. מודל החיזוי יהיה $\mathbf{X}\hat{\boldsymbol{w}} = \mathbf{X}\underbrace{\mathbf{V}^{(100)}\hat{\boldsymbol{w}}^{(100)}}_{\hat{\boldsymbol{w}}} = \widetilde{\mathbf{X}}\hat{\boldsymbol{w}}^{(100)}$ והוא <u>ליניארי</u> ביחס לנתונים המקוריים.

פיתרון שאלה 4 [15 נק']

א. [7 נק'] איזה מודל תבחרו? נמקו את בחירתכם.

מודל A

FP = 9	TN = 91
TP = 78	FN = 22

המודל לא עומד בדרישה $(TPR = 78\%)$

מודל B

FP = 4	TN = 96
TP = 90	FN = 10

המודל עומד בדרישות עם עלות:

$$cost_B = 10 \cdot x + 4 \cdot 5x = 30x$$

מודל C

FP = 1	TN = 99
TP = 79	FN = 21

המודל לא עומד בדרישה $(TPR = 79\%)$

מודל D

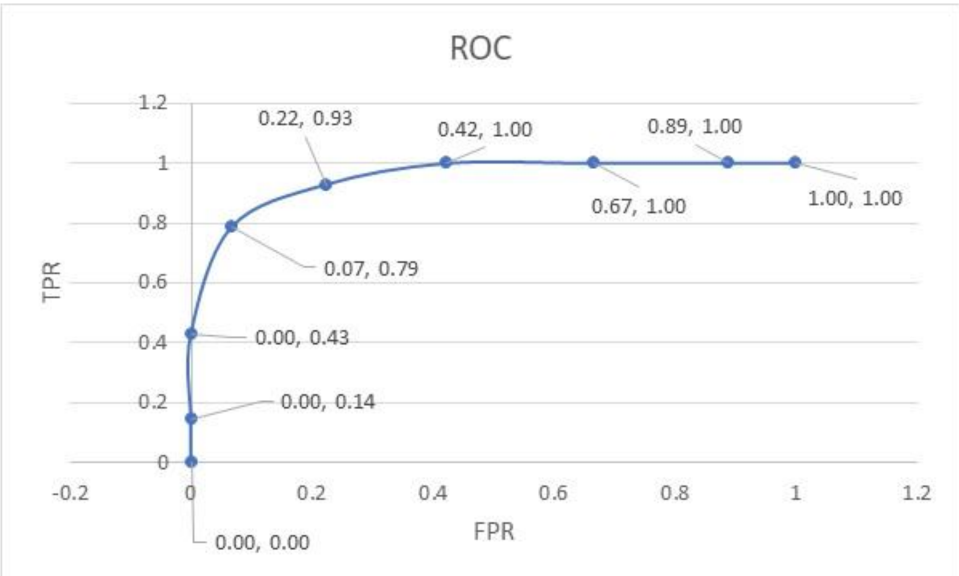
FP = 2	TN = 98
TP = 82	FN = 18

המודל עומד בדרישות עם עלות:

$$cost_B = 18 \cdot x + 2 \cdot 5x = 28x$$

ולכן התשובה היא מודל D (בטור 2 התשובה הייתה מודל A).

ב. [8 נק'] עקומת ROC מציגה את ביצועי המודל על פי שיעור החיובים האמיתיים (TPR) מול שיעור החיובים הכוזבים (FPR) כנגד סף החלטה משתנה (decision threshold)



פיתרון שאלה 5 [30 נק'] + מפתח בדיקה

א. **יכולות לעזור:** הוספת דוגמאות, גריעה של מאפיינים, הוספה של הנחות מקדימות.
הערה: גם הוספה של מאפיינים יכולה לעזור במקרים מסוימים, לכן לא ירדו נקודות למי שסימן אפשרות זו, אולם מי שלא סימן "גריעה של מאפיינים" ירדו לו נקודות.

ב. ERM הוא אלגוריתם אשר בוחר, מתוך קבוצת היפותזות, את ההיפותזה שמביאה למינימום את שגיאת האימון.

ג. דוגמאות להשלמת מאפיינים חסרים: החלפת המאפיינים בממוצע המאפיין על פני הדוגמאות הלא חסרות, אלגוריתם EM (כפי שנלמד בכיתה).

ד. להלן התשובות:

a. מקדם פירסון בתחום הממשי $[-1,1]$ – **נכון**
b. אם בוחרים היפותזה לפי חצי ראשון של הנתונים, אז השגיאה האמפירית על החצי השני הוא משערך בלתי מוטה של שגיאת ההכללה – **נכון**
c. רגרסיה לוגיסטית היא אלגוריתם לסיווג בינארי באמצעות מפריד לינארי – **נכון**
d. פונקצית הלוגריתם של הסיגמואיד היא לא קמורה ולא קעורה - **לא נכון** (היא קעורה)
e. אלגוריתם EM בכל איטרציה מוריד את פונקציית ה-Log-likelihood או משאיר אותה ללא שינוי - **לא נכון** (הוא מעלה או משאיר ללא שינוי)

ה. גודל רצוי של test set לבחירה של אחת מתוך n היפותזות - לוגריתמי ב-n

ו. עבור נתונים $x_1 \dots x_n, k$ פונקציית המטרה היא $f(C_1 \dots C_k) = \sum_{i=1}^{k} min_{\mu} \sum_{x \in C_i} ||x-\mu||^2$ כאשר האשכולות הם $C_1 \dots C_k$.

הערות: מי שכתב ביטוי שקול (למשל, במקום מינימיזציה על μ בתוך הסכום, לקח את μ להיות הממוצע על אברי C_i) קיבל את מלוא הנקודות. מי שבמקום $||x-\mu||^2$ רשם משהו כדוגמת $(x-\mu)^2$, גם זכה למלוא הניקוד.
במילים אחרות, כל מי שהראה בוודאות שהוא יודע שמדובר במרחק אוקלידי בריבוע קיבל את מלוא הנקודות.
לעומת זאת, מי שכתב ביטויים כגון $||x-\mu||$ או $(x-\mu)$ (בלי העלאה בריבוע) או פונקציות מרחק כלליות כגון $d(x,\mu)$ או $d^2(x,\mu)$ - ירדו לו נקודות.
מי שהוסיף ביטוי נירמול שונים (למשל חלוקה ב $|C_i|$ בתוך הסכום... והיו עוד דוגמאות נירמול שונות) - ירדו לו נקודות.

ז. השלמת האלגוריתם:

Step A: חישוב של מרכזי האשכולות (ממוצע) $\forall i = 1..k: \mu_i = \sum_{j:\, a[j]=i} x_j \;/\; \#\{j: a[j] = i\}$

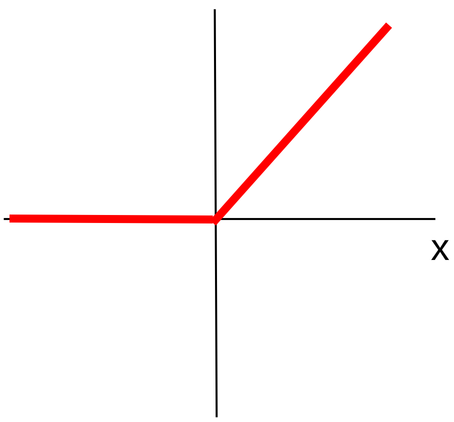
(<u>הערה</u> 1: יש המון דרכים שקולות לכתוב זאת כפסאודוקוד, כולן התקבלו)

(<u>הערה</u> 2: בעיקרון יש לבדוק שהמכנה ב Step A הוא לא 0, אבל זה "אפקט מסדר שני" ולא נדרש בבחינה)

Step B: שיוך מחדש של נקודות: $\forall j = 1..n: a[j] = argmin_{i=1..k} ||x_j - \mu_i||$

(<u>הערה</u>: כאן אפשר אבל לא חובה להעלות את המרחק בריבוע, כי זה שקול, וגם כאן התקבלו ביטויים כגון $(x_j - \mu_i)$, כלומר בלי סימן של נורמה אוקלידית.)

ח. פונקציית ReLU: $max(0, x)$



ט. פונקציית hinge loss (עבור y=+1): $max(0, 1 - z)$