מבוא למערכות לומדות (236756)

סמסטר חורף תשפ"א – 08 בפברואר 2021

מרצה: פרופ' ניר אילון

מבחן מסכם מועד א' – טור 1

הנחיות הבחינה:

- **משך הבחינה:** 2.5 שעות.
- **חומר עזר:** המבחן בחומר סגור (ללא ספרים, מחברות, דפי נוסחאות).
- מותר השימוש במחשבון רגיל בלבד.
- במבחן 7 דפים ממוספרים סה"כ, כולל עמוד זה שמספרו 1.
- במבחן 5 שאלות, יש לענות על כולן.
- יש לכתוב את תשובותיכם המנומקות על דפים בכתב יד קריא. תשובה בכתב יד שאינו קריא לא תיבדק.
- יש לכתוב את מספר תעודת הזהות שלכם <u>ואת מספר הטור</u> בראש דף התשובות הראשון שלכם.
- בתום המבחן יש לסרוק את כל דפי התשובות שלכם לפי סדרם.
- נא לכתוב רק את שהתבקשתם ולצרף הסברים קצרים עפ"י ההנחיות.

בהצלחה!

שאלה 1 [10 נק']

היזכרו במסווג Naïve Bayes עבור שני מימדים:

$$\hat{y} = \text{argmax}_{\text{y}} \Pr[y] \cdot \Pr[X_1 = x_1|y] \cdot \Pr[X_2 = x_2|y]$$

בשאלה זו נשתמש בגרסת ה-Gaussian NB ונמדל את ההסתברות בעזרת ההתפלגות הנורמלית:

$$Pr[X_j = x_j|y] = \frac{1}{\sigma_j\sqrt{2\pi}}\exp\left\{-\frac{(x_j - \mu_{yj})^2}{2\sigma_j^2}\right\}$$

נסווג וקטור דו-מימדי $\boldsymbol{x} \in \mathbb{R}^2$ לאחת מ-3 מחלקות.

את הנתונים נגדיר באמצעות מס' תעודת הזהות שלכם.

שתי הספרות הראשונות (משמאל) יגדירו את התוחלת $\boldsymbol{\mu}_1$ של המחלקה הראשונה. השתיים הבאות יגדירו את התוחלת $\boldsymbol{\mu}_2$ והבאות את $\boldsymbol{\mu}_3$. את הווקטור $\boldsymbol{x}$ יגדירו שתי הספרות הבאות במספר תעודת הזהות.

למשל, אם מספר תעודת הזהות שלכם הוא 123456789, אלה יהיו הנתונים:

$$\boldsymbol{\mu}_1 = \begin{bmatrix}1\\2\end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix}3\\4\end{bmatrix}, \boldsymbol{\mu}_3 = \begin{bmatrix}5\\6\end{bmatrix}, \boldsymbol{x} = \begin{bmatrix}7\\8\end{bmatrix}$$

שימו לב: אסור שבין ארבעת הווקטורים יהיו שני וקטורים שווים. לכן, אם יצאו לכם שני וקטורים שווים, למשל $\boldsymbol{\mu}_1 = \boldsymbol{x}$, החליפו את הווקטור המאוחר יותר (לפי הסדר לעיל) בווקטור $\begin{bmatrix}4\\7\end{bmatrix}$ (ציינו זאת במבחן).

נניח התפלגות אחידה על המחלקות – משמע $\Pr[y = 1] = \Pr[y = 2] = \Pr[y = 3]$.

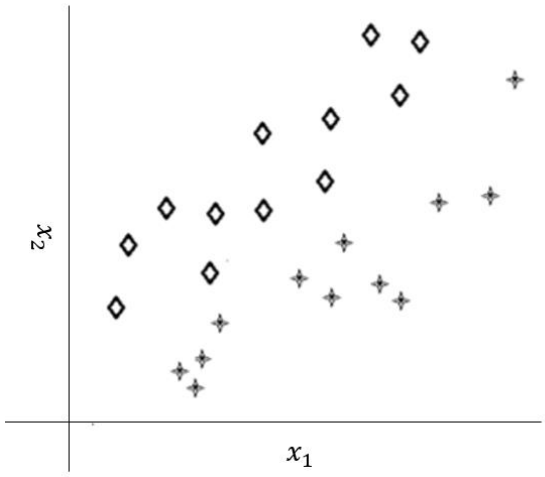
לסיום, נגדיר $\sigma_1^2 = 4,\ \sigma_2^2 = 1$.

א. [1 נק'] כתבו במחברת את $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{x}$ לפי מספר תעודת הזהות שלכם.

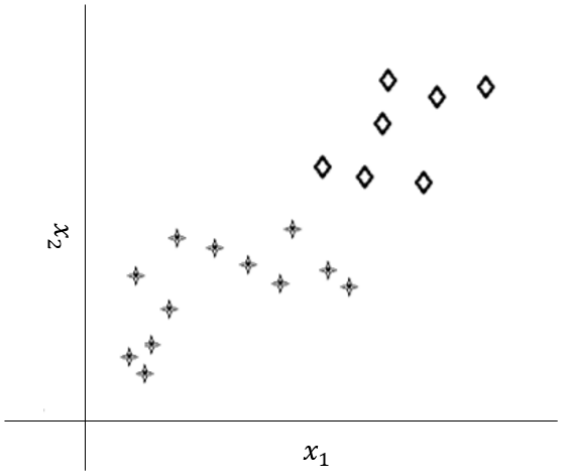
ב. [9 נק'] מצאו את הסיווג $\hat{y}$ שהמודל יחזה עבור ה-$\boldsymbol{x}$ שכתבתם. הסבירו וצרפו חישובים רלוונטיים.

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).

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.

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).

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

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

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).

שאלה 3 [25 נק']

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

לאורך כל הסעיפים בשאלה הניחו כי ברשותכם קופסה שחורה LS, שמחזירה פיתרון שלמעלה.

משמע, בהינתן $\mathbf{X}\in\mathbb{R}^{m\times d}, \boldsymbol{y}\in\mathbb{R}^m$, הקופסה מחזירה וקטור פיתרון d-מימדי $\hat{\boldsymbol{w}} = LS(\mathbf{X},\boldsymbol{y})$.

א. [4 נק'] הציעו דרך להשתמש בקופסה השחורה כדי לקבל פיתרון עבור בעיית ה-ridge regression (עם $\lambda > 0$ נתון):

$$\hat{\boldsymbol{w}}_\lambda = \text{argmin}_{\boldsymbol{w}\in\mathbb{R}^d}\|\mathbf{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2$$

משמע, עליכם להציע $\mathbf{X}', \boldsymbol{y}'$ כך שיתקיים $\hat{\boldsymbol{w}}_\lambda = LS(\mathbf{X}',\boldsymbol{y}')$.

$X' = \begin{bmatrix} X \\ \sqrt{\lambda} I_{d\times d} \end{bmatrix}$ $y' = \begin{bmatrix} y \\ 0 \end{bmatrix}$

$y = x_1 + 4x_2$

בסעיפים הבאים נתונה מטריצת דוגמאות $\mathbf{X}$ רחבה מאוד ברוחב $d = 10^5$ ובנוסף וקטור תיוגים רציפים $\boldsymbol{y}$.

$x_1, \dots, x_{10^5}$

צוות מחקר מעוניין לבצע רגרסיה ליניארית על הדאטה.

$w^* = (1, 4, 10^5, 10^5, \dots, 10^5, 0, \dots, 0)$ (100)

ב. [8 נק'] עקב אילוצי חומרה, על הצוות לבחור וקטור משקלים $\boldsymbol{w}$ דליל שבו לכל היותר 100 משקלים (איברים) שאינם 0.
הצוות מציע לפתור את הבעיה כרגיל בעזרת הקופסה השחורה $\hat{\boldsymbol{w}} = LS(\mathbf{X},\boldsymbol{y})$. לאחר מכן, הם ישמרו את 100 המשקלים הגדולים ביותר (בערך מוחלט) בווקטור $\hat{\boldsymbol{w}}$ ויאפסו את כל היתר.
תארו מקרה שבו קיימים וקטורים דלילים בעלי שגיאת אימון נמוכה, אך השיטה שמציע הצוות תיכשל ותגיע לשגיאת אימון גבוהה.

$w^* = (1, 4, 0, \dots, 0)$

ג. [5 נק'] האם רגולריזציה יכולה לעזור לפתור את הבעיה שתיארתם בסעיף הקודם? כן, L_1 רגולרי נבנה
אם כן – הסבירו כיצד. L_1 regression
אחרת – הציעו פיתרון אחר.

ד. [8 נק'] בפרוייקט אחר עם אותו דאטה, לצוות אין אילוצי חומרה. הם מצאו את הווקטור האופטימלי $\hat{\boldsymbol{w}}$, וגילו שיש לו שגיאת אימון נמוכה אבל שגיאת מבחן גבוהה. אחת החוקרות הציעה להוריד תחילה את הנתונים ממימד $d = 10^5$ למימד 100 בעזרת PCA ורק לאחר מכן לפתור בעיית LS.
כלומר, תחילה נטיל את הנתונים למימד נמוך על ידי $\tilde{\mathbf{X}} = \underbrace{\mathbf{X}}_{m\times 10^5} \cdot \underbrace{\mathbf{V}^{(100)}}_{10^5\times 100}$.
לאחר מכן, נשתמש בקופסה השחורה ונקבל $\hat{\boldsymbol{w}}^{(100)} = LS(\tilde{\mathbf{X}},\boldsymbol{y})$.

ד.1. כיצד ההצעה של החוקרת יכולה לעזור במצב של overfitting?

ד.2. האם המודל שהוצע ליניארי ביחס לייצוג המקורי $\mathbf{X}$ של הנתונים? הסבירו.

שאלה 4 [15 נק']

בשאלה זו הניחו כי בידיכם מערך נתונים המתאר מטופלים בבי"ח מקומי וברצונכם לממש מודל חיזוי לסוכרת.

א. [7 נק'] הניחו שהעלות העסקית של תוצאה חיובית כוזבת (false positive) יקרה פי 5 מהעלות של תוצאה שלילית כוזבת (false negative).

לכן נדרוש מהמודל את הדרישות הבאות:

a. שיעור רגישות (sensitivity או True positive rate) בגובה של לפחות 80%,

b. שיעור false positive rate בגובה 10% או פחות,

c. ממזער את העלות העסקית.

ה-confusion matrices הבאות מתארות את הביצועים של 4 מודלים שונים על 200 דוגמאות.

איזה מודל תבחרו? נמקו את בחירתכם.

מודל A

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

מודל B

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

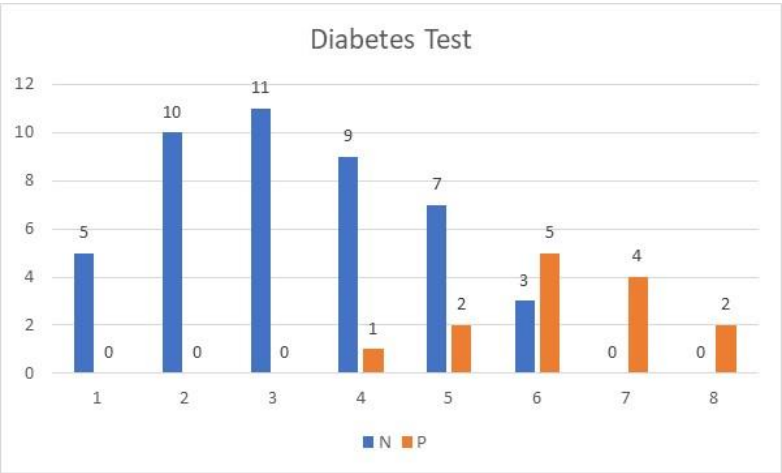
מודל C

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

מודל D

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

ב. [8 נק'] באיור הבא מובאות תוצאות ניסוי לאיתור חולי סוכרת. הציון שהמודל מחזיר הוא מספר שלם בין 1-8. העמודות הכחולות מציינות מטופלים בריאים ואילו העמודות הכתומות מציינות מטופלים עם סוכרת. ציירו את עקומת ה-ROC בהתאם לתוצאות (ציינו שמות לצירים האופקי והאנכי, חשבו וציינו את כל נקודות העקומה).



שאלה 5 [30 נק']

כתבו במחברת את התשובות לכל השאלות הבאות.

א. [6 נק'] איזו מהפעולות הבאות יכולות לעזור לפתור בעיית overfitting? (אין צורך להסביר)
(ניתן לבחור יותר מאפשרות אחת. עבור כל אפשרות, הניחו שמבצעים אותה לחוד, כלומר מבלי לשנות עוד משהו)

a. הוספת עוד דוגמאות לקבוצת האימון (כלומר הגרלה של עוד דוגמאות iid מהתפלגות הנתונים)

b. גריעה של 50 הדוגמאות האחרונות של קבוצת האימון

c. הוספת הנחות מקדימות (prior assumptions) על המודל הנילמד

d. גריעה של הנחות מקדימות (prior assumptions) על המודל הנילמד

e. הוספת מאפיינים (features) לנתונים

f. גריעה של מאפיינים (features) מהנתונים

ב. [3 נק'] הגדירו ERM (Empirical Risk Minimization)

$\underset{h \in H}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}(h(x_i) \neq y_i)$

ג. [2 נק'] תארו בקצרה שתי דרכים להשלמת מאפיינים חסרים (data imputation)

ד. [5 נק'] לכל אחת מהטענות הבאות, כתבו במחברת התשובות האם היא נכונה או לא נכונה (אין צורך להסביר).

a. מקדם פירסון (Pearson correlation) בין שני משתנים מקריים הוא בתחום הממשי $[-1,1]$.

b. נניח כי מגרילים מדגם $(x_1, y_1), (x_2, y_2), \dots (x_m, y_m)$ באופן i.i.d. מהתפלגות $\mathcal{D}$ כלשהי כאשר m זוגי, ואז קובעים את ההיפותזה h להיות פלט של אלגוריתם שמשתמש בנתונים $(x_1, y_1) \dots (x_{m/2}, y_{m/2})$ בלבד. אז השגיאה האמפירית של h על שאר הנתונים $\left(x_{\frac{m}{2}+1}, y_{\frac{m}{2}+1}\right) \dots (x_m, y_m)$ היא משערך בלתי מוטה של שגיאת ההכללה של h.

c. רגרסיה לוגיסטית (logistic regression) הוא אלגוריתם לסיווג בינארי (binary classification) באמצעות מפריד לינארי.

d. פונקציית הלוגריתם של הסיגמואיד (log-sigmoid) היא לא קמורה ולא קעורה.

e. אלגוריתם ה-EM (Expectation Maximization) בכל איטרציה מוריד את פונקציית ה-log-likelihood או משאיר את ערכה ללא שינוי.

ה. [3 נק'] בידינו אוסף היפותזות $h_1, \dots, h_n$, כל אחת היא פלט מאלגוריתם למידה כלשהו על אוסף נתונים זהה. ברצוננו לבחור אחת מבין n היפותזות אלה, ולשם כך נגריל קבוצת מבחן (Test Set) בגודל N. כיצד יש לבחור את N כפונקציה של n? בחרו את האפשרות הנכונה (הסימון $\approx$ כאן משמעו 'סדר גודל'). אין צורך להסביר.

a. פונקציית שורש $N \approx \sqrt{n}$
b. פונקציה לינארית $N \approx n$
c. פונקציה לוגריתמית $N \approx \log n$
d. פונקציה ריבועית $N \approx n^2$
e. פונקציה קבועה $N \approx 1$

ו. [3 נק'] כזכור, אלגוריתם Lloyd מיועד לאישכול מידע (Data Clustering) ביחס לפונקציית המטרה *k-means*. כתבו פונקציית מטרה זו.
הניחו שהנתונים הם $x_1, \dots, x_n \in \mathbb{R}^d$ ו-k הוא פרמטר נתון.

ז. [4 נק'] השלימו את צעדים Step A,B בפסאודו-קוד (pseudocode) של אלגוריתם Lloyd שלפניכם:
(הערה: במימוש להלן אנו מציעים איתחול אקראי לשם פשטות, אבל בעולם האמיתי עושים דברים אחרים.
בכל מקרה, האיתחול לא משפיע על צעדים A,B של האלגוריתם)

function runLloyd($x_1, \dots, x_n \in \mathbb{R}^d, k$)

Initialization:

Set cluster assignments $a[1] \dots a[n]$, each chosen randomly in $\{1..k\}$

(The $i'th$ data point is randomly assigned to cluster $a[i] \in \{1..k\}$ for $i = 1..n$)

Repeat until some stopping condition:

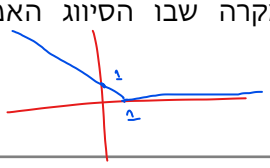
Step A: _____*(complete in solution notebook)*_____
_____*להשלים במחברת הבחינה*_____

Step B: _____*(complete in solution notebook)*_____
_____*להשלים במחברת הבחינה*_____

ח. [2 נק'] כתבו הגדרה, או ציירו שרטוט ברור של פונקציית ה-ReLU (Rectified Linear Unit) המשמשת כפונקציית אקטיבציה פופולרית ברשתות נוירונים.

max{0, x}

ט. [2 נק'] כתבו הגדרה, או ציירו שרטוט ברור של פונקציית ה-hinge-loss המשמשת <u>להגדרה של SVM</u> כפי שנילמד בכיתה (ניתן לענות עבור המקרה של נקודה עם סיווג אמיתי y=1 או עבור המקרה שבו הסיווג האמיתי y=-1, אין צורך לצייר את שני המקרים).



max{0, 1-x}