מבוא למערכות לומדות (236756)

סמסטר חורף תשפ"א – 08 במרץ 2021

מרצה: פרופ' ניר אילון

מבחן מסכם מועד ב'

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 6 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 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']

בסעיפים הבאים נתון עץ החלטה שמטרתו לסווג דוגמאות לאחת מ-10 מחלקות (classes).

העץ נבנה על ידי אלגוריתם לא ידוע ואומן על סט אימון (training set) נתון $\boldsymbol{S}^{train} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{m}$.

א. [5 נק'] נתון שבחלק מהעלים ישנן דוגמאות אימון ממספר מחלקות שונות.

א.1. מודל דטרמיניסטי:

בהינתן דוגמה $\boldsymbol{x}$ שֶׁמְמוּפָּה לעלה בו מספר מחלקות שונות, הציעו דרך לחזות לאיזו מחלקה שייכת הדוגמה.

א.2. מודל הסתברותי:

בהינתן דוגמה $\boldsymbol{x}$ שֶׁמְמוּפָּה לעלה בו מספר מחלקות שונות, הציעו דרך לחזות התפלגות על המחלקות.

משמע, עליכם להחזיר מערך D עם 10 הסתברויות, כך שבתא D_i תופיע הסתברות לכך ש-$\boldsymbol{x}$ שייכת למחלקה y_i.

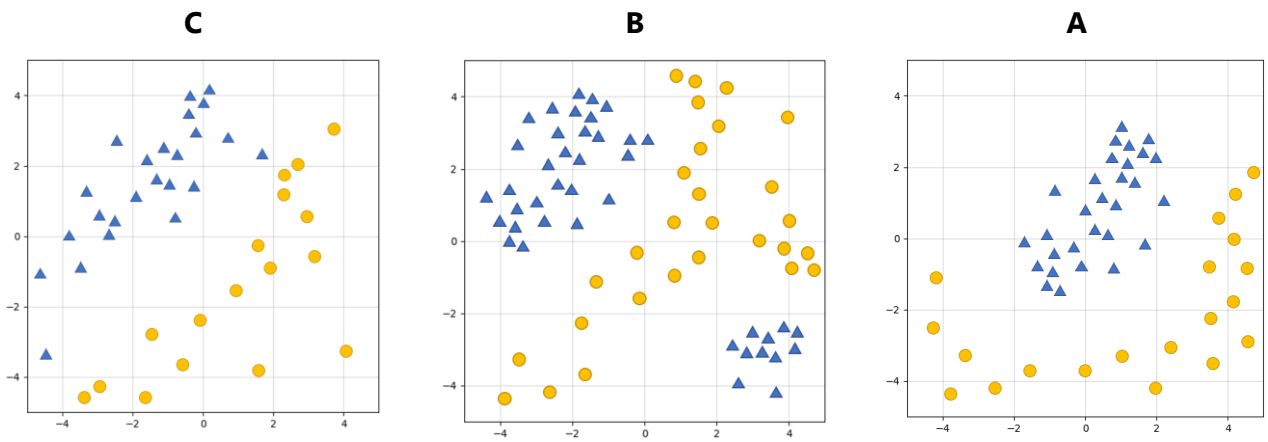
ב. [5 נק'] בבדיקה של העץ הנתון (הדטרמיניסטי) התגלה ששגיאת האימון נמוכה מאוד אך שגיאת ה-validation גבוהה.

הסבירו את התופעה והציעו דרך לשנות את העץ הקיים כדי להתמודד עם המצב.

שימו לב: לא ניתן לאמן עץ חדש אלא רק לערוך את העץ הקיים.

שאלה 2 [10 נק']

לפניכם שלושה datasets דו-מימדיים עם תיוג בינארי (משולש כחול או עיגול צהוב).



לכל dataset נתון, בחרו מהרשימה הבאה את כל המודלים שיכולים להתאים אותו באופן מושלם (שגיאת אימון 0).

i. Adaboost עם decision stump כמסווג בסיס חלש. A, B, C

ii. פרספטרון. C

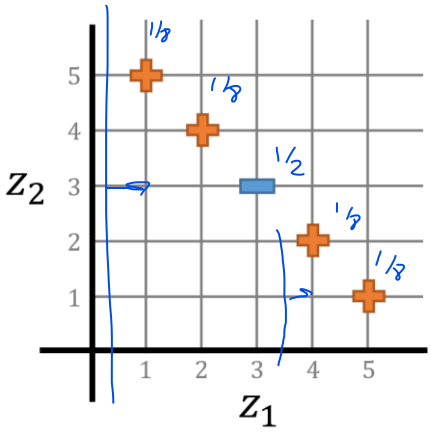
iii. SVM with degree 2 polynomial kernel. A, C, B

iv. kNN כאשר $k = 3$ (הניחו שדוגמה לא נחשבת כשכנה של עצמה). A, B,

שאלה 3 [20 נק']

נתון ה-dataset הבא, המכיל 5 דוגמאות במרחב דו-מימדי המסווגות $+1$ (כתום) או -1 (כחול).

<u>שימו לב</u>: מערכת הצירים מסומנת ב-z_1, z_2 רק כדי להבדיל בין הצירים לסימון של הדוגמאות.



נכתוב את הדוגמאות במפורש (משמאל לימין): $\boldsymbol{x}_1 = [1,5],\ \boldsymbol{x}_2 = [2,4],\ \boldsymbol{x}_3 = [3,3],\ \boldsymbol{x}_4 = [4,2],\ \boldsymbol{x}_5 = [5,1]$

$y_1 = y_2 = y_4 = y_5 = +1,\ \ y_3 = -1$

<u>תזכורת</u>: decision stump הינו עץ החלטה עם צומת שורש ושני צמתי עלים בלבד.

אם התשובה לשאלה של השורש חיובית, נסווג $+1$ (כתום), אחרת -1 (כחול).

למשל, כלל ההחלטה $z_1 \geq 4.5$, מסווג את הדוגמה הימנית ביותר ב-$(+1)$ ואת היתר ב-(-1).

נריץ Adaboost עם decision stump כמסווג בסיס חלש. כפי שהראינו בתרגול, נניח כי האלגוריתם מבצע ERM "חמדני" ובוחר בכל שלב מסווג חלש שמגיע לשגיאה הנמוכה ביותר ביחס לדאטה ולהתפלגות $D_i^{(t)}$ באותה איטרציה t.

א. [5 נק'] הציעו מסווג חלש h_1 שהאלגוריתם עשוי לבחור באיטרציה הראשונה.

$z_1 > 0$ או $z_1 < 6$

$z_2 > 0$ " $z_2 < 6$

ב. [5 נק'] לאילו דוגמאות האלגוריתם <u>יגדיל</u> את ההסתברות בהתפלגות המעודכנת לאיטרציה השנייה? נמקו.

לדוגמאות שטועות לו כיון סבירות ה (-) $(x_3 = (3,3))$

<u>תזכורת</u>: הוכחתם בתרגיל בית שהשגיאה של h_1 ביחס להתפלגות המעודכנת $D^{(2)}$ שווה בדיוק לחצי.

משמע, מתקיים $\sum_{i=1}^{5} D_i^{(2)} \cdot \mathbf{1}_{h_1(\boldsymbol{x}_i) \neq y_i} = \frac{1}{2}$.

ג. [5 נק'] השתמשו בתזכורת כדי לחשב במדויק את ההתפלגות לאיטרציה השנייה.

עליכם לכתוב במפורש את כל חמש ההסתברויות $D_1^{(2)}, \dots, D_5^{(2)}$.

$D_1^{(2)} = D_2^{(2)} = D_4^{(2)} = D_5^{(2)} = \frac{1}{8}$

$D_3^{(2)} = \frac{1}{2}$

ד. [5 נק'] הציעו מסווג חלש h_2 שהאלגוריתם עשוי לבחור באיטרציה השנייה.

$z_1 > 3.5$ או

$z_1 < 2.5$

$z_2 > 3.5$ או

$z_2 < 2.5$ או

שאלה 4 [15 נק']

עבור מודל ליניארי $y_i = \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + \varepsilon_i$, כאשר:

- $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m$ הן דוגמאות נתונות
- $\boldsymbol{w}$ הוא וקטור לא ידוע
- השגיאה מתפלגת i.i.d לפי $\varepsilon_i \sim \mathcal{N}(0,\ 1)$

בתרגיל בית 4 הראיתם שפיתרון הרגרסיה הליניארית עם רגולריזציית ℓ_1 (LASSO) שקול ל-MAP regressor תחת המודל המתואר לעיל ותחת Laplacian prior, משמע $w_j \sim \text{Laplace}(0, b)$ עבור $b > 0$.

כעת נראה שאם מניחים שונות נפרדת לכל משקל, משמע $w_j \sim \text{Laplace}(0, b_j)$ עבור $b_j > 0$, אזי ה-MAP regressor שמתקבל שקול לפיתרון בעיה שנקראת **Adaptive LASSO** ומוגדרת באופן הבא:

$$\widehat{\boldsymbol{w}}_{\boldsymbol{\lambda}}^{\text{AL}} \triangleq \underset{\boldsymbol{w} \in \mathbb{R}^{\text{d}}}{\text{argmin}} \left[\sum_{i=1}^{m} (\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - y_i)^2 + \sum_{j=1}^{d} \lambda_j |w_j| \right]$$

עבור אוסף מקדמים $\lambda_1, \dots, \lambda_d \in \mathbb{R}_{>0}$.

כפי שהסברנו, בסעיפים הבאים נניח ש-$w_j \sim \text{Laplace}(0, b_j)$ באופן בלתי-תלוי.

$$\ln\left(\prod_{i=1}^{d} p(w_i|b_i)\right) = \sum_{i=1}^{d} \ln\left(\frac{1}{2b_i} \cdot e^{-\frac{|w_i|}{b_i}}\right) = \sum_{i=1}^{d} \ln\left(\frac{1}{2b_i}\right) + \sum_{i=1}^{d} -\frac{|w_i|}{b_i}$$

א. [5 נק'] פתחו אנליטית את הביטוי הבא:

$$\ln(p(\boldsymbol{w}|b_1, \dots, b_d))$$

תזכורת: פונקציית ה-pdf של התפלגות Laplace מקיימת $p(w_j|\mu, b_j) = \frac{1}{2b_j} \exp\left\{-\frac{|w_j|}{b_j}\right\}$.

ב. [10 נק'] הראו שה-MAP regressor שמתקבל תחת ההנחות שבשאלה, שקול ל-Adaptive LASSO.

משמע, עליכם להראות שמתקיים:

$$\widehat{\boldsymbol{w}}_{\boldsymbol{\lambda}}^{\text{AL}} = \widehat{\boldsymbol{w}}_{\boldsymbol{b}}^{\text{MAP}} \triangleq \underset{\boldsymbol{w} \in \mathbb{R}^{\text{d}}}{\text{argmax}}\, p(\boldsymbol{w}|\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}, b_1, \dots, b_d)$$

ולהסביר מהי הבחירה המתאימה של $b_1, \dots, b_d$ כתלות ב-$\lambda_1, \dots, \lambda_d$.

תזכורת: ראינו בתרגול שמתקיים $\ln(p(\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}|\boldsymbol{w})) = -\frac{m}{2}\ln 2\pi - \frac{1}{2}\sum_{i=1}^{m}(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - y_i)^2$

שאלה 5 [20 נק']

ישנם שני מפעלים לייצור שבבים. במפעל הראשון ההסתברות לכשל בייצור שבב בודד הינה p_1 ובמפעל השני p_2.

השבבים נשלחים לבדיקת איכות במעבדה חיצונית בארגזים המכילים 10 שבבים כל אחד.

המעבדה בודקת את תקינות השבבים ומסמנת ב-x_i את מספר השבבים התקולים בארגז ה-i.

השבבים מתפלגים $i.i.d$ ולכן ההסתברות לקבלת k שבבים תקולים בארגז שבתוכו 10 שבבים ממפעל $j \in \{1,2\}$ מתפלגת בינומית:

$$\Pr[x_i = k \mid p_j] = \binom{10}{k} p_j^k (1-p_j)^{10-k}$$

בסעיפים הבאים השתמשו בארבע הספרות האחרונות (הימניות) של תעודת הזהות שלכם בתור x_1, x_2, x_3, x_4.

למשל, אם מספר תעודת הזהות שלכם הוא 123456789, תקבלו: $x_1 = 6, x_2 = 7, x_3 = 8, x_4 = 9$.

א. [1 נק'] כתבו במחברת הבחינה את $x_1, \dots x_4$ לפי מספר תעודת הזהות שלכם.

ב. [6 נק'] בסעיף זה <u>בלבד</u> הניחו כי $x_1, \dots x_4$ מקורם ממפעל 1. בעזרת MLE, חַשְׁבוּ (שַׁעֲרְכוּ) את p_1.

בסעיפים הבאים הניחו כי בעקבות תקלה, תוויות הסימון מכל ארגז אבדו.

כלומר, לכל ארגז i, תווית הסימון $z_i \in \{1,2\}$ שמסמנת את מפעל המקור של ארגז זה, לא ידועה (latent).

ג. [6 נק'] בסעיף זה ממשו את צעד ה-E מאלגוריתם EM על התצפית השלישית x_3 (שהגדרתם בסעיף א').

כלומר, עבור $j = 1,2$, עליכם לחשב את $Q_{3,j}^{(1)}$, ההסתברות המשוערכת (באיטרציה ה-1) שהארגז ה-3 הגיע ממפעל j.

לצורך החישוב, הניחו:

- $p_1^{(0)} = \frac{1}{\lambda+2}$, $p_2^{(0)} = \frac{1}{\lambda+3}$ כאשר λ היא הספרה הראשונה (השמאלית) של תעודת הזהות שלכם.
- ההסתברות שארגז אקראי מגיע ממפעל 1 הינה 0.4.

תזכורת נוסחת החישוב:

$$Q_{i,j}^{(1)} = \Pr[z_i = j \mid x_i ; \Theta^{(0)}] = \frac{\Pr[x_i, z_i = j \mid \Theta^{(0)}]}{\sum_{j'} \Pr[x_i, z_i = j' \mid \Theta^{(0)}]}$$

ד. [7 נק'] בשלב זה תממשו את צעד ה-M מאלגוריתם EM על מנת למצוא את $p_1^{(1)}$.

להזכירכם, עליכם למצוא את

$$p_1^{(t+1)} = \operatorname{argmax}_{p_1 \in [0,1]} F(Q, \Theta)$$

כאשר $F(Q,\Theta)$ הוא תוחלת ה-log likelihood, משמע:

$$F(Q,\Theta) = \sum_i \sum_j Q_{ij} (\log(\Pr[z_i = j]) + \log(\Pr[x_i | z_i = j]))$$

בסעיף זה הניחו את מטריצת ההסתברויות הבאה (<u>ולא</u> זו מהסעיף הקודם):

$$Q^{(1)} = \begin{pmatrix} 0.3 & 0.7 \\ 0.2 & 0.8 \\ 0.51 & 0.49 \\ 0.4 & 0.6 \end{pmatrix}$$

שאלה 6 – שאלות קצרות [25 נק']

כתבו במחברת את התשובות לכל השאלות הבאות.

א. [4 נק'] מהו ה-representer theorem? נסחו את המשפט בצורה מדויקת, בהקשר של SVM.

ב. [5 נק'] השלימו את החלק החסר בפסאודוקוד הבא ל-SGD לאימון logistic regression.

להזכירכם, בהינתן נקודות $x_1, \dots, x_m \in \mathbb{R}^d$ ותיוגים $y_1, \dots, y_m \in \{+1, -1\}$, מטרת האימון היא למצוא וקטור משקלים $w \in \mathbb{R}^d$, שמביא <u>למקסימום</u> את פונקציית המטרה הבאה:

$$f(w) = -\sum_{i=1}^{m} \ln(1 + e^{-y_i \langle x_i, w \rangle})$$

<u>הערה</u>: ניתן לכלול רק הוראות פשוטות של גישה לזיכרון ואריתמטיקה וכן לולאות ותנאים (כמו בשפת תכנות סטנדרטית כדוגמת פייתון או סי), ואין לכם "גישה" לפונקציות ספרייה כלשהן, <u>מלבד</u> פונקציית שורש ריבועי, לוגריתם והעלאה בחזקה. כמו כן, ניתן לגשת לפונקציה $uniform(a, b)$ שמגרילה מספרים שלמים בין a ל-b באופן אחיד.

SGD-for-Logistic-Regression:

Input:

training data $x_1, \dots, x_m \in \mathbb{R}^d$, $y_1, \dots, y_m \in \{-1, +1\}$

parameters $T \in \mathbb{N}$ (number of iterations) and μ (learning rate)

Output: Solution vector $w \in \mathbb{R}^d$

$w \leftarrow (0,0, \dots 0)$

for $i = 1, \dots, T$

_____

_____

_____

_____

return w

ג. [4 נק'] פרטו שני שימושים שונים של autoencoders.

ד. [4 נק'] הסבירו מהי שיטת ה-dropout בהקשר של אימון רשתות עמוקות.

הדרך המנע מחיבורים שווים מ כל אפוק

ה. [4 נק'] נתונה מחלקת היפותזות $\mathcal{H}$ על מרחב דוגמאות $\mathcal{X}$ ותת-קבוצה סופית $C \subseteq \mathcal{X}$.

הגדירו במדוייק את האמירה "$\mathcal{H}$ מנפצת (shatters) את C" (הניחו שב-$\mathcal{H}$ יש רק היפותזות בינאריות, כלומר מסווגים ל-± 1).

$\forall y_1, \dots, y_{|C|}$ $\exists h \in H: \forall x_i: h(x_i) = y_i$

ו. [3 נק'] לכל אחת מהטענות הבאות, כתבו במחברת התשובות האם היא נכונה או לא נכונה (אין צורך להסביר).

a. LDA (Latent Discriminative Analysis) היא שיטת סיווג מונחית (supervised) באמצעות מפריד לינארי.

b. הזמן שלוקח לבצע איטרציית backpropagation אחת (על מעבד יחיד, ללא מִקבול) באימון של רשת עמוקה הוא לינארי בגודל הרשת. שימו לב: הכוונה ב"גודל הרשת" היא למספר הנוירונים ועוד מספר הקשתות.

ניתן להניח שחישוב הנגזרת של פונקציית אקטיבציה לוקח זמן קבוע כלומר $O(1)$.

c. כש"מגדלים" עצי החלטה (Decision Trees), כדאי ליצור צמתים עם אנטרופיה (entropy) גבוהה.

b: