מבוא ללמידה ממוכנת – עבודה מסכמת

סמסטר א', תשפ"ד, 31.03.2024

מרצה: ד"ר דן רוזנבאום
מתרגל: שרון רוטגייזר

משך העבודה: חמישה ימים, עד ה-04.04.2024, 23:59.

הנחיות:

1. קראו היטב כל שאלה. ודאו כי אתם מבינים את השאלה לפני שתתחילו לענות עליה.
2. כתבו בכתב יד ברור וקריא / מוקלד. תשובות לא קריאות לא תיבדקנה.
3. יש לכתוב את התשובות בטופס הבחינה, במקום המיועד לכך ולהגיש PDF.
4. יש לפרט את תשובותיכם תמיד, כל עוד מלבנים גדולים עבורם.

משמעת:

אנא מלאו וחתמו על ההצהרה הבאה:

אני, _____ (שם ות"ז)

מצהיר/ה בזאת כי ההגשה היא העבודה האישית שלי. לא התייעצתי עם אנשים אחרים ולא השתמשתי במקורות אונליין. אני מודע/ת לכך שחשד להעתקה מאנשים / מודלי שפה / מקור אונליין עשוי להוביל להגנה על העבודה בשיחה מול הסגל ובהתאם לכך, לפסילת המבחן והעלאה לוועדת משמעת.

_____
חתימה

בהצלחה!

חלק ראשון – שאלות חישוביות (32 נקודות)

שאלה 1, LDA (8 נק').

נתון אוסף של 10 נקודות עם 2 מאפיינים, המשתייכים לשתי מחלקות, כאן.

C1={(4,2),(2,4),(2,3),(3,6),(4,4)}
C2={(9,10),(6,8),(9,5),(8,7),(10,8)}

א. חשבו את ה between scatter matrix, $S_B \in M_2(\mathbb{R})$.

ב. חשבו את ה within scatter matrix, $S_w \in M_2(\mathbb{R})$.

ג. מצאו את הערכים העצמיים והווקטורים העצמיים של $A = S_W^{-1}S_B$.

ד. בתוך הקישור ל desmos (מתחילת השאלה) או על גבי צילום מסך, ציירו את הקו שמתקבל מ - LDA, עליו נטיל את הנקודות והוסיפו אותו להגשה.

שאלה 2, Bayesian Risk (8 נק').

עד כה בקורס, למדנו מסווג דוגמאות בעזרת מסווג MAP, שאומר לבחור את המחלקה C שממקסמת את P(C|x). עם זאת, לפעמים להחלטות שונות מלווה סיכון. נגדיר זאת כך:

$$R(a|x) = \sum_{j=1}^{m} P(c_j|x) \cdot \lambda(a|c_j)$$

במילים: בהינתן x, בשביל לחשב את הסיכון שבלקיחת החלטה a, נעבור על פני כל המחלקות. נחשב את הסיכוי ש c_j זאת המחלקה הנכונה, ואז נכפול זאת בכמות ההפסד $\lambda(a|c_j)$ (כלומר כמה מחיר נשלם אם המחלקה הנכונה היא c_j ולקחנו החלטה a).

א. הראו כי עבור פונקציית המחיר λ המוגדרת:

$$\lambda(a|c_j) = \begin{cases} 1 & a \neq c_j \\ 0 & a = c_j \end{cases}$$

הסיכון $R(a|x) = 1 - P(a|x)$, כלומר, כדי למזער את הסיכון, צריך למקסם את $P(a|x)$, ששקול בדיוק למסווג MAP.

כעת, אנו מגיעים לחנות ורוצים לבחור איזו טלוויזיה לקנות. נתון ש P(bad)=0.2, P(fair=0.5) ו P(good)=0.3. כמו כן, נתונה טבלת ההתפלגות הבאה:

	Good	Fair	Bad
sharp picture	0.9	0.5	0.2
diminished picture	0.1	0.5	0.8

למשל, P(sharp|good)=0.9.

כמו כן, נתונה טבלת ההפסד / סיכון:

$\lambda(\alpha_i \mid w_j)$	Good	fair	Bad
Buy a TV	0	5	20
Don't buy a TV	10	5	0

למשל, אם נחליט לקנות טלוויזיה והיא גרועה, נשלם 20 (הפסד גדול).

ב. עליכם לחשב את הסיכון עבור כל החלטה, בהינתן ההתרשמות שלכם, x (האם האיכות ברורה או לא).

שאלה 3, MLE (8 נק').

ניזכר בהגדרת רגרסיה לינארית – אנחנו מניחים כי הקו אותו אנו מנסים ללמוד הוא $y = w^\top x + \varepsilon$, כאשר $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

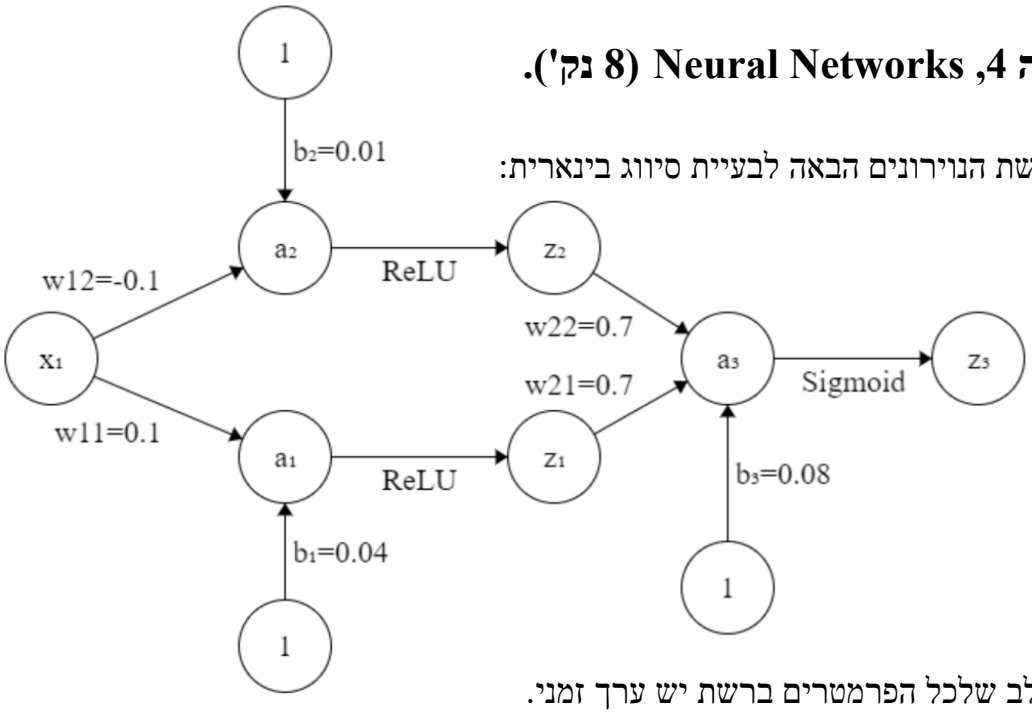
א. הראו כי אומד ה-MLE של w, זה שממקסם את ההסתברות לדגימות $\{(x_i, y_i)\}_{i=1}^n$, הוא זה שממזער את סכום ריבועי השגיאות:

$$\hat{w} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - w^\top x_i)^2$$

ב. כעת, נניח שלכל דגימה יש רק מאפיין אחד. בסעיף א', ההנחה שלכם בפתרון הייתה ש $y_i \sim \mathcal{N}(wx_i, \sigma^2)$. כעת, נניח שאנחנו מנסים לפתור בעיה שונה ובה $y_i \sim \mathcal{N}(\log(wx_i), 1)$. מצאו את אומד הנראות המקסימלי עבור w כעת.

שאלה 4, Neural Networks (8 נק').

תהי רשת הנוירונים הבאה לבעיית סיווג בינארית:



שימו לב שלכל הפרמטרים ברשת יש ערך זמני.

א. עבור $x_1=0.3$, חשבו את z_3 ופרטו כאן את החישובים.

ב. מהו הסיווג $\hat{y}$ שהרשת תתן עבור $x_1=0.3$?

ג. רשמו את תהליך ה backpropagation עבור הפרמטר b_3. שימו לב שההפסד ניתן על ידי $L(y, z_3) = (z_3 - y)^2$. בתשובתכם אל תציבו / תחשבו אף אחד מהמשקלים או הביטויים, אלא כתבו ביטוי כללי המשתמש בביטויים מהצורה $\frac{\partial \alpha}{\partial \beta}$ ומפשט אותם.

$$\frac{\partial L}{\partial b_3} =$$

ד. חזרו על סעיף ג' אבל הפעם עבור הפרמטר w_{11}.

חלק שני – שאלות פתוחות (68 נקודות)

שאלה 1, מסווגים לינאריים (27 נק').

נתונה קבוצת אימון $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$, כאשר $x_i \in \mathbb{R}^d$ ו $y_i \in \{0,1\}$.
אנחנו מניחים מודל של logistic regression ורוצים ללמוד w מתאים ל-D הנ"ל.

א. האם זה אפשרי לקבל פתרון עם נוסחה סגורה עבור w הממקסם את ה -log likelihood? אם כן, חשבו אותו ואחרת, הסבירו מדוע, ואיך כן נוכל לחשב את w?

ב. ניזכר כי עבור דגימה חדשה x, אנחנו ניתן לה y=1 אם $p(y = 1|x; w) \geq 0.5$.
הראו כי גבול ההחלטה הזה הוא בעצם מסווג לינארי. תזכורת: המודל נתון על ידי:

$$p(y = 1|x; w) = \frac{1}{1 + \exp(-w^\top x)}$$

ג. כעת, נניח שהדגימות הם בעלות מאפיינים בינאריים, כלומר $x_i \in \{0,1\}^d$. כמו כן, ידוע ש $x_i^1 = 1$ רק עבור התוויות $y_i = 1$. מה הייתם מצפים שיקרה ל w_1, המשקל הנלמד עבור המאפיין הראשון?

ד. כעת, נתון כי $D = \{(x_1, 0), (x_2, 0), (x_3, 1), (x_4, 1)\}$. מהי פונקציית הנראות של w עבור D הנ"ל?

כעת, אנו חוזרים שוב למקרה הכללי. נניח שכל דגימה היא $x_i \in \mathbb{R}^2$ ואנו ממפים את הדגימות למרחב קלט חדש, $x \to [1, x_1^2, x_2^2]$. כאשר $x = (x_1, x_2)$.

ה. כתבו ביטוי לגבול ההפרדה במרחב החדש ובעזרת פרמטרים מתאימים, $\theta = [b, w_1, w_2]^\top$. התשובה צריכה להיות באמצעות x_1, x_2, b, w_1, w_2.

ו. מהי הצורה הגיאומטרית של גבול ההפרדה, המוגדר על ידי המשוואה שלמעלה?

ז. אם אנחנו מוסיפים רגולריזציה L2 למשקלים w_1, w_2, מה קורה להם, ככל שאנחנו מגדילים את λ?
תזכורת: λ זהו ההיפר פרמטר הנותן משקל לחלק של הרגולריזציה

ח. אם אנחנו מוסיפים רגולריזציה L2 למשקלים w_1, w_2, מה קורה לצורה של גבול ההחלטה, ככל שאנחנו מגדילים את λ (ששולט בחלק של הרגולריזציה)?

לסיום, נדבר על המקרה בו אנו רוצים ללמוד C מחלקות. למודל כזה, נקרא SoftMax Regression, בגלל שמודל הסיווג הוא מהצורה:

$$p(y = c|x) = \frac{\exp(w_c^\top x)}{\sum_{d \in C} \exp(w_d^\top x)}$$

ט. נניח כי $|C| = 2$, כתבו מחדש את מודל הסיווג עבור $p(y = 1|x)$ והראו שזה שקול למודל של logistic regression.

שאלה 2, עצים (21 נק').

בשאלה זו נתעסק בעצי החלטה. נניח שהמידע איתו אנחנו מתעסקים הוא עם n דגימות אימון, כל דגימה היא ממשית עם d מאפיינים ($x_i \in \mathbb{R}^d$) ועם תווית 0/1. הפיצולים נעשים כך שימקסמו את ה information gain, ואנו בוחרים אותם בתור קו הפרדה לינארי המקביל לאחד הצירים.

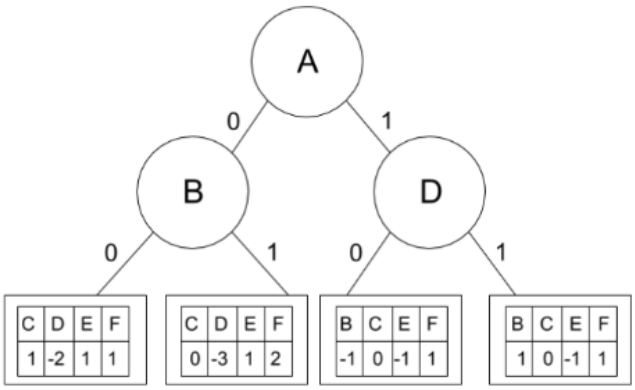
א. הוכיחו או תנו דוגמה נגדית: קיימת איזשהי התפלגות על n ערכים, כך שהאנטרופיה שלה קטנה מ 1-.

ב. הוכיחו או תנו דוגמה נגדית: בכל נתיב מהשורש לעלה, אותו המאפיין לא ייבחר פעמיים.

ג. הוכיחו או תנו דוגמה נגדית: ה information gain בשורש הוא לפחות כמו ה information gain של כל צומת אחרת בעץ.

אדם לא היה מרוצה מהדרך שבה עצים עובדים והחליט להמציא טכניקה חדשה, בשם "עץ פרספטרון". האלגוריתם של יצירת עץ נשאר אותו דבר, אבל מוגדר לעץ גובה מקסימלי. ברגע שהגענו לגובה מקסימלי עם הדגימות המתאימות, במקום שהעלה ייתן את "סיווג הרוב", מאומן מודל פרספטרון על אותן הדגימות. בדוגמה הבאה, יש לכל דגימה 6 מאפיינים – A, B, C, D, E, F. במסלול השמאלי ביותר, פיצלנו לפי A=0 ואז לפי B=0, ככה נשארנו עם ארבעה מאפיינים. לכל מאפיין יש את המשקל שלו, שנלמד עבורו במסווג הנוכחי. למשל, המשקל של C הוא 1.

בתשובותיכם, הניחו ש bias=0 בכל מודל perceptron.



ד. מה יהיה הסיווג של עץ הפרספטרון עבור הדגימה $x = [1,1,0,1,0,1]$.

ה. נכון או לא נכון: גבול ההחלטה של עץ פרספטרון יהיה לינארי.

ו. נכון או לא נכון: עבור עומק מקסימלי קטן, עצי החלטה רגילים נוטים לעשות יותר underfitting בהשוואה לעצי פרספטרון.

ז. נתונה קבוצת אימון $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$, כאשר $x_i \in \{0,1\}^3$, כלומר כל דגימה X היא נקודה 3 מימדית עם הערכים הבוליאניים X_1, X_2, X_3 בהתאמה. כמו כן, נתונה פונקציית החיזוי האמיתי שלהן $Y = (X_1 \wedge X_2) \vee (\neg X_1 \wedge \neg X_2)$.

אנחנו מנסים ללמוד מסווג $h: X \to Y$ באמצעות עץ החלטה רגיל מעומק 1.

מה אפשר לומר על ה sample complexity של הבעיה הזאת? הקפידו להשתמש בנוסחה המתאימה ביותר למקרה המתואר והסבירו את תשובתכם. בשאלה זו אין צורך בחישוב.

שאלה 3, תיאוריית PAC (20 נק').

נרצה לדבר על בעיות סיווג בינאריות.

א. נכון או לא נכון: עבור משפחת מסווגים $\mathcal{H}$, ובהנחה שזו מכילה את המסווג המושלם, ניתן לקחת קבוצה S מהתפלגות D עם כמות דגימות מספיקה, כך שהשגיאה האמיתית מעל ההתפלגות D תהיה בטווח של $\pm\varepsilon$ לכל $h \in \mathcal{H}$ בהסתברות לבחירתנו. הסבירו את תשובתכם.

ב. נכון או לא נכון: יהיו $\mathcal{H}_1, \mathcal{H}_2$ שתי מחלקות היפותזות סופיות כך ש $\mathcal{H}_1 \subset \mathcal{H}_2$. יהי h_1 המסווג המשיג את השגיאה המינימלית על קבוצת אימון S עם $\mathcal{H}_1$ ויהי h_2 המסווג המשיג את השגיאה המינימלית על אותה קבוצת האימון S עם $\mathcal{H}_2$. בגלל שמתקיים $|\mathcal{H}_2| \geq |\mathcal{H}_1|$, $L_D(h_2) \geq L_D(h_1)$. הסבירו את תשובתכם.

ג. תהי משפחת ההיפותזות $\mathcal{H}_k$, המכילה את אוסף כל k האינטרוולים <u>הלא חופפים</u> ב $\mathbb{R}$.

$$\mathcal{H}_k = \left\{ \bigcup_{i=1}^{k} [a_i, b_i] \mid a_i, b_i \in \mathbb{R}, a_1 < b_1 < a_2 < \cdots < b_k \right\}$$

חשבו את $VCdim(\mathcal{H}_k)$ ותנו הוכחה מלאה.

ד. לכל אחד מ-3 הסעיפים הבאים, ענו האם סיבוכיות הדגימות תגדל, לא תשתנה או תקטן:

i. ננסה ללמוד את הבעיה עם $\mathcal{H}^*$ במקום $\mathcal{H}$, כך ש $|\mathcal{H}^*| = |\mathcal{H}|$.

ii. הקטנת הפרמטר δ של PAC.

iii. הקטנת הפרמטר ε של PAC.