<u>הטכניון-מכון טכנולוגי לישראל</u>
הפקולטה להנדסת חשמל

אביב תשע"ט
06.10.2019

<u>מערכות לומדות - 046195</u>
<u>בחינה סופית - מועד ב'</u>

משך המבחן: <u>3 שעות</u>

1. <u>אין להשתמש בכל חומר עזר</u>, פרט לדף נוסחאות שיחולק עם הבחינה, ומחשבון.
2. יש לענות על כל השאלות.
3. משקל כל שאלה מצוין בטופס הבחינה. סך הנקודות הוא 100.
4. נא לכתוב בצורה ברורה ומסודרת.
5. יש לכלול פירוט מלא של דרך הפתרון. <u>תשובות לא מנומקות לא יזכו בניקוד</u>.

ב ה צ ל ח ה !!!

שאלה מס' 1 (36%)

שני חלקי השאלה בלתי תלויים זה בזה

חלק א'

נתון: $D=\{x_i,y_i\}_{i=1}^N$, כאשר $x_i\in\mathbb{R}^d$ ו- $y_i\in\mathbb{R}$.

נתון המודל הסטטיסטי הבא: $y_i=w^Tx_i+\varepsilon_i$, כאשר $w\in\mathbb{R}^d$ ו- $\{\varepsilon_i\}_{i=1}^N$ הינם משתנים i.i.d. מפולגים נורמלית, $\varepsilon_i\sim\mathcal{N}(0,\sigma^2)$.

ידוע מראש כי הפרמטרים לבעיה מתפלגים נורמלית, $w\sim\mathcal{N}(0,\beta^2I_d)$.

שימו לב: β ו- σ הינם פרמטרים ידועים ואין צורך לשערך אותם.

7% א. הראו כי שערוך MAP של הפרמטרים w , שקול לפתרון בעיית הרגרסיה הבאה:

$$(1)\ w_{\text{linear}}=\arg\min_w \|y-Xw\|^2+\lambda\|w\|^2$$

רשמו במפורש את $X\in\mathbb{R}^{n\times d}$, $y\in\mathbb{R}^n$ ואת הפרמטר λ כפונקציה של נתוני הבעיה.

6% ב. רשמו את הפתרון לבעיית הרגרסיה (1) .

4% ג. הסבירו את המשמעות של הפרמטר λ כתלות בפרמטרים של הבעיה. איזו בעיה הוא נועד לפתור? הסבירו זאת דרך נקודת המבט של בעיית ה- MAP ודרך נקודת המבט של בעיית הרגרסיה.

חלק ב'

הסעיפים בחלק זה בלתי תלויים בסעיפים הקודמים.

נתון: $D=\{(x_i,y_i)\}_{i=1}^N$ כאשר $x\in\mathbb{R}^d$ ו- $y\in\{1,...,K\}$.

בחלק זה של השאלה, נדון ברגרסיה של הסתברויות.

במילים אחרות, נרצה למצוא **פילוג** דיסקרטי $p(y\,|\,x)$ רציף ב- x .

לשם כך, נניח מודל פרמטרי, $p_{\text{model}}(y\,|\,x,w)$.

בנוסף, נגדיר פונקציית מרחק בין הפילוג האמפירי של ה- Data $\hat{p}_{\text{data}}$ ובין ההתפלגות p_{model} :

$$d_{KL}(\hat{p}_{\text{data}}\,\|\,p_{\text{model}})=\mathbb{E}_{y|x\sim\hat{p}_{\text{data}}}\log\frac{\hat{p}_{\text{data}}(y\,|\,x)}{p_{\text{model}}(y\,|\,x,w)}$$

8% ד. הראו שמזעור ה- KL-divergence, d_{KL} , שקול לשערוך MLE. כלומר, הראו בצורה מפורשת שהפתרונות של שתי הבעיות שקולים.

<u>המשך שאלה 1</u>:

הוצע להחסיר את האנטרופיה של p_{model} כאיבר רגולריזציה, באופן הבא:

$$w_{\text{logistic}} = \arg\min_w \sum_{i=1}^{N} \left(-\log p_{\text{model}}\left(y_i \mid x_i, w\right) + \lambda \sum_{k=1}^{K} p_{\text{model}}\left(y=k \mid x_i, w\right) \log p_{\text{model}}\left(y=k \mid x_i, w\right) \right) \quad (2)$$

עבור $\lambda > 0$.

4% ה. הסבירו במילים את המשמעות של האיבר שנוסף. מה מעודדת הרגולריזציה במקרה זה?

7% ו. כעת נניח מודל פרמטרי לוגיסטי , $p_{\text{model}}\left(y=k \mid x, w\right) = \frac{\exp\left(w_k^T x\right)}{\sum_{j=1}^{K} \exp\left(w_j^T x\right)}$.

רשמו אלגוריתם גרדיאנט בעל <u>עדכון סדרתי</u> לפתרון בעיית האופטימיזציה הנתונה ב- (2).

<u>הנחיה:</u> ראשית, רשמו את כלל העדכון כתלות ב- $\nabla_w p_{\text{model}}\left(y \mid x, w\right)$.

לאחר מכן, רשמו במפורש את למה שווה הביטוי $\nabla_w p_{\text{model}}\left(y \mid x, w\right)$

<u>פתרון:</u>

א)

$$\begin{aligned}
\arg\max_w p\left(w \mid D\right) &= \arg\max_w p\left(D \mid w\right) p\left(w\right) \\
&= \arg\max_w \log p\left(D \mid w\right) p\left(w\right) \\
&= \arg\max_w \sum_{i=1}^{N} \log p\left(x_i, y_i \mid w\right) + \log p\left(w\right) \\
&= \arg\max_w -\frac{N}{2}\log 2\pi\sigma - \sum_{i=1}^{N} \frac{1}{2\sigma^2}\left(y_i - w^T x_i\right)^2 - \sum_{i=1}^{d}\left(\frac{1}{2}\log 2\pi\beta - \frac{1}{2\beta^2} w_i^2\right) \\
&= \text{argmin}_w \sum_{i=1}^{N} \frac{1}{2\sigma^2}\left(y_i - w^T x_i\right)^2 + \frac{1}{2\beta^2} w^T w \\
&= \text{argmin}_w \sum_{i=1}^{N} \left(y_i - w^T x_i\right)^2 + \frac{\sigma^2}{\beta^2} w^T w \\
&= \text{argmin}_w \left\|y - Xw\right\|^2 + \frac{\sigma^2}{\beta^2}\left\|w\right\|^2
\end{aligned}$$

כאשר:

$$\lambda = \frac{\sigma^2}{\beta^2}$$

$$y = \left(y_1, \ldots, y_N\right)^T$$

$$X = \left(x_1, \ldots, x_N\right)^T$$

ב)
נגזור ונשווה ל-0:

$$
\begin{aligned}
&\frac{d}{dw}\left(\|y-Xw\|^2+\lambda\|w\|^2\right)=0\\
&\Leftrightarrow X^T\left(y-Xw\right)+\lambda w=0\\
&\Leftrightarrow \left(X^TX+\lambda I\right)w=X^Ty\\
&\Leftrightarrow w=\left(X^TX+\lambda I\right)^{-1}X^Ty
\end{aligned}
$$

ג)

λ הינו מקדם אשר קובע כמה חשיבות נותנים לאיבר הרגולריזציה $\|w\|^2$. $\lambda=\frac{\sigma^2}{\beta^2}$ מכמת את מידת הוודאות שלנו ב- Prior אל מול הוודאות ה- Data. ככל שהיחס $\frac{\sigma}{\beta}$ קטן יותר, אנחנו מתקרבים לבעיה המקורית של שערוך MLE ומציאת Fit ל- Data. ככל שהיחס $\frac{\sigma}{\beta}$ גדל, נחפש פתרון בעל משקולות קטנים יותר, על אף שנוסיף Bias לפתרון בעיית הרגרסיה $\arg\min_w\|y-Xw\|^2$. חיפוש פתרון בעל משקולות קטנים יותר מקטין באופן אפקטיבי את מחלקת ההשערות שבה אנו מחפשים פתרון לבעיית הרגרסיה ובכך מקטין את ה Variance של הבעיה. בנוסף, הוא מאפשר לעיתים יציבות נומרית גדולה יותר.

ד)
נרשום במפורש את תוצאת בעיית מזעור ה KL-divergence ונראה שהוא זהה לפתרון בעיית ה MLE:

$$
\begin{aligned}
\arg\min_w d_{KL}\left(\hat{p}_{\text{data}}\,\|\,p_{\text{model}}\right)&=\arg\min_w \mathbb{E}_{y|x\sim\hat{p}_{\text{data}}}\log\frac{\hat{p}_{\text{data}}\left(y\,|\,x\right)}{p_{\text{model}}\left(y\,|\,x,w\right)}\\
&=\arg\min_w \mathbb{E}_{y|x\sim\hat{p}_{\text{data}}}\left[\log\hat{p}_{\text{data}}\left(y\,|\,x\right)-\log p_{\text{model}}\left(y\,|\,x,w\right)\right]\\
&=\arg\min_w -\mathbb{E}_{y|x\sim\hat{p}_{\text{data}}}\log p_{\text{model}}\left(y\,|\,x,w\right)\\
&=\arg\max_w \mathbb{E}_{y|x\sim\hat{p}_{\text{data}}}\log p_{\text{model}}\left(y\,|\,x,w\right)\\
&=\arg\max_w \frac{1}{N}\sum_{i=1}^{N}\log p_{\text{model}}\left(y_i\,|\,x_i,w\right)\\
&=\arg\max_w p_{\text{model}}\left(y\,|\,x,w\right)\\
&=w_{MLE}
\end{aligned}
$$

ה)
ניזכר כי פונקציית האנטרופיה הינה:

$$
h\left(p_{\text{model}}\,|\,x_i,w\right)=-\sum_{i=1}^{K}p_{\text{model}}\left(y=k\,|\,x_i,w\right)\log p_{\text{model}}\left(y=k\,|\,x_i,w\right)
$$

כלומר במקרה שלנו אנחנו מנסים למזער את מינוס האנטרופיה, או במילים אחרות, למקסם את האנטרופיה. כזכור מהתרגול על עצים, אנטרופיה היא מדד לחוסר האחידות של משתנה אקראי. לכן מקסום האנטרופיה שקול להעדפת פתרונות שבהם ההתפלגות $p_{\text{model}}\left(y=k\,|\,x_i,w\right)$ רחוקה מהתפלגות דטרמיניסטית.

ו)

נסמן את פונקציית האקטיביציה הלוגיסטית:

$$p_{\text{model}}\left(y=i \mid x, w\right)=\frac{e^{w_i^T x}}{\sum_{k=1}^{K} e^{w_k^T x}}$$

$$\Rightarrow \nabla_{w_j} p_{\text{model}}\left(y=i \mid x, w\right)=\left\{\begin{array}{ll} \dfrac{e^{w_i^T x}\sum_{k=1}^{K} e^{w_k^T x}x-e^{w_i^T x}e^{w_j^T x}x}{\left(\sum_{k=1}^{K} e^{w_k^T x}\right)^2} & i=j \\ \dfrac{-e^{w_i^T x}e^{w_j^T x}x}{\left(\sum_{k=1}^{K} e^{w_k^T x}\right)^2} & i\neq j \end{array}\right\}=\left\{\begin{array}{ll} \dfrac{e^{w_i^T x}\sum_{k=1}^{K} e^{w_k^T x}-e^{w_i^T x}e^{w_j^T x}}{\left(\sum_{k=1}^{K} e^{w_k^T x}\right)^2}x & i=j \\ \dfrac{-e^{w_i^T x}e^{w_j^T x}}{\left(\sum_{k=1}^{K} e^{w_k^T x}\right)^2}x & i\neq j \end{array}\right\}$$

כאשר,

$$\nabla_{w} p_{\text{model}}\left(y=i \mid x, w\right)=\left(\nabla_{w_1} p_{\text{model}}\left(y=i \mid x, w\right),\ldots,\nabla_{w_K} p_{\text{model}}\left(y=i \mid x, w\right)\right)$$

מכאן, נחשב את הגרדיאנט:

$$\begin{aligned} \Delta w &= \nabla_{w}\left(-\log p_{\text{model}}\left(y_t \mid x_t, w\right)+\lambda\sum_{k=1}^{K} p_{\text{model}}\left(y=k \mid x_t, w\right)\log p_{\text{model}}\left(y=k \mid x_t, w\right)\right)= \\ &=-\frac{1}{p_{\text{model}}\left(y_t \mid x_t, w_t\right)}\nabla_{w} p_{\text{model}}\left(y_t \mid x_t, w\right)+\lambda\sum_{k=1}^{K}\left(1+p_{\text{model}}\left(y=k \mid x_t, w\right)\right)\nabla_{w} p_{\text{model}}\left(y=k \mid x_t, w\right) \\ &=-\frac{1}{g_{y_t}\left(x \mid w\right)}\nabla_{w} g_{y_t}\left(x \mid w\right)+\lambda\sum_{k=1}^{K}\left(1+g_k\left(x \mid w\right)\right)\nabla_{w} g_k\left(x \mid w\right) \end{aligned}$$

ולבסוף נקבל את כלל העדכון הבא:

$$w_{t+1}=w_t-\eta\Delta w_t$$

שאלה מס' 2 (38%)

בשאלה זו נתמקד ברשת עצבית היזון קדמי (feedforward) עם מטריצות משקולות $\{W_\ell\}_{\ell=1}^{L}$, כאשר $W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$, ללא איברי היסט (Bias).

מוצא כל שכבה עובר דרך פונקציית אקטיבציה לינארית: $\varphi(u) = u$. המופעלות בבעיה עם קלט $X \in \mathbb{R}^{d_0}$ ופלט $O \in \mathbb{R}^{d_L}$, עבור $d_L < d_0$. כלומר, מוצא הרשת הוא: $O = W_L W_{L-1} \cdots W_2 W_1 X$.
לדוגמא: נסתכל על הנוירון ה- j בשכבה הראשונה אז מוצאו הינו $\varphi_j(\sum_{i=1}^{d_0} w_{j,i} x_i) = \sum_{i=1}^{d_0} w_{j,i} x_i$
ועבור כל המוצאים של השכבה הראשונה נקבל $O_1 = W_1 X$ כאשר:

$$W_1 = \begin{pmatrix} w_{1,1} & \cdots & w_{1,d_0} \\ \vdots & \vdots & \vdots \\ w_{d_1,1} & \cdots & w_{d_1,d_0} \end{pmatrix}, X = \begin{pmatrix} x_1 \\ \vdots \\ x_{d_0} \end{pmatrix}$$

חלק א'

5% א. כמה שאלות לחימום:

I. כמה שכבות נסתרות יש?
II. כמה נוירונים יש בשכבה ה- ℓ ?
III. כמה פרמטרים נלמדים יש ברשת?
IV. לאחר שהרשת נלמדה, הראו כי מספיק לשמור $d_0 d_L$ פרמטרים ברשת.

7% ב. נניח שניתן לבחור את L ו- $\{W_\ell\}_{\ell=1}^{L}$ כרצוננו. האם ניתן להשתמש במשפט הקירוב האוניברסלי? אם כן, נמקו. אם לא, הגדירו מה היא משפחת הפונקציות אותה ניתן לממש באמצעות רשת זו.

10% ג. חשבו את $\frac{\partial O}{\partial W_\ell}$, לפי אלגוריתם Backpropagation בהינתן קלט X ומטריצות $\{W_\ell\}_{\ell=1}^{L}$.

חלק ב':

מכאן ואילך נניח כי הפלט הוא סקלר $d_L = 1$, וכי הרשת מאותחלת ע"י דגימה אקראית של כל משקל בצורה בלתי תלויה מפילוג תלוי שכבה עם שונות σ_ℓ^2 וממוצע אפס.

8% ד. נניח שרכיבי הקלט X נדגמו מפילוג עם ממוצע אפס ושונות 1. מצאו אתחול אקראי למשקולות עבורו גם הכניסה לכל נוירון ברשת היא עם ממוצע אפס ושונות 1.

8% ה. נסמן את יציאת שכבת הנוירונים ה- ℓ ב- $V_\ell = W_\ell W_{\ell-1} \cdots W_1 X$, ואת הגרדיאנט $G_\ell = \frac{\partial O}{\partial V_\ell}$. מצאו אתחול אקראי למשקולות עבורו רכיבי הגרדיאנט G_ℓ : $\forall \ell < L$ כולם עם ממוצע אפס ושונות 1.

פתרון:

א)

ניתן לראות מיידית:

ברשת יש L-1 שכבות נסתרות.

בשכבה הנסתרת ה-ℓ יש d_ℓ נוירונים.

בסה"כ יש $\sum_{\ell=1}^{L} d_\ell d_{\ell-1}$ פרמטרים, מכיוון שזו סכום גדלי מטריצות המשקולות.

אם נסמן $W = W_L W_{L-1} \cdots W_2 W_1 \in \mathbb{R}^{d_0 \times d_L}$, נוכל לראות שהחזאי שנלמד שקול לחזאי הלינארי $O = WX$,

ולכן נוכל לשמור בסה"כ $d_0 d_L$ פרמטרים – מספר הערכים ב-W.

ב)

מכיוון שהחזאי הנלמד הוא לינארי, ניתן לממש רק פונקציות לינאריות מהצורה: $f(X) = WX$. בנוסף ברור שגם ניתן לממש כל פונקציה לינארית מצורה זו: פשוט נקבע $W_1 = W$ ולכל שאר המטריצות $W_\ell = I$.

מכיוון שכך לא ניתן לקרב כל פונקציה רציפה עם רשת מסוג זה. מכאן, משפט הקירוב האוניברסלי לא תקף במקרה זה. הסיבה שהמשפט לא תקף היא שבמשפט יש תנאי שפונקציית האקטיבציה היא לא פולינום, בעוד שמקרה זה היא כן פולינום: $\varphi(u) = u$.

ג)

נשים לב כי המוצא $O \in \mathbb{R}^{d_L}$, כלומר ישנן d_L יציאות לרשת הנוירונים.

לכן, הנגזרת ביחס למטריצת המשקולות ה- l הינה $\frac{\partial O}{\partial W_\ell} = \left(\frac{\partial O_1}{\partial W_\ell}, \ldots, \frac{\partial O_{d_L}}{\partial W_\ell}\right)$.

מעתה, נתייחס לגרדיאנט של היציאה O_i (היציאה ה- i) בלבד.

נסמן את יציאת (וכניסת) שכבת הנוירונים ה-ℓ ב- $V_\ell = W_\ell W_{\ell-1} \cdots W_1 X$, ואת הגרדיאנט $G_{i,\ell} = \frac{\partial O_i}{\partial V_\ell}$.

המוצא של השכבה האחרונה באינדקס ה- i הינו למעשה היציאה ה- i, $O_i = v_{i,L}$, ולכן,

$$g_{i,L} = \frac{\partial O_i}{\partial v_{i,L}} = \frac{\partial O_i}{\partial O_i} = 1,$$

כאשר סימנו $g_{i,L}$ באותיות קטנות כדי לציין גודל סקלרי.

לכל שכבה אחרת, נחשב בצורה רקורסיבית:

$$G_{i,\ell-1} = \frac{\partial O_i}{\partial V_{\ell-1}} = \frac{\partial V_\ell}{\partial V_{\ell-1}} \frac{\partial O_i}{\partial V_\ell} = W_\ell^T G_{i,\ell}$$

כך שיתקיים:

$$G_{i,\ell} = W_{\ell+1}^T \cdots W_{L-1}^T W_{i,L}^T g_{i,L} = W_{\ell+1}^T \cdots W_{L-1}^T W_{i,L}^T.$$

לסיום, הנגזרת לפי המשקל היא:

$$\frac{\partial O_i}{\partial W_\ell} = \frac{\partial O_i}{\partial V_\ell} \frac{\partial V_\ell}{\partial W_\ell} = G_{i,\ell} V_{\ell-1}^T$$

כך שנקבל:

$$\frac{\partial O_i}{\partial W_\ell} = \frac{\partial O_i}{\partial V_\ell} \frac{\partial V_\ell}{\partial W_\ell} = W_{\ell+1}^T \cdots W_{L-1}^T W_{i,L}^T X^T W_1^T \cdot\cdot W_{\ell-2}^T W_{\ell-1}^T$$

כדי לוודא את נכונות החישוב, נשים לב למימדיות:

$$W_{\ell+1}^T \cdots W_{L-1}^T W_{i,L}^T = \left(W_{i,L} W_{L-1} \cdots W_{\ell+1}\right)^T \in \mathbb{R}^{d_l \times 1}$$

ובאותו אופן,

$$X^T W_1^T \cdot\cdot W_{\ell-2}^T W_{\ell-1}^T = \left(W_{\ell-1} \cdots W_1 X\right)^T \in \mathbb{R}^{1 \times d_{\ell-1}}$$

ולכן,

$$\frac{\partial O_i}{\partial W_\ell} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$$

לבסוף, ניזכר שעשינו זאת עבור כל יציאה בנפרד,
לכן, הנגזרת הכוללת הינה:

$$\frac{\partial O}{\partial W_\ell} = \left(\frac{\partial O_1}{\partial W_\ell}, \ldots, \frac{\partial O_{d_L}}{\partial W_\ell}\right) \in \mathbb{R}^{d_L \times d_\ell \times d_{\ell-1}}$$

כצפוי לנגזרת של וקטור באורך d_L במטריצה בגודל $d_\ell \times d_{\ell-1}$.

ד)
נדרוש שסכום הכניסות לכל ניורון יהיה בעל ממוצע אפס ושונות 1. נתחיל בכניסה לשכבה הנסתרת הראשונה.
הדרישה לממוצע אפס מתקיימת בכל מקרה כי $E\left[X\right] = 0$ ו- $E\left[W_1\right] = 0$

$$0 = E\left[W_1 X\right] = E\left[W_1\right] E\left[X\right]$$

מהדרישה לשונות 1 (לשם הפשטות נסמן ב- w_{ij} במקום ב-$w_{ij,1}$ את רכיבי המטריצה W_1):

$$\forall i \in 1, \ldots, d_1: \;\; 1 = \text{Var}\left(\left(W_1 X\right)_i\right) = \text{Var}\left(\sum_{j=1}^{d_0} w_{ij} x_j\right) = \text{Cov}\left(\sum_{j=1}^{d_0} w_{ij} x_j, \sum_{r=1}^{d_0} w_{ir} x_r\right) = \sum_{r=1}^{d_0} \sum_{j=1}^{d_0} \text{Cov}\left(w_{ij} x_j, w_{ir} x_r\right)$$

כאשר:

$$\begin{aligned}
\text{Cov}\left(w_{ij} x_j, w_{ir} x_r\right) &= E\left[w_{ij} x_j w_{ir} x_r\right] - E\left[w_{ij} x_j\right] E\left[w_{ir} x_r\right] \\
&= E\left[w_{ij} w_{ir}\right] E\left[x_j x_r\right] - E\left[w_{ij}\right] E\left[w_{ir}\right] E\left[x_j\right] E\left[x_r\right] \\
&= E\left[w_{ij} w_{ir}\right] E\left[x_j x_r\right]
\end{aligned}$$

כאשר במעבר האחרון השתמשנו בכך ש- $E\left[X\right] = 0$ ו- $E\left[W_1\right] = 0$. מכך שרכיבי האתחול בלתי תלויים ועם שונות σ_1^2, נקבל

$$\text{Cov}\left(w_{ij} x_j, w_{ir} x_r\right) = \delta_{jr} \sigma_1^2 \text{Var}\left(x_j\right) = \delta_{jr} \sigma_1^2$$

ולכן:

$$1 = \sum_{r=1}^{d_0} \sum_{j=1}^{d_0} \text{Cov}\left(w_{ij} x_j, w_{ir} x_r\right) = \sum_{r=1}^{d_0} \sum_{j=1}^{d_0} \delta_{jr} \sigma_1^2 = \sum_{j=1}^{d_0} \sigma_1^2 = d_0 \sigma_1^2$$

מכאן הבחירה $\sigma_1 = \frac{1}{\sqrt{d_0}}$ תיתן $1 = \text{Var}\left(\left(W_1 X\right)_i\right)$, כנדרש.

באופן דומה לשאר הנוירונים: $\forall \ell = 1, \ldots, L: \sigma_\ell = \frac{1}{\sqrt{d_{\ell-1}}}$.

ה)
לשם הגיוון, בסעיף זה נבצע את החישוב על הביטויים הסופיים, ולא שכבה-שכבה כמו בסעיף הקודם. מכיוון שכל המשקולות בלתי תלויות ועם ממוצעים שווים לאפס, נקבל:

$$E\left[G_\ell\right] = E\left[W_{\ell+1}^T \cdots W_{L-1}^T W_L^T\right] = 0$$

עכשיו, נחשב את השונות ושוב, מכיוון שכל המשקולות בלתי תלויות ועם ממוצעים שווים לאפס, נקבל:

$$\begin{aligned}\operatorname{Var}\left(g_{i,\ell}\right) &= \operatorname{Var}\left(\sum_{i_{\ell+1},\ldots,i_L} w_{ii_{\ell+1},\ell+1}\cdots w_{i_L,L}\right)\\ &= \sum_{i_{\ell+1},\ldots,i_L} \operatorname{Var}\left(w_{ii_{\ell+1},\ell+1}\right)\cdots\operatorname{Var}\left(w_{i_L,L}\right)\\ &= \sum_{i_{\ell+1},\ldots,i_L} \sigma_{\ell+1}^2\cdots\sigma_L^2 = d_{\ell+1}\cdots d_L\sigma_{\ell+1}^2\cdots\sigma_L^2\end{aligned}$$

מכאן, נקבל שבחירה $\forall\ell=1,\ldots,L:\sigma_\ell=\frac{1}{\sqrt{d_\ell}}$, תיתן לנו $\operatorname{Var}\left(g_{i,\ell}\right)=1$.

<u>שאלה מס' 3</u> (24%)

אסופת שאלות בנושאים שונים:

(שערוך)

8% א. נתונות לנו N מדידות IID $\{x_i\}_{i=1}^{N}$ כאשר x_i מגיע מההתפלגות הבאה:

$$P_X(x) = \frac{1}{\mu} e^{-\frac{1}{\mu}(x-\theta)}, \;\; x \ge \theta, \;\; \mu > 0$$

(1) מצאו את משערך MLE עבור הפרמטר μ בהנחה כי θ פרמטר ידוע.

(2) מצאו את משערך MLE עבור הפרמטר θ בהנחה כי μ פרמטר ידוע.

<u>פתרון:</u>

א. כמו בסעיף הקודם, פונקציית ה-Likelihood (הפעם כפונקציה של μ כי הוא המשתנה הלא ידוע בסעיף זה):

$$L(\mu) = \prod_{i=1}^{N} \frac{1}{\mu} e^{-\frac{1}{\mu}(x_i-\theta)} = \frac{1}{\mu^N} e^{-\frac{1}{\mu}\sum_{i=1}^{N}(x_i-\theta)} I_{\{\mu>0\}}$$

נניח כאן כי $\mu > 0$ אחרת האינדיקטור מתאפס.

$$l(\mu) = \log L(\mu) = -N \log \mu - \frac{1}{\mu}\sum_{i=1}^{N}(x_i - \theta)$$

מגזירה והשוואה לאפס נקבל

$$l'(\mu) = -\frac{N}{\mu} + \frac{1}{\mu^2}\sum_{i=1}^{N}(x_i - \theta) = 0 \Rightarrow \hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \theta)$$

הנגזרת השנייה שלילית ולכן זוהי אכן נקודת מקסימום.

ב. נכתוב את ה-Likelihood:

$$L(\theta) = \prod_{i=1}^{N} \frac{1}{\mu} e^{-\frac{1}{\mu}(x_i-\theta)} I_{\{x_i \ge \theta\}} = \frac{1}{\mu^N} e^{-\frac{1}{\mu}\sum_{i=1}^{N}(x_i-\theta)} I_{\left\{\min_i x_i \ge \theta\right\}}$$

כאשר I_A היא פונקציית אינדיקטור (מקבלת 1 אם המאורע A מתקיים ו-0 אחרת). נשים לב כי $L(\theta)$ היא פונקציה מונוטונית עולה ב-θ בתחום שבו $\theta \le \min_i x_i$. לכן, משערך הסבירות המירבית יתקבל בערך המקסימלי האפשרי עבור θ בתחום זה: $\hat{\theta}_{MLE} = \min_i x_i$.

(עצים)

6% ב. רותי רוצה לבנות עץ החלטה המבוסס על קריטריון האנטרופיה וקובע על סמך העונה, מצב הלחות והטמפרטורה האם צפוי לרדת גשם או לא. ברשותה 6 מדידות מהעבר:

עונה	מצב הלחות	טמפרטורה	האם ירד גשם?
אביב	גבוהה	נמוכה	כן
אביב	נמוכה	נמוכה	לא
סתיו	גבוהה	נמוכה	לא
סתיו	גבוהה	גבוהה	כן
קיץ	גבוהה	נמוכה	כן
קיץ	נמוכה	גבוהה	לא

מהו הקריטריון שישמש בצומת הראשונה בעץ?

<u>פתרון:</u>

נחשב את האנטרופיה שתושרה לאחר פיצול לפי כל אחד מהמאפיינים האפשריים-

עונה:

עונה	האם ירד גשם?
אביב	כן
אביב	לא
סתיו	לא
סתיו	כן
קיץ	כן
קיץ	לא

אביב: 1-/1+.

$$H(\text{rain}|\text{winter}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$$

סתיו: 1-/1+.

$$H(\text{rain}|\text{winter}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$$

קיץ: 1-/1+.

$$H(\text{rain}|\text{winter}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$$

אנטרופיה כללית בחלוקה על סמך עונה:

$$H(\text{rain}|\text{season}) = \frac{2}{6}\cdot 1 + \frac{2}{6}\cdot 1 + \frac{2}{6}\cdot 1 = 1$$

מצב הלחות:

מצב הלחות	האם ירד גשם?
גבוהה	כן
נמוכה	לא
גבוהה	לא
גבוהה	כן
גבוהה	כן
נמוכה	לא

לחות גבוהה: +3/-1

$$H(\text{rain}|\text{high humidity}) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.81$$

לחות נמוכה: +0/-2.

$$H(\text{rain}|\text{low humidity}) = 0$$

אנטרופיה כללית בחלוקה על סמך מצב הלחות:

$$H(\text{rain}|\text{humidity}) = \frac{4}{6}\cdot 0.81 + \frac{2}{6}\cdot 0 = 0.54$$

טמפרטורה:

טמפרטורה	האם ירד גשם?
נמוכה	כן
נמוכה	לא
נמוכה	לא
גבוהה	כן
נמוכה	כן
גבוהה	לא

טמפרטורה נמוכה: +2/-2.

$$H(\text{rain}|\text{low temperature}) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$$

טמפרטורה גבוהה: $+1/-1$.

$$H(\text{rain}|\text{high temperature}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$$

אנטרופיה כללית בחלוקה על סמך גובה הטמפרטורה:

$$H(\text{rain}|\text{humidity}) = 1$$

קריטריון הלחות השיג את האנטרופיה הכללית הכי נמוכה ולכן זהו הולך להיות הפיצול הראשון בעץ.

(SVM)

6% ג נתונה בעיית האופטימיזציה הבאה:

$$\min_{\tilde{w}\in\mathbb{R}^{d+1},\xi\in\mathbb{R}^n}\left(\frac{1}{2}\|\tilde{w}\|^2+C\sum_{i=1}^{n}\xi_i\right) \qquad (P)$$
$$\text{subject to} \quad : y_i\tilde{w}^T\tilde{x}_i \geq 1-\xi_i \qquad i=1,...,n$$
$$\xi_i \geq 0 \qquad i=1,...,n$$

עבור $\tilde{w}=(w,b),\ \tilde{x}_i=(x_i,1)$

הוכיחו או הפריכו את הטענה הבאה:

בעיית האופטימיזציה הנתונה שקולה לבעיית ה- Soft-SVM שנלמדה בכתה.

<u>פתרון:</u>

אין הבדל בנושא זה בין בעיית ה SVM הרגילה ובעיית ה- Soft-SVM, ולכן נדון בבעיה הרגילה.

בעיית ה- SVM הפרימאלית ניתנת לפי:

$$\begin{cases}\min\limits_{\tilde{w}\in\mathbb{R}^{m+1}}\frac{1}{2}\|\tilde{w}\|^2\\ s.t.\ \ y_i\tilde{w}^T\tilde{x}_i\geq 1\quad i=1,...,\text{n}\end{cases}=$$
$$=\begin{cases}\min\limits_{w\in\mathbb{R}^m,b\in\mathbb{R}}\frac{1}{2}\|w\|^2+\frac{1}{2}b^2\\ s.t.\ \ y_i\tilde{w}^T\tilde{x}_i\geq 1\quad i=1,...,\text{n}\end{cases}$$

כעת יש רגורלריזציה על איבר ה- BIAS בניגוד לבעיה המקורית. לכן מדובר בשתי בעיות שונות עם פתרון שונה.

נפרש בעיה זו – כעת, אנחנו רוצים למצוא מישור מפריד בעל Margin גדול ככל הניתן, אולם תוך כדי מתן העדפה לישרים מפרידים אשר עוברים בקירוב דרך הראשית (כלומר בעלי b קטן).

(PCA)

6% ד. עבור סדרת נקודות נתונות $\{x_1,\dots,x_n\}$ ב- $\mathbb{R}^2$ (בעלות ממוצע 0), חושבה מטריצת קווריאנס-המדגם הבאה:

$$P_n=\begin{pmatrix}3 & 2\\ 2 & 6\end{pmatrix}$$

(1) איזה מהווקטורים הבאים מייצג (עד כדי נירמול בקבוע) את הכיוון העיקרי (הראשון) של הנקודות הנתונות?

$$w_1=\begin{pmatrix}-2\\ 1\end{pmatrix},\ w_2=\begin{pmatrix}1\\ 1\end{pmatrix},\ w_3=\begin{pmatrix}1\\ 2\end{pmatrix}$$

(2) חשבו את שני הרכיבים הראשיים של $x=(1,0)^T$

<u>פתרון:</u>

(1

בחישוב מהיר ניתן לראות:

$$P_nw_1=\begin{pmatrix}3 & 2\\ 2 & 6\end{pmatrix}\begin{pmatrix}-2\\ 1\end{pmatrix}=\begin{pmatrix}-6+2\\ -4+6\end{pmatrix}=\begin{pmatrix}-4\\ 2\end{pmatrix}=2\begin{pmatrix}-2\\ 1\end{pmatrix}$$

מכאן, ש w_1 הינו וקטור עצמי בעל ערך עצמי 2.

$$P_nw_2=\begin{pmatrix}3 & 2\\ 2 & 6\end{pmatrix}\begin{pmatrix}1\\ 1\end{pmatrix}=\begin{pmatrix}3+2\\ 2+6\end{pmatrix}=\begin{pmatrix}5\\ 8\end{pmatrix}\neq\alpha\begin{pmatrix}1\\ 1\end{pmatrix}$$

כלומר, w_2 איננו וקטור עצמי של מטריצת הקווריאנס.

$$P_nw_3=\begin{pmatrix}3 & 2\\ 2 & 6\end{pmatrix}\begin{pmatrix}1\\ 2\end{pmatrix}=\begin{pmatrix}3+4\\ 2+12\end{pmatrix}=\begin{pmatrix}7\\ 14\end{pmatrix}=7\begin{pmatrix}1\\ 2\end{pmatrix}$$

כלומר, w_3 אף הוא וקטור עצמי, בעל ערך עצמי 7.

מאחר שערך עצמי זה המתאים ל w_3 גבוה מהערך העצמי המתאים לקטור העצמי w_1,
הרי שהוא מייצג את הכיוון העיקרי של הנקודות הנתונות.

(2

כעת, נרצה להטיל את הוקטור $x=(1\quad 0)^T$ על מערכת הצירים של הכיוונים העיקריים (PRINCIPAL COMPONENTS).

$$x_{PCA}(1) = \frac{w_3^{\ T} x}{\|w_3\|} = \frac{1}{\sqrt{5}}$$

$$x_{PCA}(2) = \frac{w_1^{\ T} x}{\|w_1\|} = \frac{-2}{\sqrt{5}}$$

$$\Rightarrow \boxed{x_{PCA} = \frac{1}{\sqrt{5}}\begin{pmatrix}1 & -2\end{pmatrix}}$$