אוניברסיטת ת"א, ביה"ס להנדסת חשמל, למידת מכונה סטטיסטית

תרגיל בית 4

תרגיל בית זה עוסק במודלים לינאריים.
התרגיל מורכב משני חלקים – חלק א' תיאורטי וחלק ב' שהינו תרגיל מחשב. יש להגיש כל חלק **בנפרד**, כל עבודה תיבדק בנפרד.

הגשה: עליכם להגיש קובץ **PDF** עבור התרגיל התיאורטי, ואילו עבור תרגיל המחשב יש להגיש קובץ **py.** עם הקוד וקובץ **PDF** עם הגרפים והתוצאות הנדרשים. יש להגיש את התרגיל התיאורטי ואת תרגיל המחשב בתאי הגשה **נפרדים** ב-moodle. חובה לציין מספר ת.ז. בקבצי ההגשה (בקוד – בהערה בתחילתו).
תזכורת: מי שמגיש בזוג, יש להגיש פעם אחת בלבד תרגיל תיאורטי ופעם אחת תרגיל מחשב (ניתן לערבב זוגות).

חלק א' - שאלות תיאורטיות:

שאלה 1
נניח את המודל שלמדנו בכיתה

$$y(\mathbf{x},\mathbf{w})=\mathbf{w}^T\phi(\mathbf{x})$$

נתון בסיס נתונים לאימון המודל, $\{\mathbf{x}_n,t_n\}_{n=1}^N$. משתמשים בקריטריון אימון מעט שונה מזה שלמדנו. מחשבים את וקטור הפרמטרים $\mathbf{w}$ שמביא למינימום את

$$E_D(\mathbf{w})=\frac{1}{2}\sum_{n=1}^N r_n\left(t_n-\mathbf{w}^T\phi(\mathbf{x}_n)\right)^2$$

כאשר $r_n>0,\ n=1,2,\ldots,N$, הם קבועים נתונים מראש (בכיתה היה לנו $r_n=1,\ n=1,2,\ldots,N$).

א. קבלנו נוסחא לוקטור הפרמטרים המשוערך, $\mathbf{w}^*$, שמשיג את המינימום של פונקציית המטרה.
ב. חיזרו על סעיף א' עבור המקרה שבו פונקצית המטרה אותה רוצים להביא למינימום היא

$$E_D(\mathbf{w})+\frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

שאלה 2
נניח את המודל הבא

$$y(\mathbf{x},\mathbf{w})=w_0+\sum_{i=1}^M w_i x_i$$

בהינתן בסיס נתונים לאימון, $\{\mathbf{x}_n,t_n\}_{n=1}^N$, משערך הריבועים הפחותים שלמדנו מביא למינימום את

$$E_D(\mathbf{w})\triangleq\frac{1}{2}\sum_{n=1}^N(t_n-y_n)^2$$

עבור $y_n=y(\mathbf{x}_n,\mathbf{w})=w_0+\sum_{i=1}^M w_i x_{n,i}$.

יהיו $\{e_{n,i}\}$ משתנים אקראיים גאוסיים, $e_{n,i}\sim N(0,\sigma^2)$, בלתי תלויים סטטיסטית, עבור שונות σ^2 ידועה. עכשיו רוצים למצוא את $\tilde{\mathbf{w}}$ שמביא למינימום את $E\{\tilde{E}_D(\mathbf{w})\}$ ($E\{\cdot\}$ מציין תוחלת ביחס ל-$\{e_{n,i}\}$) עבור

$$\tilde{E}_D\left(\mathbf{w}\right) \square \frac{1}{2}\sum_{n=1}^{N}\left(t_n-\tilde{y}_n\right)^2 \quad , \quad \tilde{y}_n = w_0+\sum_{i=1}^{M} w_i\left(x_{n,i}+e_{n,i}\right)$$

הראו ש- $\tilde{\mathbf{w}}$ המתקבל מביא למינימום את

$$E_D\left(\mathbf{w}\right)+\frac{\sigma^2 N}{2}\sum_{i=1}^{M} w_i^2$$

ז"א, זהו הפתרון של ridge regression עם פרמטר $\lambda=\sigma^2 N$.

<u>שאלה 3</u>

עבור multiclass logistic regression קיבלנו בכיתה את הביטוי הבא עבור הנגזרת של $E\left(\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_K\right)$ ביחס ל- $\mathbf{w}_j$ (מאורגן כוקטור שורה):

$$\nabla_{\mathbf{w}_j} E\left(\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_K\right)=\sum_{n=1}^{N}\left(y_{n,j}-t_{n,j}\right)\phi_n^T$$

א. רישמו את מטריצת ההסיאן (מטריצת הנגזרות השניות) של $E\left(\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_K\right)$ ביחס לפרמטרים $\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_K$. שימו לב שאם $\phi\left(\mathbf{x}\right)$ הוא וקטור ממימד M אזי מטריצת ההסיאן היא ממימד $KM\times KM$ ומורכבת מ- K^2 בלוקים שלכ"א מהם יש מימד $M\times M$. יש לקבל ביטוי עבור הבלוק ה- $\left(k,j\right)$, $\nabla_{\mathbf{w}_k}\nabla_{\mathbf{w}_j} E\left(\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_K\right)$.

ב. הראו שמטריצת ההסיאן שקיבלתם היא positive semidefinite. רמז: אחת האפשרויות היא להשתמש באי שוויון קושי שווארץ שאומר: $\left(\sum_{k=1}^{K} a_k b_k\right)^2 \le \left(\sum_{k=1}^{K} a_k^2\right)\left(\sum_{k=1}^{K} b_k^2\right)$.

<u>חלק ב' – תרגיל מחשב</u>

<u>שימו לב</u>- פתרון לחלק זה יש להגיש בנפרד, הבדיקה תהיה נפרדת מן התרגיל התיאורטי.

יש לציין בשורה הראשונה בקובץ הקוד את ת.ז ושמות המגישים, בצורת הערה.

דרישות-

- יש להשתמש בספריות לייבוא בסיס הנתונים, שרטוט גרפים ו- NumPy בלבד! אין להשתמש במודלים מוכנים.
- אין להשתמש ב- test set בתור דוגמאות אימון!
- זמן ריצת האלגוריתם צריך להיות סביר, כלומר בסדר גודל של דקות בודדות על מחשב PC סטנדרטי.
- הניקוד יחושב עפ"י יעילות חישובית, זמן ריצה, מספר לולאות מועט ככל האפשר בקוד ודיוק.

בתרגיל זה נכתוב מסווג עבור בסיס הנתונים MNIST, שמורכב מתמונות בגודל 28×28 פיקסלים (grayscale) של ספרות (0-9) שכתובות בכתב יד ומתיוגים שלהם - הספרות שנכתבו למעשה. לדוגמא, ניתן לראות 10 דוגמאות אימון מבסיס הנתונים:

וקטור התיוגים של דוגמאות האימון: $[1,9,6,5,8,1,4,4,1,8]$.

על מנת להקל על החישובים, נבצע התאמה של וקטור התיוגים לצורת one-hot, כלומר עבור המאגר שלנו, נהפוך כל תיוג יחיד לוקטור באורך 10, כך שהאיבר באינדקס של התיוג האמיתי יהיה '1', וכל שאר האיברים יהיו '0'.

למשל:

עבור התיוג 5, הוקטור המתאים יהיה $[0,0,0,0,0,1,0,0,0,0]$ (זיכרו שהספרה הראשונה היא 0).

עבור התיוג 0, נקבל $[1,0,0,0,0,0,0,0,0,0]$ וכן הלאה.

כפי שנלמד בכיתה, נרצה למזער את ה- cross entropy loss: $E(\mathbf{w}_0,...,\mathbf{w}_9)=-\sum_{n=1}^{N}\sum_{k=0}^{9}t_{n,k}\ln y_{n,k}$

כאשר N הינו מספר דוגמאות האימון. $t_{n,k}$ מייצג את התיוג (האמיתי) עבור הדוגמא ה- n, ו- $y_{n,k}$ הינו החיזוי עבור תיוג הדוגמא ה- n והוא מוגדר ע"י $y_{n,k}=\frac{e^{\mathbf{w}^T{}_k\mathbf{x}_n}}{\sum_j e^{\mathbf{w}^T{}_j\mathbf{x}_n}}$ (כאשר $\mathbf{x}_n$ הוא התמונה כאשר משטחים אותה לוקטור בתוספת של 1 בסופה, וכל וקטור משקולות, $j=0,1,\ldots,9$, $\mathbf{w}_j$, כולל היסט (bias) בסופו). כפי שראיתם בהרצאה, הגרדיאנט מחושב ע"י

$$\nabla_{\mathbf{w}_j}E(\mathbf{w}_0,\mathbf{w}_1,...,\mathbf{w}_9)=\sum_{n=1}^{N}(y_{n,j}-t_{n,j})\mathbf{x}_n \ .$$

<u>הדרכה לפתרון:</u>

1. יבאו את בסיס הנתונים MNIST. ניתן להיעזר בקוד לדוגמא שפורסם ב-moodle.
2. שטחו את דוגמאות בבסיס הנתונים מתמונות לוקטורים חד מימדיים באורך 785 כך שהאיבר האחרון הוא 1 עבור כל הדוגמאות.

3. בנו את המטריצה - $\mathbf{X}=\begin{bmatrix}\mathbf{x}_1^T\\ \vdots \\ \mathbf{x}_N^T\end{bmatrix}$ ששורותיה מורכבות מדוגמאות מבסיס הנתונים. הוקטורים $\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N\}$ הם באורך $28\times 28+1=785$, כאשר כל וקטור מייצג דוגמה מבסיס הנתונים (תוספת של 1 לטובת bias כפי שצוין למעלה).
4. הפרידו את בסיס הנתונים ל-3 קבוצות: 60% דוגמאות ל- training set, 20% דוגמאות ל-validation set ו-20% דוגמאות ל-test set. ניתן להיעזר בקוד שפורסם ב-moodle ובו דוגמא להפרדה ל-2 קבוצות.
5. אתחלו וקטורי משקולות אקראיים, $\mathbf{w}_0,...,\mathbf{w}_9$, באורך 785. **זכרו**, ישנו איבר bias בכל וקטור, והוא האיבר האחרון.
6. נסחו את פונקציית השגיאה (loss) בכתיב מטריצי.
7. מזערו אותה באמצעות אלגוריתם gradient descent שנלמד בכיתה:
   a. בכל איטרציה, בצעו עדכון: $\forall j\in\{0,...,9\}:\mathbf{w}_j^{(r+1)}=\mathbf{w}_j^{(r)}-\eta\nabla E(\mathbf{w}_j^{(r)})$.
   b. עבור כל איטרציה, חשבו את ה-loss על ה-training set.
   c. עבור כל איטרציה, חשבו את דיוק המודל על ה-validation set:

$$\frac{\#Correct\ classifications\ on\ validation\ set}{Size\ of\ validation\ set}\cdot 100\%$$

   d. התחילו הרצות עם $\eta=0.01$. אם אין התכנסות, שנו את ערכו והריצו את הקוד מחדש.
   e. הפסיקו את הריצה עם תנאי עצירה המבטיח התכנסות של הדיוק המתקבל על ה-validation set, כלומר הדיוק אינו משתנה הרבה בין האיטרציות.

הגישו את סעיפים 8-10 בקובץ PDF:

8. ציירו גרף של ערך ה-loss המתקבל על ה-training set כפונקציה של מספר האיטרציה.
9. ציירו גרף של דיוק המודל על ה- validation set כפונקציה של מספר האיטרציה.
10. רשמו את הדיוק המתקבל על כל אחת מהקבוצות (training/validation/test) לאחר האיטרציה האחרונה.

נשים לב לנקודות הבאות:

1. החיזוי של הדוגמא ה- n מתקבל ע"י $\arg\max_j y_{n,j}$.

2. ניתן לרשום $\nabla_{\mathbf{w}_j}E(\mathbf{w}_0,\mathbf{w}_1,...,\mathbf{w}_9)=\mathbf{X}^T\left(\mathbf{y}_j-\mathbf{t}_j\right)$.

3. על מנת להאיץ את הריצה, יש לממש את וקטורי המשקולות בתור מטריצה, ולבצע את כל החישובים עם מטריצה זו.