Intro to Machine Learning: Assignment #3

April 30, 2023

Dor Bourshan
315780122

Theory Questions

1 Step-size Perceptron.

נאמר שהאלגוריתם Perceptron מבצע את העדכון הבא בעת בשגיאה ($\hat{y}_t \neq y_t$):

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \eta_t y_t \mathbf{x}_t$$

נניח שהדטא מופרד עם מרווח $\gamma > 0$ וכי מתקיים $\|\mathbf{x}_t\|_2 = 1$ לכל t [1]. בנוסף נניח כי האלגוריתם מבצע M שגיאות וכולן מתרחשות בתחילתו.

נראה שעבור $\eta_t = \frac{1}{\sqrt{t}}$ מספר השגיאות של האלגוריתם חסום ע"י $\mathcal{O}\left(\frac{4}{\gamma^2}\ln\left(\frac{1}{\gamma}\right)\right)$ כאשר אנו מניחים כי $\gamma < 1$ אחרת אין לחסם משמעות.

מכך שהדטא מופרד, קיים מסווג מושלם שנסמנו $\mathbf{w}^*$, ונוכל להניח בלי הגבלת הכלליות כי $\|\mathbf{w}_t\|_2 = \frac{1}{2}$.

לכל $1 \le t \le M$ האלגוריתם מבצע טעות, ולכן:

$$\langle \mathbf{w}_{t+1}, \mathbf{w}^* \rangle = \langle \mathbf{w}_t + \eta_t y_t \mathbf{x}_t, \mathbf{w}^* \rangle = \langle \mathbf{w}_t, \mathbf{w}^* \rangle + \langle \eta_t y_t \mathbf{x}_t, \mathbf{w}^* \rangle = \langle \mathbf{w}_t, \mathbf{w}^* \rangle + \eta_t \langle y_t \mathbf{x}_t, \mathbf{w}^* \rangle \ge \langle \mathbf{w}_t, \mathbf{w}^* \rangle + \frac{1}{\sqrt{t}}\gamma$$

כלומר עבור $t > M$ מתקיים כי

$$\langle \mathbf{w}_{t+1}, \mathbf{w}^* \rangle \ge \gamma \cdot \sum_{t=1}^{M} \frac{1}{\sqrt{t}} \ge \gamma\sqrt{M} \tag{1}$$

בנוסף, מתקיים כי לכל $1 \le t \le M$

$$\|\mathbf{w}_{t+1}\|_2^2 = \|\mathbf{w}_t + \eta_t y_t \mathbf{x}_t\|_2^2 = \|\mathbf{w}_t\|_2^2 + 2\langle \mathbf{w}_t, \eta_t y_t \mathbf{x}_t \rangle + \|\eta_t y_t \mathbf{x}_t\|_2^2$$

נשים לב כי לכל $1 \le t \le M$ האלגוריתם מבצע טעות ולכן $\langle \mathbf{w}_t, y_t \mathbf{x}_t \rangle < 0$ ולכן $2\langle \mathbf{w}_t, \eta_t y_t \mathbf{x}_t \rangle < 0$ כלומר אנו מקבלים כי

$$\|\mathbf{w}_{t+1}\|_2^2 = \|\mathbf{w}_t\|_2^2 + 2\langle \mathbf{w}_t, \eta_t y_t \mathbf{x}_t \rangle + \|\eta_t y_t \mathbf{x}_t\|_2^2 \le \|\mathbf{w}_t\|_2^2 + \frac{1}{t}\|\mathbf{x}_t\|_2^2 \overset{1}{=} \|\mathbf{w}_t\|_2^2 + \frac{1}{t}$$

כלומר עבור $t' > M$ מתקיים כי

$$\|\mathbf{w}_t\|_2^2 \le \sum_{p=1}^{M} \frac{1}{p} = H_M \le 1 + \ln(M) \tag{2}$$

אם כן עבור $t > M$ מתקיים כי:

$$\gamma\sqrt{M} \le \langle \mathbf{w}_t, \mathbf{w}^* \rangle \underset{\text{Cauchy–Schwarz}}{\overset{1}{\le}} \|\mathbf{w}_t\|_2 \|\mathbf{w}^*\|_2 \overset{2}{\le} \sqrt{\frac{1}{4} + \frac{1}{4}\ln(M)}$$

אם כן

$$\gamma\sqrt{M} \leq \sqrt{\frac{1}{4}+\frac{1}{4}\ln\left(M\right)}$$

- אם $M=0$ ברור כי מתקיים כי $M \leq \frac{4}{\gamma^2}\ln\left(\frac{1}{\gamma}\right)$.
- אם $M=1$ מתקיים כי $\gamma \leq \frac{1}{4}$ ולכן $1 < 16\ln\left(4\right) \leq \frac{4}{\gamma^2}\ln\left(\frac{1}{\gamma}\right)$
- עבור $M \geq 2$:

  עבור $x \geq 2$ מתקיים כי

$$1+\ln\left(x\right) \leq 4\ln\left(x\right)$$

כיוון ש

$$1 < \ln\left(8\right) \leq \ln\left(x^3\right) = 4\ln\left(x\right)-\ln\left(x\right)$$

ולכן

$$\frac{1}{4}+\frac{1}{4}\ln\left(x\right) \leq \ln\left(x\right)$$

אם כן מתקיים כי

$$\gamma\sqrt{M} \leq \sqrt{\frac{1}{4}+\frac{1}{4}\ln\left(M\right)} \leq \sqrt{\ln\left(M\right)}$$

ולכן

$$\gamma^2 M \leq \ln\left(M\right)$$

ומהרמז מתקיים כי

$$M \leq \frac{2}{\gamma^2}\ln\left(\frac{1}{\gamma^2}\right) = \frac{4}{\gamma^2}\ln\left(\frac{1}{\gamma^2}\right)$$

לכן סה"כ מתקיים כי $\frac{4}{\gamma^2}\ln\left(\frac{1}{\gamma^2}\right)$ חוסם את M כנדרש.

$\square$

2 Convex functions.

2.1 section (a)

תהיי $f:\mathbb{R}^n \to \mathbb{R}$ פונקציה קמורה, $A \in \mathbb{R}^{n\times n}$ ו $\mathbf{b} \in \mathbb{R}^n$. נראה כי $g(\mathbf{x}) = f(A\mathbf{x}+\mathbf{b})$ היא פונקציה קמורה.

ניזכר כי מטריצה A מגדירה העתקה לינארית, כלומר A היא פונקציה $T_A : \mathbb{R}^n \to \mathbb{R}^n$. אם כן נסמן את הקבוצה הבאה :

$$C = \{A\mathbf{x}+\mathbf{b} \mid \mathbf{x} \in \mathbb{R}^n\} \subseteq \mathbb{R}^n$$

C קבוצה קמורה. יהיי $\lambda \in (0,1)$, ויהיו $\mathbf{c}_1, \mathbf{c}_2 \in C$, קיימים $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ כך ש $\mathbf{c}_i = A\mathbf{x}_i + \mathbf{b}$ עבור $i \in \{1,2\}$.

$$\lambda\mathbf{c}_1 + (1-\lambda)\mathbf{c}_2 = \lambda(A\mathbf{x}_1+\mathbf{b}) + (1-\lambda)(A\mathbf{x}_2+\mathbf{b}) = A\lambda\mathbf{x}_1 + A(1-\lambda)\mathbf{x}_2 + \mathbf{b} = A(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) + \mathbf{b}$$

ברור כי $\mathbf{y} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in \mathbb{R}^n$ ולכן $A\mathbf{y}+\mathbf{b} \in C$ כנדרש.

אם כן מתקיים כי C קבוצה קמורה ולכל $\mathbf{c}_1, \mathbf{c}_2 \in C$ ולכל $\lambda \in (0,1)$, קיימים $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ כך ש $\mathbf{c}_i = A\mathbf{x}_i + \mathbf{b}$ עבור $i \in \{1,2\}$ וכן $A(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) + \mathbf{b} \in C$. אם כן

$$\begin{aligned} g(\lambda\mathbf{c}_1 + (1-\lambda)\mathbf{c}_2) &= f(A(\lambda\mathbf{c}_1 + (1-\lambda)\mathbf{c}_2) + \mathbf{b}) = f(\lambda(A\mathbf{c}_1+\mathbf{b}) + (1-\lambda)(A\mathbf{c}_2+\mathbf{b})) \\ &\le \lambda f(A\mathbf{c}_1+\mathbf{b}) + (1-\lambda) f(A\mathbf{c}_2+\mathbf{b}) \end{aligned}$$

כלומר

$$g(\lambda\mathbf{c}_1 + (1-\lambda)\mathbf{c}_2) \le \lambda f(A\mathbf{c}_1+\mathbf{b}) + (1-\lambda) f(A\mathbf{c}_2+\mathbf{b}) = \lambda g(\mathbf{c}_1) + (1-\lambda) g(\mathbf{c}_2)$$

כנדרש מפונקציה קמורה.

□

2.2 section (b)

תהיינה m פונקציות קמורות $f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})$ כך ש $f_i : \mathbb{R}^d \to \mathbb{R}$. נגדיר פונקציה חדשה $g : \mathbb{R}^d \to \mathbb{R}$:

$$g(\mathbf{x}) = \max_i f_i(\mathbf{x})$$

נראה כי g פונקציה קמורה.

ראשית $\mathbb{R}^d$ קבוצה קמורה (מהיותה שדה).

יהיו $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$. ויהיי $\lambda \in (0,1)$. מתקיים כי

$$\begin{aligned} g(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) &= \max_i f_i(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \leq \max_i (\lambda f_i(\mathbf{x}_1) + (1-\lambda) f_i(\mathbf{x}_2)) \\ &\leq \lambda \max_i f_i(\mathbf{x}_1) + (1-\lambda) \max_i f_i(\mathbf{x}_2) \end{aligned}$$

כלומר

$$g(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \leq \lambda \max_i f_i(\mathbf{x}_1) + (1-\lambda) \max_i f_i(\mathbf{x}_2) = \lambda g(\mathbf{x}_1) + (1-\lambda) g(\mathbf{x}_2)$$

כנדרש מפונקציה קמורה.

□

2.3 section (c)

תהיי $\ell_{log} : \mathbb{R} \to \mathbb{R}$ פונקציית log loss המוגדרת ע"י

$$\ell_{log}(z) = \log_2\left(1 + e^{-z}\right)$$

נוכיח תחילה כי ℓ_{log} היא פונקציה קמורה.

תחילה $\mathbb{R}$ היא קבוצה קמורה (מהיותה שדה). נסמן לשם הנוחות

$$\begin{aligned} g(z) &= 1 + e^{-z} \\ f(x) &= \log_2(x) \end{aligned}$$

אם כן

$$\ell_{log} = h = f \circ g$$

ניזכר במשפט שהוכח בקורס חדוו"א

משפט: אם f גזירה פעמיים בקטע, והנגזרת השנייה אי שלילית בכל הקטע, אז הפונקציה קמורה בקטע.

נגזור את h פעמיים ונראה כי הנגזרת אי שלילית ונסיק כי פונקציית log loss היא פונקציה קמורה.

אם כן לפי כלל השרשרת:

$$h^{'} = f^{'}\left(g\left(x\right)\right)g^{'}\left(x\right) = -\frac{e^{-x}}{\left(1+e^{-x}\right)\ln\left(2\right)}$$

ושוב

$$h^{''} = \frac{1}{\ln\left(2\right)}\left(\frac{e^{-x}\left(1+e^{-x}\right)-e^{-2x}}{\left(1+e^{-x}\right)^{2}}\right) = \frac{1}{\ln\left(2\right)}\left(\frac{e^{-x}\left(\left(1+e^{-x}\right)-e^{-x}\right)}{\left(1+e^{-x}\right)^{2}}\right) = \frac{1}{\ln\left(2\right)}\left(\frac{e^{-x}}{\left(1+e^{-x}\right)^{2}}\right) > 0$$

לפיכך ℓ_{log} קמורה כנדרש.

□

כעת נסיק כי $f : \mathbb{R}^{d} \to \mathbb{R}$ המוגדרת ע"י

$$f\left(\mathbf{w}\right) = \ell_{log}\left(y\mathbf{w}\cdot\mathbf{x}\right)$$

היא פונקציה קמורה ביחס ל $\mathbf{w}$.

ראשית $\mathbb{R}^{d}$ קבוצה קמורה (מהיותה שדה).

יהיי $\lambda \in (0,1)$. ויהיו $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^{d}$

$$\begin{aligned} f\left(\lambda\mathbf{w}_1 + (1-\lambda)\mathbf{w}_2\right) &= \ell_{log}\left(\left\langle\lambda\mathbf{w}_1 + (1-\lambda)\mathbf{w}_2, y\mathbf{x}\right\rangle\right) = \ell_{log}\left(\left\langle\lambda\mathbf{w}_1 + (1-\lambda)\mathbf{w}_2, y\mathbf{x}\right\rangle\right) \\ &= \ell_{log}\left(\lambda y\left\langle\mathbf{w}_1,\mathbf{x}\right\rangle + (1-\lambda)y\left\langle\mathbf{w}_2,\mathbf{x}\right\rangle\right) \end{aligned}$$

מכך שהראנו כי ℓ_{log} פונקציה קמורה וכן $\left\langle\mathbf{w}_i,\mathbf{x}\right\rangle \in \mathbb{R}$,

$$\begin{aligned} f\left(\lambda\mathbf{w}_1 + (1-\lambda)\mathbf{w}_2\right) &= \ell_{log}\left(\lambda y\left\langle\mathbf{w}_1,\mathbf{x}\right\rangle + (1-\lambda)y\left\langle\mathbf{w}_2,\mathbf{x}\right\rangle\right) \le \lambda\ell_{log}\left(y\left\langle\mathbf{w}_1,\mathbf{x}\right\rangle\right) + (1-\lambda)\ell_{log}\left(y\left\langle\mathbf{w}_2,\mathbf{x}\right\rangle\right) \\ &= \lambda f\left(\mathbf{w}_1\right) + (1-\lambda)f\left(\mathbf{w}_2\right) \end{aligned}$$

כנדרש מפונקציה קמורה.

□

3 Ranking.

נניח כי $\mathcal{X} \subseteq \mathbb{R}^d$ וניתנת קבוצת אימון של n רשימות של k איברים, לכל רשימה ניתן וקטור label. במפורש

$$S = \left\{ \left(\left(\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_k^i \right), \mathbf{y}^i \right) \right\}_{i=1}^n$$

כך שלכל $\mathbf{y}^i \in \mathbb{R}^k$ $1 \leq i \leq n,$ מושם ערך לכל ערך ב $\left(\mathbf{x}_1^i, \ldots, \mathbf{x}_k^i\right) = \left(\bar{\mathbf{x}}^i\right)$.

המטרה היא ללמוד פונקציית ranking $h : \mathcal{X}^k \to \mathbb{R}^k$ שמבצעת rank נכון לרשימת האיברים שמגיעים מ S.

מגדירים את loss $Kendall - Tau$ בין שני וקטורי rank : $\mathbf{y}', \mathbf{y}$ ע"י :

$$\Delta \left(\mathbf{y}', \mathbf{y} \right) = \frac{2}{k\left(k-1\right)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \mathbf{1} \left\{ sgn \left(y'_j - y'_r \right) \neq sgn \left(y_j - y_r \right) \right\}$$

נניח שאנו מעוניינים ללמוד פונקציית rank לינארית, כלומר פונקציה מהצורה

$$h_{\mathbf{w}} \left(\left(\mathbf{x}_1, \ldots, \mathbf{x}_k \right) \right) = \left(\left\langle \mathbf{w}, \mathbf{x}_1 \right\rangle, \ldots, \left\langle \mathbf{w}, \mathbf{x}_k \right\rangle \right)$$

עבור איזשהו $\mathbf{w} \in \mathbb{R}^d$, והמטרה היא למנמם את Kendall-Tau loos על S

$$\sum_{i=1}^{n} \Delta \left(h_{\mathbf{w}} \left(\bar{\mathbf{x}}^i \right), \mathbf{y}^i \right)$$

כיוון שפונקציה זו קשה לאופטימיזציה, אנו נשתמש בפונקציית hinge loss

$$\sum_{i=1}^{n} \ell \left(h_{\mathbf{w}} \left(\bar{\mathbf{x}} \right), \mathbf{y} \right)$$

כאשר

$$\ell \left(h_{\mathbf{w}} \left(\bar{\mathbf{x}} \right), \mathbf{y} \right) = \frac{2}{k\left(k-1\right)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \max \left\{ 0, 1 - sgn \left(y_j - y_r \right) \left\langle \mathbf{w}, \mathbf{x}_j - \mathbf{x}_r \right\rangle \right\}$$

3.1 section (a)

תחילה נוכיח כי פונקציית hinge loss לעיל היא פונקציה קמורה ביחס ל $\mathbf{w}$.

אנחנו נראה כי $1 - sgn \left(y_j - y_r \right) \left\langle \mathbf{w}, \mathbf{x}_j - \mathbf{x}_r \right\rangle$ היא פונקציה קמורה ביחס ל $\mathbf{w}$ וע"י שימוש ב 2.2 נסיק כי $\max \left\{ 0, 1 - sgn \left(y_j - y_r \right) \left\langle \mathbf{w}, \mathbf{x}_j - \mathbf{x}_r \right\rangle \right\}$ פונקציה קמורה (שכן פונקציית האפס היא פונקציה קמורה). כדי להשלים את התמונה אנו נצטרך להוכיח כי סכום של פונקציות קמורות היא פונקציה קמורה.

אם כן ראשית נוכיח כי סכום של פונקציות קמורות היא פונקציה קמורה באינדוקציה.

- **בסיס :** נניח כי g, f הן פונקציות קמורות על הקבוצה הקמורה C ונסמן $h = f + g$. יהיי $\lambda \in (0,1)$ ויהיו $\mathbf{c}_1, \mathbf{c}_2 \in C$

$$h\left(\lambda \mathbf{c}_1 + (1-\lambda)\mathbf{c}_2\right) = f\left(\lambda \mathbf{c}_1 + (1-\lambda)\mathbf{c}_2\right) + g\left(\lambda \mathbf{c}_1 + (1-\lambda)\mathbf{c}_2\right) \leq \lambda\left(f\left(\mathbf{c}_1\right) + g\left(\mathbf{c}_1\right)\right) + (1-\lambda)\left(f\left(\mathbf{c}_2\right) + g\left(\mathbf{c}_2\right)\right) = \lambda h\left(\mathbf{c}_1\right) + (1-\lambda) h\left(\mathbf{c}_2\right)$$

כנדרש

- **צעד :** נניח כי הנ"ל נכון עבור n ונראה עבור $n+1$. יהיו $f_1, \ldots, f_{n+1}$ פונקציות קמורות על הקבוצה הקמורה C : נגדיר $h = \sum_i^{n+1} f_i$. מהנחת האינדוקציה $\sum_i^n f_i$ היא פונקציה קמורה, ובדומה להוכחה מקרה הבסיס $\sum_i^n f_i + f_{n+1}$ כלומר $h = \sum_i^{n+1} f_i$ קמורה כנדרש.

כעת נראה כי $f(\mathbf{w}) = 1 - sgn\left(y_j - y_r\right)\langle \mathbf{w}, \mathbf{x}_j - \mathbf{x}_r \rangle$ היא פונקציה קמורה ביחס ל $\mathbf{w}$. נסמנה

ראשית $\mathbf{w} \in \mathbb{R}^d$ ו $\mathbb{R}^d$ קבוצה קמורה (מהיותה שדה). יהיי $\lambda \in (0,1)$ ויהיו $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$

$$\begin{aligned}
f\left(\lambda \mathbf{w}_1 + (1-\lambda)\mathbf{w}_2\right) =& 1 - sgn\left(y_j - y_r\right)\langle \lambda \mathbf{w}_1 + (1-\lambda)\mathbf{w}_2, \mathbf{x}_j - \mathbf{x}_r \rangle \\
&= 1 - sgn\left(y_j - y_r\right)\left(\langle \lambda \mathbf{w}_1, \mathbf{x}_j - \mathbf{x}_r \rangle + \langle (1-\lambda)\mathbf{w}_2, \mathbf{x}_j - \mathbf{x}_r \rangle\right) \\
&= 1 - \lambda \cdot sgn\left(y_j - y_r\right)\langle \mathbf{w}_1, \mathbf{x}_j - \mathbf{x}_r \rangle + (1-\lambda) sgn\left(y_j - y_r\right)\left(\langle \mathbf{w}_2, \mathbf{x}_j - \mathbf{x}_r \rangle\right) \\
&= \lambda + 1 - \lambda - \lambda \cdot sgn\left(y_j - y_r\right)\langle \mathbf{w}_1, \mathbf{x}_j - \mathbf{x}_r \rangle - (1-\lambda) sgn\left(y_j - y_r\right)\left(\langle \mathbf{w}_2, \mathbf{x}_j - \mathbf{x}_r \rangle\right) \\
&= \lambda\left(1 - \cdot sgn\left(y_j - y_r\right)\langle \mathbf{w}_1, \mathbf{x}_j - \mathbf{x}_r \rangle\right) + (1-\lambda)\left(1 - sgn\left(y_j - y_r\right)\left(\langle \mathbf{w}_2, \mathbf{x}_j - \mathbf{x}_r \rangle\right)\right) \\
&= \lambda f\left(\mathbf{w}_1\right) + (1-\lambda) f\left(\mathbf{w}_2\right)
\end{aligned}$$

כנדרש מפונקציה קמורה.

מסקנה : $\ell\left(h_{\mathbf{w}}\left(\bar{\mathbf{x}}\right), \mathbf{y}\right)$ פונקציה קמורה כנדרש.

□

3.2 section (b)

נוכיח כי פונקציית hinge loss לעיל חוסמת מלמעלה את Kendall-Tau loss

יהיי $\mathbf{w} \in \mathbb{R}^d$ וכן $\bar{\mathbf{x}} \in \mathcal{X}^k$ ו $\mathbf{y} \in \mathbb{R}^k$

נתבונן ב

$$\begin{aligned}
\max\left\{0, 1 - sgn\left(y_j - y_r\right)\langle \mathbf{w}, \mathbf{x}_j - \mathbf{x}_r \rangle\right\} &= \max\left\{0, 1 - sgn\left(y_j - y_r\right)\left(y'_j - y'_r\right)\right\} \\
&= \begin{cases} t(y) & sgn\left(y_j - y_r\right)\left(y'_j - y'_r\right) < 1 \\ 0 & otherwise \end{cases} \\
&= \begin{cases} t(y) & sgn\left(y_j - y_r\right) sgn\left(y'_j - y'_r\right)\left|y'_j - y'_r\right| < 1 \\ 0 & otherwise \end{cases} \\
&= \begin{cases} 1 + p(y) & sgn\left(y_j - y_r\right) \neq sgn\left(y'_j - y'_r\right) \\ r(y) & sgn\left(y_j - y_r\right) = sgn\left(y'_j - y'_r\right) \wedge \left|y'_j - y'_r\right| < 1 \\ 0 & otherwise \end{cases}
\end{aligned}$$

כאשר $p(y), r(y)$ פונקציות אי שליליות, ולכן מתקיים כי

$$\left(\begin{cases}1+p(y) & sgn(y_j-y_r)\neq sgn\left(y'_j-y'_r\right)\\ r(y) & sgn(y_j-y_r)=sgn\left(y'_j-y'_r\right)\wedge\left|y'_j-y'_r\right|<1\\ 0 & otherwise\end{cases}\right)\leq\left(\begin{cases}1 & sgn\left(y'_j-y'_r\right)\neq sgn(y_j-y_r)\\ 0 & otherwise\end{cases}\right)$$

אבל נשים לב כי

$$\mathbf{1}\left\{sgn\left(y'_j-y'_r\right)\neq sgn(y_j-y_r)\right\}=\begin{cases}1 & sgn\left(y'_j-y'_r\right)\neq sgn(y_j-y_r)\\ 0 & otherwise\end{cases}$$

כלומר מתקיים כי

$$\mathbf{1}\left\{sgn\left(y'_j-y'_r\right)\neq sgn(y_j-y_r)\right\}\leq\max\{0,1-sgn(y_j-y_r)\langle\mathbf{w},\mathbf{x}_j-\mathbf{x}_r\rangle\}$$

ולכן מתקיים :

$$\Delta\left(h_{\mathbf{w}}\left(\bar{\mathbf{x}}^i\right),\mathbf{y}^i\right)\leq\ell\left(h_{\mathbf{w}}(\bar{\mathbf{x}}),\mathbf{y}\right)$$

כנדרש

□

3.3 section (c)

נניח כי הדטא מופרד במרווח $\gamma>0$ (כלומר קיים $\mathbf{w}^*\in\mathbb{R}^d$ ו $\gamma>0$ כך ש $sgn(y_j-y_r)\langle\mathbf{w}^*,\mathbf{x}_j-\mathbf{x}_r\rangle\geq\gamma$ לכל $1\leq i\leq n$ ולכל $1\leq j<r\leq k$).

נראה כי אם נמנמם את hinge loss יוביל להיפותזה שמנמנמת את Kendall-Tau loss

מהנתון מתקיים כי קיים $\mathbf{w}^*\in\mathbb{R}^d$ $sgn(y_j-y_r)\langle\mathbf{w}^*,\mathbf{x}_j-\mathbf{x}_r\rangle\geq\gamma>0$ ולכן קיים $\mathbf{w}^{**}=\frac{1}{\gamma}\mathbf{w}^*$ כלומר מתקיים כי $sgn(y_j-y_r)\langle\mathbf{w}^{**},\mathbf{x}_j-\mathbf{x}_r\rangle=sgn(y_j-y_r)\left\langle\frac{1}{\gamma}\mathbf{w}^*,\mathbf{x}_j-\mathbf{x}_r\right\rangle\geq1>0$

כלומר מתקיים כי $1-sgn(y_j-y_r)\langle\mathbf{w}^{**},\mathbf{x}_j-\mathbf{x}_r\rangle\leq0$

ולכן

$$\max\{0,1-sgn(y_j-y_r)\langle\mathbf{w}^{**},\mathbf{x}_j-\mathbf{x}_r\rangle\}=0$$

כלומר למנמם את hinge loss על S פירושו

$$\sum_{i=1}^{n}\ell\left(h_{\mathbf{w}^{**}}(\bar{\mathbf{x}}),\mathbf{y}\right)=0$$

מ 3.2 אנו מקבלים כי

$$0 \leq \sum_{i=1}^{n} \Delta\left(h_{\mathbf{w}^{**}}\left(\bar{\mathbf{x}}^{i}\right), \mathbf{y}^{i}\right) \leq \sum_{i=1}^{n} \ell\left(h_{\mathbf{w}^{**}}\left(\bar{\mathbf{x}}\right), \mathbf{y}\right) = 0$$

כלומר

$$\sum_{i=1}^{n} \Delta\left(h_{\mathbf{w}^{**}}\left(\bar{\mathbf{x}}^{i}\right), \mathbf{y}^{i}\right) = 0$$

עבור $\mathbf{w}^{**}$ שנמצא ע"י המנמום של $\sum_{i=1}^{n} \ell\left(h_{\mathbf{w}^{**}}\left(\bar{\mathbf{x}}\right), \mathbf{y}\right)$.

כלומר ע"י מנמום $\sum_{i=1}^{n} \ell\left(h_{\mathbf{w}^{**}}\left(\bar{\mathbf{x}}\right), \mathbf{y}\right)$ הצלחנו למנמם (עבור אותה פונקציה) את $\sum_{i=1}^{n} \Delta\left(h_{\mathbf{w}^{**}}\left(\bar{\mathbf{x}}^{i}\right), \mathbf{y}^{i}\right)$

כנדרש.

$\square$

4 Gradient Sedcent on Smooth Functions.

הגדרה : נאמר שפונקציה $f:\mathbb{R}^n\to\mathbb{R}$ המקיימת $f\in C^1$ היא $\beta-\text{smooth}$ אם לכל $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$ מתקיים

$$f\left(\mathbf{y}\right)\leq f\left(\mathbf{x}\right)+\nabla f\left(\mathbf{x}\right)^{\top}\left(\mathbf{y}-\mathbf{x}\right)+\frac{\beta}{2}\|\mathbf{x}-\mathbf{y}\|^{2}$$

תהיי $f:\mathbb{R}^n\to\mathbb{R}$ פונקציה $\beta-\text{smooth}$ ואי שלילית.

נניח כי gradient descent algorithm מבוצע על f עם $\eta>0$, כלומר

$$\mathbf{x}_{t+1}\leftarrow\mathbf{x}_t-\eta\nabla f\left(\mathbf{x}_t\right)$$

ונניח כי אנו מתחילים בנקודה $\mathbf{x}_0$כלשהי.

נראה כי אם מתקיים ש $\eta<\frac{2}{\beta}$ אז מתקיים ש

$$\lim_{t\to\infty}\|\nabla f\left(\mathbf{x}_t\right)\|=0$$

אם כן, מהיות f פונקציה $\beta-\text{smooth}$ ואי שלילית מתקיים עבור $\mathbf{x}_{t+1},\mathbf{x}_t$ כי :

$$f\left(\mathbf{x}_{t+1}\right)\leq f\left(\mathbf{x}_t\right)+\nabla f\left(\mathbf{x}_t\right)^{\top}\left(\mathbf{x}_{t+1}-\mathbf{x}_t\right)+\frac{\beta}{2}\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^{2}$$

נזכור כי $\mathbf{x}_{t+1}\leftarrow\mathbf{x}_t-\eta\nabla f\left(\mathbf{x}_t\right)$ ולכן

$$f\left(\mathbf{x}_{t+1}\right)-f\left(\mathbf{x}_t\right)\leq\nabla f\left(\mathbf{x}_t\right)^{\top}\left(-\eta\nabla f\left(\mathbf{x}_t\right)\right)+\frac{\beta}{2}\|\eta\nabla f\left(\mathbf{x}_t\right)\|^{2}$$

כלומר

$$f\left(\mathbf{x}_{t+1}\right)-f\left(\mathbf{x}_t\right)\leq-\eta\|\nabla f\left(\mathbf{x}_t\right)\|^{2}+\frac{\beta}{2}\eta^{2}\|\nabla f\left(\mathbf{x}_t\right)\|^{2}$$

כלומר

$$f\left(\mathbf{x}_{t+1}\right)-f\left(\mathbf{x}_t\right)\leq\|\nabla f\left(\mathbf{x}_t\right)\|^{2}\left(\frac{\beta}{2}\eta^{2}-\eta\right)$$

נזכור את ההנחה כי $0<\eta<\frac{2}{\beta}$ ולכן מתקיים כי

$$\eta^{2}\frac{\beta}{2}-\eta<0$$

או לחלופין

$$^{1}\ \ \eta-\eta^{2}\frac{\beta}{2}>0$$

ולכן

$$\left(\frac{\beta}{2}\eta^2-\eta\right)^{-1}f\left(\mathbf{x}_{t+1}\right)-f\left(\mathbf{x}_t\right)\geq\|\nabla f\left(\mathbf{x}_t\right)\|^2\geq 0$$

הנ"ל נכון לכל t, ננצל זאת על מנת לקבל סכום טלסקופי

$$+\begin{cases}\left(\frac{\beta}{2}\eta^2-\eta\right)^{-1}f\left(\mathbf{x}_{t+1}\right)-f\left(\mathbf{x}_t\right)\geq\|\nabla f\left(\mathbf{x}_t\right)\|^2\\ \left(\frac{\beta}{2}\eta^2-\eta\right)^{-1}f\left(\mathbf{x}_{t}\right)-f\left(\mathbf{x}_{t-1}\right)\geq\|\nabla f\left(\mathbf{x}_{t-1}\right)\|^2\\ \vdots\\ \left(\frac{\beta}{2}\eta^2-\eta\right)^{-1}f\left(\mathbf{x}_{1}\right)-f\left(\mathbf{x}_0\right)\geq\|\nabla f\left(\mathbf{x}_0\right)\|^2\end{cases}$$

ונקבל

$$\left(\eta-\frac{\beta}{2}\eta^2\right)^{-1}\left(f\left(\mathbf{x}_0\right)-f\left(\mathbf{x}_{t+1}\right)\right)\geq\sum_{t=0}^{T}\|\nabla f\left(\mathbf{x}_t\right)\|^2$$

כאשר נשים לב כי מהיותה של f אי שלילית וכן $\left(\eta-\frac{\beta}{2}\eta^2\right)^{-1}>0$ [1] מתקיים כי

$$\left(\eta-\frac{\beta}{2}\eta^2\right)^{-1}f\left(\mathbf{x}_0\right)\geq\sum_{t=0}^{T}\|\nabla f\left(\mathbf{x}_t\right)\|^2$$

כלומר לכל T

$$positive-const\geq\sum_{t=0}^{T}\|\nabla f\left(\mathbf{x}_t\right)\|^2$$

נסמן

$$S_T=\sum_{t=0}^{T}\|\nabla f\left(\mathbf{x}_t\right)\|^2$$

מכך ש $\|\nabla f\left(\mathbf{x}_t\right)\|^2\geq 0$ זו סדרה מונוטונית עולה, והראנו כי חסומה, ולכן מתכנסת. כלומר

$$\sum_{t=0}^{\infty}\|\nabla f\left(\mathbf{x}_t\right)\|^2=L$$

ממשפט בקורס חדוו"א אנו יודעים שכיוון ש $\|\nabla f\left(\mathbf{x}_t\right)\|^2\geq 0$ וכן הטור מתכנס, מתקיים כי

$$\lim_{t\to\infty}\|\nabla f\left(\mathbf{x}_t\right)\|^2=0$$

נשים לב כי

$$\|\nabla f\left(\mathbf{x}_t\right)\| = \sqrt{\|\nabla f\left(\mathbf{x}_t\right)\|^2} \geq 0$$

ולכן ממשפט בקורס חדוו"א :

משפט : אם f רציפה ומתקיים כי $\lim_{n\to\infty} a_n = L$ אזי $\lim_{n\to\infty} f\left(a_n\right) = f\left(L\right)$

ולכן, כיוון ש $f\left(x\right) = \sqrt{x}$ פונקציה רציפה מתקיים כי

$$\lim_{t\to\infty} \|\nabla f\left(\mathbf{x}_t\right)\| = 0$$

כנדרש.

$\square$

Programming Assignment

5 SGD for Hinge loss.

5.1 (a)

כדי לבחור את η_0 הטוב ביותר, ניתן לאלגוריתם ערכים שהחלו ב

$$\left(10^{-5}, 10^{-4}, \ldots, 10^{4}, 10^{5}\right)$$

η_0 הטוב ביותר שהתקבל הוא $\eta_0 = 1$ עם accuracy$= 0.979$

נתמקד באזור בשבו התקבלה התוצאה הטובה ביותר, והאלגוריתם הורץ שנית (ברזולוציית קפיצות קטנה יותר)

$$(-9999, -9899, \ldots, 9801, 9901)$$

η_0 הטוב ביותר שהתקבל הוא $\eta_0 = 1$ עם accuracy$= 0.977$

נתמקד באזור בשבו התקבלה התוצאה הטובה ביותר, והאלגוריתם הורץ שלישית (ברזולוציית קפיצות קטנה יותר)

$$(-499, -489, \ldots, 481, 491)$$

η_0 הטוב ביותר שהתקבל הוא $\eta_0 = 1$ עם accuracy$= 0.979$

לאחר שהובן כח הערכים מתכנסים ל $\eta_0 = 1$ הורץ בפעם הרביעית על

$$(0.5, 0.51, \ldots, 1.49, 1.5)$$

η_0 הטוב ביותר שהתקבל הוא $\eta_0 = 1$ עם accuracy$= 0.981$

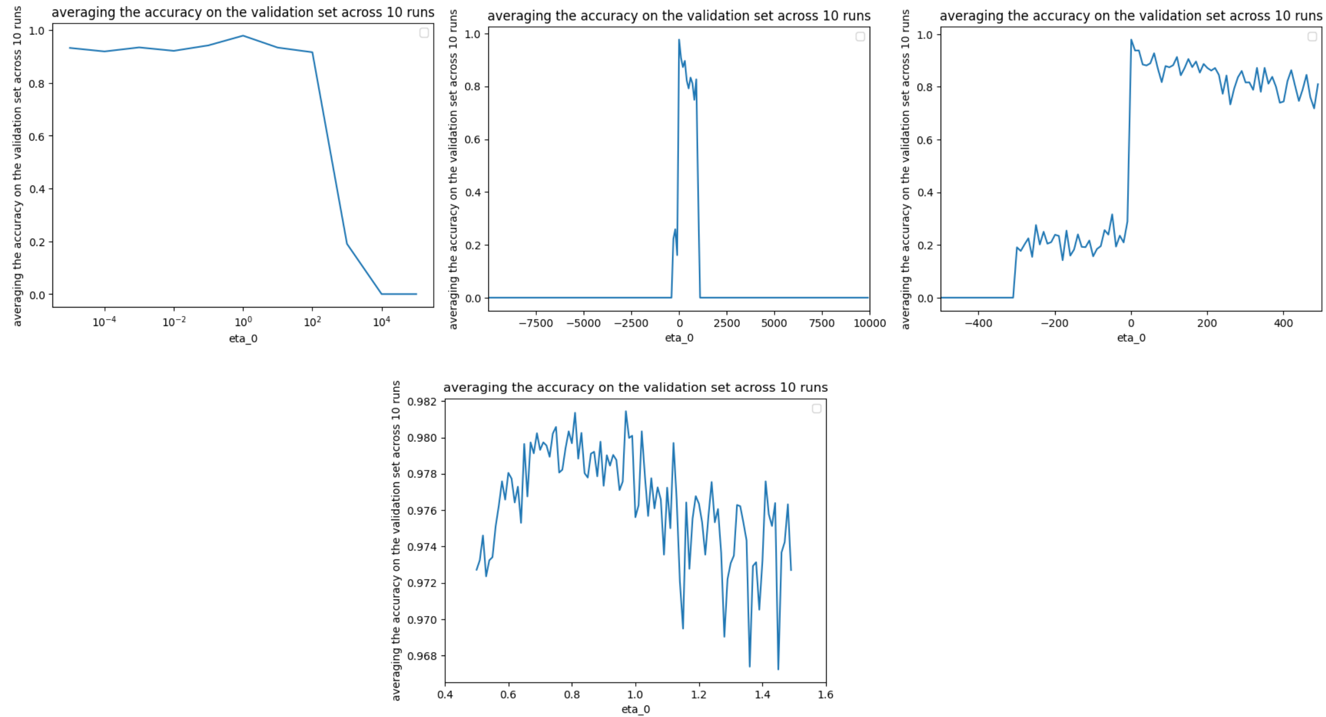

Figure 1: accuracy as function of η_0

5.2 (b)

כדי לבחור את C הטוב ביותר, ניתן לאלגוריתם ערכים שהחלו ב

$$\left(10^{-5}, 10^{-4}, \ldots, 10^{4}, 10^{5}\right)$$

C הטוב ביותר שהתקבל הוא $C = 10^{-4}$ עם accuracy$= 0.986$

נתמקד באזור בשבו התקבלה התוצאה הטובה ביותר, והאלגוריתם הורץ שנית (ברזולוציית קפיצות קטנה יותר)

$$(-9999.9999, -9899.9999, \ldots, 9800.0001, 9900.0001)$$

C הטוב ביותר שהתקבל הוא $C = 0.001$ עם accuracy$= 0.987$

נתמקד באזור בשבו התקבלה התוצאה הטובה ביותר, והאלגוריתם הורץ שלישית (ברזולוציית קפיצות קטנה יותר)

$$(-499.9999, -489, \ldots, 480.0001, 490.0001)$$

C הטוב ביותר שהתקבל הוא $C = 1$ עם accuracy$= 0.986$

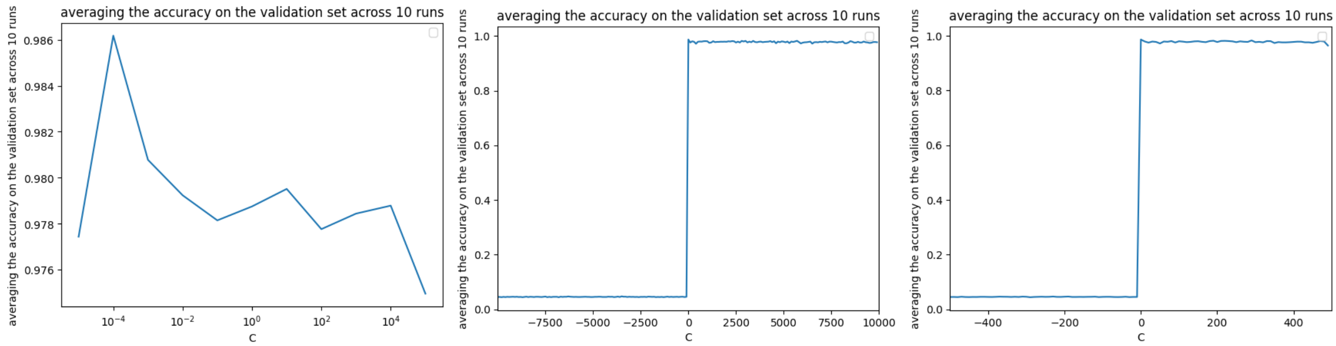

Figure 2: accuracy as function of C

5.3 (c)

ניתן לראות כי הפיקסלים הלבנים באמצע יכולים להעיד על אמצע הספרה 8 והפיקסלים בצבעי לבן - אפור בצדדים מעידים על הספרה 8 גם כן.

הפיקסלים השחורים מעידים על הספרה 0.

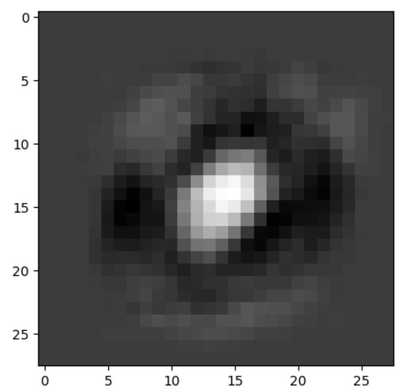

Figure 3: $\mathbf{w}$ as an image

5.4 (d)

על סט הבדיקה התקבל :

$$accuracy = 0.992835$$

6 SGD for log-loss.

6.1 (a)

כדי לבחור את η_0 הטוב ביותר, ניתן לאלגוריתם ערכים שהחלו ב

$$\left(10^{-5}, 10^{-4}, \ldots, 10^{4}, 10^{5}\right)$$

η_0 הטוב ביותר שהתקבל הוא $\eta_0 = 10^{-5}$ עם accuracy$= 0.955$

נתמקד באזור בשבו התקבלה התוצאה הטובה ביותר, והאלגוריתם הורץ שנית (ברזולוציית קפיצות קטנה יותר)

$$(-9999, -9899, \ldots, 9801, 9901)$$

η_0 הטוב ביותר שהתקבל הוא $\eta_0 = 10^{-5}$ עם accuracy$= 0.959$

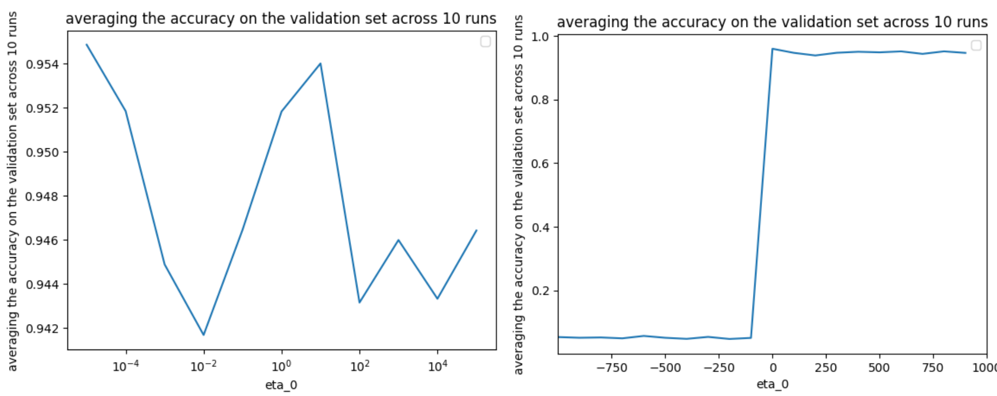

Figure 4: accuracy as function of η_0

6.2 (b)

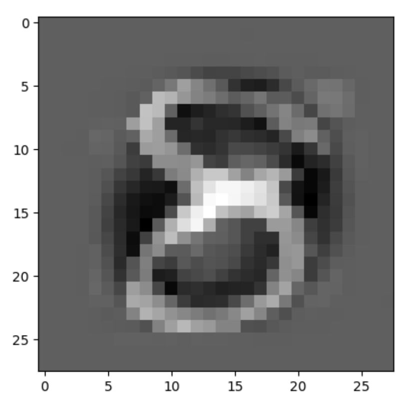

Figure 5: $\mathbf{w}$ as an image

על סט הבדיקה התקבל :

$$accuracy = 0.973$$

6.3 (c)

ניתן הסבר :

- ההגעה לערך המינימלי של פונקציית log-loss ע"י אלגוריתם SGD מתבצע מהר, ולכן $\mathbf{w}$ "נלכד" בערך המינימלי ולכן $\|\mathbf{w}_t\|_2$ אינו משתנה לאחר האיטרציות הראשונות יחסית ל $20,000$.
- בכל שלב η קטן, ולכן השינוי ב$\|\mathbf{w}_t\|_2$, שתלוי בו, יקטן ככל ש t גדל.

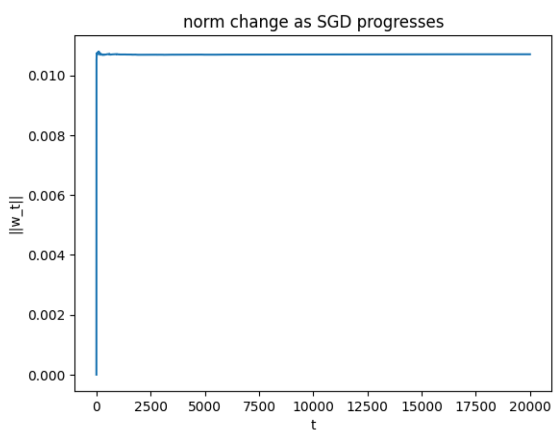

Figure 6: $\|\mathbf{w}_t\|_2$ as function of t