ML HW-3 208935916 :ת.ז.

1. (15 points) Step-size Perceptron.

פתרון: נניח כי w^* הוא מפריד לינארי של המידע ונניח $\|w^*\|=1$.

בנוסף, $\gamma > 0$ הוא השוליים של w^*, ולכל t מתקיים

$$y_t \cdot w^* \cdot x_t = \frac{|w^* \cdot x_t|}{\|w^*\|} = dist(x_t, hyperplane(w^*)) \geq \gamma$$

לאחר כל טעות בזמן t, מתקיים:

$$w_{t+1} \cdot w^* = (w_t + \eta_t y_t x_t) \cdot w^* = w_t \cdot w^* + \eta_t \cdot y_t \cdot x_t \cdot w^* \geq$$

$$w_t \cdot w^* + \eta_t \gamma = w_t \cdot w^* + \frac{1}{\sqrt{t}} \gamma$$

אז, לאחר m טעויות אצורות בהתחלה, נקבל:

$$w_{t+1} \cdot w^* \geq w_m \cdot w^* + \frac{1}{\sqrt{m}} \cdot \gamma \geq \dots \geq \gamma \sum_{i=1}^{m} \frac{1}{\sqrt{i}} \geq \gamma \int_1^{m+1} \frac{1}{\sqrt{x}} dx$$

$$= \gamma (2\sqrt{m+1} - 2) \underset{**}{\geq} \gamma \sqrt{em}$$

** - עבור m מספיק גדול -

$$\lim_{m \to \infty} \frac{\sqrt{em}}{2\sqrt{m+1} - 2} = \frac{\sqrt{e}}{2} < 1$$

בנוסף, לאחר כל טעות בזמן t נקבל:

$$\|w_{t+1}\|^2 \leq \|w_t\|^2 + \frac{1}{t} \leq \dots \leq \sum_{i=1}^{m} \frac{1}{i} \leq 1 + \int_1^m \frac{1}{x} dx = 1 + \ln(m) = \ln(e \cdot m)$$

אז אם נסכם נקבל:

$$\gamma \sqrt{em} \leq w_{t+1} \cdot w^* \leq \|w_{t+1}\| \cdot \|w^*\| = \|w_t\| \leq \sqrt{\ln(em)}$$

$$0 < \gamma \sqrt{em} \leq \sqrt{\ln(em)} \rightarrow \gamma^2 em \leq \ln(em) \rightarrow em \leq \frac{1}{\gamma^2} \ln(em) \Leftarrow$$

ולפי שימוש ברמז נקבל:

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

2. **(15 points) Convex functions.**

(a) Let $f : \mathbb{R}^n \to \mathbb{R}$ a convex function, $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. Show that, $g(\mathbf{x}) = f(A\mathbf{x}+b)$ is convex.

<u>פתרון</u>:

נזכיר תחילה כי $\mathbb{R}^n$ הוא קבוצה קמורה.

יהיו $x_1, x_2 \in \mathbb{R}^n$, ויהי $\alpha \in [0,1]$. אז מתקיים:

$$g(\alpha x_1 + (1-\alpha)x_2) = f(A(\alpha x_1 + (1-\alpha)x_2) + b) \underset{\text{קמורה } f}{\le}$$
$$\alpha f(Ax_1+b) + (1-\alpha) f(Ax_2+b) = \alpha g(x_1) + (1-\alpha) g(x_2)$$

כנדרש $\square$

(b) Consider m convex functions $f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})$, where $f_i : \mathbb{R}^d \to \mathbb{R}$. Now define a new function $g(\mathbf{x}) = \max_i f_i(\mathbf{x})$. Prove that $g(\mathbf{x})$ is a convex function. (Note that from (a) and (b) you can conclude that the hinge loss over linear classifiers is convex.)

<u>פתרון</u>: יהיו $x_1, x_2 \in \mathbb{R}^d$ ו-$\alpha \in [0,1]$. מתקיים:

$$g(\alpha x_1 + (1-\alpha)x_2) = \max_i f_i(\alpha x_1 + (1-\alpha)x_2) \underset{\text{קמירות לכל } f_i}{\le}$$

$$\max_i \alpha f_i(x_1) + (1-\alpha) f_i(x_2) \le \max_i \alpha f_i(x_1) + \max_i (1-\alpha) f_i(x_2)$$

$$= \alpha g(x_1) + (1-\alpha) g(x_2)$$ כנדרש $\square$

(c) Let $\ell_{log} : \mathbb{R} \to \mathbb{R}$ be the log loss, defined by

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

Show that ℓ_{log} is convex, and conclude that the function $f : \mathbb{R}^d \to \mathbb{R}$ defined by $f(\mathbf{w}) = \ell_{log}(y\mathbf{w} \cdot \mathbf{x})$ is convex with respect to $\mathbf{w}$.

פתרון:

- ℓ_{log} גזירה פעמיים
- $$\ell'_{log}(z) = \ln(2)\cdot\frac{-e^{-z}}{1+e^{-z}}$$

אז:
$$\ell''_{log} = \ln(2)\cdot\frac{e^{-z}(1+e^{-z}) - e^{-z}\cdot e^{-z}}{(1+e^{-z})^2} = \ln(2)\,\frac{e^{-z}\cdot(1+e^{-z}-e^{-z})}{(1+e^{-z})^2}$$

$$= \ln(2)\,\frac{e^{-z}}{(1+e^{-z})^2} \geq 0$$

נזכר במשפט הבא - עבור פונ' f גזירה פעמיים ($f:\mathbb{R}\to\mathbb{R}$) מתקיים - f קמורה $\iff f''(x)\geq 0$.

אז נסיק כי ℓ_{log} קמורה. כעת - עבור $x\in\mathbb{R}^d$ ו- $y\in\mathbb{R}$

נראה כי - $f_{x,y}:\mathbb{R}^d\to\mathbb{R}$ המקיימת $f(w)=\ell_{log}(y\cdot w\cdot x)$ קמורה

ביחס ל w. מתקיים - $yw\cdot x = w\cdot(yx) = Aw$ עבור $A=(yx)^T$.

מסעיף א' $\Leftarrow$ $\ell_{log}(Aw)=f(w)$ קמורה.

3. **(20 points) Ranking**.

(a) Prove that the hinge loss described above for the ranking objective is convex in **w**.

נביט בפונ' hinge loss.

$$\ell(h_w(\bar{x}), y) = \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \max\{0, 1 - \text{sgn}(y_j - y_r)\, w \cdot (x_j - x_r)\}$$

נראה כי ℓ קמורה ביחס ל-w.

יהיו $w_1, w_2 \in \mathbb{R}^d$ ויהי $\alpha \in [0,1]$. מתקיים:

$$\ell(h_{\alpha w_1 + (1-\alpha) w_2}(\bar{x}), y) = \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \max\{0, 1 - \text{sgn}(y_j - y_r)(\alpha w_1 + (1-\alpha) w_2)(x_j - x_r)\}$$

$$= \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \max\{0, \alpha(1 - \text{sgn}(y_j - y_r) w_1 (x_j - x_r)) + (1-\alpha)(1 - \text{sgn}(y_j - y_r) \cdot w_2 (x_j - x_r))\}$$

$$\le \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \max\{0, \alpha(1 - \text{sgn}(y_j - y_r) w_1 \cdot (x_j - x_r))\} + \max\{0, (1-\alpha)(1 - \text{sgn}(y_j - y_r) w_2 (x_j - x_r))\}$$

$$= \alpha \ell(h_{w_1}(\bar{x}), y) + (1-\alpha) \ell(h_{w_2}(\bar{x}), y)$$

- כאשר $\alpha, 1-\alpha$ אי שליליים, ומסכום מקסימום האי-שוויון בפונ' המקסימום.

אזי, ℓ קמורה ביחס ל-w.

(b) Prove that the hinge loss upper-bounds the Kendall-Tau loss, i.e. that $\Delta(h_{\mathbf{w}}(\bar{\mathbf{x}}), \mathbf{y}) \leq \ell(h_{\mathbf{w}}(\bar{\mathbf{x}}), \mathbf{y})$ for all $\mathbf{w} \in \mathbb{R}^d, \bar{\mathbf{x}} \in \mathcal{X}^k, \mathbf{y} \in \mathbb{R}^k$.

<u>פתרון</u>: נזכיר בפונ' Kendall-Tau:

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

יהי $w \in \mathbb{R}^d$, $\bar{x} \in \mathcal{X}^k$, $y \in \mathbb{R}^k$. מתקיים:

$$\Delta(h_w(\bar{x}), y) = \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \mathbb{1}\{\operatorname{sgn}(h_w(\bar{x})_j - h_w(\bar{x})_r) \neq \operatorname{sgn}(y_j - y_r)\}$$

$$= \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \mathbb{1}\{\operatorname{sgn}(w \cdot x_j - w \cdot x_r) \neq \operatorname{sgn}(y_j - y_r)\}$$

$$= \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \mathbb{1}\{\operatorname{sgn}(w(x_j - x_r)) \neq \operatorname{sgn}(y_j - y_r)\}$$

$$\overset{\bullet}{\leq} \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \mathbb{1}\{\operatorname{sgn}(w(x_j - x_r)) \operatorname{sgn}(y_j - y_r) \in \{0, -1\}\}$$

$$\overset{\bullet\bullet}{\leq} \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \max\{0, 1 - \operatorname{sgn}(y_j - y_r) w \cdot (x_j - x_r)\} = \ell(h_w(\bar{x}), y)$$

- הבטנו גם במקרים בהם אני הנגזרים שווים ל-0.
- אם $\mathrm{sgn}(w(x_j-x_r))\,\mathrm{sgn}(y_j-y_r)=-1 \Rightarrow$

$$-\mathrm{sgn}(y_j-y_r)\,w\cdot(x_j-x_r)\ge 0$$

$$\Rightarrow 1 \le \max\{0,\,1-\mathrm{sgn}(y_j-y_r)\,w\cdot(x_j-x_r)\}.$$

במקרים התמרים – לא הפחתנו את הסכום ע"י לקיחת 0 בפונ המקס'.

$\square$

4. **(15 points) Gradient Descent on Smooth Functions.** We say that a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is β-smooth if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

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

In words, β-smoothness of a function f means that at every point $\mathbf{x}$, f is upper bounded by a qaudratic function which coincides with f at $\mathbf{x}$.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a β-smooth and non-negative function (i.e., $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$). Consider the (non-stochastic) gradient descent algorithm applied on f with constant step size $\eta > 0$:

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

Assume that gradient descent is initialized at some point $\mathbf{x}_0$. Show that if $\eta < \frac{2}{\beta}$ then

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

(Hint: Use the smoothness definition with points $\mathbf{x}_{t+1}$ and $\mathbf{x}_t$ to show that $\sum_{t=0}^{\infty} \|\nabla f(\mathbf{x}_t)\|^2 < \infty$ and recall that for a sequence $a_n \geq 0$, $\sum_{n=1}^{\infty} a_n < \infty$ implies $\lim_{n\to\infty} a_n = 0$. Note that f is not assumed to be convex!)

פתרון:

כיון f β-Smooth. $\Leftarrow$ $\forall x,y \in \mathbb{R}^n$

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{\beta}{2}\|x-y\|^2$$

נציב $x = x_t$, $y = x_{t+1}$ ונקבל ש- $x_t - x_{t+1} = \eta \nabla f(x_t)$

אזי לכל $t \geq 0$:

$$0 \leq \eta \|f(x_t)\|^2 = \nabla f(x_t) \cdot (\eta \nabla f(x_t)) = \nabla f(x_t)(x_t - x_{t+1})$$

$$\leq f(x_t) - f(x_{t+1}) + \frac{\beta}{2}\|x_t - x_{t+1}\|^2$$

נסכום:

$$\sum_{t=0}^{T} \|f(x_t)\|^2 \leq \frac{1}{\eta} \sum_{t=0}^{T} f(x_t) - f(x_{t+1}) + \frac{\beta}{2}\|x_t - x_{t+1}\|^2$$

$$= \frac{1}{\eta}(f(x_0) - f(x_{T+1})) + \frac{\beta}{2\eta}\sum_{t=0}^{\infty}\|\eta \nabla f(x_t)\|^2$$

$$\Leftarrow \quad \left(1 - \frac{\beta\eta}{2}\right)\sum_{t=0}^{T}\|\nabla f(x_t)\|^2 \leq \frac{1}{\eta}(f(x_0) - f(x_{T+1}))$$

$\Leftarrow$ אזי כ $\eta < \frac{2}{\beta}$ ו- $0 < 1 - \frac{\beta\eta}{2}$, ו f אי שלילית $\Leftarrow f(x_{T+1}) \geq 0$

$$\sum_{t=0}^{T}\|\nabla f(x_t)\|^2 \leq \frac{1}{1-\frac{\beta\eta}{2}} \cdot \frac{1}{\eta}(f(x_0) - f(x_{T+1})) \leq \frac{1}{1-\frac{\beta\eta}{2}} \cdot \frac{1}{\eta} f(x_0) \Leftarrow$$

ועבור ערך $T\to\infty$ נקבל: $\sum_{t=0}^{\infty}\|\nabla f(x_t)\|^2 \le \frac{1}{1-\frac{\beta\eta}{2}}\cdot\frac{1}{\eta}f(x_0)<\infty$

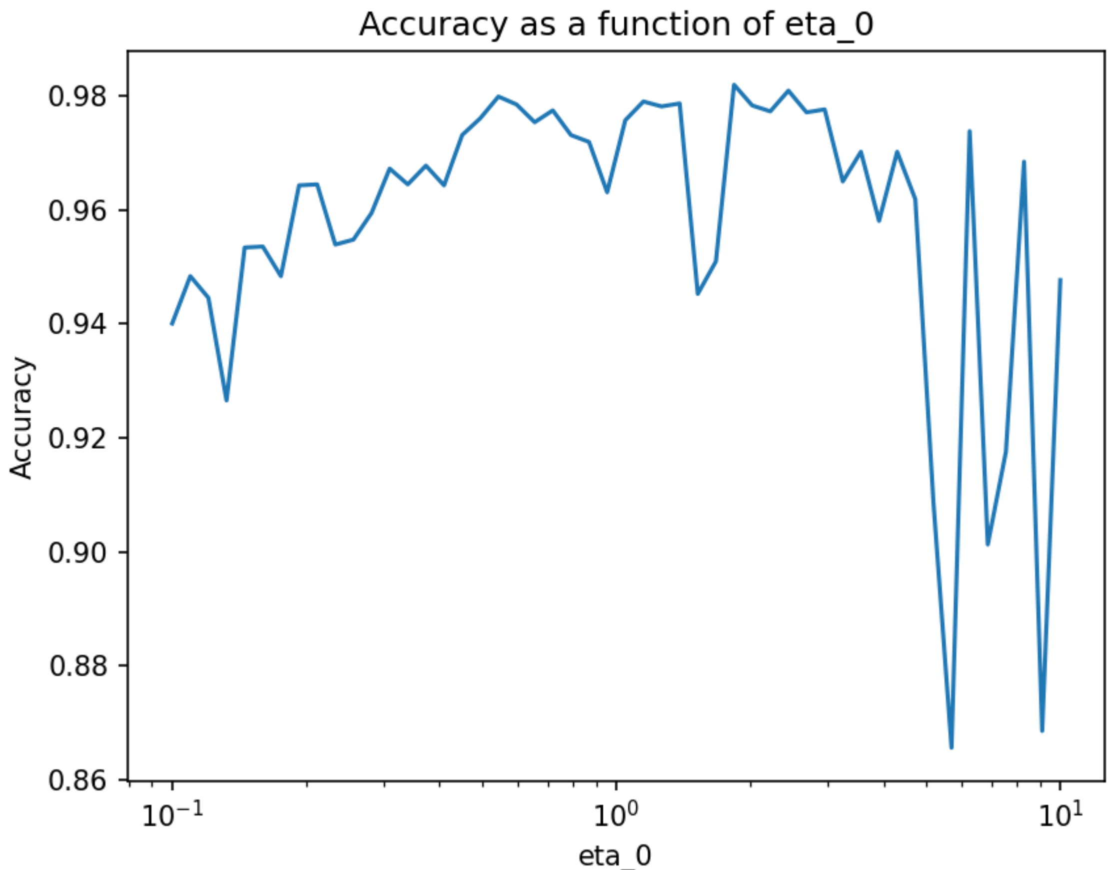
$\Leftarrow$ $\lim_{t\to\infty}\|\nabla f(x_t)\|^2=0$ וגם $\lim_{t\to\infty}\|\nabla f(x_t)\|=0$

□ כנדרש.

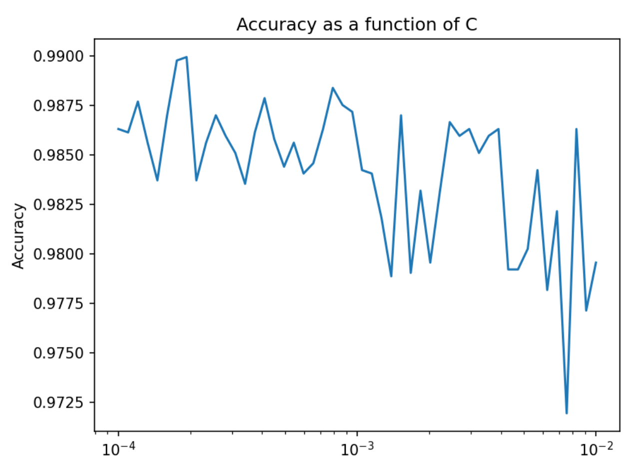
חלק תכנותי:

שאלה 1:

סעיף א:

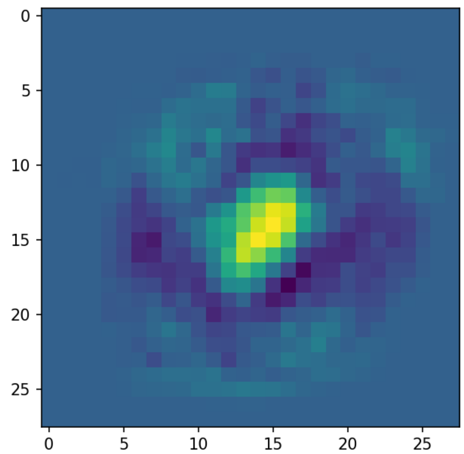


סעיף ב:



סעיף ג:

הסבר אינטואיטיבי – נשייך את הצבעים הבהירים לספרה 8, ואת הצבעים הכהים לספרה 0. הסימון הצהוב מבטא **דגש במרכז** (כלומר את התוכן שמופיע במרכז הספרה 8 לעומת הספרה 0). מנגד הסימון הכהה מבטא **דגש בצדדים** בצורה מעגלית, כלומר את התוכן של המעטפת של 0 לעומת המעטפת של הספרה 8.

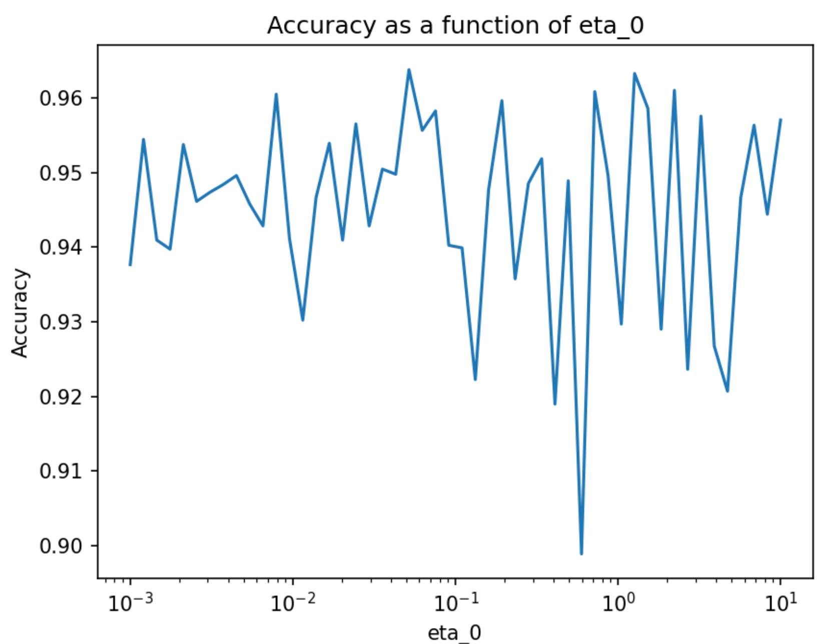


סעיף ד:

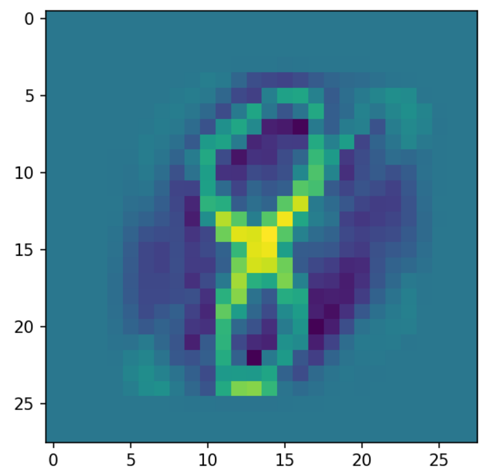
The accuracy of the best classifier on the test set is: 0.9923234390992836

שאלה 2:

סעיף א':

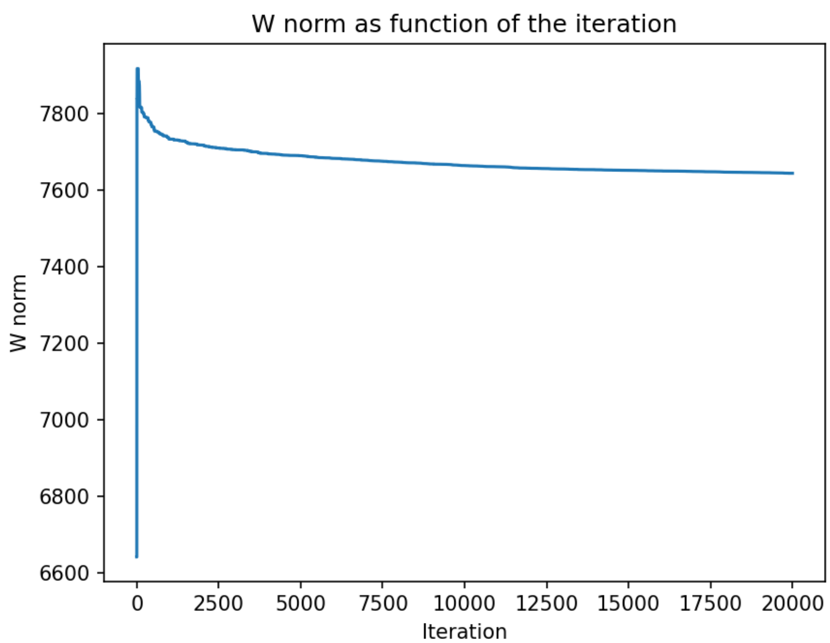


סעיף ב׳:



The accuracy of the best classifier on the test set is: 0.981064483111566

סעיף ג:



הסבר:

ככל ש-SGD מתקדם בתהליך – הנורמה קטנה (אמנם נשארת די יציבה). ההסבר לכך הוא שהאלגוריתם מגיע במהירות גבוהה יחסית למסווג טוב, ולאחר הגעה זאת ממשיך ומבצע רק שינויים קטנים (טיובים).