תרגיל בית מספר 3

מבוא ללמידה חישובית

בנימין אסטרליס ותומר סמרה

315636761 321164923
tomersamara@mail.tau.ac.il esterlis@mail.tau.ac.il

הכתובת של קבצי הקוד+ התמונות בצבע:specific/a/home/cc/students/math/esterlis/Downloads/ML-HW3

1 חלק תיאורטי

1. ראשית, ניזכר בעובדה המכיתה כי ה $VCdim$ של נוירון בודד הינו d+1 לפי $Radons' \ theorem$.
נרצה להראות שה $VCdim$ של רשת נוירונים היא לפחות ה $VCdim$ של נוירון בודד:
לשם כך נגדיר את רשת הנוירונים הבאה,אשר תתנהג בדיוק כמו נוירון בודד:
את הערך של הנוירוון הראשון בשכבה הראשונה נעביר האלה לכל שאר הנוירונים ע"י מכפלה של השכבה הראשונה במטריצת היחידה, והפעלת $sign$ קורדינאטה-קורדינאטה.
נעשה זאת על כל השכבות מהשנייה ועד השכבה ה $L-1$.
לבסוף בשכבה ה L נעביר כפלט את הסימן של הנוירון הראשון בשכבה ה $L-1$,
ולכן קיבלנו כי הנוריון הבודד בשכבה האחרונה מחזיר לנו את הסימן של הנוירון הראשון שבכבה הראשונה.

הראנו שהרשת שלנו מתנהגת בדיוק כמו ניירון בודד ולכן כל מה שניתן לסווג ע"י נוירון בודד ניתן לסווג ע"י הרשת שבנינו,
ולכן ה $VCdim$ של רשתות נוירונים $\leq$ מה $VCdim$ של נוירון בודד.

2. (א) מכוון שההפונקציית המטרה היא פונקציה מ $\mathcal{R}^d \to \mathcal{R}^d$ אזי מרחב ההשערות שלנו הוא כל הפונקציות מהסוג הזה,
כעת לפי הגדרת $\mathcal{H}$ אנו מקבלים פונקציות מ $\mathcal{R}^d \to \mathcal{R}$ ולכן בעצם אנו צריכים לשרשר את $\mathcal{H}$, d פעמים ולכן נקבל כי מחלקת ההיפותיזות שלנו היא: $\mathcal{H}^d$

נחסום כעת את קצב הגדילה, לפי $Lemma \ 7.2$ והחישוב שעשינו בתרגול מתקבל כי קצב הגדילה של נוירון בודד הוא: $\leq (\frac{em}{d+1})^{d+1}$
ולכן לפי $Lemma \ 7.3$ מתקבל כי החסם הוא:

$$\pi_{\mathcal{H}^d}(m) \leq (\pi_{\mathcal{H}}(m))^d \leq ((\frac{em}{d+1})^{d+1})^d = (\frac{em}{d+1})^{d\cdot(d+1)}$$

(ב) מכוון ש $\mathcal{C}$ הינה מחלקה של כל רשתות הנוירונים עם L שכבות ו d נוירונים, ובסעיף הקודם הראנו כי כל שכבה הינה מהצורה הבאה: $\mathcal{H}^d$
ולכן מתקבל כי המחלקה הנ"ל היא: $\mathcal{H}^{\mathcal{H}^{d^{\cdot^{\cdot^{\mathcal{H}^d}}}}}$, כאשר בחזקה יש $\mathcal{L}-1$ פעמים $\mathcal{H}^d$.
נחסום כעת את קצב הגדילה,לפי הנימוקים של התרגיל הקודם מתקבל כי קצב הגדילה של נוירון בודד הוא: $\leq (\frac{em}{d+1})^{d+1}$
ולכן לפי $Lemma \ 7.4$ מתקבל כי החסם הוא:

$$\pi_{\mathcal{C}}(m) \leq \pi_{layer1}(m) \cdot ... \cdot \pi_{layer\mathcal{L}-1}(m) \cdot \pi_{layer\mathcal{L}}(m) \leq (\frac{em}{d+1})^{d(d+1)(\mathcal{L}-1)+(d+1)}$$

(ג) ניתן לראות כי יש לנו $\mathcal{L}-1$שכבות בעלות d נוירונים בכל שכבה, בנוסף לכך יש לנו בשכבה האחרונה רק נוירון אחד, ובכל נוירון יש לנו $d+1$ולכן בסה"כ נקבל כי מספר הפרמטרים הדרוש הוא:

$$N = (d+1)\cdot((\mathcal{L}-1)\cdot d+1) = d\cdot(d+1)\cdot(\mathcal{L}-1)+(d+1)$$

(ד) בונוס, מחוסר זמן לצערי לא היה זמן לעשות

(ה) בעקבות ההוכחת שהראנו עד הלום מתקיים:

$$2^{VCdim(\mathcal{L})} = \pi_c(m) \leq \left(\frac{em}{d+1}\right)^{d\cdot(d+1)\cdot(\mathcal{L}+1)+(d+1)} = \left(\frac{em}{d+1}\right)^N \leq (em)^N$$

וכאשר נציב $VCdim(\mathcal{C}) = m$, ובעקבות מה שהראתי החלק הקודם של שאלה וסעיף d מתקבל כי $m \leq 2Nlog_2(eN)$ ולכן בסה"כ $VCdim(\mathcal{C}) \leq 2Nlog_2(eN)$.

3. (א) האלגוריתם:
האלגוריתם SGD החדש פועל באותו האופן של האלגוריתם הרגיל, רק בשלב של בחירת הנקודה נבחר את הנקודה באופן הבא:

$$y_{t+1} = x_t - \nabla f(x_t)$$
$$x_{t+1} = \Pi_{\mathcal{K}}(y_{t+1})$$

ואת $\Pi_{\mathcal{K}}(x)$ באופן הבא:

$$\Pi_{\mathcal{K}}(x)=\begin{cases}x & x\in\mathcal{K}\\ \frac{x}{\|x\|}\cdot R & otherwise\end{cases}$$

שזה בעצם לבחור את הנקודה הכי קרוב לנקודה x אשר בקבוצה הקמורה $\mathcal{K}$,המעגל ברדיוס R.

(ב) בונוס_בני

(ג) ראשית ניתן לראות כי בהוכחה שעשינו בכיתה עד שיווין (12) לא מתעסקים באלגוריתם SGD עצמו ולכן החלק הזה בהוכחה נשאר גם פה ומתקבל כי:

$$\mathbb{E}\left[(\bar{W})-f(w^{\star})\right]\le\frac{1}{T}\sum_{i}\mathbb{E}\left[V_t(W_t-w^{\star})\right]$$

מכיוון שאנו משתמשים ב SGD עם $projection$ מהדרך שבה אנו מקדמים את W_t מתקבל ש $W_{t+1}=\Pi_{\mathcal{K}}\left(W_t-\eta V_t\right)$'
ולכן מתקיים $\|W_{t+1}-w^{\star}\|=\|\Pi_{\mathcal{K}}\left(W_t-\eta V_t\right)-w^{\star}\|$
לפי סעיף ב' ומכיוון ש $\mathcal{K}$ הינה קמורה מתקיים

$$\begin{aligned}\|W_{t+1}-w^{\star}\|_2^2&\le\|W_t-\eta V_t-w^{\star}\|_2^2=\\&=\|W_t-w^{\star}\|_2^2+\eta^2\|V_t\|_2-2\eta\left(W_t-w^{\star}\right)V_t\end{aligned}$$

והעברת אגפים תיתן לנו :

$$\left(W_t-w^{\star}\right)V_t\le\frac{\|W_t-w^{\star}\|_2^2-\|W_{t+1}-w^{\star}\|_2^2}{2\eta}+\frac{\eta^2\|V_t\|_2}{2}$$

וכעת המשך ההוכחה הוא זהה לחלוטין להוכחה שראינו בכית החל מסעיף 1

Bonus a Deserve I
stabbing without week the through it Making For
fork a with someone

(ד) נראה כי הטענה אינה נכונה בעזרת דוגמא נגדית:
נבחר $\forall x,z\ K_2(x,z):=(x,z)$ ואת K_1 באופן שרירותי ולכן נקבל כי $\forall x,z\ K(x,z)=\frac{K_1(x,z)}{0}$ ומכוון שחלוקה ב0 אינה מוגדרת אזי גם K אינו מוגדר ולכן בפרט K אינו גרעיון מוגדר חיובית.

(ה) ידוע לנו ש $K=aK_1$ ולכן מכיוון ש K_1 מוגדרת חיובית לכל וקטור v מתקיים $v^T\cdot K_1\cdot v>0$
ולכן מתקיים $v^T\cdot K\cdot v=v^T\cdot aK_1\cdot v>0$ קיבלנו ש K מוגדרת חיובית.

4. (א)

(ב) ראשית נסמן:

$$j^{\star}=\underset{j\in[K]}{argmax}\left(w_j\cdot x_i-w_{y_i}\cdot x_i+\chi_{(y_i\neq j)}\right)$$

כעת נגזור את l:
על מנת לחשב הדרוש נפצל למקרים הבאים:

i. $\underline{j\neq j^{\star}\wedge j\neq y_i}$:
מכוון ש l אינה תלוייה כלל בj נקבל כי הנגזרת היא 0 במקרה זה.

ii. $:j \neq j^\star \wedge j = y_i$

מכוון ש l תלוי בj נקבל

$$l(w_1, ..., w_K, x_i, y_i) = w_{j^\star} \cdot x_i - w_j \cdot x_i + \chi_{j \neq j^\star}$$

ולכן גזירה לפי j נקבל כי הנגזרת היא: $-x_i$.

iii. $:j = j^\star \wedge j \neq y_i$

מכוון שl תלוי בj נקבל כי:

$$l(w_1, ..., w_K, x_i, y_i) = w_j \cdot x_i - w_{y_i} \cdot x_i + \chi_{(j \neq y_i)} = w_j \cdot x_i - w_{y_i} \cdot x_i + 1$$

ולכן גזירה לפי j נקבל כי הנגזרת היא: x_i.

iv. $:j = j^\star \wedge j = y_i$

מכוון שl תלוי בj נקבל כי:

$$l(w_1, ..., w_K, x_i, y_i) = w_j \cdot x_i - w_j \cdot x_i + \chi_{(j \neq y_i)} = w_j \cdot x_i - w_j \cdot x_i + 1$$

ולכן גזירה לפי j נקבל כי הנגזרת היא: $x_i - x_i = 0$.

ולכן בסה"כ:

$$\nabla_{w_j} l = \begin{cases} -x_i & j \neq j^\star \wedge j = y_i \\ x_i & j = j^\star \wedge j \neq y_i \\ 0 & otherwise \end{cases}$$

ולכן ניתן לחשב את הsub descent gradient באופן הבא:

$$\nabla_{w_j} f = w_j + \nabla_{w_j} l = \begin{cases} w_j - C' x_i & j \neq j^\star \wedge j = y_i \\ w_j + C' x_i & j = j^\star \wedge j \neq y_i \\ w_j & otherwise \end{cases}$$

כאשר $C' = \frac{C}{m}$.

וכעת נתאר האלגוריתם $Stochastic\, subgradient\, descent$ כאשר $C, T, \eta, \{x_i\}_{i=1}^m, \{y_i\}_{i=1}^m$ נתונים לנו:

i. נאתחל את $\left\{ \forall 1 \leq i \leq k \quad w_{i0} = 0 \right.$

ii. לכל $t \in [T]$ נבצע את הבאים:

א'. נבחר באופן אקראי עם התפלגות אחידה $i \in [m]$

ב'. נחשב $j^\star$ באופן הבא:'

$$\underset{j \in [k]}{argmax} (w_j \cdot x_i - w_{y_i} \cdot x_i + \chi_{(j \neq y_i)})$$

ג'. לכל $j \in [k]$ נבצע את הבאים:

ד'. $w_{j,t+1} = (1 - \eta) \cdot w_{j,t}$

ה'. אם $:j^\star \neq y_i$

ו'. $w_{j^\star,t+1} = w_{j^\star,t} - \eta \cdot C \cdot x_i$

ז'. $w_{j^\star,t+1} = w_{j^\star,t} + \eta \cdot C \cdot x_i$

iii. ונחזיר $(w_{(1,T)}, ..., w_{(k,T)})$

והסיווג של x ל$\{1, ..., k\}$ יתצבע ע" החישוב:

$$\underset{j \in [k]}{argmax} \langle w_j, x \rangle$$

5. על מנת שנשנה את האלגוריתם הנוכחי שיתמש ב$Kernel$ נשים לב כי בכל שלב באלגוריתם $stochastic\ subgradient\ descent$ ה$w_{j,t}$ הינו צריוף לינארי של הדגימות שהיו קודם לכן ולכן מקבל כי: $w_{jt} = \sum_{i=1}^{t} \beta_{ji} \cdot x_i$

כאשר $\beta_{ji} = (1-\eta)^{t-1} \cdot A_{ji}\eta \cdot C$ כאשר $\alpha_{ij} = \begin{cases} 1 & j^\star \neq y_i \wedge j = y_i \\ -1 & j^\star \neq y_i \wedge j = j^\star \\ 0 & otherwise \end{cases}$, ו$j^\star$ הינו התקבל בשלב הi.

מכוון שβ_{ji} משתנה במהלך הריצה ולפי מה שראינו בתרגיל הקודם עבור $Kernel\ Perceptron$ ולכן מתקיים כי:

$$\langle w_{jt}, x\rangle = \left\langle \sum_{i=1}^{t} \beta_{ji} \cdot x_i, x \right\rangle = \sum_{i=1}^{t} \beta_{ji} \langle x \cdot x\rangle$$

<u>והאלגוריתם SGD אשר משתמש ב $Kernel$:</u>

(א) נגדיר מטרציה חדשה $A_{k\times m}$ להיות מטרצית אפסים שתייצג את המקדמים α_{ji}

i. נחשב את $j^\star$ באופן הבא:

$$j^\star = \underset{j\in[k]}{argmax} \left(\sum_{i=1}^{t-1} \left[\beta_{ji} \cdot K(x_i, x_t) - \beta_{y_t i} \cdot K(x_i, x_t)\right] + \chi_{(j\neq y_t)} \right)$$

ii. כעת אם $j^\star \neq y_t$ נציב: (בתאים האחרים נשאיר 0)

א׳. $A_{j^\star t} = -1$

ב׳. $A_{y_t,t} = 1$

ג׳. כעת נעדכן את $\beta_{ji} = (1-\eta)^{t-1} \cdot A_{ji}\eta \cdot C \quad \forall \begin{cases} 1 \leq i \leq m \\ 1 \leq j \leq k \end{cases}$

iii. כעת נגדיר מטרציה חדשה $B_{k\times m}$ כאשר $B_{ji} = \beta_{ji} \quad \forall \begin{cases} 1 \leq i \leq m \\ 1 \leq j \leq k \end{cases}$ כאשר β_{ji} הינו לאחר העדכון.

לבסוף סיווג הנקודה החדשה x יתבצע ע"י החישוב: $\underset{j\in[k]}{argmax} \left(\sum_{i=1}^{m} \beta_{ji} \cdot K(x_i, x) \right)$

6. נוכיח את הטענה באינדוקציה על d

(א) **<u>צעד בסיס ($\mathbf{d=1}$)</u>**:
צריך כי $h : \{0,1\} \to \{0,1\}$ היה עץ החלטות מגובה 2 ולכן נגדיר אותו באופן הבא:
עץ החלטות שלנו יהיה בעל שורש יחיד מצורה: $[(x_1 = 0)?]$ ולכן זה עץ מגובה 2, והוא מקיים את התנאים הדרושים

(ב) **<u>צעד אינדוקציה ($\mathbf{d=1}$)</u>**:
נניח כי מתקיים עבור $d = n$ ונוכיח עבור $n+1$ ולכן נגדיר את עץ ההחלטות T באופן הבא:

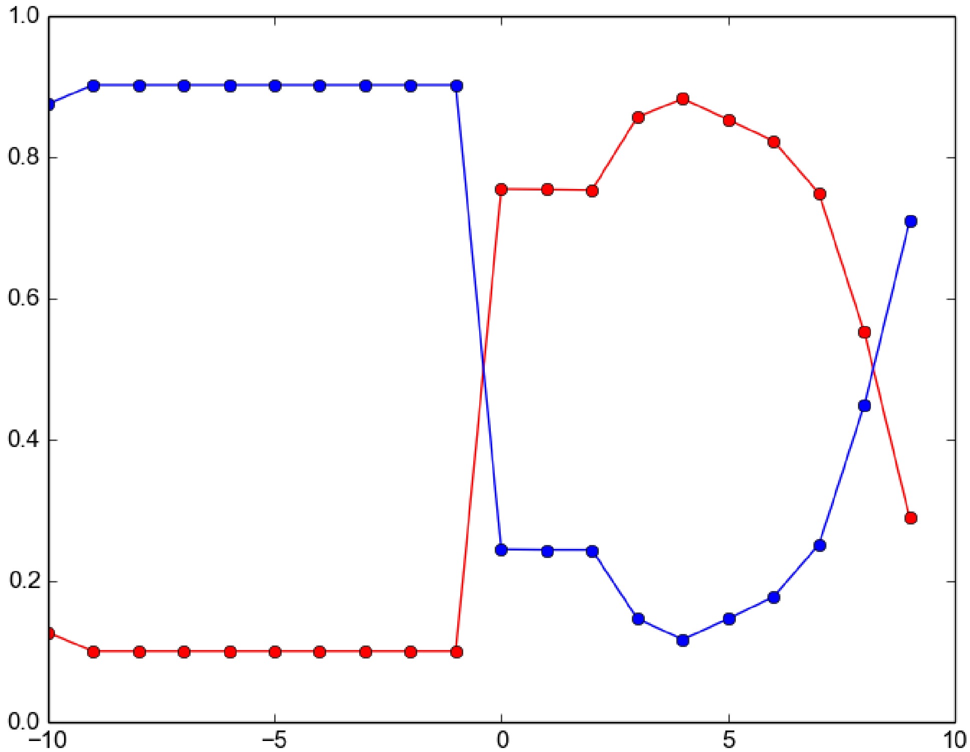
i. נבנה את העץ מגודל $n+1$ עבור הקורדינאטות $x_1, ..., x_n$, מכוון שיש לנו n קורדניאטות מתקבל כי תנאי ההנחה מתקיימים וניתן לבנות עץ החלטות T' מגודל $n+1$ כשבשכבה הi נשאל את השאלה $x_i = 0?$ ונתקדם לפיה.

ii. כעת נוסיף לעץ T' שכבה חדשה, לכל עלה נוסיף צומת חדשה על מנת לסווג גם לפי הקורדיניאטה האחרונה, x_{n+1}, כך שצומת תיהיה מהצורה: $(x_{n+1} = 0)?$

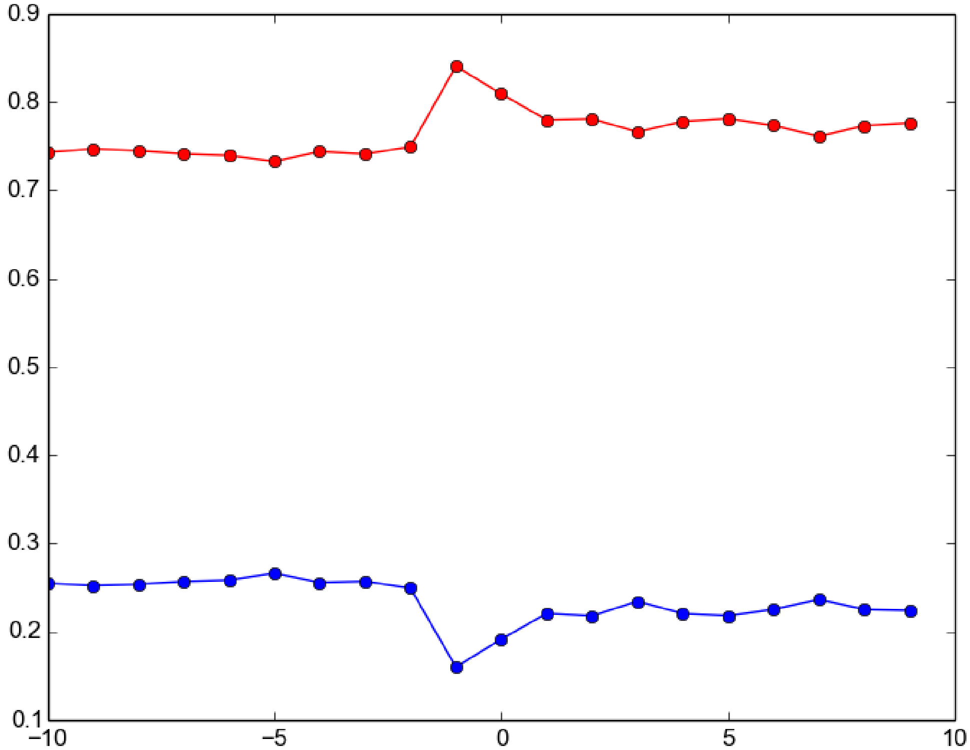
(ג) כעת קיבלנו את העץ T אשר הוא מסווג מהצורה: $h:\{0,1\}^{n+1}\rightarrow\{0,1\}$, בעקבות בניית T מתקבל כי העץ הוא מגול $n+1+1=n+2$ כדרוש.
כעת צריך לנתץ את הקבוצה $S=\{0,1\}^d$ ונקבל כי ה$VCdim\leq 2^d$.
ולכן עבור הקבוצה S הנ"ל נראה כי לכל $v\in\{0,1\}^{2^d}$,סיווג כלשהו, קיימת פונצקיה $h\in\mathcal{H}$ כך ש $h(s_i)_{s_i\in S}=v_i$, כאשר $\mathcal{H}$ הינה מחלקת עצי ההחלטה.
יהי $v\in\{0,1\}^{2^d}$ ונבנה לו את הפונציקה f,שתשרה את הסיווג, באופן הבא: $h(s_i)=(v)_i$ כאשר $h:\{0,1\}^{2^d}\rightarrow\{0,1\}$.
ניתן לבנות פונקציה כזו מכוון שבחלק הקודם של השאלה הוכחנו כי קיים עץ אשר משרה את הפוקנציה הזו, ולכן הראנו כי זה מתקיים לכל $v\in\{0,1\}^d$ ו $\Pi_{\mathcal{K}}(S)=2^{2^d}$ולכן מתקבל כי $VCdim\leq 2^d$.

2 חלק מעשי

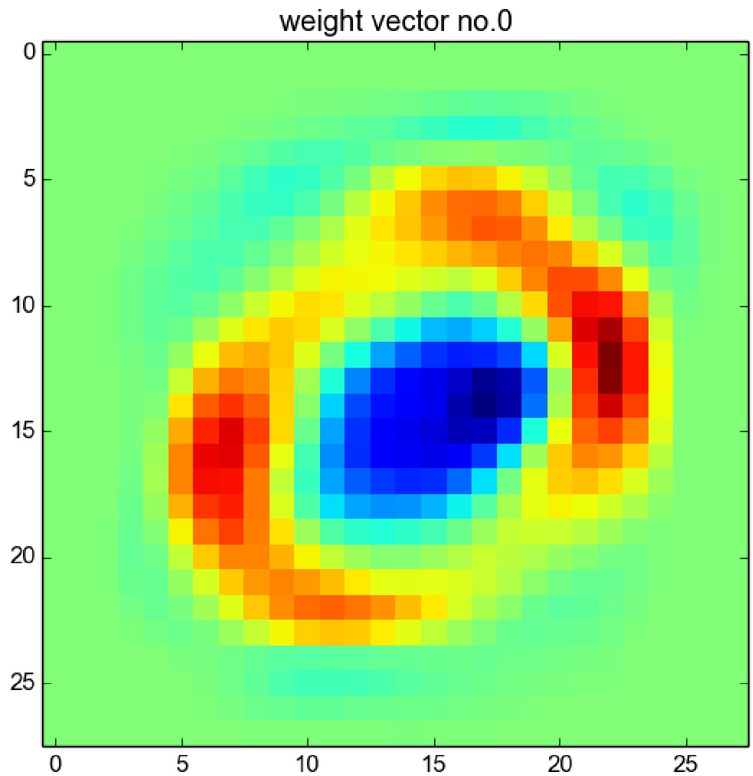
1. (א) ניתן לראות על פי התוצאות שη הטובה ביותר מתקבלת ב10^{-6}.
הC האופטימלי מתקבל ב0.1.
והדיוק הוא:0.9075



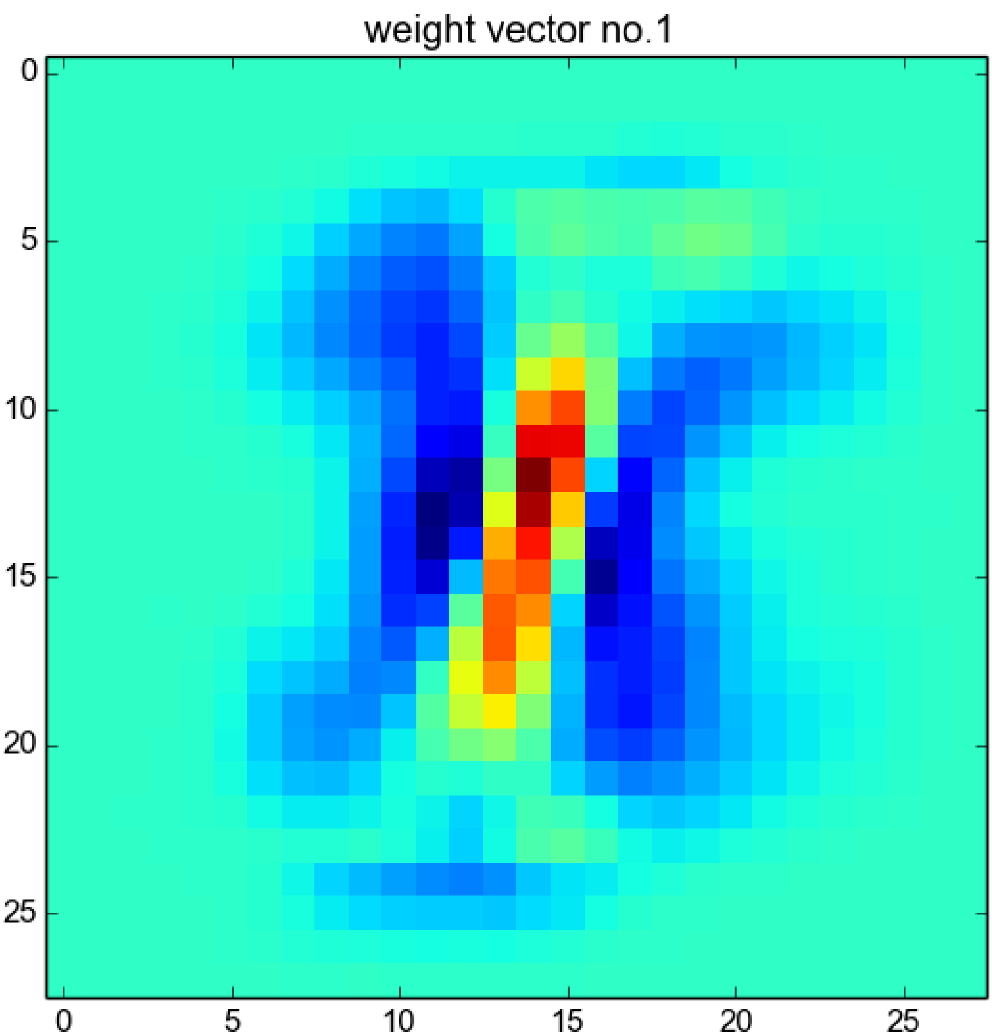
i.

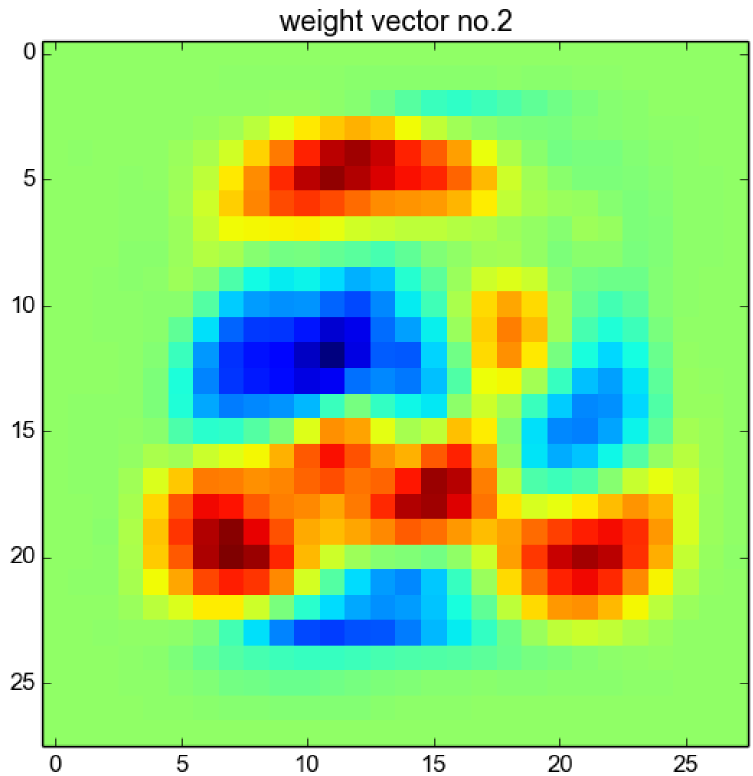


ii.

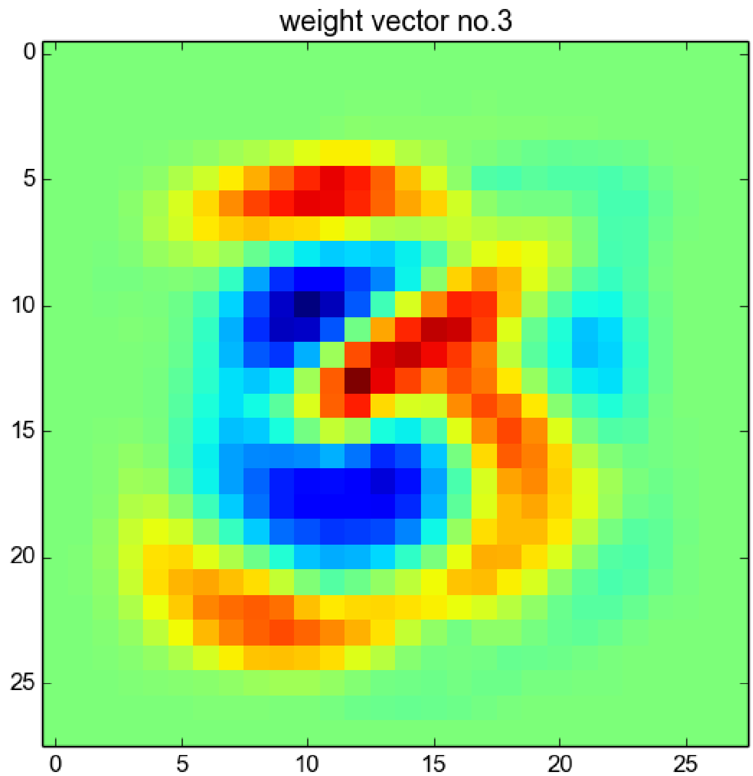


(ב) i.

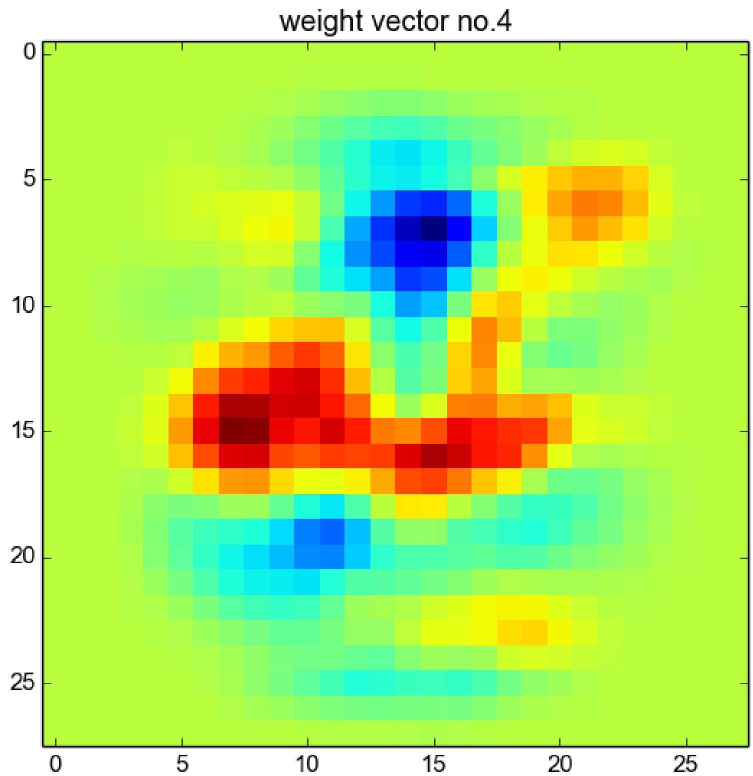

weight vector no.1
0
5
10
15
20
25
0
5
10
15
20
25



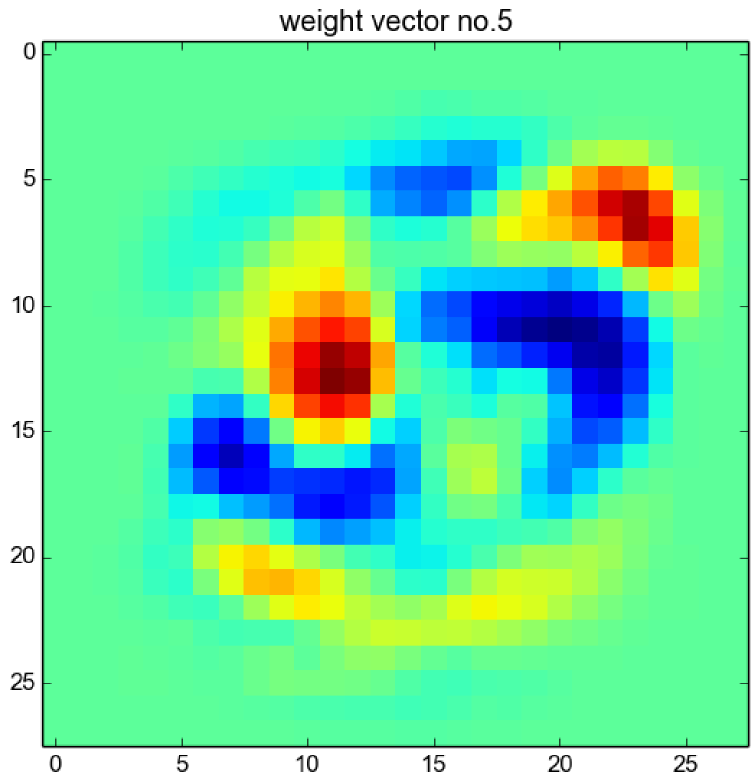
iii.



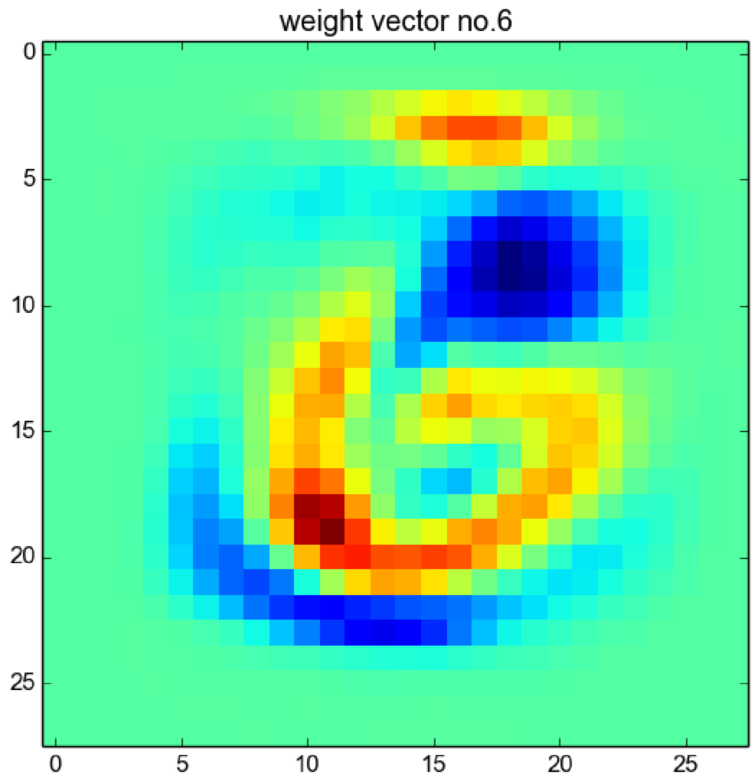
iv.



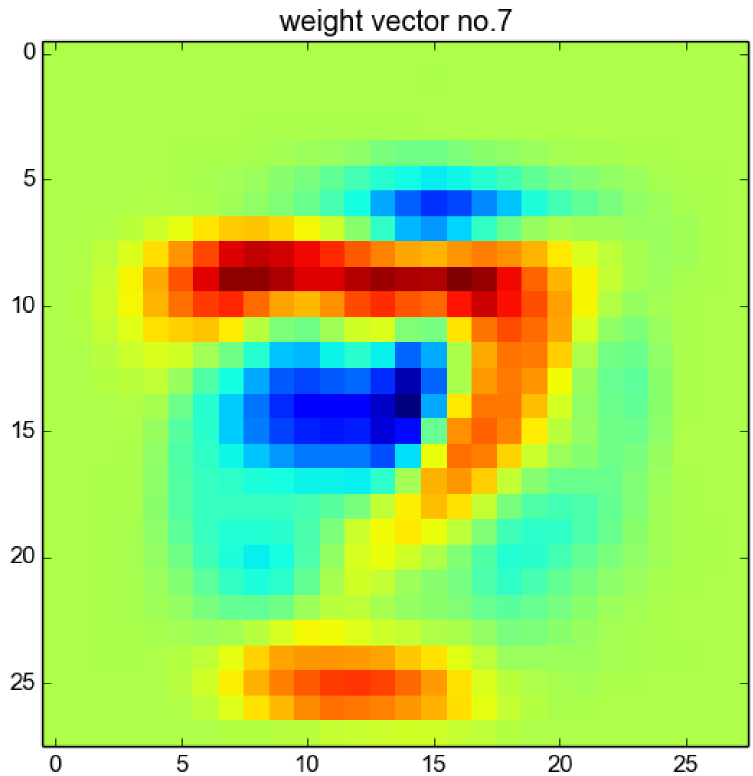
v.



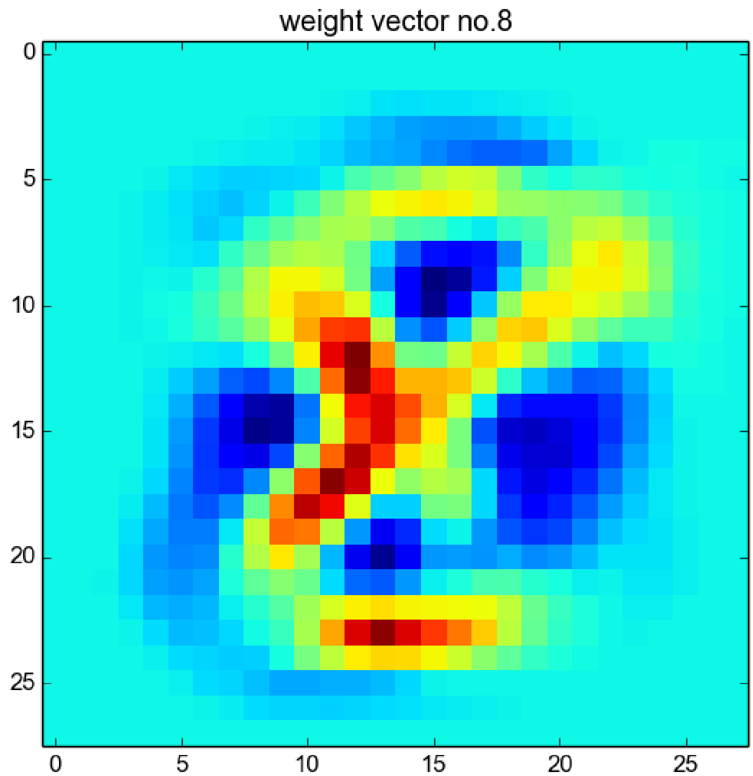
vi.

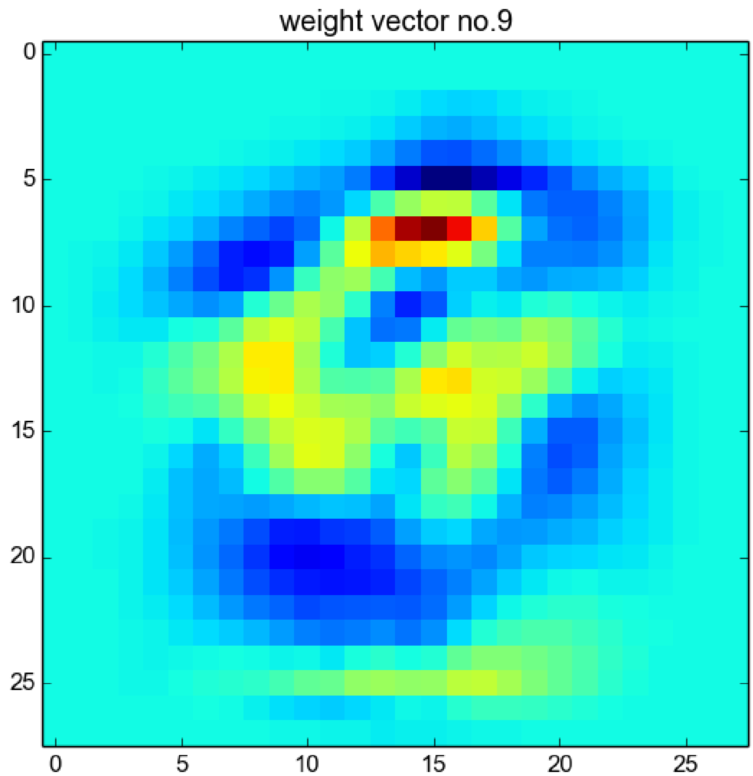


vii.



viii.


weight vector no.8
0
5
10
15
20
25
0
5
10
15
20
25



x.

(ג) הדיוק שהתקבל לנו על ה$test\ set$ הינו 0.9167

2. (א) הηהטוב ביותר התקבל ב10^{-3} עבור $T = 500$ ובהינתן η זה מתקבל כי $C = 10^9$.
הדיוק שהתקבל עבור η זו הינו: 0.8224 ועבור C זה הינו: 0.998 על הדגימה ועל ה$training\ set$הוא: 0.8407.

(ב) הדיוק על ה$test\ set$ הוא: 0.7945.