פתרון תרגיל מספר 3 ־ מערכות לומדות

שם: מיכאל גרינבאום, **ת.ז:** 211747639

13 במאי 2020

1. **צ"ל:** $h_{\mathcal{D}} = \text{argmax}_{y\in\{-1,1\}}\mathbb{P}(x \mid y)\cdot\mathbb{P}(y)$

הוכחה:

יהי x, תחילה נשים לב כי לכל $k\in\{-1,1\}$ מתקיים

$$\mathbb{P}(x \mid y=k)\cdot\mathbb{P}(y=k) = \frac{\mathbb{P}(y=k \mid x)\cdot\mathbb{P}(x)}{\mathbb{P}(y=k)}\cdot\mathbb{P}(y=k) = \mathbb{P}(y=k \mid x)\cdot\mathbb{P}(x)$$

ולכן

$$\begin{aligned}\mathbb{P}(x) &= \mathbb{P}(x \mid y=1)\cdot\mathbb{P}(y=1) + \mathbb{P}(x \mid y=-1)\cdot\mathbb{P}(y=-1)\\ &= \mathbb{P}(x)\cdot\left[\mathbb{P}(y=1 \mid x) + \mathbb{P}(y=-1 \mid x)\right]\\ &\Rightarrow \boxed{1 = \mathbb{P}(y=1 \mid x) + \mathbb{P}(y=-1 \mid x)}\end{aligned}$$

נחלק ל2 מקרים:

(א) $h_{\mathcal{D}}(x) = 1$, כלומר $\mathbb{P}(y=1 \mid x) \geq \frac{1}{2}$ ולכן

$$\frac{\mathbb{P}(x \mid y=1)\cdot\mathbb{P}(y=1)}{\mathbb{P}(x)} = \mathbb{P}(y=1 \mid x) \geq \frac{1}{2} \Rightarrow \boxed{\mathbb{P}(x \mid y=1)\cdot\mathbb{P}(y=1) \geq \frac{1}{2}\cdot\mathbb{P}(x)}$$

עתה מהמשוואה שקיבלנו לפני המקרים נקבל כי

$$\begin{aligned}&1 = \mathbb{P}(y=1 \mid x) + \mathbb{P}(y=-1 \mid x) \geq \frac{1}{2} + \mathbb{P}(y=-1 \mid x)\\ &\Rightarrow \frac{\mathbb{P}(x \mid y=-1)\cdot\mathbb{P}(y=-1)}{\mathbb{P}(x)} = \mathbb{P}(y=-1 \mid x) \leq \frac{1}{2}\\ &\Rightarrow \mathbb{P}(x \mid y=-1)\cdot\mathbb{P}(y=-1) \leq \frac{1}{2}\cdot\mathbb{P}(x) \leq \mathbb{P}(x \mid y=1)\cdot\mathbb{P}(y=1)\end{aligned}$$

ולכן נקבל כי $\text{argmax}_{y\in\{-1,1\}}\mathbb{P}(x \mid y)\cdot\mathbb{P}(y) = 1 = h_{\mathcal{D}}(x)$ מההגדרה, כנדרש

(ב) עתה אם $h_{\mathcal{D}}(x) = -1$, כלומר $\mathbb{P}(y=1 \mid x) < \frac{1}{2}$ מהמשוואה שקיבלנו לפני המקרים נקבל כי

$$1 = \mathbb{P}(y=1 \mid x) + \mathbb{P}(y=-1 \mid x) < \frac{1}{2} + \mathbb{P}(y=-1 \mid x) \Rightarrow \boxed{\frac{1}{2} < \mathbb{P}(y=-1 \mid x)}$$

לכן

$$\frac{\mathbb{P}(x \mid y=1)\cdot\mathbb{P}(y=1)}{\mathbb{P}(x)} = \mathbb{P}(y=1 \mid x) < \frac{1}{2} \Rightarrow \boxed{\mathbb{P}(x \mid y=1)\cdot\mathbb{P}(y=1) < \frac{1}{2}\cdot\mathbb{P}(x)}$$

וגם

$$\frac{\mathbb{P}(x\mid y=-1)\cdot\mathbb{P}(y=-1)}{\mathbb{P}(x)}=\mathbb{P}(y=-1\mid x)\geq\frac{1}{2}\Rightarrow\mathbb{P}(x\mid y=-1)\cdot\mathbb{P}(y=-1)\geq\frac{1}{2}\cdot\mathbb{P}(x)$$

ולכן

$$\mathbb{P}(x\mid y=1)\cdot\mathbb{P}(y=1)<\frac{1}{2}\cdot\mathbb{P}(x)\leq\mathbb{P}(x\mid y=-1)\cdot\mathbb{P}(y=-1)$$

ולכן נקבל כי $h_{\mathcal{D}}(x)=-1=\text{argmax}_{y\in\{-1,1\}}\mathbb{P}(x\mid y)\cdot\mathbb{P}(y)$ מההגדרה,

מ.ש.ל.☺

2. **צ"ל:** $h_{\mathcal{D}}=\text{argmax}_{y\in\{-1,1\}}\delta_y(x)$

הוכחה:

לפי הסעיף הקודם נקבל כי

$$\begin{aligned}
h_{\mathcal{D}}&=\text{argmax}_{y\in\{-1,1\}}\mathbb{P}(x\mid y)\cdot\mathbb{P}(y)\\
&\overset{x|y\sim\mathcal{N}(\mu_y,\Sigma)}{=}\text{argmax}_{y\in\{-1,1\}}\frac{1}{\sqrt{(2\pi)^d\cdot\det(\Sigma)}}\cdot e^{-\frac{1}{2}(x-\mu_y)^T\cdot\Sigma^{-1}\cdot(x-\mu_y)}\cdot\mathbb{P}(y)\\
&\overset{\frac{1}{\sqrt{(2\pi)^d\cdot\det(\Sigma)}}>0}{=}\text{argmax}_{y\in\{-1,1\}}e^{-\frac{1}{2}(x-\mu_y)^T\cdot\Sigma^{-1}\cdot(x-\mu_y)}\cdot\mathbb{P}(y)\\
&=\text{argmax}_{y\in\{-1,1\}}e^{-\frac{1}{2}(x-\mu_y)^T\cdot\Sigma^{-1}\cdot(x-\mu_y)}\cdot e^{\ln(\mathbb{P}(y))}=\text{argmax}_{y\in\{-1,1\}}e^{-\frac{1}{2}(x-\mu_y)^T\cdot\Sigma^{-1}\cdot(x-\mu_y)+\ln(\mathbb{P}(y))}\\
&=\text{argmax}_{y\in\{-1,1\}}\ln\left(e^{-\frac{1}{2}(x-\mu_y)^T\cdot\Sigma^{-1}\cdot(x-\mu_y)+\ln(\mathbb{P}(y))}\right)\\
&=\text{argmax}_{y\in\{-1,1\}}\left[-\frac{1}{2}(x-\mu_y)^T\cdot\Sigma^{-1}\cdot(x-\mu_y)+\ln(\mathbb{P}(y))\right]\\
&=\text{argmax}_{y\in\{-1,1\}}\left[-\frac{1}{2}\left[x^T\cdot\Sigma^{-1}\cdot x-\mu_y^T\cdot\Sigma^{-1}\cdot x-x^T\cdot\Sigma^{-1}\cdot\mu_y+\mu_y^T\cdot\Sigma^{-1}\cdot\mu_y\right]+\ln(\mathbb{P}(y))\right]\\
&\overset{x^T\cdot\Sigma^{-1}\cdot x\text{ constant}}{=}\text{argmax}_{y\in\{-1,1\}}\left[\frac{1}{2}\left[\mu_y^T\cdot\Sigma^{-1}\cdot x+x^T\cdot\Sigma^{-1}\cdot\mu_y-\mu_y^T\cdot\Sigma^{-1}\cdot\mu_y\right]+\ln(\mathbb{P}(y))\right]\\
&\overset{(\mu_y^T\cdot\Sigma^{-1}\cdot x)^T=x^T\cdot[\Sigma^{-1}]^T\cdot\mu_y=x^T\cdot\Sigma^{-1}\cdot\mu_y}{=}\text{argmax}_{y\in\{-1,1\}}\left[\frac{1}{2}\left[2\cdot x^T\cdot\Sigma^{-1}\cdot\mu_y-\mu_y^T\cdot\Sigma^{-1}\cdot\mu_y\right]+\ln(\mathbb{P}(y))\right]\\
&=\text{argmax}_{y\in\{-1,1\}}\left[x^T\cdot\Sigma^{-1}\cdot\mu_y-\frac{1}{2}\mu_y^T\cdot\Sigma^{-1}\cdot\mu_y+\ln(\mathbb{P}(y))\right]=\text{argmax}_{y\in\{-1,1\}}\delta_y(x)
\end{aligned}$$

כלומר קיבלנו כי $\boxed{\text{argmax}_{y\in\{-1,1\}}\delta_y(x)=h_{\mathcal{D}}}$, כנדרש

מ.ש.ל.☺

3. **צ"ל:** $\mu_y,\Sigma,\mathbb{P}(y)$

הוכחה:

תחילה בתור עזר נחשב את לכל $y\in\{-1,1\}$ נחשב את כמות הדגימות שהמחלקה היא y:

$$\boxed{s_y=\sum_{i=1}^{m}1_{y_i=y}}$$

עתה נשים לב כי $\mathbb{P}(y)$ זה מספר הדגימות שה $y_i=y$ מסך כל הדגימות, כלומר

$$\boxed{\mathbb{P}(y)=\frac{\sum_{s\in S}1_{s=y}}{|S|}=\frac{1}{|S|}\cdot s_y=\frac{s_y}{|S|}=\frac{s_y}{m}}$$

בהרצאה 1 ראינו שהדרך לחשב תוחלת היא $\frac{1}{m}\sum_{i=1}^{m} x_i$ ולכן, נעשה באופן דומה רק עבור $y_i = y$ (רק אלה שמקיימים את התנאי) ונקבל כי הנוסחה היא

$$\boxed{\mu_y = \frac{\sum_{i=1}^{m} x_i \cdot 1_{y_i=y}}{s_y}}$$

ועתה נחשב את מטריצת ה־ Σ לכל y, ולכן לפי הנוסחא שראינו בהרצאה 1, כשהיא נכתבת בכתיב מטריצי נקבל

$$\boxed{\Sigma_y = \frac{1}{s_y - 1} \cdot \sum_{i=1}^{m} [x_i - \mu_y] \cdot [x_i - \mu_y]^T \cdot 1_{y_i=y}}$$

עתה נוכל לומר כי אין תלות בין מחלקות שונות ונעשה ממוצע משוכלל על ה־ Covariance Matrix שחישבנו כדי לחשב את ה־ $Covariance$ הכללי (הרעיון לחישוב מטריצת ה־ cov בעזרת ממוצע משוקלל של cov המחלקות הוא רעיון די מוכר ולאחר חיפוש באינטרנט קוראים לזה $pool - covariance$, ובעזרת שיטה זאת, תוצאות LDA בשאלה $10 - 11$ יוצאות טובות יותר ב־ $5 - 6\%$ מאשר החישוב שראינו בהרצאה 1)

$$\begin{aligned}\Sigma &= \frac{(s_{-1} - 1) \cdot \Sigma_{-1} + (s_1 - 1) \cdot \Sigma_1}{(s_1 - 1) + (s_{-1} - 1)} = \frac{(s_{-1} - 1) \cdot \Sigma_{-1} + (s_1 - 1) \cdot \Sigma_1}{s_1 + s_{-1} - 2} \\ &= \frac{(s_{-1} - 1) \cdot \Sigma_{-1} + (s_1 - 1) \cdot \Sigma_1}{m - 2} = \frac{1}{m-2} \cdot \left[\sum_{y \in \{-1,1\}} \cdot \left[\sum_{i=1}^{m} [x_i - \mu_y] \cdot [x_i - \mu_y]^T \cdot 1_{y_i=y}\right]\right]\end{aligned}$$

כלומר קיבלנו נוסחא לחישוב $\boxed{\Sigma = \frac{1}{m-2} \cdot \left[\sum_{y \in \{-1,1\}} \cdot \left[\sum_{i=1}^{m} [x_i - \mu_y] \cdot [x_i - \mu_y]^T \cdot 1_{y_i=y}\right]\right]}$, אפשר לפשט ולהגיע לביטוי $\boxed{\Sigma = \frac{1}{m-2} \cdot \left[\sum_{i=1}^{m} [x_i - \mu_{y_i}] \cdot [x_i - \mu_{y_i}]^T\right]}$,

נשים לב שנוסחא זאת מאוד דומה לנוסחא המקורית חוץ מהעובדה שמנרמלים ב־ $m - k$ כאשר יש k מחלקות ומכל דגימה מחסרים את תוחלת המחלקה המתאימה לה, וזאת הכללה ל־ $Covariance$ שראינו בהרצאה הראשונה.

מ.ש.ל.☺

4. **צ"ל:** מה לקטלג כ־ $positive$ ומה לקתלג כ־ $negative$

הוכחה:

לדעתי עדיף נגדיר $positive$ כספאם ו־ $negative$ כלא ספאם. הסיבה לכך היא לעודד שלא תקרה השגיאה $type - I$ error שבה קיטלגנו משהו כספאם כשהוא לא ספאם. הסיבה שזה בעייתי, כי במקרה של שגיאה זאת, המשתמשים צריכים ללכת לחפש את המייל בספאם שהוא בדיוק מה שמערכת הספאם באה למנוע, ממיילים מפריעים להופיע במיילים של המשתמש.

מ.ש.ל.☺

5. **צ"ל:** לכתוב את $hard - SVM$ כבעיית QP קנונית

הוכחה:

תחילה נגדיר $\mathbb{R}^{d+1} \ni \begin{bmatrix} w \\ b \end{bmatrix} = v$ כאשר $w \in \mathbb{R}^d, b \in \mathbb{R}$.

עתה נגדיר $\boxed{Q = \begin{bmatrix} 2 \cdot I_d & 0_{d\times1} \\ 0 & 0 \end{bmatrix}}$, כלומר מטריצה אלכסונית שיש 2 באלכסון, חוץ מהאיבר האחרון שהוא 0, נשים לב כי

$$\frac{1}{2} \cdot v^T \cdot Q \cdot v = \frac{1}{2} \cdot \begin{bmatrix} w \\ b \end{bmatrix}^T \cdot \begin{bmatrix} 2 \cdot I_d & 0_{d\times1} \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} w \\ b \end{bmatrix} = \frac{1}{2} \begin{bmatrix} w & b \end{bmatrix} \cdot \begin{bmatrix} 2 \cdot w \\ 0 \cdot b \end{bmatrix} = \frac{1}{2} \cdot \left(2 \cdot \|w\|^2 + 0 \cdot \|b\|^2\right) = \|w\|^2$$

עתה נגדיר $\boxed{a=(0,\ldots,0)\in\mathbb{R}^{d+1}}$ ונקבל כי $a^T\cdot v=0$ מההגדרה ולכן

$$\frac{1}{2}\cdot v^T\cdot Q\cdot v+a^T\cdot v=\|w\|^2+0=\|w\|^2$$
$$\Rightarrow\boxed{\text{argmin}_{v\in\mathbb{R}^{d+1}}\left[\frac{1}{2}\cdot v^T\cdot Q\cdot v+a^T\cdot v\right]=\text{argmin}_{w\in\mathbb{R}^d,b\in\mathbb{R}}\|w\|^2}$$

עתה נשאר להגדיר את ההגבלות, נגדיר $\boxed{d=(-1,\ldots,-1)\in\mathbb{R}^m}$,
ונגדיר $a_i=[y_i\cdot x_i,y_i]\in\mathbb{R}^{d+1}$, ונגדיר מטריצה $A\in\mathbb{R}^{m\times(d+1)}$ באופן הבא: $\boxed{A_{i,j}=-a_{i,j}}$, נשים לב כי

$$[A\cdot v]_j=\langle -a_j,v\rangle=\left\langle-\begin{bmatrix}y_j\cdot x_j\\ y_j\end{bmatrix},\begin{bmatrix}w\\ b\end{bmatrix}\right\rangle=-y_j\cdot\left\langle\begin{bmatrix}x_j\\ 1\end{bmatrix},\begin{bmatrix}w\\ b\end{bmatrix}\right\rangle=-y_j\cdot(\langle x_j,w\rangle+b)$$

ולכן נשים לב כי האילוץ ה־ j בהצגה הקנונית היא

$$[A\cdot v]_j\le[d]_j\Rightarrow-[A\cdot v]_j\ge-[d]_j$$
$$y_j\cdot(\langle x_j,w\rangle+b)=-[A\cdot v]_j\ge-[d]_j=1$$

שהיא בדיוק האילוץ ה־ j של הבעיה המקורית. אז האילוצים שווים וגם המזעור שווה ולכן הבעיות שקולות עם ה־ Q,A,a,d,v שהוגדרו והבעיה היא

$$\begin{matrix}\text{argmin}_{v\in\mathbb{R}^{d+1}}\left[\frac{1}{2}\cdot v^T\cdot Q\cdot v+a^T\cdot v\right] & & \text{argmin}_{w\in\mathbb{R}^d,b\in\mathbb{R}}\|w\|^2\\ & = \quad = & \\ \text{s.t }A\cdot v\le d & & \text{s.t }y_j\cdot(\langle x_j,w\rangle+b),\forall j\end{matrix}$$

מ.ש.ל.☺

6. **צ"ל:** שוויון $argmin$
הוכחה:
תחילה נסתכל על $\text{argmin}_{w,\{\xi_i\}}\left[\frac{\lambda}{2}\|w\|^2+\frac{1}{m}\sum_{i=1}^n\xi_i\right]$ s.t. $y_i\cdot\langle x_i,w\rangle\ge1-\xi_i\wedge\xi_i\ge0$
נשים לב שאפשר לכתוב את התנאי $y_i\cdot\langle x_i,w\rangle\ge1-\xi_i$ בתור $\xi_i\ge1-y_i\cdot\langle x_i,w\rangle$.
לכן אפשר להחליף את ההגבלות על ξ_i ב־ $\xi_i\ge\max\{0,1-y_i\cdot\langle x_i,w\rangle\}=l^{hinge}(y_i\cdot\langle x_i,w\rangle)$,
כלומר אפשר לכתוב את הנוסחא באופן הבא:

$$\text{argmin}_{w,\{\xi_i\}}\left[\frac{\lambda}{2}\|w\|^2+\frac{1}{m}\sum_{i=1}^n\xi_i\right]\text{ s.t. }\xi_i\ge l^{hinge}(y_i\cdot\langle x_i,w\rangle)$$

עתה מהיות ואנחנו רוצים למזער את ξ_i, נקבל כי $\xi_i=l^{hinge}(y_i\cdot\langle x_i,w\rangle)$ במינימום, ולכן ניתן לכתוב את הבעיה באופן הבא $\text{argmin}_w\left[\frac{\lambda}{2}\|w\|^2+\frac{1}{m}\sum_{i=1}^n l^{hinge}(y_i\cdot\langle x_i,w\rangle)\right]$, ולכן קיבלנו כי

$$\text{argmin}_w\left[\frac{\lambda}{2}\|w\|^2+\frac{1}{m}\sum_{i=1}^n l^{hinge}(y_i\cdot\langle x_i,w\rangle)\right]$$
$$=\text{argmin}_{w,\{\xi_i\}}\left[\frac{\lambda}{2}\|w\|^2+\frac{1}{m}\sum_{i=1}^n\xi_i\right]\text{ s.t. }y_i\cdot\langle x_i,w\rangle\ge1-\xi_i\wedge\xi_i\ge0$$

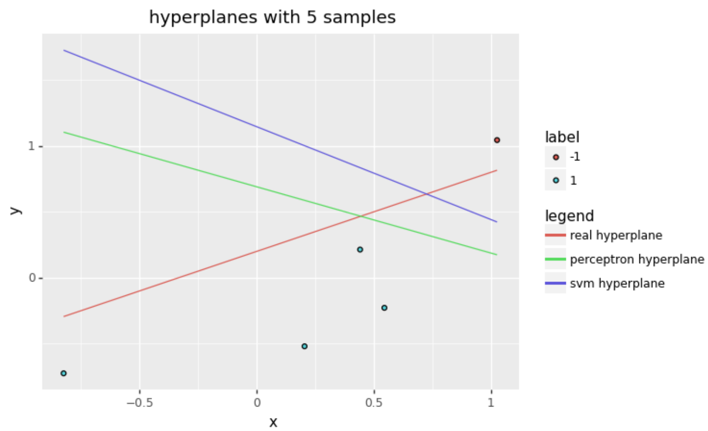
כנדרש

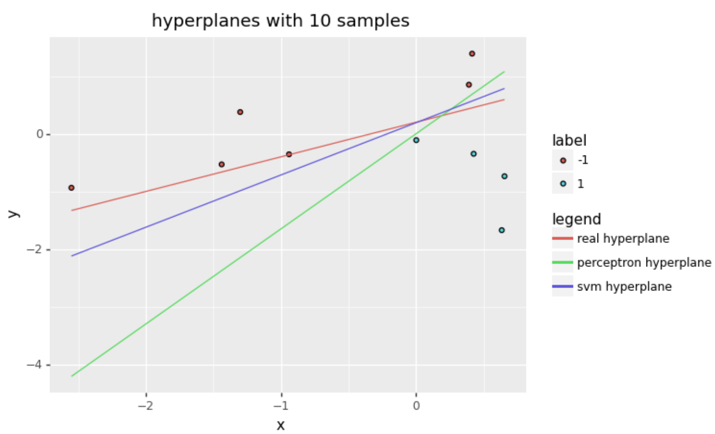
מ.ש.ל.☺

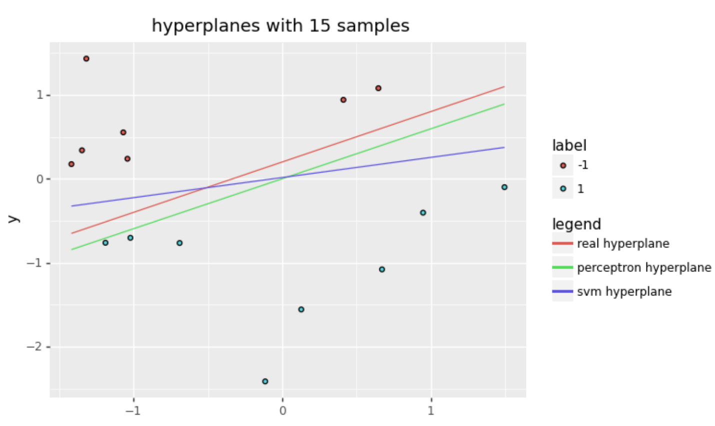
7. **בוצע**

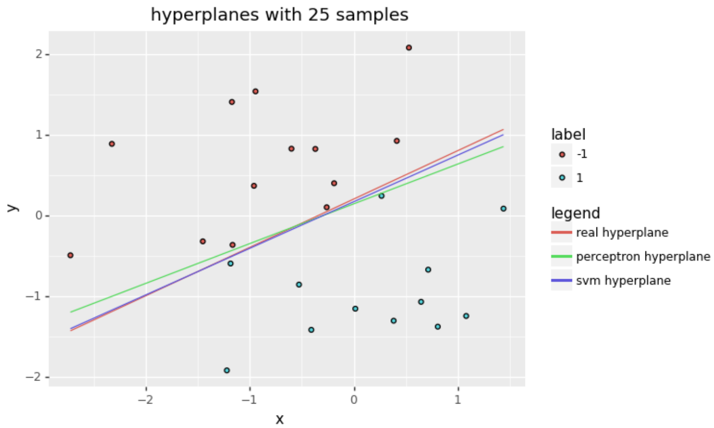
8. **בוצע**

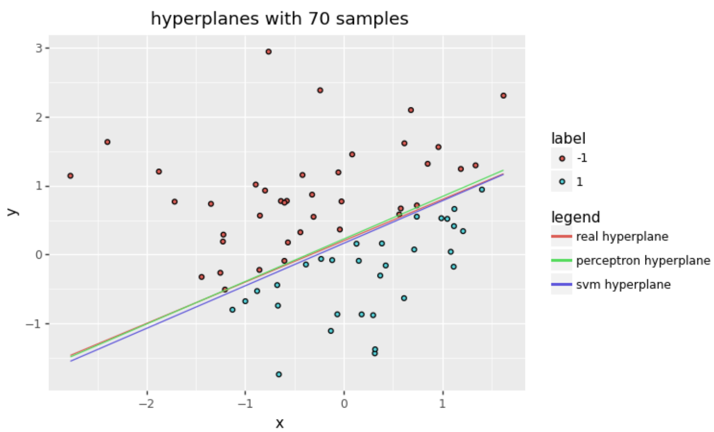
9. **צ"ל:** הייפר פליינז לנקודות שנוצרו
הוכחה:









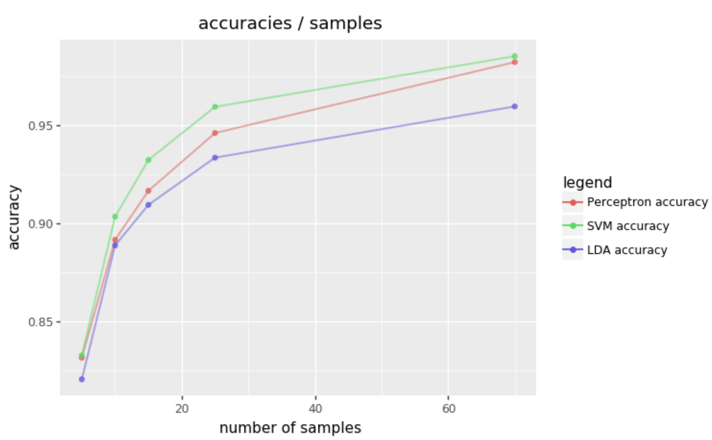


אפשר לראות שככל שיש יותר נקודות, המודלים מתקרבים יותר ויותר לפונקציה האמיתית, ואפשר לשים לב להתנהגות svm עם ה$margin$ (לדוגמא עם ה15 דגימות אפשר לראות התנהגות זאת)

מ.ש.ל.☺

10. **צ"ל:** הצלחה \ מודל

הוכחה:



מ.ש.ל.☺

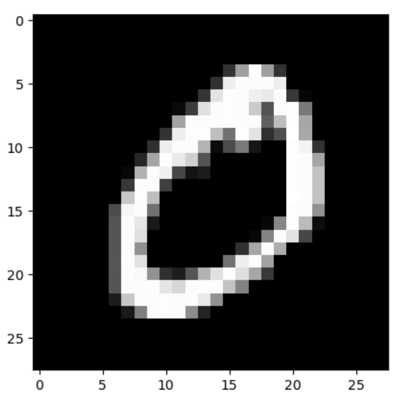
11. **צ"ל:** הצלחה \ מודל

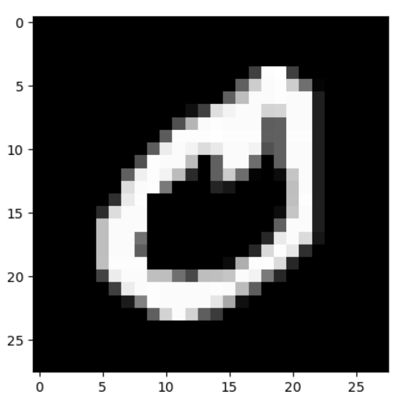
הוכחה:

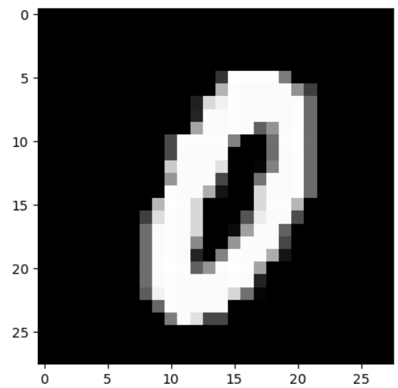
אפשר לראות שכל המודלים מצליחים בהסתברות די גבוהה אפילו עם מספר מצומצם של נקודות, שזה לא כל כך מפתיע כי דאגנו שההנחות של כל המודלים יתקיימו, התפלגות נורמלית וגם ניתן להפרדה על ידי $half-space$.
תחילה אפשר לראות ש־ SVM מצליח יותר מ־ $Percetron$ כי הוא לא בוחר סתם $half-space$ אלא אחד ספציפי שממקסם את ה־ $margin$.
אפשר לראות ש־ $SVM, Perceptron$ מצליחים יותר מ־ LDA כי ההנחות שלהם שהן מאוד מגבילות באפו כלליי (שהמודל ניתן להפרדה על ידי $half-space$) מתקיימות ולכן במקרה ספציפי זה, אלגוריתמים האלה מתפקדים מעולה.
הערה: אלגוריתם ה$Perceptron$ בכלל לא יסיים לרוץ אם ההנחה לא תתקיים (לפי האלגוריתם בשאלה 7)

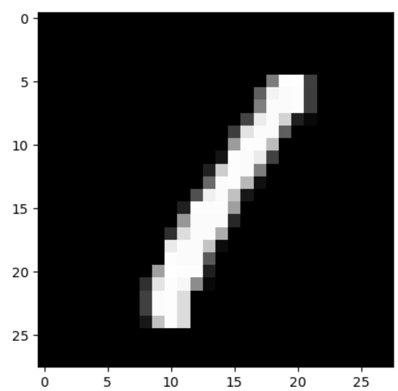
מ.ש.ל.☺

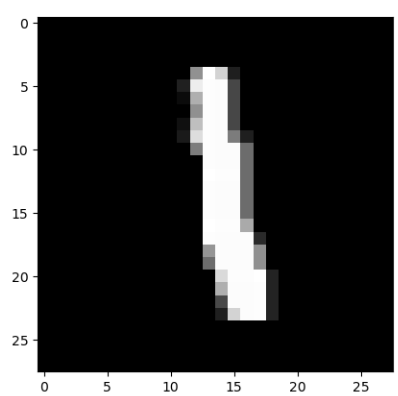
12. **תמונות:**

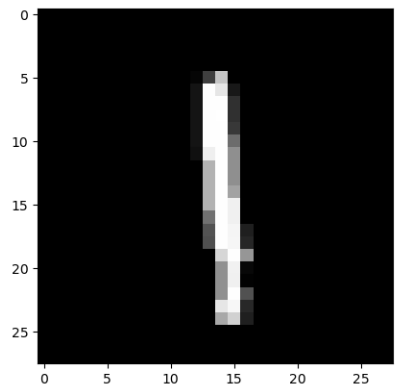





0
5
10
15
20
25
0 5 10 15 20 25


0
5
10
15
20
25
0 5 10 15 20 25

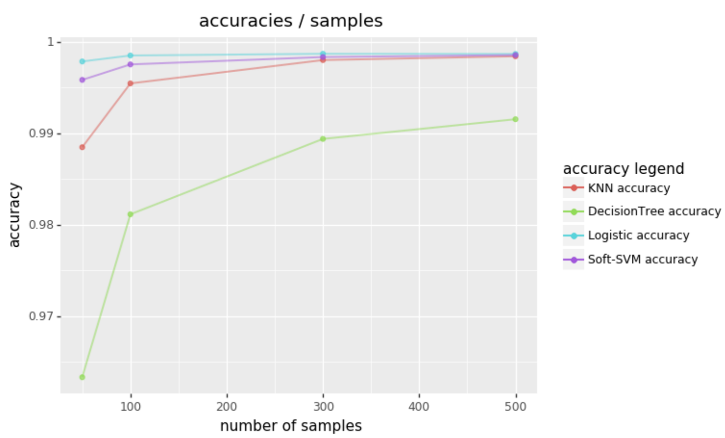

0
5
10
15
20
25
0 5 10 15 20 25

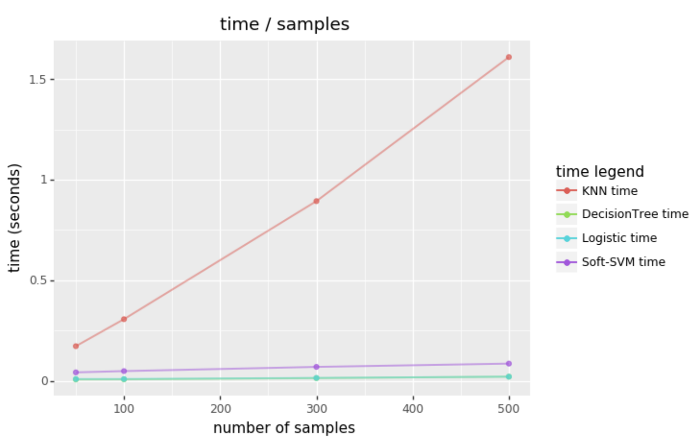


ניתן לראות הרבה דרכים שונות לכתוב מספרים אבל אפשר לראות בבירור איזה מספר מופיע באיזו תמונה.

13. **בוצע**

14. **צ"ל:** הצלחה \ מודל
הוכחה:





תחילה נשים לב לתוצאות המדהימות שיש לכל הקלסיפיירים, כולם הצליחו בהסתברות $98\%+$ כמעט בכל המקרים.
אפשר לראות ש־ knn (אפילו עם k קטן) לוקח **הרבה** יותר זמן משאר האלגוריתמים ולכן רוב הסיכויים שלא היינו בוחרים להשתמש בו, אם רצינו תשובה מיידית או לרוץ על הרבה מידע.
בחרתי ב־ KNN ש־ $k = 3$ מכיוון שרציתי k קטן לריצה מהירה ושיהיה אי זוגי בשביל שלא יהיה מצב של שוויון.
תחילה ניסיתי להגביל ב־ $decisionTree$ בהתחלה וזה הוביל לתוצאות די גרועות יחסית לאחרים, בסוף החלטתי להשאיר את הקבועים הדיפולטיביים וזה הוביל לתוצאות ממש טובות
בחרתי ב־ $Soft - SVM$ ב־ $C = 0.01$ (ככל ש־ C קטן יותר אז האלגוריתם הוא יותר $soft$) וראיתי שהוא נותן תוצאות מעולות, אז השארתי אותו כך.

מ.ש.ל.☺