תרגיל 3 - IML - מאי ביבי

שאלה 1:

מסווג Bayes optimal:

$$\forall x \in \mathcal{X} \ h_{\mathcal{D}}(x) = \begin{cases} +1 & \Pr(y=1 \mid x) \geq \frac{1}{2} \\ -1 & \text{otherwise} \end{cases}$$

נרצה להראות

$$h_{\mathcal{D}}(x) = \underset{y\in\{\pm1\}}{\text{argmax}} \Pr(x \mid y) \Pr(y)$$

מכלל בייס מתקיים:

$$\underset{y\in\{\pm1\}}{\text{argmax}} \Pr(x \mid y) \Pr(y) = \underset{y\in\{\pm1\}}{\text{argmax}} \Pr(y \mid x) \Pr(x)$$

$$= \text{argmax}_y \left\{\Pr(y=1 \mid x) \Pr(x), \Pr(y=-1 \mid x) \Pr(x)\right\} \ (*)$$

כיוון שהמאורעות $y=1$, $y=-1$ זרים ומהווים חלוקה של מרחב ההסתברות, מנוחסאת ההסתברות השלמה

$$\Pr(y=1 \mid x) \Pr(x) + \Pr(y=-1 \mid x) \Pr(x) = \Pr(x)$$

כיוון ש- $\Pr(x) > 0$ (אחרת $\Pr(y \mid x)$ לא מוגדר), נחלק בו:

$$\Longrightarrow \Pr(y=1 \mid x) + \Pr(y=-1 \mid x) = 1$$

ולכן בהכרח או $\Pr(y=1 \mid x) \geq \frac{1}{2}$ או $\Pr(y=-1 \mid x) \geq \frac{1}{2}$. כלומר לעשות max במקרה הזה שקול לבדיקה $\Pr(y=1 \mid x) \geq \frac{1}{2}$, ולכן נקבל

$$(*) = \begin{cases} 1 & \Pr(y=1 \mid x) \geq \frac{1}{2} \\ -1 & \text{else} \end{cases} = h_{\mathcal{D}}(x)$$

שאלה 2:

פונקצית הצפיפות:

$$f(\mathbf{x} \mid y) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}-\mu_y\right)^\top \Sigma^{-1}\left(\mathbf{x}-\mu_y\right)\right\}$$

נרצה להראות שאם היינו יודעים את μ_{-1}, μ_{+1} אז

$$h_{\mathcal{D}}(x) = \underset{y\in\{\pm1\}}{\text{argmax}} \delta_y(\mathbf{x})$$

כאשר

$$\delta_y(\mathbf{x}) = \mathbf{x}^\top \Sigma^{-1} \mu_y - \frac{1}{2} \mu_y^\top \Sigma^{-1} \mu_y + \ln \Pr(y) \quad y \in \{\pm1\}$$

מהשאלה הקודמת אנחנו יודעים כי $h_{\mathcal{D}}(\mathbf{x}) = \underset{y\in\{\pm1\}}{\operatorname{argmax}} \Pr(\mathbf{x} \mid y)\Pr(y)$.

$$h_{\mathcal{D}}(\mathbf{x}) = \underset{y\in\{\pm1\}}{\operatorname{argmax}} \Pr(\mathbf{x} \mid y)\Pr(y) \overset{1}{=} \underset{y\in\{\pm1\}}{\operatorname{argmax}} f(\mathbf{x} \mid y)\Pr(y) \overset{2}{=} \underset{y\in\{\pm1\}}{\operatorname{argmax}} \ln\left(f(\mathbf{x} \mid y)\Pr(y)\right)$$

$$= \underset{y\in\{\pm1\}}{\operatorname{argmax}} \ln f(\mathbf{x} \mid y) + \ln\Pr(y)$$

$$= \underset{y\in\{\pm1\}}{\operatorname{argmax}} \ln\left(\frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}-\mu_y\right)^\top \Sigma^{-1}\left(\mathbf{x}-\mu_y\right)\right\}\right) + \ln\Pr(y)$$

$$= \underset{y\in\{\pm1\}}{\operatorname{argmax}} \ln\left(\frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}}\right) - \frac{1}{2}\left(\mathbf{x}-\mu_y\right)^\top \Sigma^{-1}\left(\mathbf{x}-\mu_y\right) + \ln\Pr(y)$$

$$= \underset{y\in\{\pm1\}}{\operatorname{argmax}} -\ln\sqrt{(2\pi)^d \det(\Sigma)} - \frac{1}{2}\left(\mathbf{x}-\mu_y\right)^\top \Sigma^{-1}\left(\mathbf{x}-\mu_y\right) + \ln\Pr(y)$$

$$\overset{3}{=} \underset{y\in\{\pm1\}}{\operatorname{argmax}} -\frac{1}{2}\left(\mathbf{x}-\mu_y\right)^\top \Sigma^{-1}\left(\mathbf{x}-\mu_y\right) + \ln\Pr(y)$$

$$\overset{4}{=} \underset{y\in\{\pm1\}}{\operatorname{argmax}} -\frac{1}{2}\left(\mathbf{x}^\top-\mu_y^\top\right) \Sigma^{-1}\left(\mathbf{x}-\mu_y\right) + \ln\Pr(y)$$

$$= \underset{y\in\{\pm1\}}{\operatorname{argmax}} -\frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mathbf{x} + \frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mu_y + \frac{1}{2}\mu_y^\top\Sigma^{-1}\mathbf{x} - \frac{1}{2}\mu_y^\top\Sigma^{-1}\mu_y + \ln\Pr(y)$$

$$\overset{5}{=} \underset{y\in\{\pm1\}}{\operatorname{argmax}} \frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mu_y + \frac{1}{2}\mu_y^\top\Sigma^{-1}\mathbf{x} - \frac{1}{2}\mu_y^\top\Sigma^{-1}\mu_y + \ln\Pr(y)$$

$$\overset{6}{=} \underset{y\in\{\pm1\}}{\operatorname{argmax}} \mathbf{x}^\top\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^\top\Sigma^{-1}\mu_y + \ln\Pr(y) = \underset{y\in\{\pm1\}}{\operatorname{argmax}} \delta_y(\mathbf{x})$$

כאשר:

1 ממונוטוניות האינטגרל
2 ממונוטוניות $\ln$
3 $-\ln\sqrt{(2\pi)^d\det(\Sigma)}$ קבוע ולא משפיע על argmax
4 $\left(\mathbf{x}-\mu_y\right)^\top = \left(\mathbf{x}^\top - \mu_y^\top\right)$ מתכונות transpose
5 $-\frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mathbf{x}$ קבוע ולא משפיע על argmax
6 $\mathbf{x}^\top\Sigma^{-1}\mu_y = \mu_y^\top\Sigma^{-1}\mathbf{x}$ - כי Σ אלכסונית כי המשתנים בלתי תלויים והשונות המשותפת שלהם 0.

שאלה 3:

נשתמש באומדים על מנת להעריך את ההסתברויות: (כאשר $\mathbf{1}_{y_i=y}$ הוא המציין של המאורע $\{y_i = y\}$)
את $\Pr(y)$ נקבע כאחוז הפעמים שהוא מופיע ב־$y_1, ..., y_m$

$$\Pr(y) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{y_i=y}$$

עבור התוחלות של $\mathbf{x} \mid y$ נחשב את ממוצע כל הדגימות המתויגות y:

$$\mu_y = \frac{\sum_{i=1}^{m} \mathbf{x}_i \mathbf{1}_{y_i=y}}{\sum_{i=1}^{m} \mathbf{1}_{y_i=y}}$$

עבור מטריצת השונות, נשמש באומד המוכר הבא (מויקיפדיה):

$$\Sigma = \frac{1}{m}\sum_{y\in\{\pm1\}}\sum_{i\in[m]:y_i=y}\left(\mathbf{x}_i-\mu_y\right)^\top\left(\mathbf{x}_i-\mu_y\right)$$

שאלה 4:

יש לנו שתי שגיאות: סיווג אימייל ספאם כלא ספאם וסיווג אימייל לא ספאם כספאם. השגיאה היותר חמורה היא סיווג אימייל לא ספאם כספאם - כי אז האימייל החשוב לא יקרא. סיווג אימייל ספאם כלא ספאם הוא פחות חמור כי בסך הכל יגיע אימייל ספאם ליעד. לכן ה־negative יהיה not-spam, ואז ה־false-negative תהיה השגיאה הפחות חמורה, וה־positive יהיה ה־spam, כך שה־false-positive חמור יותר.

שאלה 5:

נכתוב את בעיית ה־Hard-SVM הבאה בצורה הקנונית של QP:

$$
\begin{aligned}
&= \underset{(\mathbf{w},b)}{\operatorname{argmin}} \|\mathbf{w}\|^2 \quad \text{s.t. } \forall i, y_i\left(\langle \mathbf{w}, x_i\rangle + b\right) \geqslant 1 \\
&= \underset{(\mathbf{w},b)}{\operatorname{argmin}} \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix}^T I \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} \quad \text{s.t. } \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \cdot \begin{pmatrix} - & x_1 & - & 1 \\ & \vdots & & \vdots \\ - & x_m & - & 1 \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} \geqslant \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\
&= \underset{(\mathbf{w},b)}{\operatorname{argmin}} \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix}^T I \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} \quad \text{s.t. } \begin{pmatrix} y_1 x_1^1 & \cdots & y_1 x_d^1 & 1 \\ \vdots & \ddots & \vdots & \vdots \\ y_m x_1^m & \cdots & y_m x_d^m & 1 \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} \geqslant \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\
&= \underset{(\mathbf{w},b)}{\operatorname{argmin}} \frac{1}{2} \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix}^T 2 \cdot I \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} + \overrightarrow{0}^T \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} \quad \text{s.t. } \begin{pmatrix} y_1 x_1^1 & \cdots & y_1 x_d^1 & 1 \\ \vdots & \ddots & \vdots & \vdots \\ y_m x_1^m & \cdots & y_m x_d^m & 1 \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} \geqslant \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\
&= \underset{(\mathbf{w},b)}{\operatorname{argmin}} \frac{1}{2} \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix}^T 2 \cdot I \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} + \overrightarrow{0}^T \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} \quad \text{s.t. } -\begin{pmatrix} y_1 x_1^1 & \cdots & y_1 x_d^1 & 1 \\ \vdots & \ddots & \vdots & \vdots \\ y_m x_1^m & \cdots & y_m x_d^m & 1 \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} \leq \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix}
\end{aligned}
$$

וקיבלנו שהשורה האחרונה היא בצורת QP קנונית עם

$$
\mathbf{v} = \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix}, \; Q = 2 \cdot I, \; \mathbf{a} = \overrightarrow{0}
$$

$$
A = -\begin{pmatrix} y_1 x_1^1 & \cdots & y_1 x_d^1 & 1 \\ \vdots & \ddots & \vdots & \vdots \\ y_m x_1^m & \cdots & y_m x_d^m & 1 \end{pmatrix}, \; \mathbf{d} = \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix}
$$

שאלה 6:

נרצה להראות שהבעיה הבאה

$$
\arg\min_{\mathbf{w}} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \ell^{\text{hinge}}\left(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle\right),
$$

$$
\text{where } \ell^{\text{hinge}}(a) = \max\{0, 1-a\}
$$

שקולה לבעיית ה־soft SVM כפי שהגדרנו:

$$
\arg\min_{\mathbf{w},\{\xi_i\}} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \xi_i \text{ s.t. } \forall_i, y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geqslant 1 - \xi_i \text{ and } \xi_i \geqslant 0
$$

נגדיר את ξ_i עבור $i = 1, ..., m$ להיות:

$$
\xi_i := \begin{cases} 0 & y_i \langle \mathbf{x}_i, \mathbf{w} \rangle > 1 \\ 1 - y_i \langle \mathbf{x}_i, \mathbf{w} \rangle & \text{otherwise} \end{cases}
$$

נראה כי ξ_i עומד באילוצים $\xi_i \geqslant 0$ ו־ $y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geqslant 1 - \xi_i \iff \xi_i \geqslant 1 - y_i \langle \mathbf{x}_i, \mathbf{w} \rangle$:
אם $y_i \langle \mathbf{x}_i, \mathbf{w} \rangle > 1$ אז $\xi_i = 0$ וכן מתקיים $1 - y_i \langle \mathbf{x}_i, \mathbf{w} \rangle < 0$ ולכן גם האילוץ השני מתקיים.
אחרת, $1 - y_i \langle \mathbf{x}_i, \mathbf{w} \rangle \geq 0$, ואנו מגדירים $\xi_i = 1 - y_i \langle \mathbf{x}_i, \mathbf{w} \rangle$ וזה בפרט עומד באילוץ השני.
נשים לב שהגדרנו את ξ_i להיות המספר המינימלי שעומד באילוצים ־ אם $y_i \langle \mathbf{x}_i, \mathbf{w} \rangle > 1$, 0 הוא המינ' ש־ξ_i יכול להיות מהאילוץ $\xi_i \geqslant 0$, ואחרת, מהאילוץ השני המספר המינ' עבור ξ_i הוא $1 - y_i \langle \mathbf{x}_i, \mathbf{w} \rangle$.

כעת נראה כי $\xi_i = \ell^{\text{hinge}} \left(y_i \langle \mathbf{x}_i, \mathbf{w} \rangle \right)$:

$$\ell^{\text{hinge}} \left(y_i \langle \mathbf{x}_i, \mathbf{w} \rangle \right) = \max \{0, 1 - y_i \langle \mathbf{x}_i, \mathbf{w} \rangle\} = \begin{cases} 0 & y_i \langle \mathbf{x}_i, \mathbf{w} \rangle > 1 \\ 1 - y_i \langle \mathbf{x}_i, \mathbf{w} \rangle & \text{otherwise} \end{cases} = \xi_i$$

ולכן

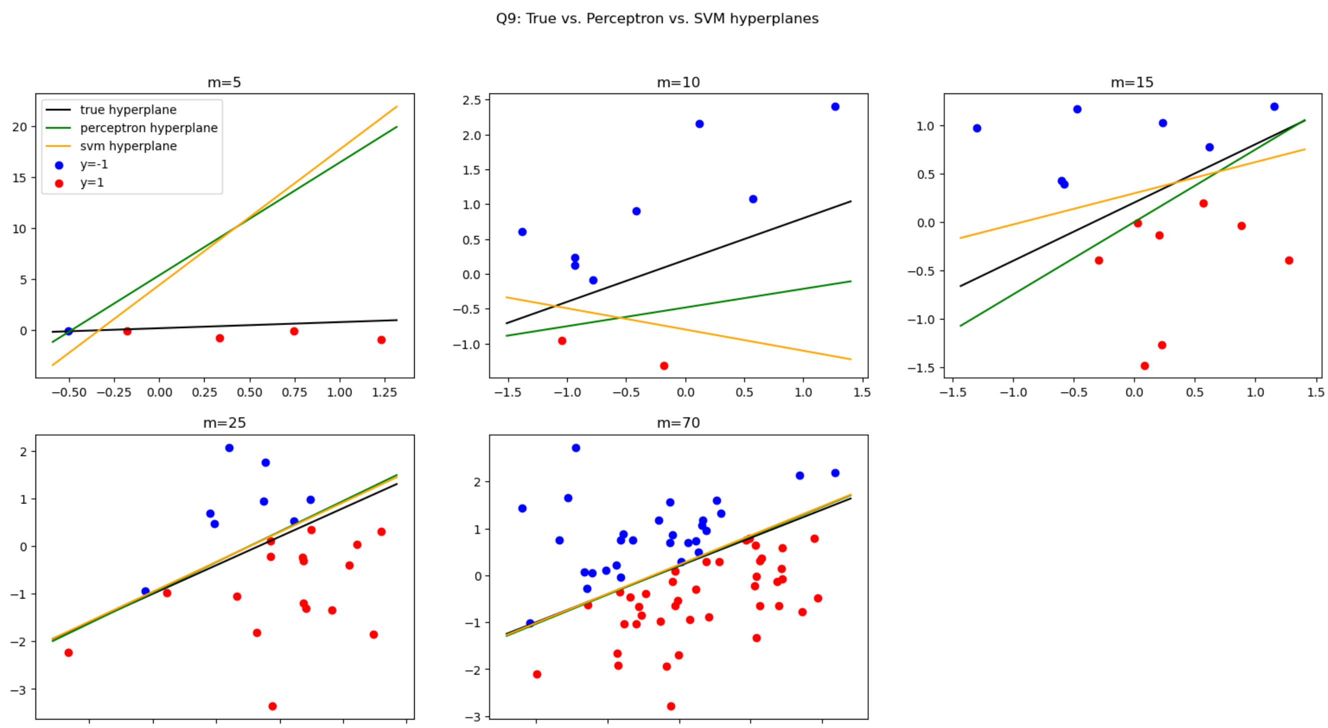
$$\frac{1}{m} \sum_{i=1}^{m} \ell^{\text{hinge}} \left(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \right) = \frac{1}{m} \sum_{i=1}^{m} \xi_i$$

הגדרנו ξ_i מינימליים שעומדים באילוצים, וזה מתאים לבעיה המקורית בה אנו רוצים למזער אותם. ממה שהראנו קודם מתקיים

$$\arg \min_{\mathbf{w}, \{\xi_i\}} \frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i = \arg \min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^{m} \ell^{\text{hinge}} \left(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \right)$$

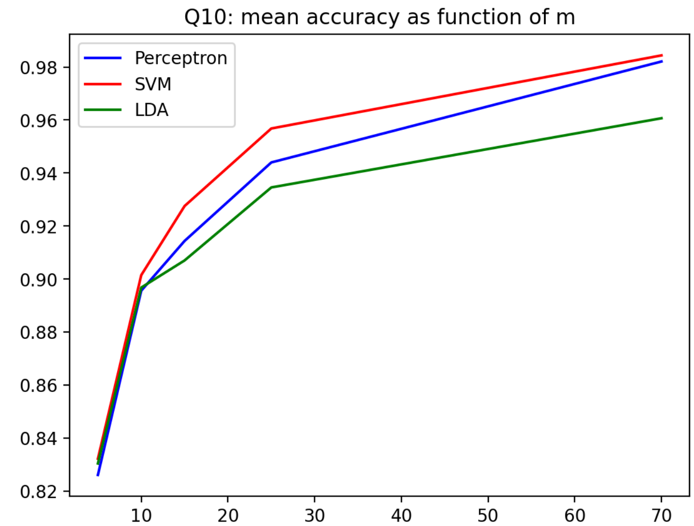
ולכן הבעיות שקולות.

שאלה 9:



שאלה 10:

גרף של mean accuracy כפונקציה של m עבור שלושת המסווגים perceptron, SVM, LDA:



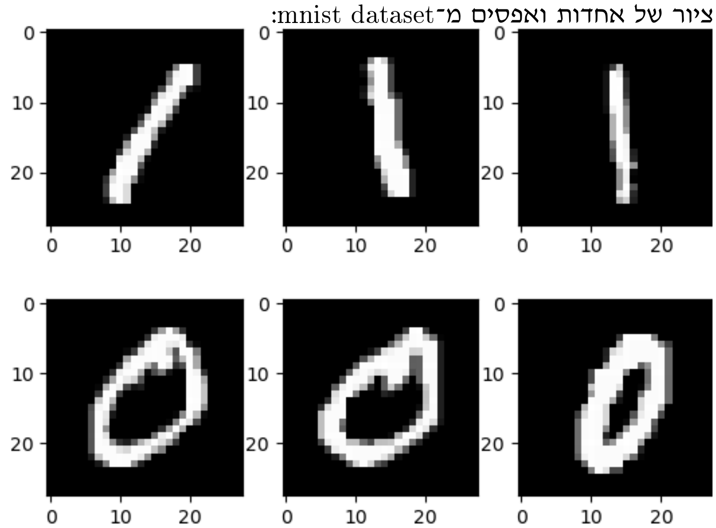
ניתן לראות שה־SVM הוא בעל ה־mean accuracy הכי גבוה, וה־LDA הכי נמוך

שאלה 11:

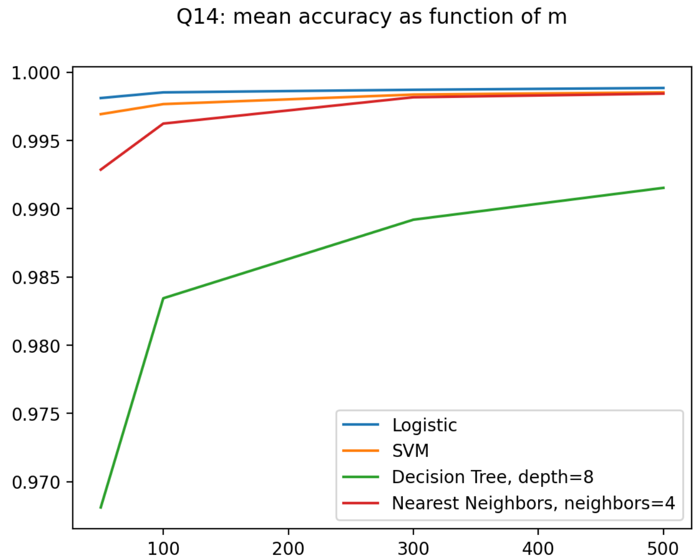
כיוון ש־LDA מניח הנחה שגויה לגבי התפלגות הדגימות (הוא מניח ש־X, y באים מהתפלגות משותפת וכן ש־X מגיע משני גאוסיאניים, בעוד שדגמנו רק את X באופן בלתי תלוי וגאוסייני, ואינו תלוי ב־y) רואים שהוא מסווג הכי פחות טוב (accuracy נמוך יותר). SVM ו־ Perceptron לא מניחים הנחה שגויה כזו ולכן מסווגים טוב יותר (accuracy גבוה יותר). ניתן לראות ש־SVM סיווג טוב יותר וזאת כי הוא בוחר על־מישור (קו ההפרדה) עם שול מקסימלי, בעוד שה־perceptron בוחר על־מישור כלשהו שעובד.

שאלה 12:

ציור של אחדות ואפסים מ־mnist dataset:



שאלה 14:



רואים כי שככל שמספר הדגימות עולה, Logistic, SVM, Nearset Neighbors מסווגים פחות או יותר עם אותו accuracy גבוה, לעומתם Decision tree קצת פחות טוב.

זמני הריצה:

	m=50	m=100	m=300	m=500
Logistic	0.007669	0.008070	0.011957	0.015337
SVM	0.039158	0.047437	0.069210	0.083393
Decision Tree, depth=8	0.003963	0.004811	0.010244	0.017135
Nearest Neighbors, neighbors=4	0.180960	0.298545	0.854777	1.470554

ניתן לראות שככל שמספר הדגימות עולה כך גם זמן הריצה. בנוסף ניתן לראות ש־Logistic, Decision Tree לוקחים בערך אותו זמן, SVM לוקח קצת יותר ו־Nearest Neighbors לוקח הרבה יותר זמן.
לגבי Nearest Neighbors: ראינו בכיתה שמסווג זה פחות יעיל כי עבור כל דגימה שהוא מסווג (ב־predict) צריך לעבור על כל ה־training set ולמצוא את K השכנים הכי קרובים. לחלופין, גם אם ב־fit הוא מכניס את ה־training set למבנה נתונים שנותן לו יעילות ב־predict (נאמר בהרצאה שניתן לעשות זאת) זה תהליך שלוקח זמן ולכן ההבדל בזמנים.