תרגיל 3 - IML - מאי ביבי

שאלה 1:

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

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 $\frac{1}{2}\geq$, 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 $h_{\mathcal{D}}(\mathbf{x})=\underset{y\in\{\pm1\}}{\text{argmax}}\Pr(\mathbf{x}\mid y)\Pr(y)$.

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

$$
\begin{aligned}
&= \underset{y\in\{\pm1\}}{\operatorname{argmax}} \ln f(\mathbf{x} \mid y) + \ln \Pr(y) \\
&= \underset{y\in\{\pm1\}}{\operatorname{argmax}} \ln\left(\frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}-\mu_y\right)^\top \Sigma^{-1}\left(\mathbf{x}-\mu_y\right)\right\}\right) + \ln \Pr(y) \\
&= \underset{y\in\{\pm1\}}{\operatorname{argmax}} \ln\left(\frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}}\right) - \frac{1}{2}\left(\mathbf{x}-\mu_y\right)^\top \Sigma^{-1}\left(\mathbf{x}-\mu_y\right) + \ln \Pr(y) \\
&= \underset{y\in\{\pm1\}}{\operatorname{argmax}} - \ln\sqrt{(2\pi)^d \det(\Sigma)} - \frac{1}{2}\left(\mathbf{x}-\mu_y\right)^\top \Sigma^{-1}\left(\mathbf{x}-\mu_y\right) + \ln \Pr(y) \\
&\overset{3}{=} \underset{y\in\{\pm1\}}{\operatorname{argmax}} - \frac{1}{2}\left(\mathbf{x}-\mu_y\right)^\top \Sigma^{-1}\left(\mathbf{x}-\mu_y\right) + \ln \Pr(y) \\
&\overset{4}{=} \underset{y\in\{\pm1\}}{\operatorname{argmax}} - \frac{1}{2}\left(\mathbf{x}^\top-\mu_y^\top\right) \Sigma^{-1}\left(\mathbf{x}-\mu_y\right) + \ln \Pr(y) \\
&= \underset{y\in\{\pm1\}}{\operatorname{argmax}} - \frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mathbf{x} + \frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mu_y + \frac{1}{2}\mu_y^\top\Sigma^{-1}\mathbf{x} - \frac{1}{2}\mu_y^\top\Sigma^{-1}\mu_y + \ln \Pr(y) \\
&\overset{5}{=} \underset{y\in\{\pm1\}}{\operatorname{argmax}} \frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mu_y + \frac{1}{2}\mu_y^\top\Sigma^{-1}\mathbf{x} - \frac{1}{2}\mu_y^\top\Sigma^{-1}\mu_y + \ln \Pr(y) \\
&\overset{6}{=} \underset{y\in\{\pm1\}}{\operatorname{argmax}} \mathbf{x}^\top\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^\top\Sigma^{-1}\mu_y + \ln \Pr(y) = \underset{y\in\{\pm1\}}{\operatorname{argmax}} \delta_y(\mathbf{x})
\end{aligned}
$$

כאשר:

1 ממונוטוניות האינטגרל

2 ממונטוניות 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 $\mathbf{x}^\top\Sigma^{-1}\mu_y = \mu_y^\top\Sigma^{-1}\mathbf{x}$ **- לכתוב הסבר אחר כך**

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 (אומד מוכר מקורס הסתברות):

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

עבור מטריצת השונות, נשמש באומד המוכר הבא:

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

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:

להסביר

$$\begin{array}{ll} & \arg\min_{(\boldsymbol{w},b)} \|\boldsymbol{w}\|^2 \\ \text{s.t.} & \forall i, y_i\left(\langle \boldsymbol{w}, x_i\rangle + b\right) \geqslant 1 \\ & \arg\min_{(\boldsymbol{w},b)} \begin{pmatrix} \boldsymbol{w} \\ b \end{pmatrix}^T \boldsymbol{I} \begin{pmatrix} \boldsymbol{w} \\ b \end{pmatrix} \\ = & \begin{pmatrix} (& y_1 x_1 &) & y_1 \\ \vdots & \vdots & & \\ (& y_m x_m &) & y_m \end{pmatrix} \begin{pmatrix} \boldsymbol{w} \\ b \end{pmatrix} \geqslant \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\ \text{s.t.} & \begin{pmatrix} \boldsymbol{w} \\ b \end{pmatrix} + \mathbf{0}^T \begin{pmatrix} \boldsymbol{w} \\ b \end{pmatrix} \\ & \arg\min_{(\boldsymbol{w},b)} \frac{1}{2} \begin{pmatrix} \boldsymbol{w} \\ b \end{pmatrix}^T [2\cdot \boldsymbol{I}] \begin{pmatrix} (& y_1 x_1 &) & y_1 \\ \vdots & & \vdots & \\ (& y_m x_m &) & y_m \end{pmatrix} \begin{pmatrix} \boldsymbol{w} \\ b \end{pmatrix} \leqslant \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix} \end{array}$$

שאלה 6:

נרצה להראות שהבעיה הבאה

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

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

שקולה לבעית ה־soft SVM כפי שהגדרנו:

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

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

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

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

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

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

ולכן

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

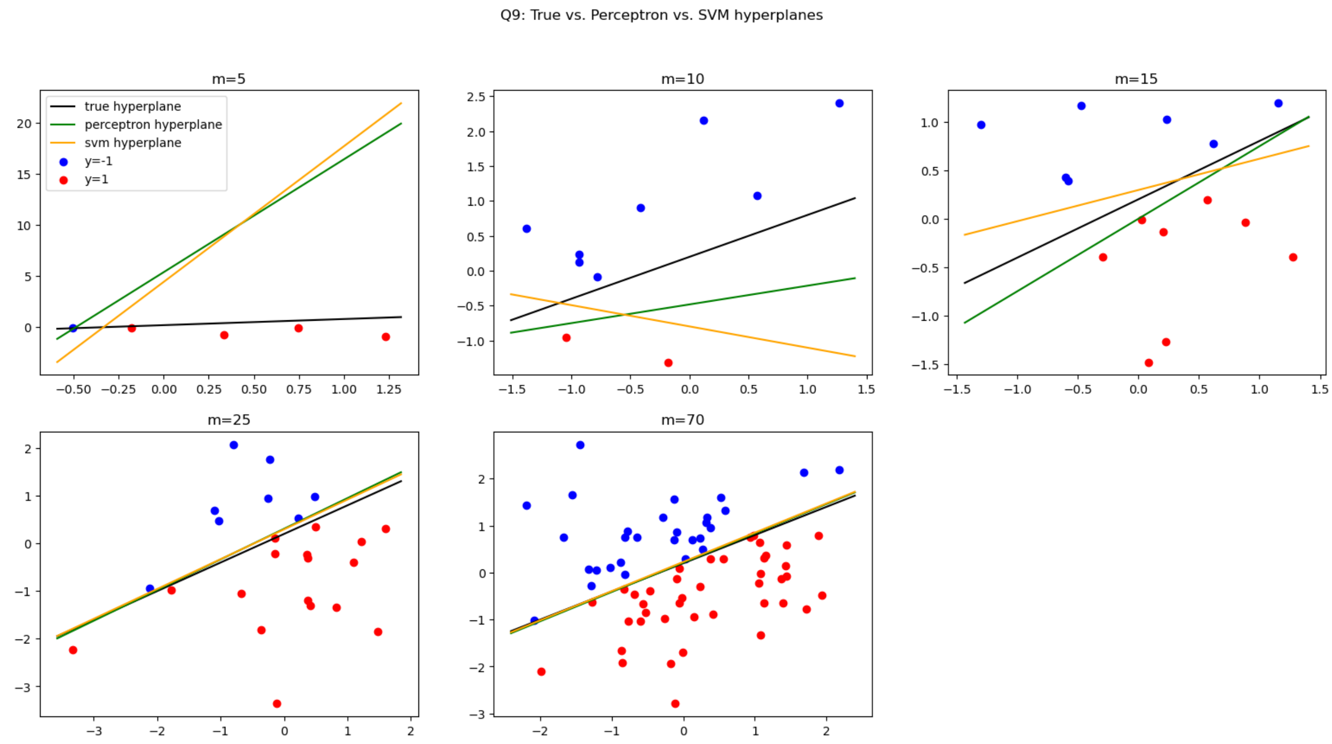
הגדרנו ξ_i מינימליים שעומדים באילוצים, וזה מתאים לבעיה המקורית בה אנו רוצים למזער אותם. ממה שהראנו קודם מתקיים

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

ולכן הבעיות שקולות.

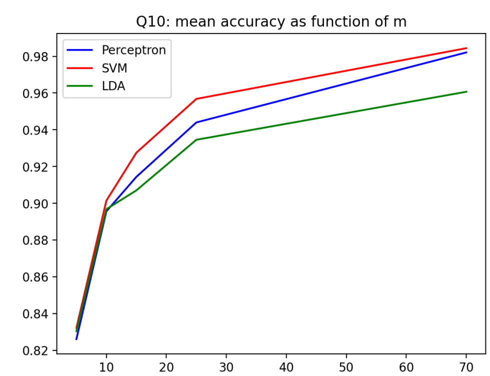
שאלה 9:

גרף של:



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

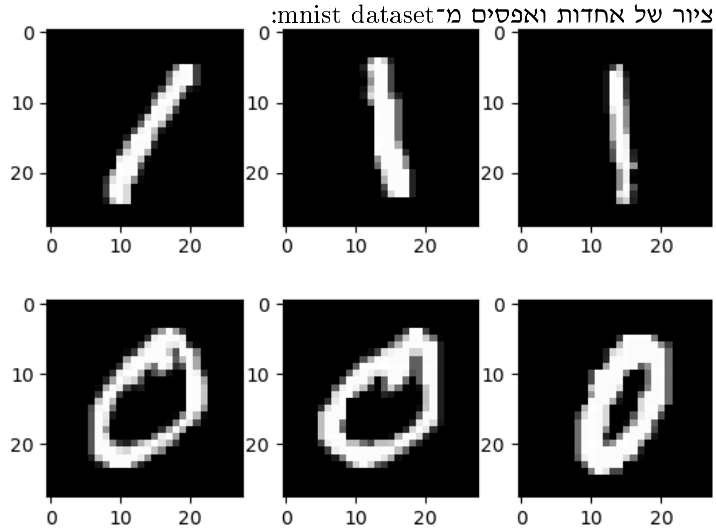
שאלה 11:

כיוון ש־LDA מניח הנחה שגויה לגבי התפלגות הדגימות (הוא מניח ש־X, y באים מהתפלגות משותפת ותלויים אחד בשני, 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.

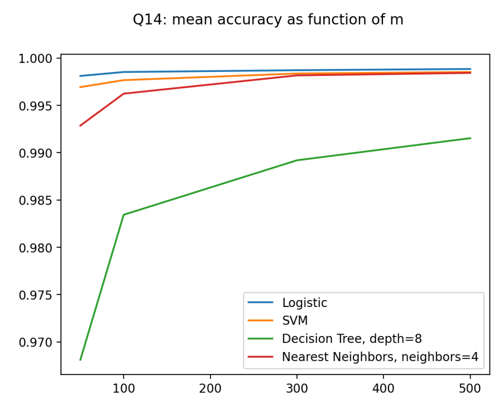
שאלה 12:

ציור של אחדות ואפסים מ־mnist dataset:



שאלה 14:

א:



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