[illegible]
304350503

עקרונות הלמידה - תרגיל בית 3

1) ראשית, נזכר בנוסחאת ב"ס ונוסחאת ההסתברות השלמה מהקורס בהסתברות:

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

$$\mathbb{P}(y|x) = \mathbb{P}(y=1|x) + \mathbb{P}(y=-1|x)$$

בשימוש בנוסחאות הללו נוכל לבחון מהו h_D:

$$\underset{y=\{\pm1\}}{\operatorname{argmax}}\ \mathbb{P}(x|y)\cdot\mathbb{P}(y) \underset{\text{חוק ב"ס}}{=} \underset{y=\{\pm1\}}{\operatorname{argmax}}\ \frac{\mathbb{P}(y|x)\cdot\mathbb{P}(x)}{\mathbb{P}(y)}\cdot\mathbb{P}(y)$$

$$= \underset{y=\{\pm1\}}{\operatorname{argmax}}\ \mathbb{P}(y|x)\cdot\mathbb{P}(x) \underset{\text{הסתברות שלמה}}{=} \underset{y=\{\pm1\}}{\operatorname{argmax}}\ \big(\mathbb{P}(y=1|x) + \mathbb{P}(y=-1|x)\big)\cdot\mathbb{P}(x)$$

כעת נבחין בין שני מקרים אפשריים:

אם $\mathbb{P}(y=1|x) \geq \frac{1}{2}$ אזי $\mathbb{P}(y=-1|x) < \frac{1}{2}$

עפ"י הגדרת מדד הסתברות כאשר יש רק שני אפשרויות ל-y (±1), אם-כן נקבל:

$$\underset{y=\{\pm1\}}{\operatorname{argmax}}\ \big(\mathbb{P}(y=1|x) + \mathbb{P}(y=-1|x)\big)\cdot\mathbb{P}(x) = \underline{1}$$

במקרה השני, $\mathbb{P}(y=1|x) < \frac{1}{2}$ וגם $\mathbb{P}(y=-1|x) \geq \frac{1}{2}$ ונקבל:

$$\underset{y=\{\pm1\}}{\operatorname{argmax}}\ \big(\mathbb{P}(y=1|x) + \mathbb{P}(y=-1|x)\big)\cdot\mathbb{P}(x) = \underline{\underline{-1}}$$

קיבלנו כי ההצגה שקולה לזו של h_D כנדרש.

(2

$$\underset{y=\{\pm1\}}{\operatorname{argmax}}\ \delta_y(x) = \underset{y=\{\pm1\}}{\operatorname{argmax}}\ x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + \ln \mathbb{P}(y)$$

כעת, נשתמש גם בחוק הסתברות מותנית:

$$\mathbb{P}(y) = f_Y(y) = \frac{f_X(x)\, f_{Y|X=x}(y)}{f_{X|Y=y}(x)}$$

נציב ונקבל:

$$= \underset{y=\{\pm1\}}{\operatorname{argmax}}\ x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + \ln\left(\frac{f_X(x)\, f_{Y|X=x}(y)}{f_{X|Y=y}(x)}\right)$$

נפעיל חוקי $\ln$:

$$= \underset{y=\{\pm1\}}{\operatorname{argmax}}\ x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + \ln(f_X(x)) + \ln(f_{Y|X=x}(y)) - \ln(f_{X|Y=y}(x))$$

כעת, נשים לב כי הביטוי $\ln(f_X(x))$ הינו קבוע ולכן להשמטה:

$$= \underset{y=\{\pm1\}}{\operatorname{argmax}}\ x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + \ln(f_{Y|X=x}(y)) - \ln(f_{X|Y=y}(x))$$

כעת, נציב את הביטוי הנתון עבור $\ln(f_{X|Y=y}(x))$ ונקבל:

$$= \underset{y=\{\pm1\}}{\operatorname{argmax}}\ x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + \ln(f_{Y|X=x}(y)) - \ln\left(\frac{1}{\sqrt{\;}}\exp\left\{-\frac{1}{2}(x-\mu_y)^T\Sigma^{-1}(x-\mu_y)\right\}\right)$$

נשים-לב כי הביטוי $\ln\left(\frac{1}{\sqrt{\;}}\right)$ הינו קבוע וניתן להשמטה:

$$= \underset{y=\{\pm1\}}{\operatorname{argmax}}\ x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + \ln(f_{Y|X=x}(y)) + \frac{1}{2}(x-\mu_y)^T\Sigma^{-1}(x-\mu_y)$$

$$= \underset{y=\{\pm1\}}{\operatorname{argmax}}\ \ln(f_{Y|X=x}(y)) + x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + \left(\frac{1}{2}x^T\Sigma^{-1} - \frac{1}{2}\mu_y^T\Sigma^{-1}\right)(x-\mu_y)$$

$$= \underset{y=\{\pm1\}}{\operatorname{argmax}}\ \ln(f_{Y|X=x}(y)) + x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + \frac{1}{2}x^T\Sigma^{-1}x - \frac{1}{2}\mu_y^T\Sigma^{-1}x - \frac{1}{2}x^T\Sigma^{-1}\mu_y + \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y$$

עבור, הביטוי $\frac{1}{2}x^T\Sigma^{-1}x$ הינו קבוע:

$$= \underset{y=\{\pm1\}}{\arg\max}\ \ln(f_{Y|X=x}(y)) + \frac{1}{2}x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}x$$

בשילוב כי Σ^{-1} הינה מטריצה סימטרית נקבל:

$$= \underset{y=\{\pm1\}}{\arg\max}\ \ln(f_{Y|X=x}(y)) + \frac{1}{2}x^T\Sigma^{-1}\mu_y - \frac{1}{2}x^T\Sigma^{-1}\mu_y$$

$$= \underset{y=\{\pm1\}}{\arg\max}\ \ln(f_{Y|X=x}(y))$$

ממונוטוניות פונקציית ה-$\ln$ נקבל:

$$= \underset{y=\{\pm1\}}{\arg\max}\ \mathbb{P}(y|x) = \underset{y=\{\pm1\}}{\arg\max}\ \mathbb{P}(x|y)\cdot\mathbb{P}(y) = h_D$$

עפ"י משפט בייס

ע"פ הגדרת שאלה 1

כנדרש.

3) בהמשך לקורס בסטטיסטיקה, אומד בלתי-מוטה לתוחלת צריך להיות הממוצע ולכן נגדיר את האומדים לתוחלות:

$$\hat{\mu}_{+1} = \frac{1}{|S_{+1}|}\cdot\sum_{i\in S_{+1}} x_i \quad ו \quad \hat{\mu}_{-1} = \frac{1}{|S_{-1}|}\cdot\sum_{i\in S_{-1}} x_i$$

בהמשך לכך, אומד בלתי-מוטה לשונות יהיה:

$$\hat{\Sigma} = \frac{1}{m-1}\sum_{i=1}^{m}(x_i-\bar{x})(x_i-\bar{x})^T$$

לבסוף, עבור ההסתברות $\mathbb{P}(y)$ נוכל להיעזר בהתפלגות ברנולי עם פרמטר p כך ש:

$$p = \frac{|S_{+1}|}{m}$$

4) כשאנו מנסים לסנן מיילים ע"פ האם מדובר בספאם או לא נוכל להבחין בין שני טעויות:

הראשונה, שאנו מסווגים מייל מסוים שהוא ספאם כאשר במציאות הוא אינו כזה ולכן "מפספס" מייל שיכול להיות חשוב לנו.

השנייה, שאנו מסווגים מייל מסוים שהוא אינו ספאם אבל במציאות הוא כן ספאם ולמרות זאת הגיע לתיבתנו.

אי-לכך, הטעות שאנחנו נרצה להימנע ממנה כמה שיותר היא הראשונה ובשביל זה נסמן את $1+$ להיות התווית "לספאם" ואת -1 להיות התווית "לא ספאם".

5) ראשית נזכיר כי מתקיים: $\|w\|^2 = w^T w = w^T I_n w$

מעבר לכך, נשים לב שכל קונסטנטה $y_i \langle w, x_i \rangle$ מוצגת ע"י $w^T (x_i y_i)$ ולכן מאפשרים להציג את המשוואה כמו שהיא ולכן נגדיר את $a=0$ ונקבל:

$$\underset{(w,b)}{\operatorname{argmin}} \|w\|^2 \text{ s.t. } \forall i,\ y_i(\langle w, x_i \rangle + b) \geq 1$$

$$= \underset{(w,b)}{\operatorname{argmin}} \begin{pmatrix} w & b \end{pmatrix} I_n \begin{pmatrix} w \\ b \end{pmatrix} \text{ s.t. } \begin{bmatrix} x_1 y_1 & 1 \\ \vdots & \vdots \\ x_m y_m & 1 \end{bmatrix} \begin{pmatrix} w \\ b \end{pmatrix} \geq \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

נציב כעת לצורה הקנונית ונקבל:

$$= \underset{(w,b)}{\operatorname{argmin}} \frac{1}{2} \overbrace{\begin{pmatrix} w & b \end{pmatrix}}^{V^T} \overbrace{2I_n}^{Q} \overbrace{\begin{pmatrix} w \\ b \end{pmatrix}}^{V} + \overbrace{0}^{a} \overbrace{\begin{pmatrix} w \\ b \end{pmatrix}}^{V}$$

$$\text{s.t. } \underbrace{\begin{bmatrix} -x_1 y_1 & -1 \\ \vdots & \vdots \\ -x_m y_m & -1 \end{bmatrix}}_{A} \underbrace{\begin{pmatrix} w \\ b \end{pmatrix}}_{V} \leq \underbrace{\begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix}}_{d}$$

6) ראשית, נשים-לב כי השוני בין הביטויים נובע מהאיבר השני בסכום ולא מהאיבר הראשון ($\frac{\lambda}{2}\|w\|^2$) הזהה לשניהם.

ע"פ הביטוי הראשון מתקיים כי $y_i \langle w, x_i \rangle \ge 1 - \xi_i$

ע"י העברת אגפים נקבל:

$$\xi_i \ge 1 - y_i \langle w, x_i \rangle$$

מהכלל השני, נתון כי $\xi_i \ge 0$ לכל i ולכן נוכל להגדיר:

$$\xi_i \ge \max\{0, 1 - y_i \langle w, x_i \rangle\}$$

נרצה למזער את האיבר השני בסכום ולכן נוכל לסמן $a = y_i \langle w, x_i \rangle$ ונקבל:

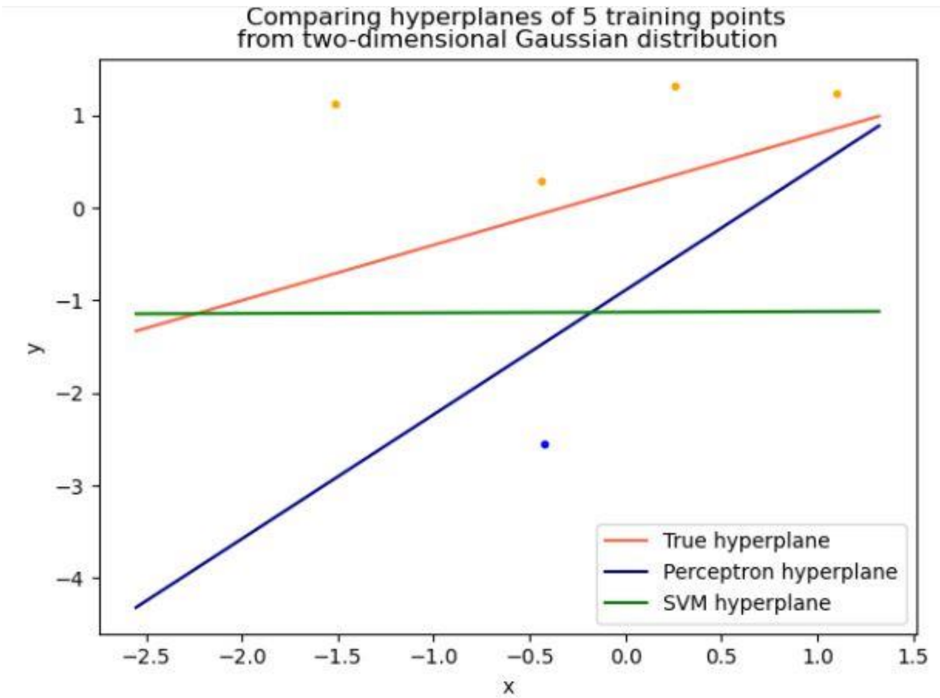
$$\underset{(w)}{\arg\min} \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} \ell^{hinge}(a)$$

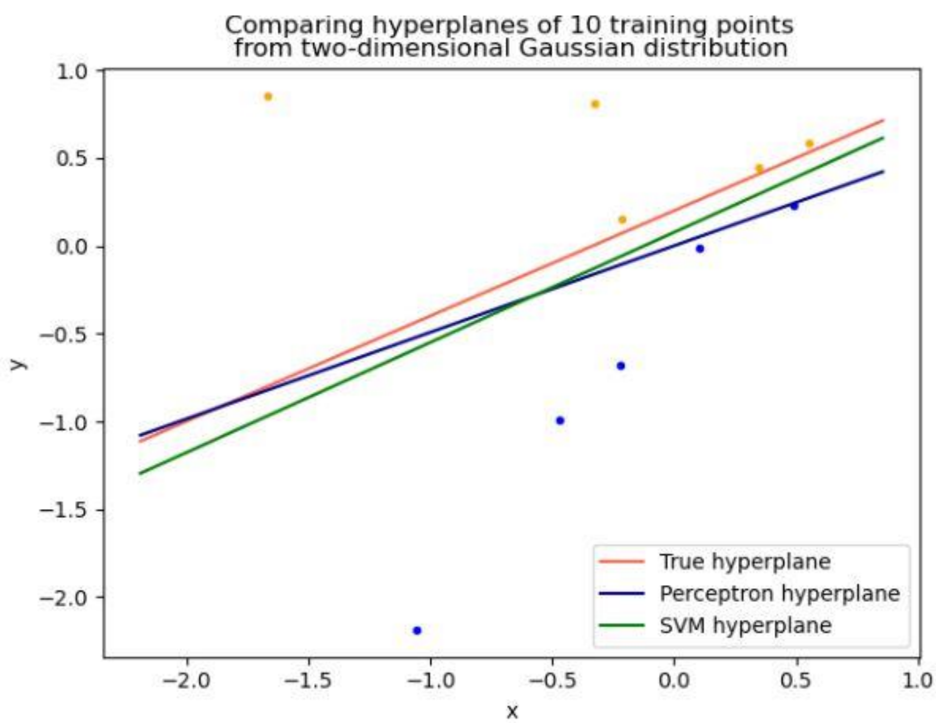
כנדרש.

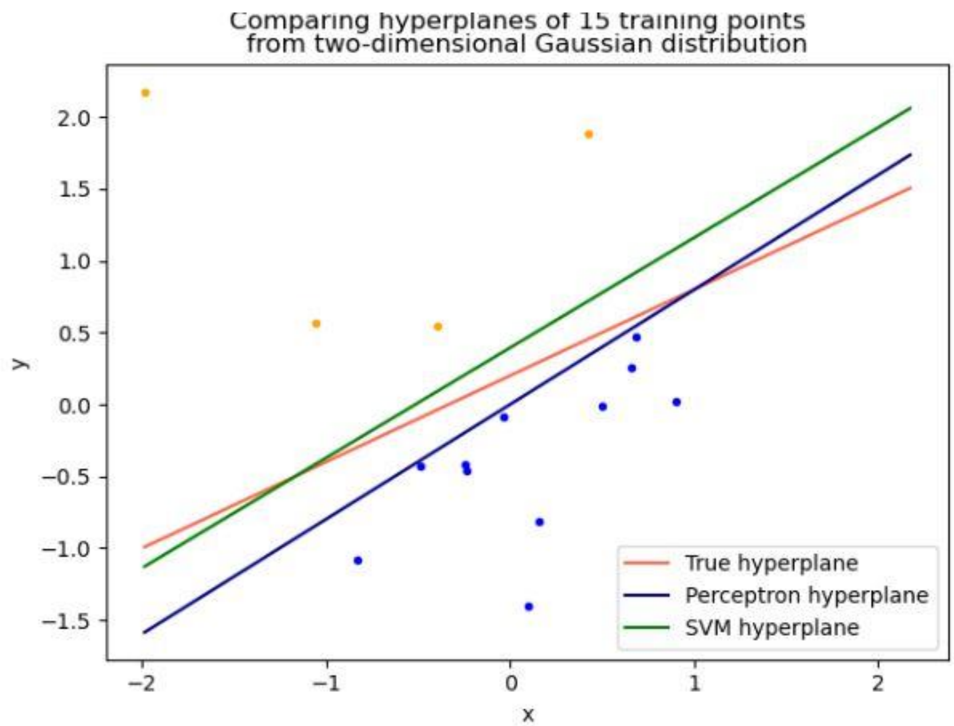
לירון גרשוני
308350503

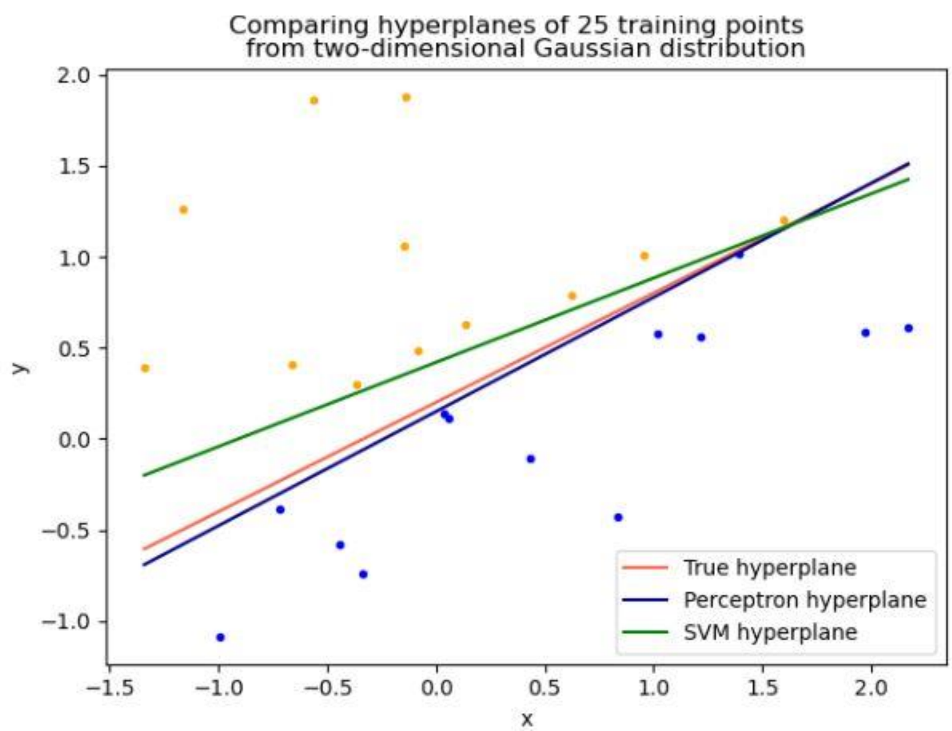
מערכות לומדות – המשך תרגיל בית 3

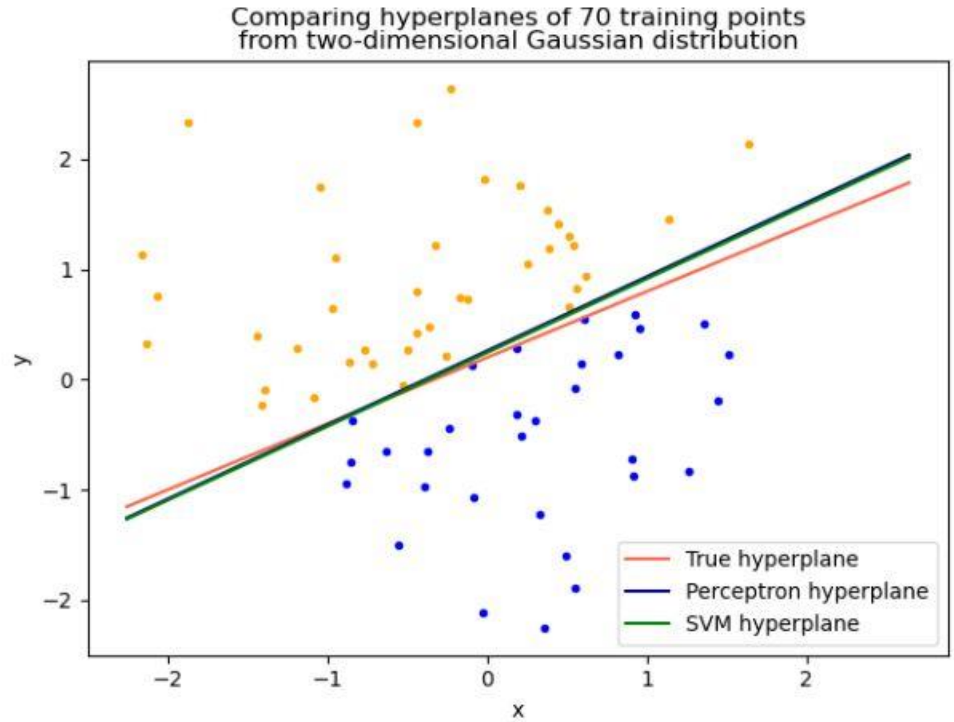
9)



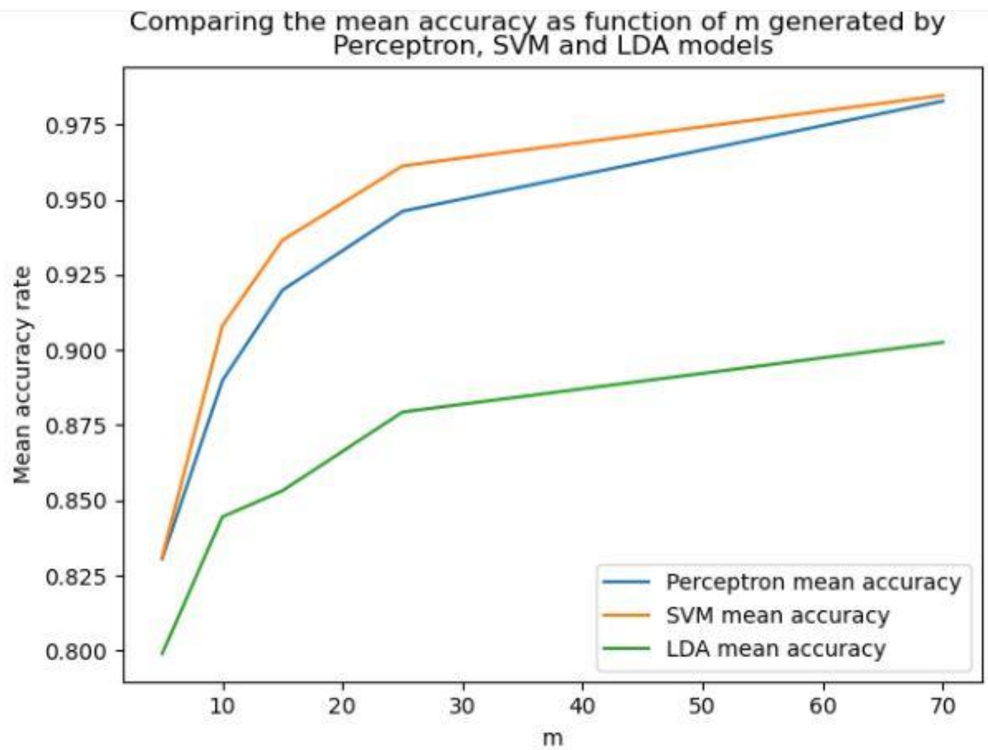



Comparing hyperplanes of 15 training points
from two-dimensional Gaussian distribution
y
x
True hyperplane
Perceptron hyperplane
SVM hyperplane


Comparing hyperplanes of 25 training points
from two-dimensional Gaussian distribution
y
x
True hyperplane
Perceptron hyperplane
SVM hyperplane


Comparing hyperplanes of 70 training points
from two-dimensional Gaussian distribution
y
x
True hyperplane
Perceptron hyperplane
SVM hyperplane

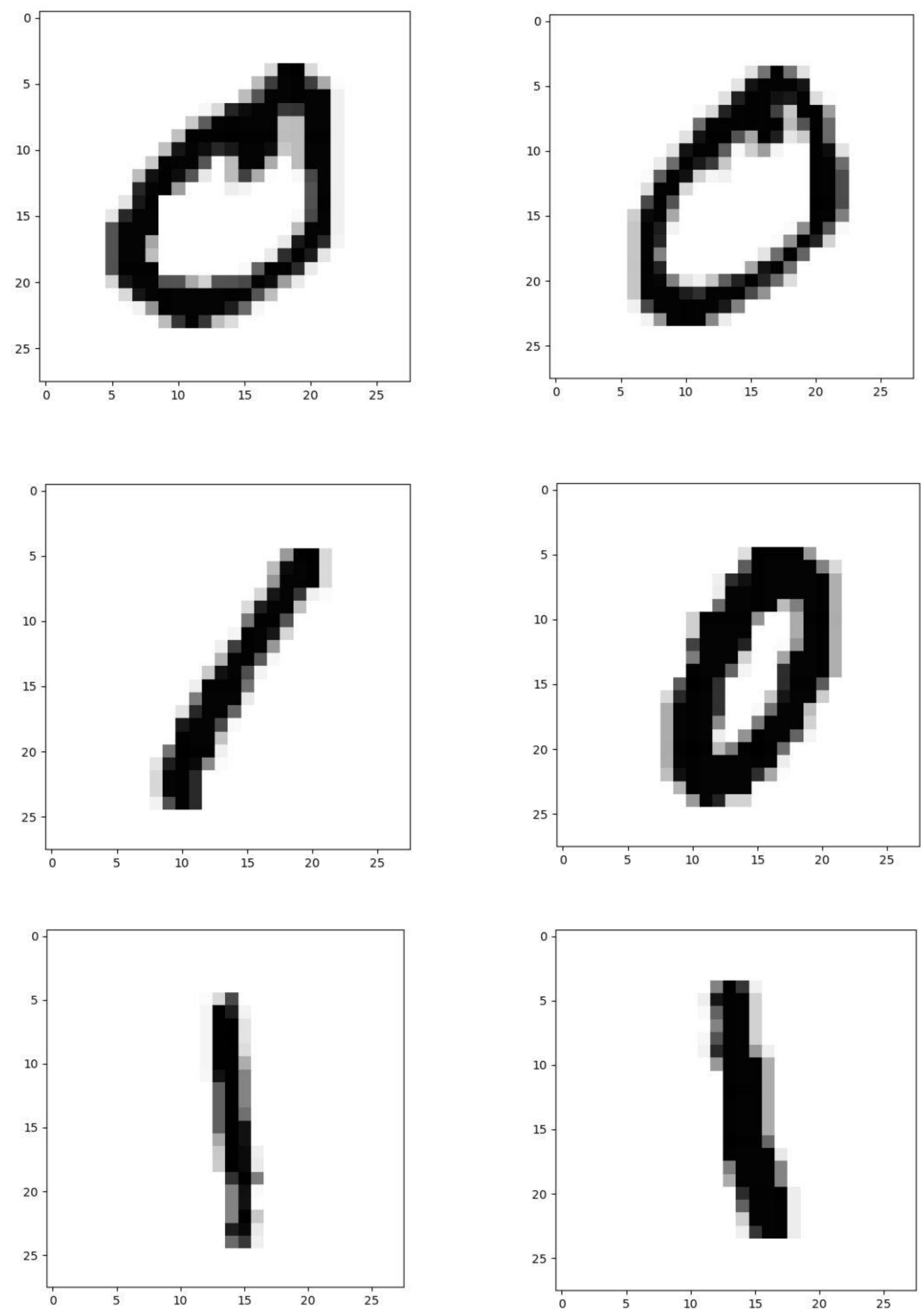
(10


Comparing the mean accuracy as function of m generated by
Perceptron, SVM and LDA models
Mean accuracy rate
m
Perceptron mean accuracy
SVM mean accuracy
LDA mean accuracy

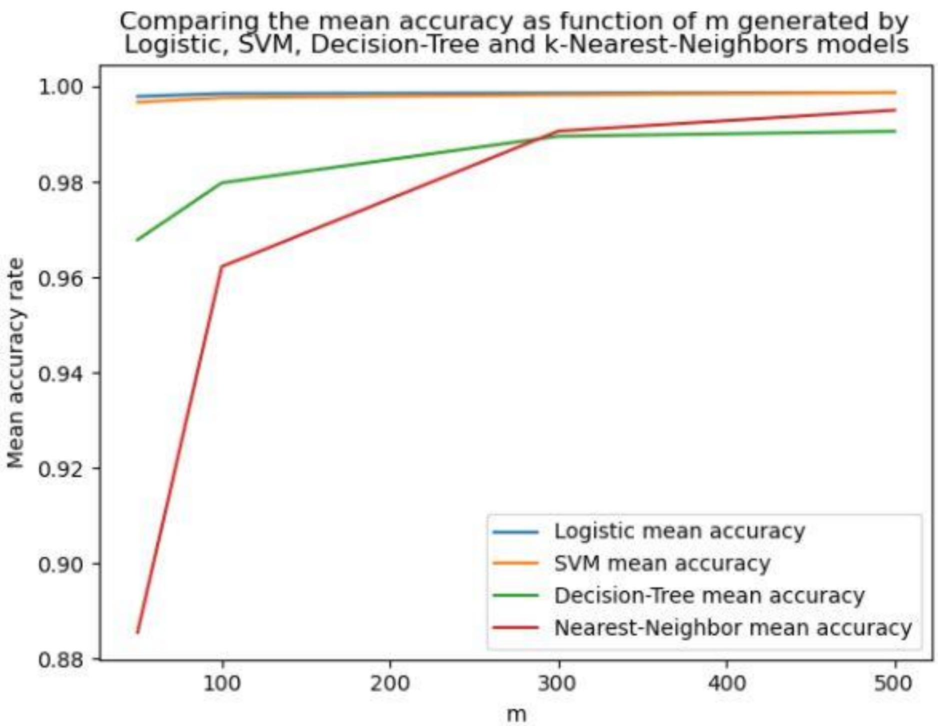
11) המסווג בעל התוצאות הטובות ביותר הוא מסווג SVM עבור ערכי ה-m שנבדקו אך עם זאת נראה שעבור m מספיק גדול, המסווג מסוג Perceptron מתכנס אליו בעוד שהמסווג מסוג LDA נותר הרבה מאחור ברמת הדיוק שלו מה שמעורר את השאלה עבור איזה m נתחיל לראות אותו מתכנס לאחרים.

אני חושב שהשוני בינהם נובע מכך שהמסווג מסוג SVM לוקח בחשבון שלו גם את השוליים (margins) עבור טעויות ולכן למעשה מוסיף נדבך שמדייק את התוצאות שלו עוד יותר. בהמשך אליו, המסווג מסוג Perceptron מהווה מתחרה טוב שכן הדאטה שאנו עובדים איתו ניתן להפרדה לינארית. לעומת זאת, וכפי שלמדנו, המסווג מסוג LDA מניח הנחות מקדימות על ההתפלגות ממנה הגיעו הנקודות וכן על מטריצת השונויות המשותפות ולכן, כאשר איננו יכולים באמת להניח זאת על הדאטה בו השתמשנו נקבל תוצאה שאינה מדויקת.

12)



(14



<u>מדידת זמני ריצה:</u>

<u>עבור m = 50:</u>
Logistic = 0.3725855
SVM = 1.8642795
Decision Tree = 0.4754824
KNearest Neighbors = 3.1184359

<u>עבור m = 100:</u>
Logistic = 0.8059226
SVM = 2.2935955
Decision Tree = 0.40316
KNearest Neighbors = 11.2997653

<u>עבור m = 300:</u>
Logistic = 1.343318
SVM = 3.1968263
Decision Tree = 0.5366957
KNearest Neighbors = 32.4929515

<u>עבור m = 500:</u>
Logistic = 1.7737247
SVM = 4.0327846
Decision Tree = 1.1298895
KNearest Neighbors = 68.4276903