# מערכות לומדות תרגיל 3

עמית בסקין 312259013

1. לכל $x \in \mathcal{X}$:

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

צ.להוכיח:

$$h_{\mathcal{D}}(\mathrm{x}) = \underset{y \in \{\pm 1\}}{\operatorname{argmax}} \Pr(\mathrm{x} \mid y) \Pr(y)$$

הוכחה:

ראשית נשים לב כי:

$$\Pr(\mathrm{x} \mid y)\Pr(y) = \frac{\Pr(\mathrm{x} \wedge y)}{\Pr(y)} \cdot \Pr(y) = \Pr(\mathrm{x} \wedge y)$$

כלומר בעצם צריך להוכיח ש־

$$h_{\mathcal{D}}(\mathrm{x}) = \underset{y \in \{\pm 1\}}{\operatorname{argmax}} \Pr(\mathrm{x} \wedge y)$$

יהא $\mathrm{x} \in \mathcal{X}$. נחלק למקרים:

- $\Pr(y=1 \mid \mathrm{x}) \geq \frac{1}{2}$, קרי:

$$\frac{\Pr(\mathrm{x} \wedge y = 1)}{\Pr(\mathrm{x})} \geq \frac{1}{2}$$

$$\Pr(\mathrm{x} \wedge y = 1) \geq \frac{1}{2}\Pr(\mathrm{x})$$

ובפרט:

$$\frac{\Pr(\mathrm{x} \wedge y = -1)}{\Pr(\mathrm{x})} \leq \frac{1}{2}$$

$$\Pr(\mathrm{x} \wedge y = -1) \leq \frac{1}{2}\Pr(\mathrm{x})$$

כלומר:

$$\Pr(\mathrm{x} \wedge y = 1) \geq \Pr(\mathrm{x} \wedge y = -1)$$

ולכן:

$$\underset{y\in\{\pm1\}}{\operatorname{argmax}} \Pr\left(\mathrm{x}\wedge y\right)=1$$

ומאחר שכל המעברים הם אםם אזי ש־ $\Pr\left(y=1\mid \mathrm{x}\right)\geq\frac{1}{2}$ אםם $\underset{y\in\{\pm1\}}{\operatorname{argmax}} \Pr\left(\mathrm{x}\wedge y\right)=1$.

מש"ל מקרה ראשון.

- $\Pr\left(y=1\mid \mathrm{x}\right)\leq\frac{1}{2}$, קרי:

$$\frac{\Pr\left(\mathrm{x}\wedge y=1\right)}{\Pr\left(\mathrm{x}\right)}\leq\frac{1}{2}$$

$$\Pr\left(\mathrm{x}\wedge y=1\right)\leq\frac{1}{2}\Pr\left(\mathrm{x}\right)$$

ובפרט:

$$\frac{\Pr\left(\mathrm{x}\wedge y=-1\right)}{\Pr\left(\mathrm{x}\right)}\geq\frac{1}{2}$$

$$\Pr\left(\mathrm{x}\wedge y=-1\right)\geq\frac{1}{2}\Pr\left(\mathrm{x}\right)$$

כלומר:

$$\Pr\left(\mathrm{x}\wedge y=1\right)\leq\Pr\left(\mathrm{x}\wedge y=-1\right)$$

ולכן:

$$\underset{y\in\{\pm1\}}{\operatorname{argmax}} \Pr\left(\mathrm{x}\wedge y\right)=-1$$

ומאחר שכל המעברים הם אםם אזי ש־ $\Pr\left(y=1\mid \mathrm{x}\right)\leq\frac{1}{2}$ אםם $\underset{y\in\{\pm1\}}{\operatorname{argmax}} \Pr\left(\mathrm{x}\wedge y\right)=-1$.

מש"ל מקרה שני.

מש"ל

2. נניח ש־ $\mathcal{X}=\mathbb{R}^d$ וש־ $\mathrm{x}\mid y\sim\mathcal{N}\left(\mu_y,\Sigma\right)$ עבור $\mu_y\in\mathbb{R}^d$ ו־ $\Sigma\in\mathbb{R}^d$, קרי פונקציית הצפיפות היא:

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

צ.להראות ש־

$$h_{\mathcal{D}}\left(\mathrm{x}\right)=\operatorname{argmax}\left\{-\frac{1}{2}\mathrm{x}^t\Sigma^{-1}\mathrm{x}-\frac{1}{2}\mu_y^t\Sigma^{-1}\mu_y+\ln\left(\Pr\left(y\right)\right)\mid y\in\{\pm1\}\right\}$$

הוכחה:

לפי נוסחת בייס:

$$\Pr\left(y\mid \mathrm{x}\right)\cdot\Pr\left(\mathrm{x}\right)=f\left(\mathrm{x}\mid y\right)\cdot\Pr\left(y\right)$$

קרי:

$$\Pr\left(y\mid \mathrm{x}\right)=\frac{f\left(\mathrm{x}\mid y\right)\cdot\Pr\left(y\right)}{\Pr\left(\mathrm{x}\right)}$$

נסמן $c=\Pr\left(\mathrm{x}\right)$ כקבוע כי מספר זה לא מושפע מ־ $y$. אז:

$$\Pr\left(y\mid \mathrm{x}\right)=c\cdot f\left(\mathrm{x}\mid y\right)\cdot\Pr\left(y\right)$$

ונציב את $f\left(\mathrm{x}\mid y\right)$:

$$\Pr\left(y\mid \mathrm{x}\right)=\frac{c}{\sqrt{\left(2\pi\right)^{d}\det\left(\Sigma\right)}}\Pr\left(y\right)\exp\left(-\frac{1}{2}\left(\mathrm{x}-\mu_{y}\right)^{t}\Sigma^{-1}\left(\mathrm{x}-\mu_{y}\right)\right)$$

ונסמן $c'=\frac{c}{\sqrt{\left(2\pi\right)^{d}\det\left(\Sigma\right)}}$ כקבוע כי לא תלוי ב־ $y$:

$$\Pr\left(y\mid \mathrm{x}\right)=c'\Pr\left(y\right)\exp\left(-\frac{1}{2}\left(\mathrm{x}-\mu_{y}\right)^{t}\Sigma^{-1}\left(\mathrm{x}-\mu_{y}\right)\right)$$

$$\ln\left(\Pr\left(y\mid \mathrm{x}\right)\right)=\ln\left(c'\right)+\ln\left(\Pr\left(y\right)\right)-\frac{1}{2}\left(\mathrm{x}-\mu_{y}\right)^{t}\Sigma^{-1}\left(\mathrm{x}-\mu_{y}\right)=$$

$$=\ln\left(c'\right)+\ln\left(\Pr\left(y\right)\right)-\frac{1}{2}\mathrm{x}^{t}\Sigma^{-1}\mathrm{x}-\frac{1}{2}\mu_{y}^{t}\Sigma^{-1}\mu_{y}+x^{t}\Sigma^{-1}\mu_{y}$$

ונסמן:

$$c''=\ln\left(c'\right)-\frac{1}{2}\mathrm{x}^{t}\Sigma^{-1}\mathrm{x}$$

כי לא תלויים ב־ $y$:

$$\Pr\left(y\mid \mathrm{x}\right)=x^{t}\Sigma^{-1}\mu_{y}-\frac{1}{2}\mu_{y}^{t}\Sigma^{-1}\mu_{y}+\ln\left(\Pr\left(y\right)\right)+c''$$

עתה, בסעיף הקודם ראינו ש־

$$h_{\mathcal{D}}\left(\mathrm{x}\right)=\underset{y\in\{\pm1\}}{\operatorname{argmax}}\Pr\left(\mathrm{x}\mid y\right)\Pr\left(y\right)$$

אבל למקסם את $\Pr\left(\mathrm{x}\mid y\right)\Pr\left(y\right)$ זה כמו למקסם את $\ln\left(\Pr\left(\mathrm{x}\mid y\right)\Pr\left(y\right)\right)$ קרי:

$$\ln\left(\Pr\left(\mathrm{x}\mid y\right)\right)+\ln\left(\Pr\left(y\right)\right)$$

קרי את:

$$x^{t}\Sigma^{-1}\mu_{y}-\frac{1}{2}\mu_{y}^{t}\Sigma^{-1}\mu_{y}+2\ln\left(\Pr\left(y\right)\right)+c''$$

ומאחר ש־ $c''$ לא תלוי ב־ $y$ אזי שזה כמו למקסם את:

$$x^{t}\Sigma^{-1}\mu_{y}-\frac{1}{2}\mu_{y}^{t}\Sigma^{-1}\mu_{y}+2\ln\left(\Pr\left(y\right)\right)$$

מש"ל

3. נתונה קבוצת דגימות $S=\{(\mathrm{x}_1,y_1),\dots,(\mathrm{x}_m,y_m)\}$. צ.להעריך את $\mu_{+1},\mu_{-1},\Sigma,\Pr(y)$ בהתבסס על $S$.

פיתרון:

נניח בה"כ שקיים $l\le m$ כך שלכל $i\in[l]$ מתקיים $y_i=1$.

נסמן:

$$S_{+1}=\{\mathrm{x}_i\mid i\in[l]\}$$

$$S_{-1}=\{\mathrm{x}_i\mid i\in[m]\setminus[l]\}$$

נסמן ב־ $X_{+1}$ ו־ $X_{-1}$ את המטריצות שהשורות שלהן הן הוקטורים ב־ $S_{+1}$ ו־ $S_{-1}$ בהתאמה.

נסמן ב־ $\mu_{+1}$ ו־ $\mu_{-1}$ את וקטורי הממוצעים של הוקטורים ב־ $S_{+1}$ ו־ $S_{-1}$ בהתאמה.

יהא $\mu$ וקטור הממוצעים של $X$, ונסמן ב־ $\widehat{X}$ את $X$ ממורכזת לפי $\mu$.

נסמן ב־ $X_{+1}^0$ ו־ $X_{-1}^0$ את מטריצות הוקטורים $X_{+1}$ ו־ $X_{-1}$ בהתאמה, ממורכזים לפי $\mu_{+1}$ ו־ $\mu_{-1}$, בהתאמה.

נסמן ב־ $\Sigma_{+1}$ ו־ $\Sigma_{-1}$ את מטריצות השונות המשותפת של $S_{+1}$ ו־ $S_{-1}$ בהתאמה, קרי:

$$\Sigma_{+1}=\frac{1}{n_{+1}}\left(X_{+1}^0\right)^t X_{+1}^0$$

$$\Sigma_{-1}=\frac{1}{n_{-1}}\left(X_{-1}^0\right)^t X_{-1}^0$$

ואז מטריצת השונות המשותפת הכוללת מחושבת על ידי:

$$\Sigma_{ij}=\frac{1}{m}\left(n_{+1}\left(\Sigma_{+1}\right)_{ij}+n_{-1}\left(\Sigma_{-1}\right)_{ij}\right)$$

נסמן ב־ $n_{+1}$ ו־ $n_{-1}$ את $|S_{+1}|$ ו־ $|S_{-1}|$ בהתאמה.

נסמן $p_{+1}=\Pr(+1)$ ו־ $p_{-1}=\Pr(-1)$ ונעריך:

$$p_{+1}=\frac{n_{+1}}{m}$$

$$p_{-1}=\frac{n_{-1}}{m}$$

מש"ל

4. ישנן שתי טעויות אפשריות בסיווג ספאם:

- לסווג הודעה כספאם למרות שהיא לא.

- לסווג כלא ספאם הודעה שהיא כן ספאם.

הטעות שאנחנו הכי לא רוצים לעשות זה לסווג הודעה כספאם למרות שהיא לא. זו טעות חמורה יותר מאשר לסווג כלא ספאם הודעה שהיא כן ספאם. אכן, אם נסווג הודעה כלא ספאם למרות שהיא ספאם, אז המשתמש ייאלץ למרבה הצער להיתקל בספאם למרות שהיה מעדיף שלא. לעומת זאת, אם ישנה הודעה שהיא לא ספאם אבל סווגה כספאם, כנראה שהמשתמש יפספס את ההודעה, וזאת על אף שפוטנציאלית מדובר בהודעה חשובה שהמשתמש צריך לקרוא, והנזק שעלול להיגרם למשתמש כתוצאה מפספוס ההודעה, עשוי להיות גדול בהרבה מלקרוא הודעת ספאם שהיה מעדיף שלא לקרוא.

אם כן, אנחנו רוצים שהשגיאה החמורה תהיה שגיאה מסוג ראשון, קרי FP, ולכן ניתן לסיווג הודעה כספאם את הלייבל positive, כלומר אם נסווג הודעה שהיא לא ספאם ב־ positive, אבל טעינו False, אז קיבלנו שגיאה מסוג ראשון, כפי שרצינו.

5. הצורה הקאנונית של תכנית ריבועית היא:

$$\underset{\mathrm{v}\in\mathbb{R}^n}{\operatorname{argmin}}\left(\frac{1}{2}\mathrm{v}^tQ\mathrm{v}+\mathrm{a}^t\mathrm{v}\right)$$

כך ש־

$$A\mathrm{v}\leq\mathrm{d}$$

עבור:

$$Q\in\mathbb{R}^{n\times n},A\in\mathbb{R}^{m\times n},\mathrm{a}\in\mathbb{R}^n,\mathrm{d}\in\mathbb{R}^m$$

נתבונן בבעיית ה־ Hard-SVM:

$$\underset{(\mathrm{w},b)}{\operatorname{argmin}}\ \|\mathrm{w}\|^2$$

כך שלכל $i$:

$$y_i\left(\langle\mathrm{w},\mathrm{x}_i\rangle+b\right)\geq 1$$

צ.לכתוב אותה כתכנית ריבועית קאנונית.

פיתרון:

נניח שיש לנו $m$ דגימות $(\mathrm{x}_i,y_i)$ כאשר כל $\mathrm{x}_i$ הוא מסדר $n-1$.

נסמן ב־ $A'$ מטריצה $m\times n-1$ שהשורות שלה הם הוקטורים $y_i\mathrm{x}_i$.

נסמן ב־ $A''$ את המטריצה שמתקבלת מ־ $A'$ על ידי הוספת עמודה שהכניסות בה הן $y_i$, וניקח $A=-A''$.

נסמן $Q'=I_{(n-1)\times(n-1)}$, ונתאר את $Q$ כמטריצת בלוקים שנתונים באלכסון הראשי: הבלוק הראשון הוא $Q'$ והבלוק השני הוא הסקלר 0.

את a ניקח להיות $0_n$ ואת d ניקח להיות $(-1)_m$

אז התכנית הריבועית שלנו היא:

$$\underset{\mathrm{v}\in\mathbb{R}^n}{\operatorname{argmin}}\left(\frac{1}{2}\mathrm{v}^tQ\mathrm{v}+\mathrm{a}^t\mathrm{v}\right)$$

קרי:

$$Q=2\begin{pmatrix}1&&&\\&\ddots&&\\&&1&\\&&&0\end{pmatrix},\mathrm{v}=\mathrm{w}=\begin{pmatrix}w_1\\\vdots\\w_n\\b\end{pmatrix},\mathrm{a}=\begin{pmatrix}0\\\vdots\\0\end{pmatrix},A=-\begin{pmatrix}y_1&y_1x_{11}&\cdots&y_1x_{1(n-1)}\\\vdots&\vdots&\ddots&\vdots\\y_m&y_mx_{m1}&\cdots&y_mx_{m(n-1)}\end{pmatrix},\mathrm{d}=-\begin{pmatrix}1\\\vdots\\1\end{pmatrix}$$

כלומר:

$$\frac{1}{2}\mathrm{v}^tQ\mathrm{v}+\mathrm{a}^t\mathrm{v}=$$

$$=\frac{1}{2}\begin{pmatrix}w_1&\cdots&w_n\end{pmatrix}2\begin{pmatrix}1&&&\\&\ddots&&\\&&1&\\&&&0\end{pmatrix}\begin{pmatrix}b\\w_1\\\vdots\\w_n\end{pmatrix}+\begin{pmatrix}0&\cdots&0\end{pmatrix}\begin{pmatrix}w_1\\\vdots\\w_n\\b\end{pmatrix}=$$

$$=\begin{pmatrix}w_1&\cdots&w_n\end{pmatrix}\begin{pmatrix}w_1\\\vdots\\w_n\end{pmatrix}=\|\mathrm{w}\|^2$$

כאשר:

$$A\mathrm{v}\leq\mathrm{d}$$

קרי:

$$-\begin{pmatrix}y_1&y_1x_{11}&\cdots&y_1x_{1(n-1)}\\\vdots&\vdots&\ddots&\vdots\\y_m&y_mx_{m1}&\cdots&y_mx_{m(n-1)}\end{pmatrix}\begin{pmatrix}b\\w_1\\\vdots\\w_n\end{pmatrix}\leq-\begin{pmatrix}1\\\vdots\\1\end{pmatrix}$$

$$\begin{pmatrix}\langle y_1\mathrm{x}_1,\mathrm{w}\rangle+y_1b\\\vdots\\\langle y_m\mathrm{x}_m,\mathrm{w}\rangle+y_mb\end{pmatrix}\geq\begin{pmatrix}1\\\vdots\\1\end{pmatrix}$$

$$\begin{pmatrix}y_1\langle\mathrm{w},\mathrm{x}_1\rangle+y_1b\\\vdots\\y_m\langle\mathrm{w},\mathrm{x}_m\rangle+y_mb\end{pmatrix}\geq\begin{pmatrix}1\\\vdots\\1\end{pmatrix}$$

$$\begin{pmatrix}y_1\left(\langle\mathrm{w},\mathrm{x}_1\rangle+b\right)\\\vdots\\y_m\left(\langle\mathrm{w},\mathrm{x}_m\rangle+b\right)\end{pmatrix}\geq\begin{pmatrix}1\\\vdots\\1\end{pmatrix}$$

וקיבלנו שלכל $i$ צריך להתקיים:

$$y_i\left(\langle\mathrm{w},\mathrm{x}_i\rangle+b\right)\geq1$$

כפי שרצינו.

מש"ל.

6. נתבונן בבעיית ה־ Soft-SVM:

$$\underset{\mathrm{w},\{\xi_i\}}{\operatorname{argmin}}\left(\frac{1}{2}\lambda\left\|\mathrm{w}\right\|^2+\frac{1}{m}\sum_{i=1}^{m}\xi_i\right)$$

כך שלכל $i$:

$$y_i\left\langle \mathrm{w},\mathrm{x}_i\right\rangle \geq 1-\xi_i$$

$$\xi_i\geq 0$$

צ.להוכיח שהיא שקולה לבעייה:

$$\underset{\mathrm{w}}{\operatorname{argmin}}\left(\frac{1}{2}\lambda\left\|\mathrm{w}\right\|^2+\frac{1}{m}\sum_{i=1}^{m}l^{hinge}\left(y_i\left\langle \mathrm{w},\mathrm{x}_i\right\rangle\right)\right)$$

כאשר:

$$l^{hinge}\left(y_i\left\langle \mathrm{w},\mathrm{x}_i\right\rangle\right)=\max\left\{0,1-y_i\left\langle \mathrm{w},\mathrm{x}_i\right\rangle\right\}$$

הוכחה:

מתקיים:

$$\sum_{i=1}^{m}l^{hinge}\left(y_i\left\langle \mathrm{w},\mathrm{x}_i\right\rangle\right)=\sum_{i,y_i\left\langle \mathrm{w},\mathrm{x}_i\right\rangle<1}\left(1-y_i\left\langle \mathrm{w},\mathrm{x}_i\right\rangle\right)$$

כלומר הבעיה הנ"ל שקולה ל־

$$\underset{\mathrm{w}}{\operatorname{argmin}}\left(\frac{1}{2}\lambda\left\|\mathrm{w}\right\|^2+\frac{1}{m}\sum_{i,y_i\left\langle \mathrm{w},\mathrm{x}_i\right\rangle<1}\left(1-y_i\left\langle \mathrm{w},\mathrm{x}_i\right\rangle\right)\right)$$

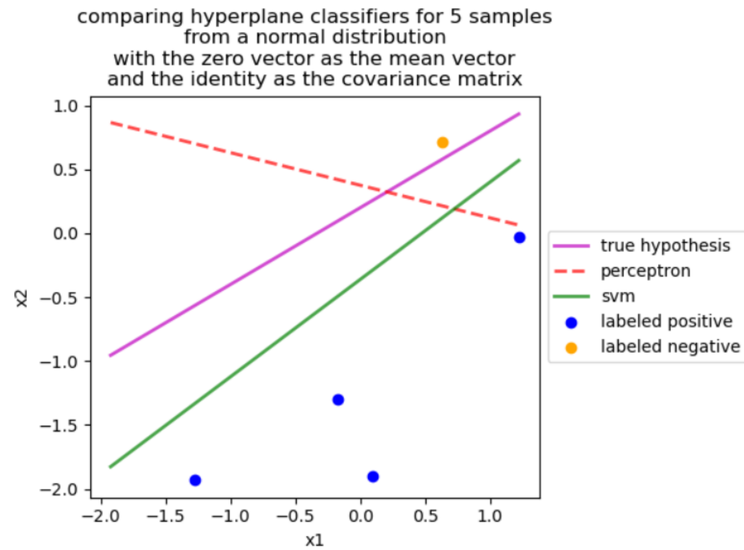
עתה:

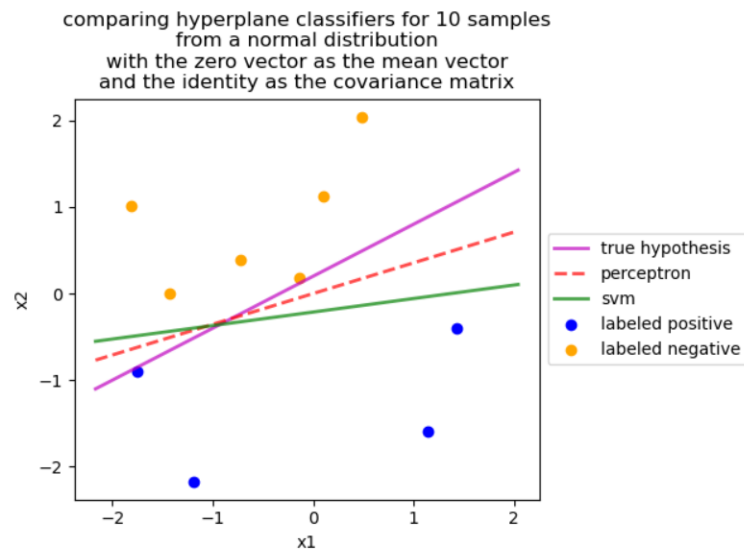
$$y_i\left\langle \mathrm{w},\mathrm{x}_i\right\rangle \geq 1-\xi_i \implies \xi_i\geq 1-y_i\left\langle \mathrm{w},\mathrm{x}_i\right\rangle$$

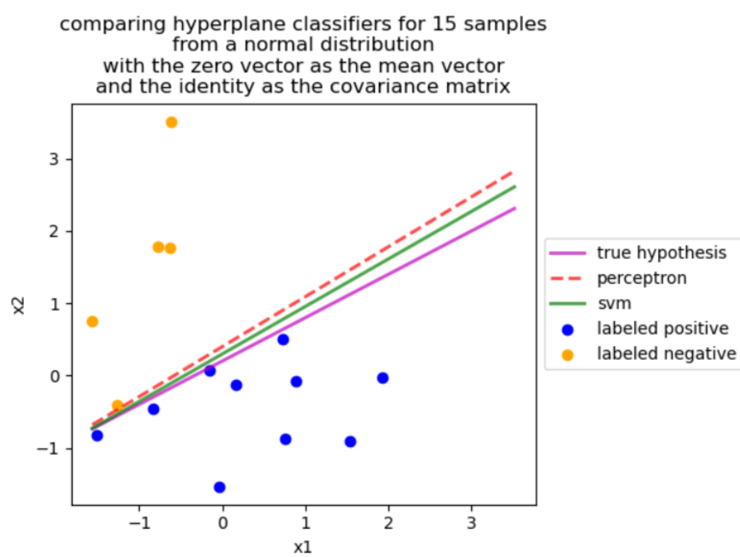
ולכן כשאנחנו ממזערים $\sum_{i=1}^{m}\xi_i$ אנחנו למעשה גם ממזערים את $1-y_i\left\langle \mathrm{w},\mathrm{x}_i\right\rangle$ ולכן המינימומים של שתי הבעיות מתקבלים באותן נקודות.

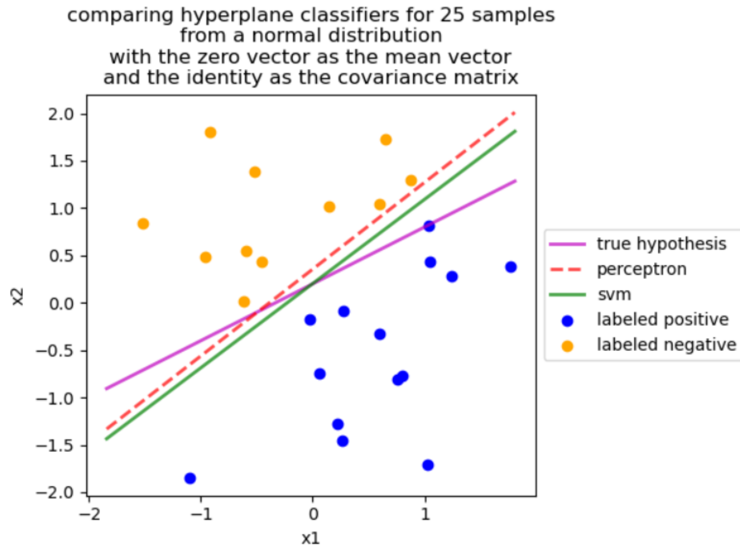
מש"ל

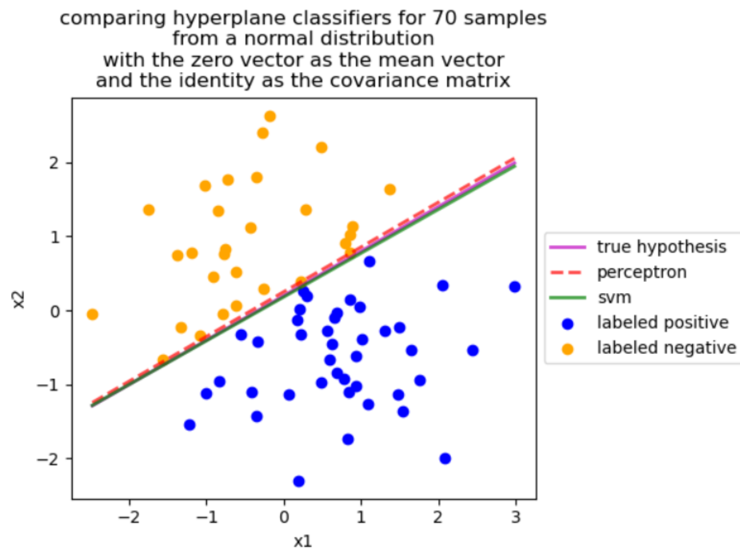
.10 הגרפים להלן:



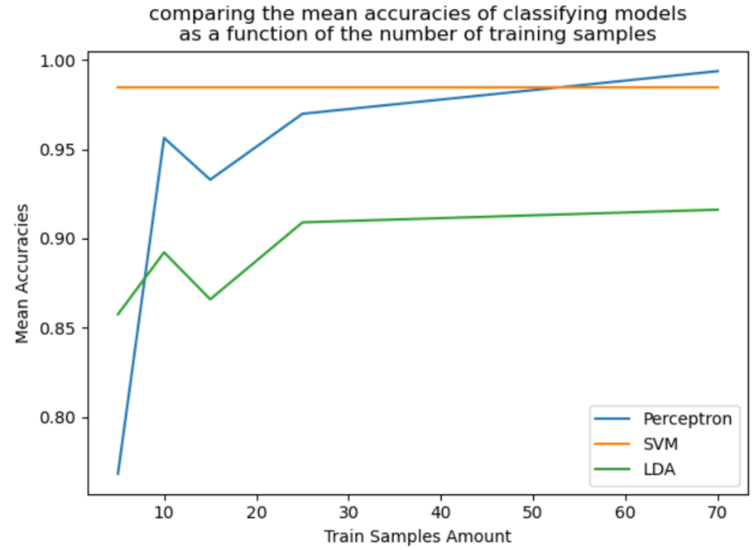




comparing hyperplane classifiers for 25 samples
from a normal distribution
with the zero vector as the mean vector
and the identity as the covariance matrix
x1
x2
true hypothesis
perceptron
svm
labeled positive
labeled negative


comparing hyperplane classifiers for 70 samples
from a normal distribution
with the zero vector as the mean vector
and the identity as the covariance matrix
x1
x2
true hypothesis
perceptron
svm
labeled positive
labeled negative

10. הגרף להלן:



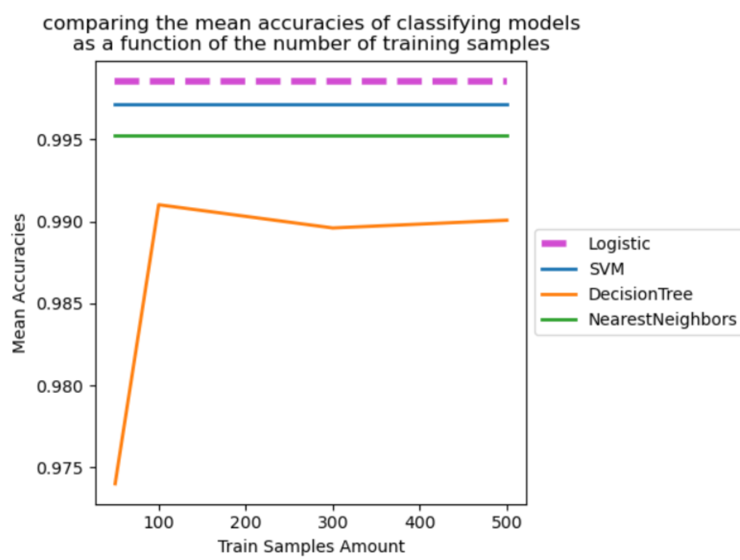

11. בממוצע, ל־ SVM יש את הביצוע הטוב ביותר, אף על פי שבאזור ה־ 55 דגימות, הפרספטרון עוקף אותו.

הסיבה לכך היא ש־ SVM מוצא את הקו שממקסם את את רוחב השוליים, מה שנותן לו מעין "טווח ביטחון" עבור הדגימות הבאות.

אכן, ככל שהשוליים צרים יותר כך גדלה ההסתברות לשגיאה, ולכן שוליים מקסימליים ממזערים את ההסתברות לכך (קרי שטח גדול יותר שמאפשר תנודות של הדגימות).

14. הגרף להלן:

עבור המימוש של Logistic, SVM, DecisionTree לקחתי את אותו המימוש מהקובץ models.py כאשר עבור עומק מקסימלי של DesicionTree לקחתי עומק של 5. כמו כן, עבור NearestNeighbors לקחתי את המימוש של NearestCentroid שעבור כל דגימה לוקח את הלייבל של קבוצת הדגימות שהצנטרואיד הממוצע שלהם הוא הקרוב ביותר לדגימה שמסתכלים עליה.




זמני הריצה להלן:

```
Logistic, 50 training samples: 0.17 seconds
Logistic, 100 training samples: 0.19 seconds
Logistic, 300 training samples: 0.19 seconds
Logistic, 500 training samples: 0.17 seconds
SVM, 50 training samples: 3.58 seconds
SVM, 100 training samples: 3.56 seconds
SVM, 300 training samples: 3.53 seconds
SVM, 500 training samples: 3.53 seconds
DecisionTree, 50 training samples: 0.37 seconds
DecisionTree, 100 training samples: 0.39 seconds
DecisionTree, 300 training samples: 0.37 seconds
DecisionTree, 500 training samples: 0.36 seconds
NearestNeighbors, 50 training samples: 0.47 seconds
NearestNeighbors, 100 training samples: 0.48 seconds
NearestNeighbors, 300 training samples: 0.47 seconds
NearestNeighbors, 500 training samples: 0.47 seconds
```

ניתן לראות שבאחוזים אלגוריתם ה־ SVM לוקח משמעותית יותר זמן משאר האלגוריתמים.

מכאן ניתן לשער שייתכן כי הדבר טמון בכך שמקסום השוליים יותר קשה לחישוב משאר החישובים שהאלגוריתמים האחרים עושים.