יובל דלוס 305211880

מבוא למערכות לומדות – תרגיל 3

1. If we knew $\mathcal{D}$, our best predictor would have been assigning the class with the higher probability:

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

where the probability is over $\mathcal{D}$. This classifier is known as the **Bayes Optimal** classifier.

Show that

$$h_{\mathcal{D}} = \underset{y\in\{\pm1\}}{\operatorname{argmax}} \Pr(\mathbf{x}|y)\Pr(y).$$

1. ראשית ניזכר בנוסחת בייס ונוסחת ההסתברות השלמה:

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

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

נשתמש בנוסחאות ונפתח את ההגדרה עבור h_D:

$$\underset{y=\{\pm1\}}{argmax} P(x \mid y) \cdot P(y) \overset{\substack{bayes \\ rule}}{\cong} argmax \frac{P(y \mid x) \cdot P(x)}{P(y)} \cdot P(y) = argmax\, P(y \mid x) \cdot P(x) \overset{\text{הסתברות שלמה}}{\cong}$$

$$= argmax\left(P(y=1 \mid x) + P(y=-1 \mid x)\right) \cdot P(x)$$

כעת נבחין בין שני מקרים אפשריים:

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

לפי הגדרת מרחב הסתברות כאשר יש רק שתי אפשרויות ל- y, נקבל חלוקה שלמה שלו ונקבל כי:

$$argmax\left(P(y=1 \mid x) + P(y=-1 \mid x)\right) \cdot P(x) = 1$$

במקרה השני $P(y=1 \mid x) < \frac{1}{2}$ אזי $P(y=-1 \mid x) \geq \frac{1}{2}$ ונקבל:

$$argmax\left(P(y=1 \mid x) + P(y=-1 \mid x)\right) \cdot P(x) = -1$$

קיבלנו שההגדרה שיולה לכל h_D כנדרש.

יובל דלוס 305211880

2. Assume that $\mathcal{X} = \mathbb{R}^d$ and that $\mathbf{x}|y \sim \mathcal{N}(\mu_y, \Sigma)$ for some mean vector $\mu_y \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ (that is, the covariance matrix Σ is the same for both $y \in \{\pm 1\}$, but the expectation μ_y is different for each $y \in \{\pm 1\}$). In other words,

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

where f is the density function for the multivariate normal distribution. Show that in this case, if we knew μ_{+1}, μ_{-1} and Σ then the Bayes Optimal classifier is

$$h_{\mathcal{D}}(\mathbf{x}) = \underset{y\in\{\pm 1\}}{\operatorname{argmax}}\, \delta_y(\mathbf{x}),$$

where δ_{+1} and δ_{-1} are functions $\mathbb{R}^d \to \mathbb{R}$ given by

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

.2

$$h_D(x) = argmax\ \delta_y(x) = argmax\ \ x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\,\Sigma^{-1}\,\mu_y + \ln P(y)$$

וכעת, מנוסחאת בייס :

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

נציב ונקבל :

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

נפעיל חוקי ln :

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

כעת שנשים לב כי הביטוי $\ln\big(f_X(x)\big)$ הינו קבוע ונוכל להשמיטו :

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

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

$$= argmax\ \ x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\,\Sigma^{-1}\,\mu_y + \ln\left(f_{Y|X=x}(y)\right) - \ln\left(\frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}}\exp\left\{-\frac{1}{2}\left(x-\mu_y\right)^T\Sigma^{-1}\left(x-\mu_y\right)\right\}\right)$$

נשים לב כי הביטוי $\ln\left(\frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}}\right)$ קבוע ונוכל להשמיטו :

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

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

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

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

ומכיוון שהביטוי $\frac{1}{2}x^T\Sigma^{-1}x$ הוא קבוע:

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

נשים לב כי הביטוי Σ^{-1} הינו מטריצה סימטרית ונקבל:

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

$$= argmax\ \ \ln\left(f_{Y|X=x}(y)\right)$$

וממונוטוניות פונקציית ה- ln נקבל:

$$= argmax\ \ P(y\mid x) \overset{\substack{bayes\\ rule}}{\cong} P(x\mid y)\cdot P(y) \overset{\text{לפי שאלה 1}}{\cong} h_D$$

כנדרש.

3. In practice, we don't know μ_{+1}, μ_{-1}, Σ and $\Pr(y)$. In order to turn the above into a classifier, given a training set $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$, we need to estimate them. Write your formula for estimating μ_{+1}, μ_{-1}, Σ and $\Pr(y)$ based on S.

3. בקורס הסתברות למדנו כי אומד בלתי מוטה לתוחלת צריך להיות הממוצע, ובהתאם נגדיר את האומדים הנ"ל:

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

ולכן האומד הבלתי תלוי לשונות יהיה:

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

לבסוף עבור הסתברות $P(y)$ נוכל לבחור בהסתברות ברנולי עם פרמטר p כך ש-

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

יובל דלוס 305211880

Spam

[Relevant material - Lecture 3]

4. You are building a spam filter - a classifier that receives an email and decides whether it's a spam message or not. What are the two kinds of errors that your classifier could make? Which of them is the error we really don't want to make? Which of the labels {spam, not-spam} should be the **negative** label and which should be the **positive** label, if we want the false-positive error (Type-I error) to be the error we really don't want to make?

4. כשאנו רוצים לסנן מיילים לפי הפרמטר האם הם ספאם או לא נצטרך להבחין בין שתי עלויות:
הראשונה, כשאנו קובעים על מייל מסוים שהוא ספאם כאשר בפועל הוא אינו כזה, לכן אנו "מפספסים" משהו שהיה יכול להיות חשוב לנו.
השניים, שאנו מדווחים על מייל מסוים שהוא תקין אבל בפועל הוא ספאם, ובזבזנו עליו את הזמן שלנו.
אם כך, הטעות שנרצה להימנע ממנה יותר היא הטעות הראשונה, תיוג מייל חשוב כספאם, לכן נסמן את +1 להיות התגית "ספאם" ו- -1 לתגית "אינו ספאם".

SVM- Formulation

[Relevant material - Recitation 4]

5. The canonical form of a Quadratic Program (QP) is:

$$\underset{\mathbf{v}\in\mathbb{R}^n}{\operatorname{argmin}}\left(\frac{1}{2}\mathbf{v}^\top Q\mathbf{v}+\mathbf{a}^\top\mathbf{v}\right)$$
$$\text{s.t. } A\mathbf{v}\leqslant\mathbf{d},$$

where $Q\in\mathbb{R}^{n\times n}, A\in\mathbb{R}^{m\times n}, \mathbf{a}\in\mathbb{R}^n, \mathbf{d}\in\mathbb{R}^m$ are fixed vectors and matrices.

Write the Hard-SVM problem as a QP problem in canonical form. Specifically, using the Hard-SVM problem formulation

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

what are the values of $Q, A, \mathbf{a}, \mathbf{d}$ that express this problem as a QP in canonical form?

Why is it interesting? QP software solvers take QP in canonical form. To use a QP solver, you'll need to express the SVM problem as QP in canonical form as above.

5. ראשית, ניזכר כי מתקיים $||w||^2=w^Tw=w^TI_nw$,
מעבר לכך, נשים לב שכל קורדינאטה $y_i\langle w,x_i\rangle$ מיוצגת ע"י $(x_iy_i)^T$ ואנו מנסים למזער את השגיאה כמה שיותר ולכן נבחר $a=0$ ונקבל:

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

$$=argmin_{(w,b)}(w\ \ b)I_n\binom{w}{b}\ s.t\begin{bmatrix}x_1y_1 & 1\\ \vdots & \vdots\\ x_my_m & 1\end{bmatrix}\binom{w}{b}\geq\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$$

נמיר זאת לצורה הקנונית ונקבל:

$$= argmin_{(w,b)} \frac{1}{2} \overbrace{(w \;\; b)}^{V^T} \overbrace{2I_n}^{Q} \overbrace{\begin{pmatrix} w \\ b \end{pmatrix}}^{V} + \overbrace{0}^{a} \overbrace{\begin{pmatrix} w \\ b \end{pmatrix}}^{V} \;\; s.t$$

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

6. In the Soft-SVM we defined the problem:

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

Show that this problem is equivalent to the problem (namely that these problem have the same solutions)

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

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

6. ראשית, נשים לב כי השוני בין הביטויים נובע מהאיבר השני בחיבור ולא מהאיבר הראשון $\left(\frac{\lambda}{2}\|w\|^2\right)$ הזהה בשניהם. על פי הביטוי הראשון מתקיים כי

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

ועי״י העברת אגפים נקבל:

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

מצד שני, נתון כי $\xi_i \geq 0$ לכל i, ולכן נוכל להגדיר:

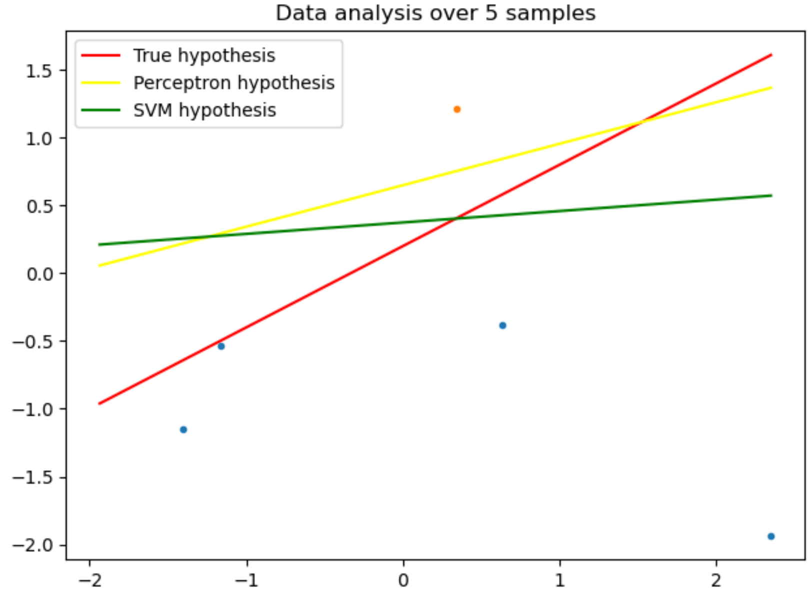
$$\xi_i \geq \max\{0, 1 - y_i\langle w, x_i\rangle\}$$

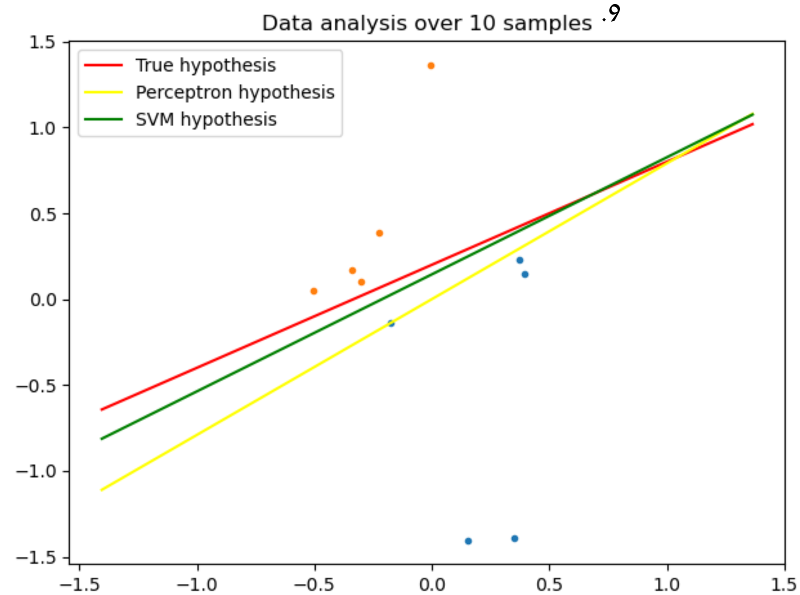
נרצה למזער את האיבר השני בביטוי ולכן נוכל לבחור $a = y_i\langle w, x_i\rangle$ ונקבל:

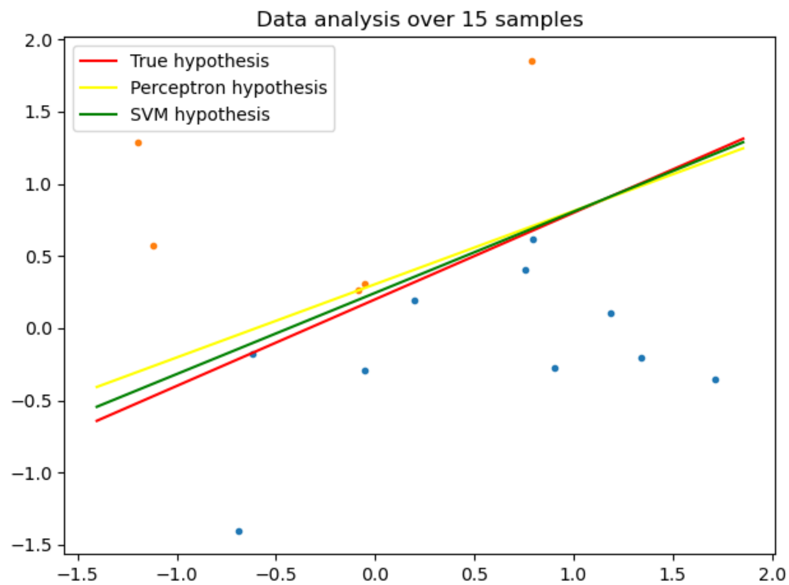
$$argmin_w \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} \ell^{hinge}(a)$$

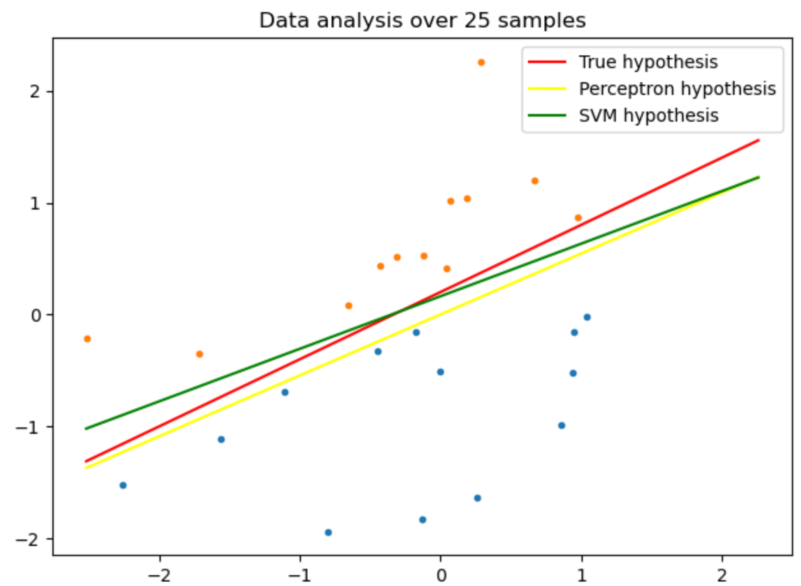
כנדרש.

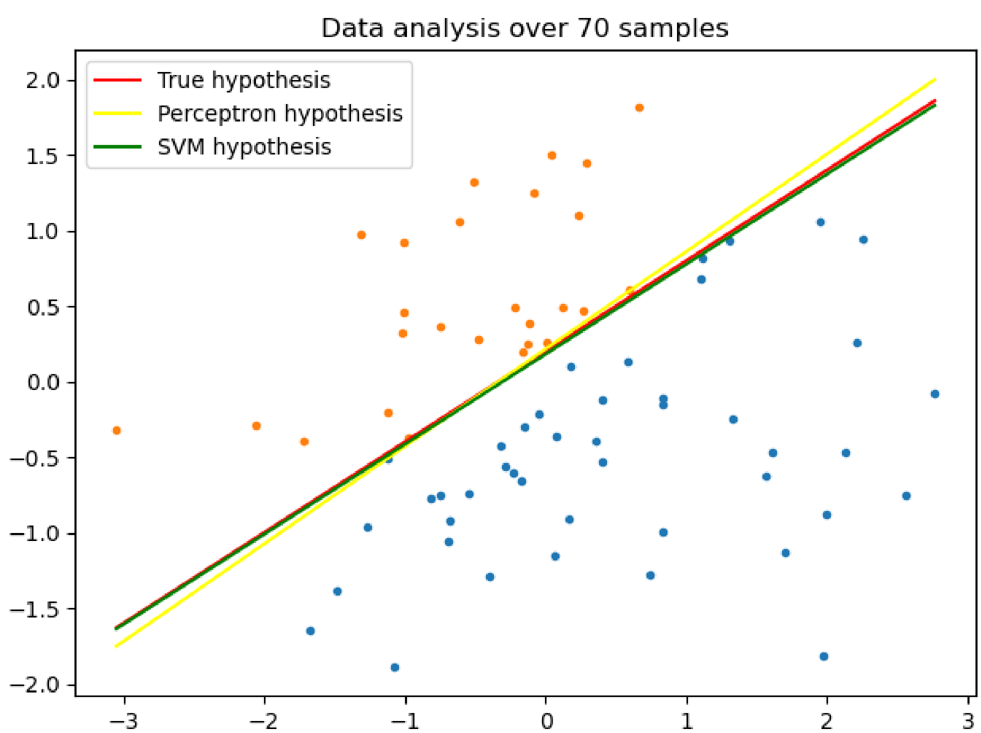
9.


Data analysis over 5 samples
True hypothesis
Perceptron hypothesis
SVM hypothesis

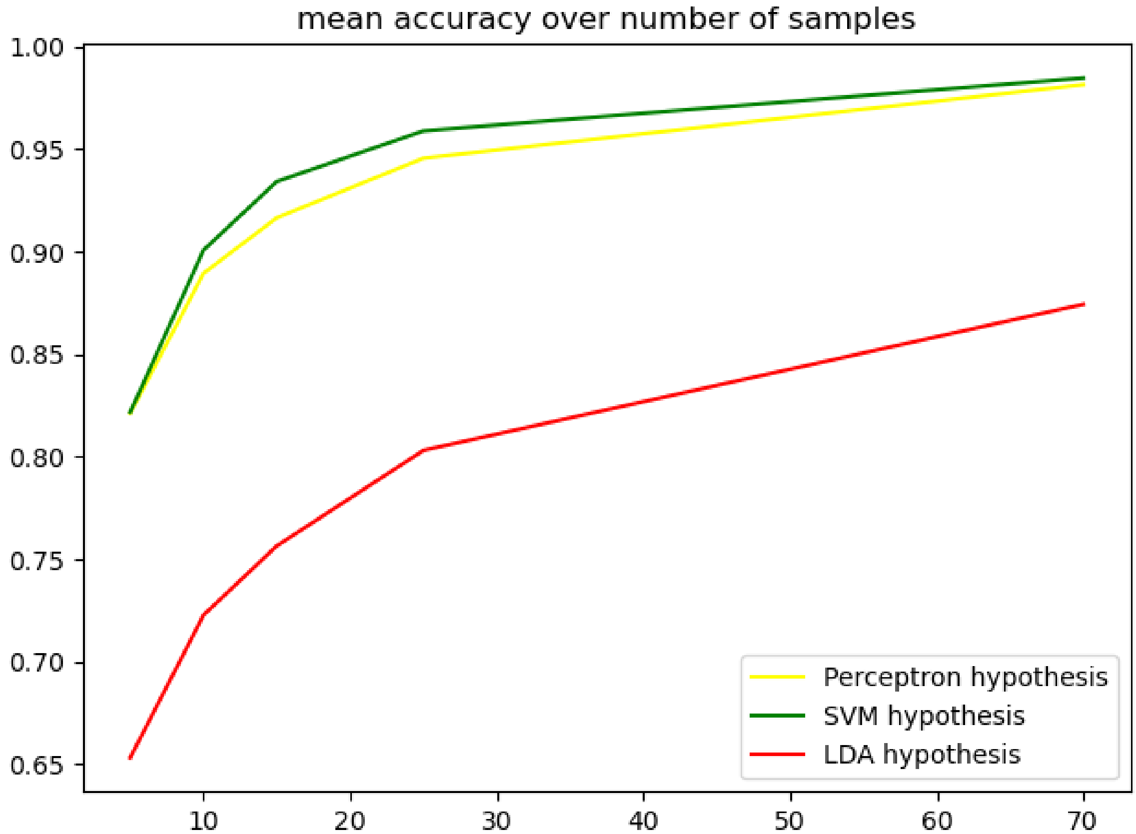

Data analysis over 10 samples
True hypothesis
Perceptron hypothesis
SVM hypothesis


Data analysis over 15 samples
True hypothesis
Perceptron hypothesis
SVM hypothesis

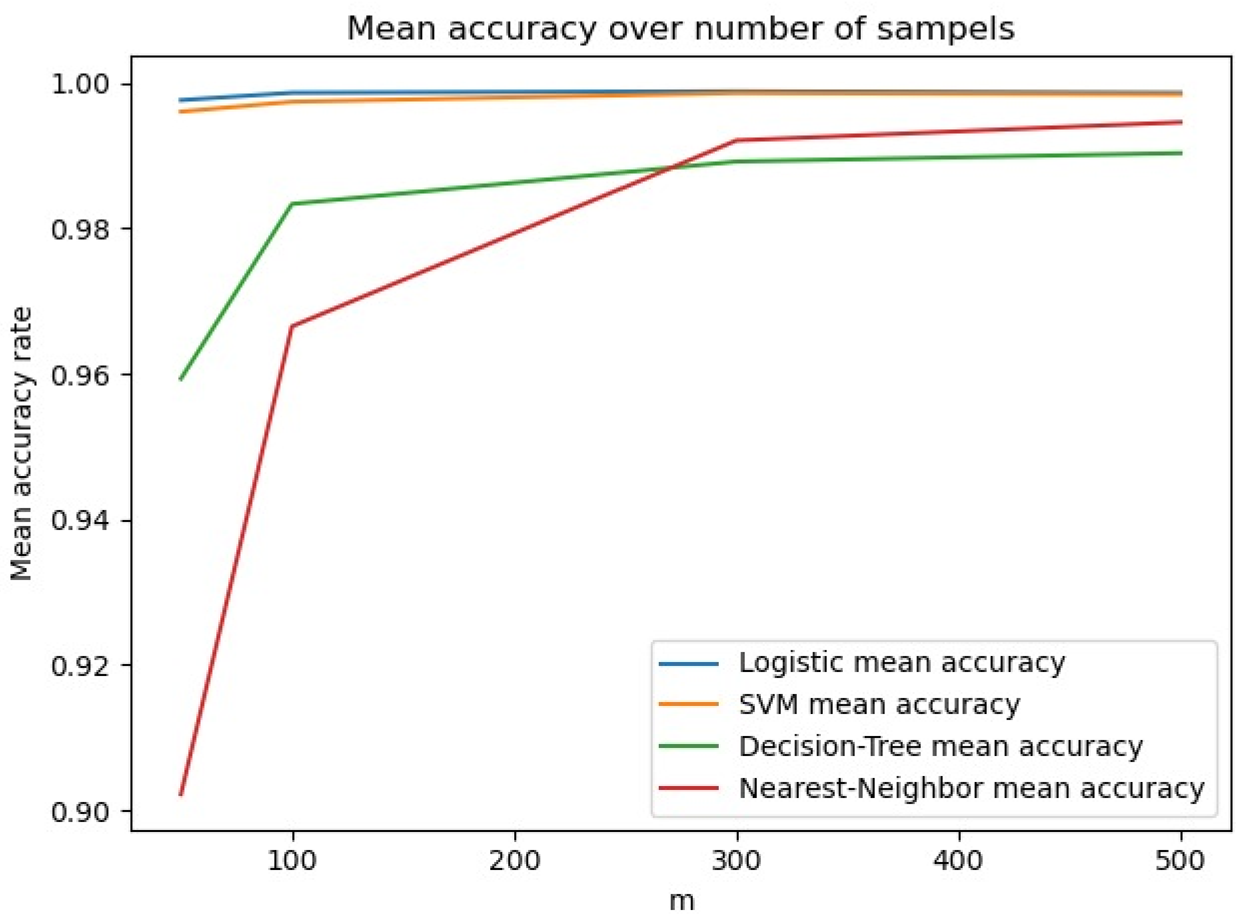

Data analysis over 25 samples
True hypothesis
Perceptron hypothesis
SVM hypothesis


Data analysis over 70 samples
True hypothesis
Perceptron hypothesis
SVM hypothesis

.10


mean accuracy over number of samples
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
10
20
30
40
50
60
70
Perceptron hypothesis
SVM hypothesis
LDA hypothesis

.12


Mean accuracy over number of sampels
Mean accuracy rate
1.00
0.98
0.96
0.94
0.92
0.90
100
200
300
400
500
m
Logistic mean accuracy
SVM mean accuracy
Decision-Tree mean accuracy
Nearest-Neighbor mean accuracy