IML 67577
תרגיל 3
יהונתן פלג
312203031

$X=\mathbb{R}^d \quad Y=\{\pm 1\}$

נניח התפלגות D על $X\times\{\pm1\}$ שהדגימות נלקחות ממנה באופן i.i.d

(1)

$\forall x\in X \quad h_D(x)\begin{cases}1 & P(y=1|x)\geq\frac{1}{2}\\ -1 & o.w\end{cases}$

צ.ל

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

לפי נוסחת בייס: $P(y|x)P(x)=P(x|y)P(y)$ אזי

$\underset{y\in\{\pm1\}}{\arg\max}\, P(x|y)P(y) \underset{\text{מבייס}}{=} \underset{y\in\{\pm1\}}{\arg\max}\, P(y|x)P(x) = \underset{y\in\{\pm1\}}{\arg\max}\, P(y|x)P(x)$

כאשר המעבר האחרון נובע מכך ש $0\leq P(x)$ והוא קבוע עבור כל בחירת y. נשים לב ש $(y=0|x)$ ו- $(y=1|x)$ מאורעות משלימים לכן $P(y=1|x)\geq\frac{1}{2} \Rightarrow P(y=0|x)<\frac{1}{2}$

מכאן

$\underset{y\in\{\pm1\}}{\arg\max}\, P(x|y)P(y) = \begin{cases}y & P(y=1|x)\geq\frac{1}{2}\\ y & P(y=0|x)<\frac{1}{2}\end{cases} = \begin{cases}1 & P(y=1|x)\geq\frac{1}{2}\\ -1 & o.w\end{cases}$

$=h_D(x)$

מ.ש.ל

(2) נניח $X=\mathbb{R}^d$ ו- $X|y\sim N(\mu_y,\Sigma)$ כאשר לכל y וקטור $\mu_y\in\mathbb{R}^d$

אזי...

$$f(x|y)=\frac{1}{\sqrt{(2\pi)^d\det(\Sigma)}}e^{-\frac{1}{2}(x-\mu_y)^T\Sigma^{-1}(x-\mu_y)}$$

נניח כי ידועים לנו μ_{+1},μ_{-1},Σ

אזי ה-classifier האופטימלי של בייס הוא

$h_D(x)=\underset{y\in\{\pm1\}}{\arg\max}\,\delta_y(x)$

כאשר δ_{+1},δ_{-1} הם פונק' $\mathbb{R}^d\to\mathbb{R}$ כלומר

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

$$h_0(x) = \underset{y\in\{\pm 1\}}{\operatorname{argmax}}\ \delta_y(x) \overset{(1)}{=} \underset{y\in\{\pm 1\}}{\operatorname{argmax}}\ P(X|y)P(y)$$

נוכל [illegible] מונוטוני — ln

$$\ln(h_0) = \underset{y\in\{\pm 1\}}{\operatorname{argmax}} \ln P(X|y)P(y) =$$

$$= \underset{y\in\{\pm 1\}}{\operatorname{argmax}} \left(\ln(P(X|y)) + \ln(P(y))\right) =$$

$$= \underset{y\in\{\pm 1\}}{\operatorname{argmax}} \left(\ln\left(\Big(\underbrace{\frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}}}_{\text{const}}\Big) e^{-\frac{1}{2}(X-\mu_y)^T\Sigma^{-1}(X-\mu_y)}\right) + \ln(P(y))\right) =$$

$$= \underset{y\in\{\pm 1\}}{\operatorname{argmax}} \left(-\tfrac{1}{2}(X-\mu_y)^T\Sigma^{-1}(X-\mu_y) + \ln(P(y))\right)$$

$$= \text{-''-} \left(-\tfrac{1}{2}X^T\Sigma^{-1}X + \tfrac{1}{2}\mu_y^T\Sigma^{-1}X + \tfrac{1}{2}X^T\Sigma^{-1}\mu_y - \tfrac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + \ln P(y)\right)$$

$$= \text{-''-} \left(X^T\Sigma^{-1}\mu_y - \tfrac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + \ln P(y)\right) =$$

$$= \underset{y\in\{\pm 1\}}{\operatorname{argmax}}\ \delta_y(x)$$

$\underline{\text{E.e.d}}$

$\times$

$S = (x_1, y_1), \dots, (x_m, y_m)$ ③

נעריך את הגדלים:

$$\hat{P}(y) = \frac{1}{m}\sum_{i=1}^{m} \delta(y_i = y)$$

$$\hat{\Sigma} = \frac{1}{m}\sum_{y\in\{\pm 1\}}\sum_{i=1}^{m} (x_i - \mu_y)(x_i - \mu_y)^T$$

$$\hat{\mu}_y = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_y)(x_i - \mu_y)^T$$

(4) שתי השגיאות שנמדדות אינן הן:

1) סיווג הודעה אמיתית כספאם

2) סיווג הודעה ספאם כאמיתית.

סיווג הודעה אמיתית כספאם היא שגיאה חמורה לא נרצה שיקרה

ולכן נרצה שהשגיאה תהיה ה false positive, כאשר נסווג

את ההודעות לפי $\begin{pmatrix} \text{spam-positive} \\ y=1 \end{pmatrix}, \begin{pmatrix} \text{real-negative} \\ y=-1 \end{pmatrix}$

(5)

$$QP: \quad \underset{v\in\mathbb{R}^n}{\operatorname{argmin}} \left(\frac{1}{2} v^T Q v + a^T v\right) \quad s.t \quad Av \leq d,$$

$$Q \in \mathbb{R}^{n\times n}, A \in \mathbb{R}^{m\times m}, d \in \mathbb{R}^m \quad - \text{fixed}$$

נכתוב את הבעיה לפי צורת Hard-SVM:

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

נציין:

$$\operatorname{argmin} \begin{pmatrix} w \\ b \end{pmatrix}^T I \begin{pmatrix} w \\ b \end{pmatrix} \quad s.t \quad \begin{pmatrix} y_1 x_1 & 1 \\ \vdots & \vdots \\ y_m x_m & 1 \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} \geq \underbrace{\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}_{\text{size } m}$$

$$= \operatorname{argmin} \frac{1}{2} \begin{pmatrix} w \\ b \end{pmatrix}^T [2I] \begin{pmatrix} w \\ b \end{pmatrix} \quad s.t \quad -\begin{bmatrix} y_1 x_1 & 1 \\ \vdots & \vdots \\ y_m x_m & 1 \end{bmatrix} \begin{pmatrix} w \\ b \end{pmatrix} \leq -\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

קיבלנו משוואה מהצורה שרצינו עם:

$$A = -\begin{bmatrix} y_1 x_1 \\ \vdots \\ y_m x_m \end{bmatrix} \quad v = \begin{pmatrix} w \\ b \end{pmatrix} \quad a = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \quad Q = 2I$$

Soft-SVM Problem:

$$\underset{w,\{\xi_i\}}{\operatorname{argmin}} \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m}\xi_i \quad s.t \quad \forall i, y_i\langle w,x_i\rangle \geq 1-\xi_i \text{ and } \xi_i \geq 0$$

ניתן לכתוב שקול לבעיה:

$$\underset{w}{\operatorname{argmin}} \frac{\lambda}{2}\|w\|^2 + 1/m\sum_{i=1}^{m} \ell^{hinge}(y_i\langle w,x_i\rangle),$$

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

מכך ש ξ_i מקיים ואנחנו מינימום: $\xi_i \geq 1-y_i\langle w,x_i\rangle$

$$\xi_i \geq \ell^{hinge}(y_i\langle w,x_i\rangle)$$

נניח בשלילה: $\xi_i \neq \ell^{hinge}(y_i\langle w,x_i\rangle) \Rightarrow \xi_i > \ell^{hinge}(y_i\langle w,x_i\rangle) \geq 0$

מונוטוניות ℓ^{hinge}:

$$\ell^{hinge}(\xi_i) \geq \ell^{hinge}(1-y_i\langle x_i,w\rangle)$$

$$y_i\langle w,x_i\rangle > 1-y_i\langle w,x_i\rangle$$

$\Downarrow$

סתירה למינימליות הפתרון

$$\Downarrow \quad \xi_i = \ell^{hinge}(y_i\langle w,x_i\rangle)$$

ומשוויון נובע שקילות המשוואות. מ.ש.ל.

Practical Questions:

Q7:

Implemented in *models.py*

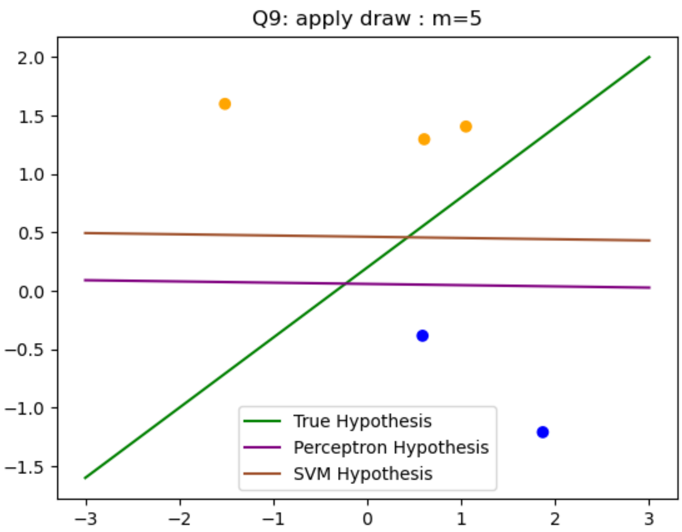
Q8:

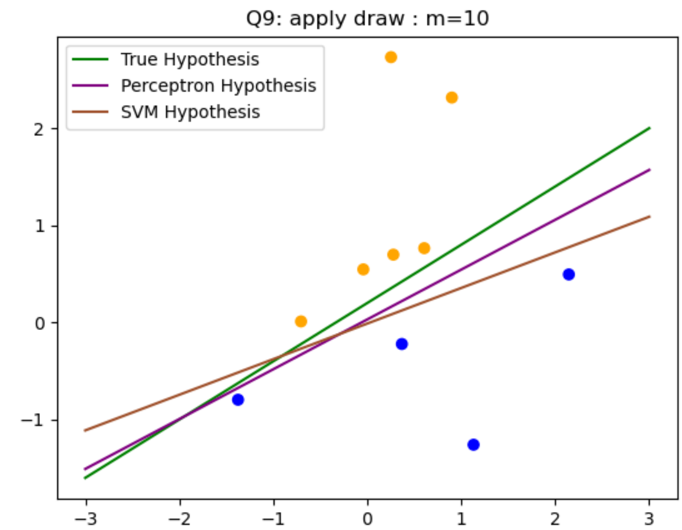
Implemented in *comparison.py*

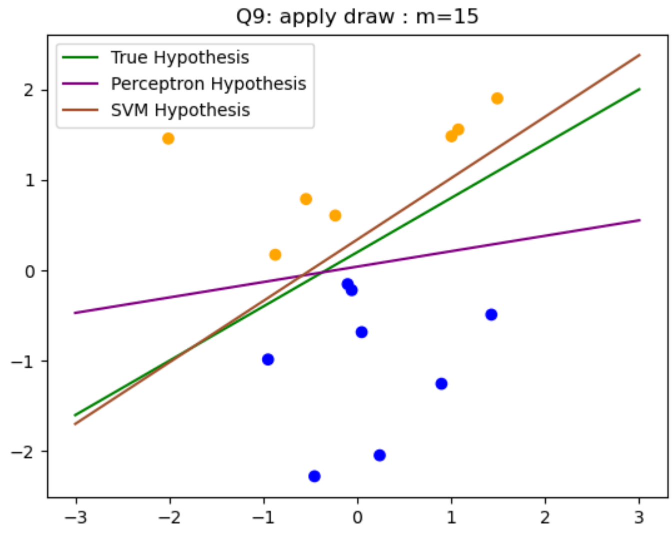
Q9:

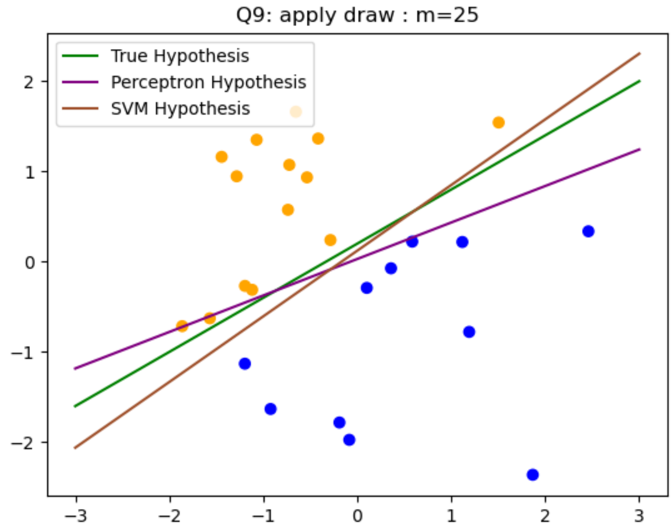
The code that generates the graph is in the function *"apply_draw"*

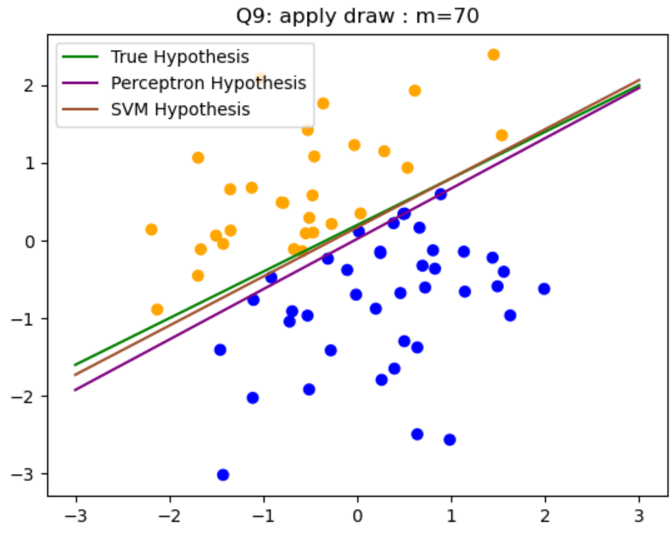
The requested Graphs:





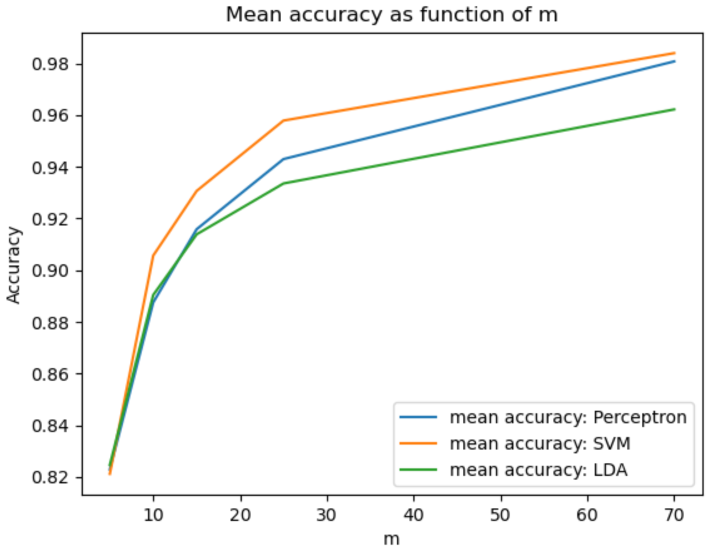






Q10:

The requested Graph:



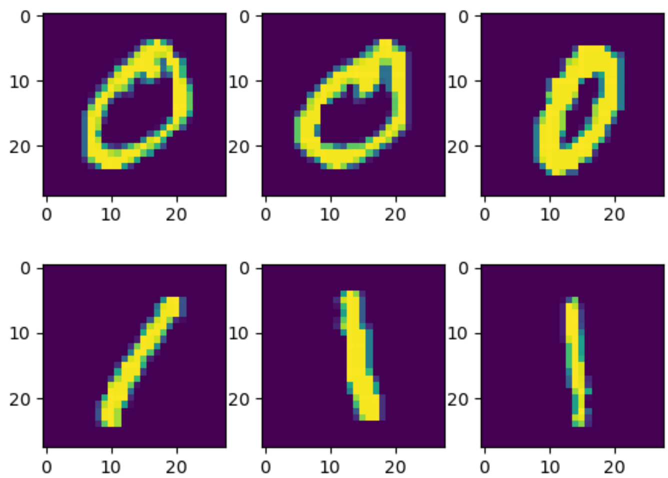
Q11:

The SVM model did better in this instance.
This can be explained by the fact that this model is based on maximizing the margins between the labels of the actual samples and not on guessing the actual model. Perceptron is a close second, and the LDA did worse, probably because it wrongly assumes the true distribution is made of 2 gaussians.

Q12:

implemented in fumction *play_with_ds*

Requested images:

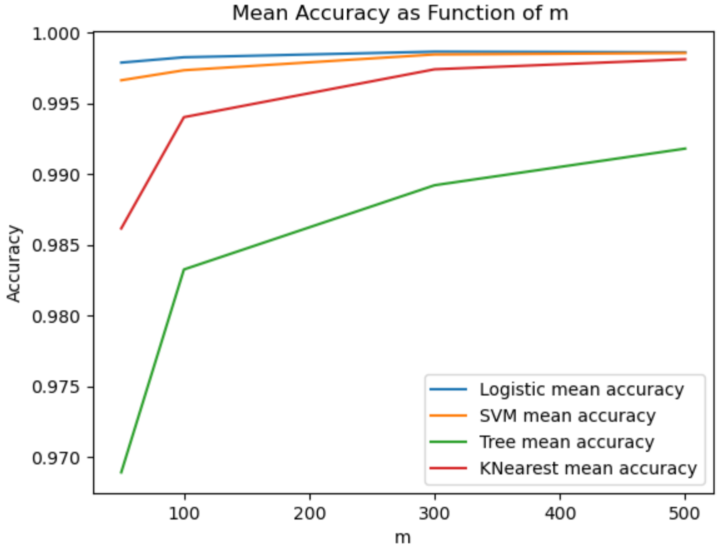


Q13:

implemented in function *rearrange_data*

Q14:

Requested graph




```
Logistic time:[0.0096577  0.01057609 0.01440191 0.02400233]
SVM time:[0.03295195 0.04133796 0.05709656 0.07117558]
Tree time:[0.01100782 0.01535079 0.0314739  0.05162932]
KNearest time:[0.23499504 0.2456953  0.27822554 0.30706699]
```

We see that the decision tree is the fastest in terms of running time (on average), while K-nearest neighbors is the slowest.
This is caused by the nature of calculations of K-nearest neighbors which is heavier than others due to its recursive action.