Introduction to machine learning 67577
שחר נחום 313586877
תרגיל 3:

שאלה 1:

1. נתון:

$$likelihood\big(h,(x,y)\big) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\big(y-h(x)\big)^2}{2\sigma^2}\right)$$

כעת, בהינתן m דגימות בלתי תלויות, נגדיר את ה $Empirical\ Loss$ שלנו:

$$L_s(h) = \frac{1}{m}\prod_{i=1}^{m} likelihood\big(h,(x,y)\big) = \frac{1}{m}\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\big(y-h(x)\big)^2}{2\sigma^2}\right)$$

כלומר נרצה למקסם את הפונקציה:

$$h^* = argmax_{h\in H}L_s(h) = argmax_{h\in H}\frac{1}{m}\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\big(y-h(x)\big)^2}{2\sigma^2}\right)$$

2. נמצא את המקסימום ע״י השוויונות הבאים:

$$\min_{\mathrm{h}\in H}\left\{\frac{1}{m}\sum_{i=1}^{m}(h(x_i)-y_i)^2\right\} = \min_{\mathrm{h}\in H}\left\{\sum_{i=1}^{m}(h(x_i)-y_i)^2\right\} =$$

$$\max_{h\in H}\left\{\exp\left(-\sum_{i=1}^{m}(h(x_i)-y_i)^2\right)\right\} = \max_{h\in H}\left\{\exp\left(-\sum_{i=1}^{m}\frac{\big(y_i-h(x_i)\big)^2}{2\sigma^2}\right)\right\}$$

$$= \max_{h\in H}\left\{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^m\exp\left(-\sum_{i=1}^{m}\frac{\big(y_i-h(x_i)\big)^2}{2\sigma^2}\right)\right\}$$

שאלה 3:

1. נשים לב כי ההסתברויות $\mathbb{P}(h_t(x)=1)$ ו $\mathbb{P}(y=1)$ הן ב״ת, ולכן:

$$TPR = \mathbb{P}(h_t(x)=1 \mid y=1) = \frac{\mathbb{P}(h_t(x)=1)\cdot\mathbb{P}(y=1)}{\mathbb{P}(y=1)}$$

היות וזו התפלגות אחידה על הקטע [0,1] מתקיים כי:

$$\mathbb{P}(h_t(x)=1) = \frac{1-t}{1} = 1-t$$

ולכן:

$$\mathbb{P}(h_t(x)=1)\cdot\mathbb{P}(y=1) = (1-t)\cdot\mathbb{P}(y=1)$$

כעת, נרצה שיתקיים:

$$TRP = p$$

ולכן,

$$\mathbb{P}(h_t(x) = 1 \mid y = 1) = \frac{(1-t)\cdot \mathbb{P}(y=1)}{\mathbb{P}(y=1)} = 1 - t = p$$

כלומר,

$$t = 1 - p$$

2. $FPR(t) = \mathbb{P}(\ h_t(x) = 1 \mid y = 0)$, כעת היות וההסתברויות $\mathbb{P}(h_t(x) = 1)$ ו $\mathbb{P}(y = 1)$ הן ב״ת:

$$FPR(t) = \mathbb{P}(\ h_t(x) = 1 \mid y = 0) = \frac{\mathbb{P}(h_t(x) = 1)\cdot \mathbb{P}(y=0)}{\mathbb{P}(y=0)} = \mathbb{P}(h_t(x) = 1)$$
$$= 1 - t$$

ראינו כי כאשר $TRP(t) = p$ מתקיים כי $t = 1 - p$, ולכן נקבל כי :

$$FPR(1-p) = p$$

נשים לב שהיות ומתקיים לנו $TPR(t) = FPR(t)$, נקבל כי ה $ROC - curve$ שלנו יהיה ישר לינארי בין הנקודות $(1,1)$,$(0,0)$, כלומר $y = x$ בטווח $[0,1]$.

<u>שאלה 4:</u>

1. <u>רשות</u>

2. נרצה להוכיח כי פונקציית ה $ROC - curve$ היא פונקציה מונוטונית עולה חלש של TPR.

$$h_t(x) = \begin{cases} 1, & if\ h(x) \ge t \\ 0, & otherwise \end{cases}$$

נשים לב כי לכל $\varepsilon > 0$:

$$h_{t+\varepsilon}(x) = \begin{cases} 1, & if\ h(x) \ge t + \varepsilon \\ 0, & otherwise \end{cases}$$

כעת:

$$FPR_t = \mathbb{P}(h_t(x) = 1 \mid y = 0) = \frac{\mathbb{P}(h_t(x) = 1)\cdot \mathbb{P}(y=0)}{\mathbb{P}(y=0)} = \mathbb{P}(h_t(x) = 1)$$

$$FPR_{t+\varepsilon} = \mathbb{P}(h_{t+\varepsilon}(x) = 1 \mid y = 0) = \frac{\mathbb{P}(h_t(x) = 1)\cdot \mathbb{P}(y=0)}{\mathbb{P}(y=0)} = \mathbb{P}(h_{t+\varepsilon}(x) = 1)$$

נשים לב כי המאורע $h_{t'}(x) = 1$ מוכל במאורע $h_t(x) = 1$, שהרי $t' = t + \varepsilon \ge t$, ולכן כל איבר שמתויג 1 בפונקציה $h_{t'}(x)$, בהכרח יתויג 1 בפונקציה $h_t(x) = 1$, ולכן נקבל כי:

$$\mathbb{P}(h_{t'}(x) = 1) \le \mathbb{P}(h_t(x) = 1)$$

ולכן $FPR_t \ge FPR_{t+\varepsilon}$. כלומר הפונקציה $FPR(t)$ היא פונקציה מונוטונית יורדת חלש.

באותו האופן, נקבל כי $TPR_t \ge TPR_{t+\varepsilon}$.

כעת, פונקציית ה $ROC - curve$ היא פונקציה $\left(FPR(t), TPR(t)\right)$, ולכן ככל שנעלה את רכיב ה x, כלומר את $FPR(t)$, בפועל, נוריד את הערך של t, כפי שראינו לעיל. ולכן כתוצאה מכך גם יגדל/ יישאר זהה (שהרי $FPR(t)$ היא מונוטונית עולה חלש) רכיב ה- y שלנו, כלומר, $FPR(t)$, **ולהיפך**.
לכן פונקציית ה $ROC - curve$ היא פונקציה מונוטונית עולה חלש ביחס ל $FPR(t)$.

<u>שאלה 5:</u>

1. נתון כי $(x; 1)$ הוא הוספה של הקורדינטה ה1 $d + 1$ בווקטור ה – d מימדי x, וערכה הוא 1.

עלינו לכתוב את ה $Bayes\ optimal\ classifier$ כך ש $h_{\mathcal{D}} = sign(\ \langle (x; 1), w\rangle$.

$$h_{\mathcal{D}}(x) = argmax_y \mathcal{D}[y \mid x\ ] = argmax_y \left(\ x^T \Sigma^{-1} \mu_y - \frac{1}{2}\left(\mu_y\right)^T \Sigma^{-1}\mu_y + \log\left(\pi_y\right)\right)$$

כעת, היות ומתקיים $y = \{\pm 1\}$ נוכל להשתמש בסימן של ההפרש בין הביטויים, ובכך לקבל את התיוג המתאים.
כלומר:

$$h_{\mathcal{D}}(x) = sign\left(x^T \Sigma^{-1}\mu_1 - \frac{1}{2}(\mu_{-1})^T\ \Sigma^{-1}\mu_1 + \log(\pi_1) - \left(x^T\Sigma^{-1}\mu_{-1} - \frac{1}{2}(\mu_1)^T\ \Sigma^{-1}\mu_{-1} + \log(\pi_{-1})\right)\right)$$

$$= sign\left((x^T\Sigma^{-1})(\mu_1 - \mu_{-1}) - \frac{1}{2}(\mu_1)^T\ \Sigma^{-1}\mu_1 + \log(\pi_1) - \frac{1}{2}(\mu_{-1})^T\ \Sigma^{-1}\mu_{-1} - \log(\pi_{-1})\right)$$

נסמן:

$$sign\left(x^T\ \overbrace{\Sigma^{-1}(\mu_1 - \mu_{-1})}^{\boldsymbol{w_1 \dots w_d}} \overbrace{- \frac{1}{2}(\mu_1)^T\ \Sigma^{-1}\mu_1 + \log(\pi_1) + \frac{1}{2}(\mu_{-1})^T\ \Sigma^{-1}\mu_{-1} - \log(\pi_{-1})}^{\boldsymbol{w_{d+1}}}\right)$$

$$sw_1 \dots w_d = \Sigma^{-1}\left(\mu_1 - \mu_{-1}\right)$$

$$w_{d+1} = -\ \tfrac{1}{2}(\mu_1)^T\ \Sigma^{-1}\mu_1 + \log(\pi_1) + \tfrac{1}{2}(\mu_{-1})^T\ \Sigma^{-1}\mu_{-1} - \log(\pi_{-1})$$

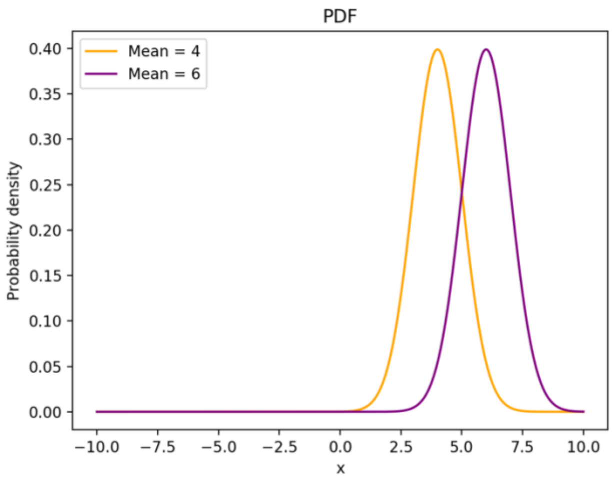
ונקבל:

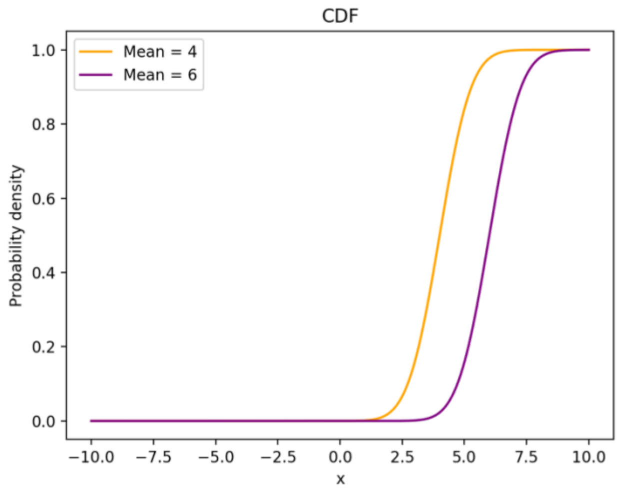
$$= sign(\langle (x; 1), w\rangle)$$

כנדרש.

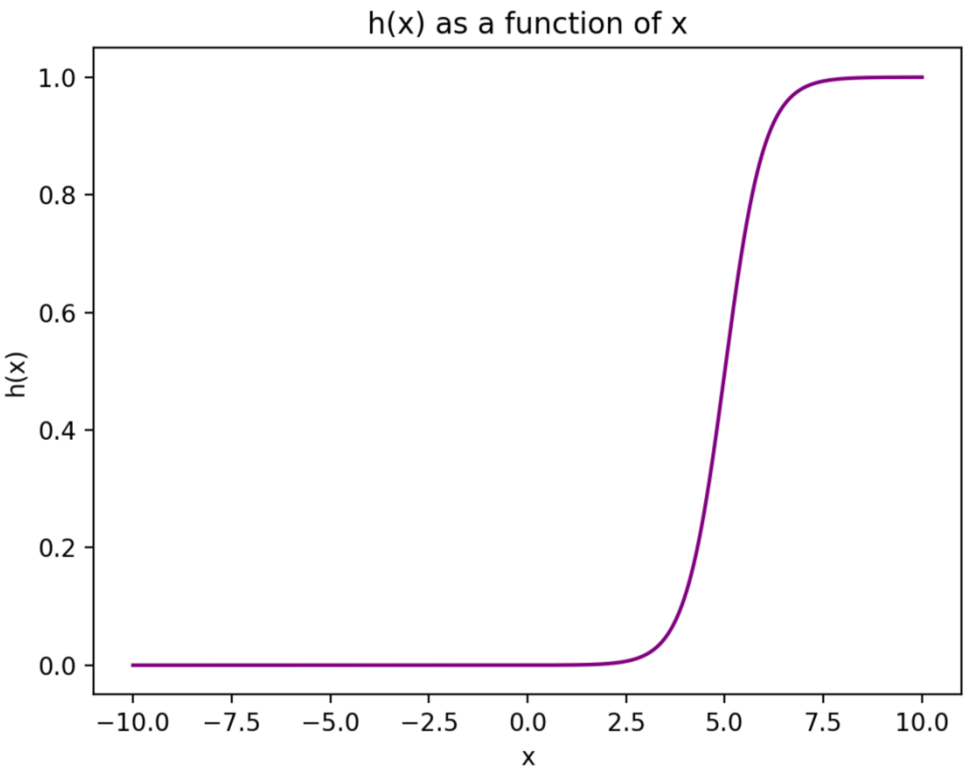
2.

i. סעיף ראשון:

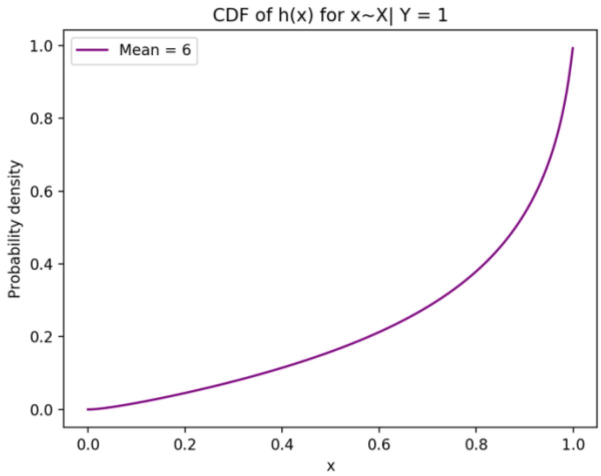


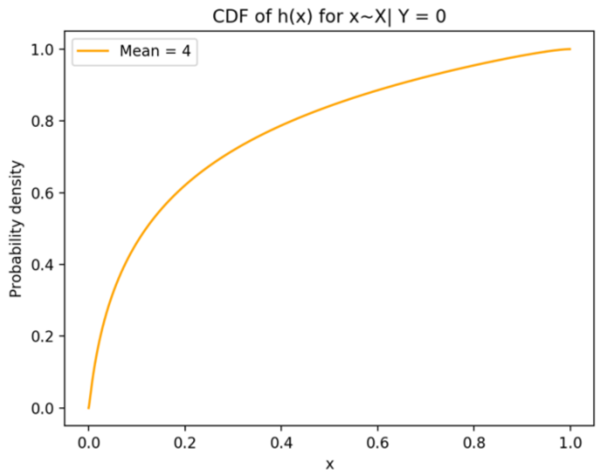


ii. סעיף שני:

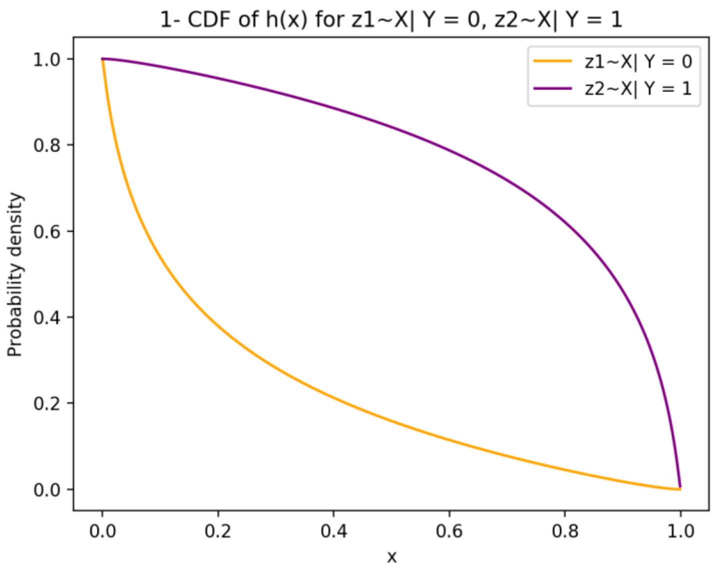


iii. סעיף שלישי:





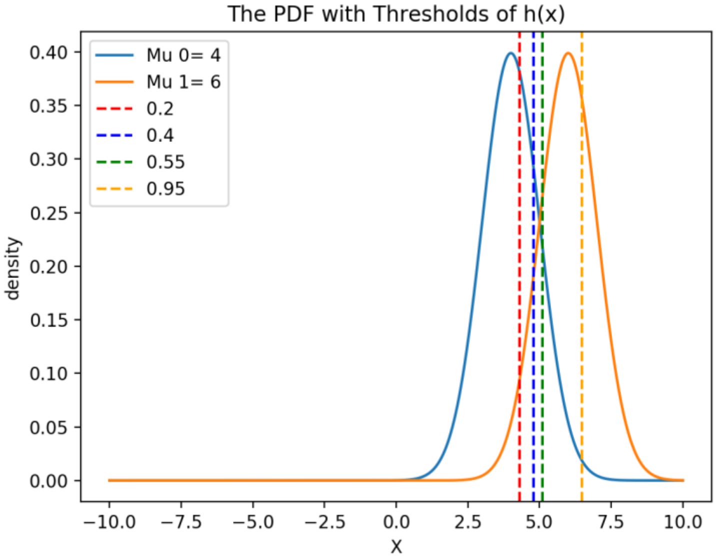
iv.



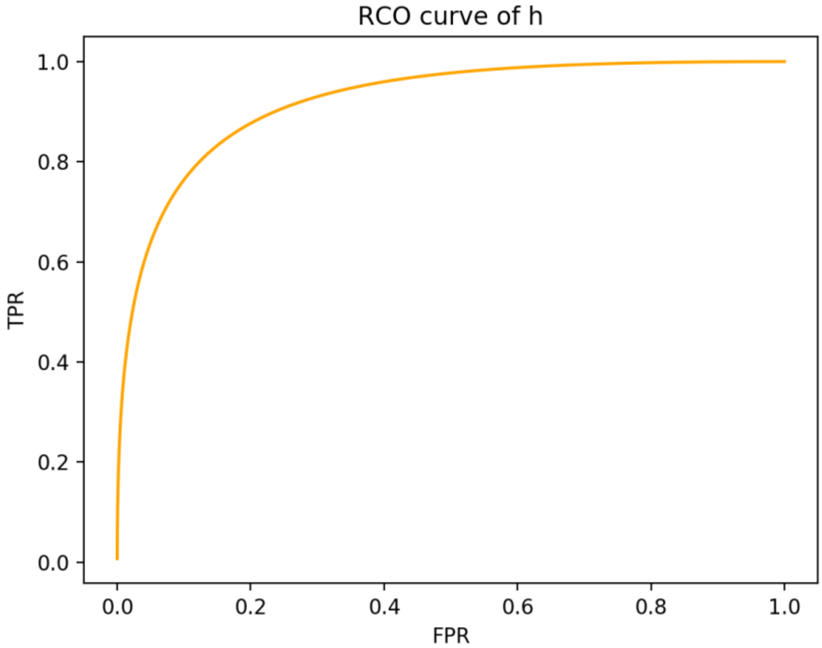
v. נחשב את הערכים:

t	TPR	FPR
0.2	0.95	0.38
0.4	0.88	0.2
0.55	0.81	0.137
0.95	0.31	0.008

vi.

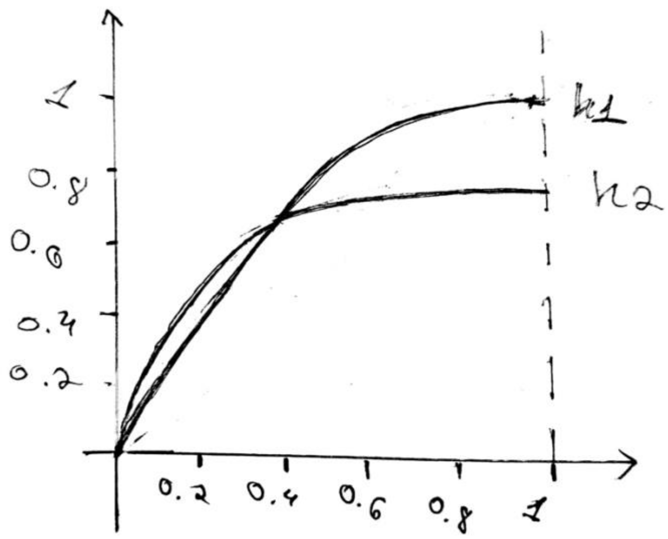


vii. הגרף:



<u>שאלה 6:</u>

1. הטענה אינה נכונה. דוגמא נגדית:

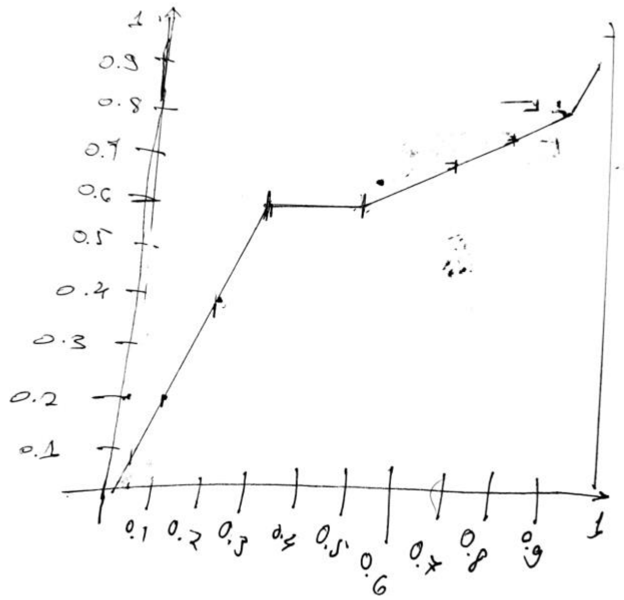


נשים לב שבעבור h_1 השטח שלה מתחת לגרף גדול מהשטח של הפונקציה h_2 , אך קיים מצב שבו נעדיף להשתמש בפונקציה האדומה -לכל $True\ Positive\ Rate$ קטן מ 0.4.

2. הטענה נכונה.

נשים לב שלכל פונקציה h_1 אשר בכל נקודה הערך שלה גדול יותר מהערך של h_2, נקבל שבכל נקודה אשר ה FPR שלהן זהה, ערך ה TPR של הפונקציה h_1 גדול יותר משל h_2, ולכן לא קיימת סיבה שנעדיף להשתמש ב h_2 על פני h_1.

3. ציור של הגרף :



נשים לב לפי הגרף ולפי התיאור שקיבלנו בשאלה כי היחס שבטווח $z \in [0.3 - 0.5]$ הינו הטווח שבו היחס הוא אופטימלי שהרי אין אנחנו משיגים TPR בעוד שכן אנו מעלים את ה rate של ה FP.

4. נשים לב שהתשובה היא חיובית בעבור LDA, ושלילית בעבור רגרסיה לוגיסטית.

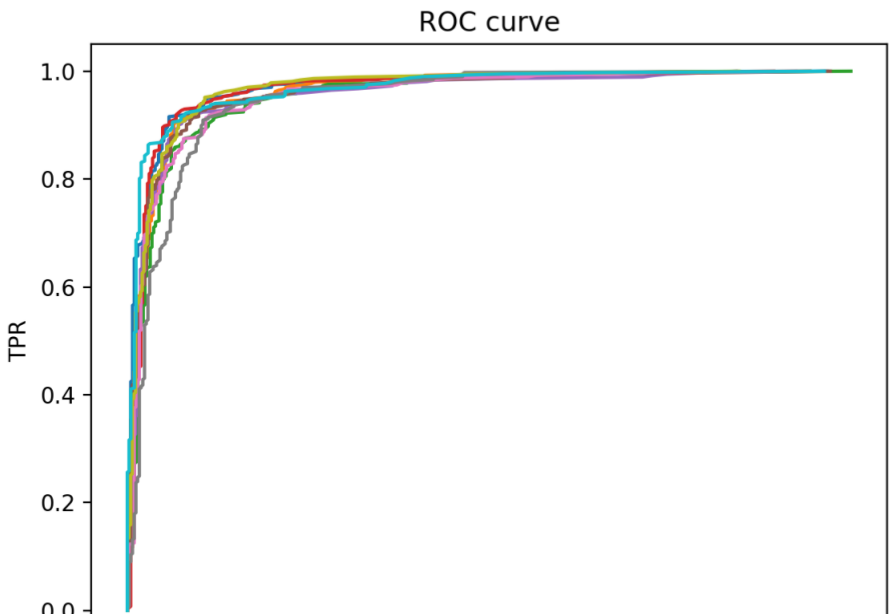
LDA:
כאשר מתקיים: $D(X \mid Y = y) = \mathcal{N}(\mu_y, \Sigma)$, אנו מנסים למקסם את ה-$Base\ optimal\ classiphier$ ובכך לחזות בצורה הטובה ביותר את ההתפלגות של $X \times Y$.

רגרסיה לוגיסטית:
תחת ההנחה ש $Y \mid X_i = x_i \sim Ber(p_i)$ אנחנו ממקסמים את פונקציית הנראות, ואנחנו לא מנסים לחזות את ההתפלגות של $X \times Y$ אלא להתאים מודל לנתונים הקיימים.

<u>שאלה 7:</u>

a. (בקוד)

b. גרף:



c. נראה את התוצאות בטבלה הבאה:

Logistic error	K=1	K=2	K=5	K=10	K=100	
0.071000	0.179000	0.209000	0.202000	0.224000	0.266000	1
0.078000	0.071000	0.215000	0.203000	0.216000	0.283000	2
0.072000	0.163000	0.192000	0.196000	0.208000	0.242000	3
0.072000	0.203000	0.228000	0.209000	0.210000	0.272000	4
0.064000	0.175000	0.181000	0.180000	0.211000	0.276000	5
0.066000	0.180000	0.200000	0.194000	0.217000	0.269000	6
0.068000	0.184000	0.202000	0.198000	0.217000	0.266000	7
0.069000	0.160000	0.193000	0.176000	0.189000	0.264000	8
0.075000	0.193000	0.214000	0.201000	0.214000	0.254000	9
0.080000	0.165000	0.178000	0.185000	0.207000	0.270000	10

נשים לב שככל שנבחר k יותר נמוך נקבל שגיאה נמוכה יותר, שהרי כך האלגוריתם תלוי בפחות נקודות. מצד שני, כאשר נסתמך על מעט נקודות, הרי שכל נקודה שהיא רעש, תשפיע במידה חזקה על החיזוי שלנו.

d. העמודות אותן בחרתי לדגום הן העמודות

```
[1,4,21,20,10]
```

הערכים העצמיים של מטריצות השונות:

```
QDA eigen values [2.34662675 1.06605307 0.01836405 0.40950138 0.39851353]
QDA eigen values [2.44570972 1.43387879 0.50487374 0.12692149 0.05456667]
LDA eigen values [1.47063543 1.28487599 1.1345723  0.44454097 0.03350132]
```

הממוצעים:

```
QDA errors:
test - 0.288200
train -  0.286643

LDA errors:
test -  0.256800
train -  0.254096
```