IML תרגיל 3

2.1 (1) צ"ל:

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

יכול להציג כ-

$$\arg\min \frac{1}{2}v^TQv+a^Tv$$
$$s.t \quad Av\leq d$$

<u>הוכחה:</u>

נגדיר: $\|w\|^2=w^TIw$

נשים לב:

$$y_i(\langle w,x_i\rangle+b)\geq 1 \iff y_iw^Tx_i+y_ib\geq 1 \iff$$
$$\iff y_ix_i^Tw+y_ib\geq 1$$

לכן, אם נסמן

$$A' = \begin{bmatrix} (\quad y_1 X_1^T \quad) & y_1 \\ (\quad y_2 X_2^T \quad) & y_2 \\ \vdots & \vdots \\ (\quad y_n X_n^T \quad) & y_n \end{bmatrix}, \quad V = \begin{pmatrix} \begin{pmatrix} w \end{pmatrix} \\ b \end{pmatrix}$$

נקבל כי:

$$\begin{bmatrix} y_1 X_1^T w + y_1 b \\ y_2 X_2^T w + y_2 b \\ \vdots \\ y_n X_n^T w + y_n b \end{bmatrix} = A'V \geq 1 \iff$$

$$\iff -A'V \leq -1$$

לכן נגדיר $V = V$, $A = -A'$, $d = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$

וקיים אי-שוויון הדרוש אם $Av < d$

כעת, נסמן ב- n את המימד של w.

כיוון ש $\|w\|^2 = w^T w = w^T I_n w$

נגדיר $M_{n+1\times n+1} \ni Q = \begin{bmatrix} \begin{bmatrix} I_n \end{bmatrix} \\ 0 \end{bmatrix}$

ונקח $a=0_V$ ונקבל כי $\operatorname{argmin}\|w\|^2$

שקול ל- $\operatorname{argmin} V^TQV + a^TV$

<u>כנדרש</u>.

② ראשית, נשים לב כי אם מתקיים $y_i\langle w,x_i\rangle \geq 1$

אז בחירת ה ξ_i אופטימלית ויתן תוצאה מינימלית

מי בחר את הסימן הקשה הזה??

עבור $\sum_i \xi_i$ יהיה $\xi_i=0$, לכן עבור i שמקיים כזה מתקיים ש

$$\xi_i = l^{hinge}\left(y_i\langle w,x_i\rangle\right)$$

אחרת, אם $y_i\langle w,x_i\rangle<1$ אזי לפי האילוץ מתקיים

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

$$\iff -y_i\langle w,x_i\rangle+1 \leq \xi_i \iff 1-y_i\langle w,x_i\rangle \leq \xi_i \iff l^{hinge}\left(y_i\langle w,x_i\rangle\right) \leq \xi_i$$

לכן, ערך ה ξ_i שנותן ערך מינימלי ל- $\sum_i \xi_i$

הוא כאשר $\xi_i = \ell^{hinge}\left(y_i\langle w,x_i\rangle\right)$.

כלומר, הראינו כי בכל מקרה ערכי ה ξ_i שיתנו ערך מינימלי ל- $\frac{1}{m}\sum \xi_i$ הינו רק $\xi_i = \ell^{hinge}\left(y_i\langle w,x_i\rangle\right)$

לכן בעיית האופטימיזציה של soft SVM שקולה לבעיה:

$$\underset{w,\{\xi_i\}}{\operatorname{argmin}} \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_i \ell^{hinge}\left(y_i\langle w,x_i\rangle\right)$$

כנדרש.

$\blacksquare$

(2.2)

3 (a) נרצה להגדיר שראינו בתרגול, נוסיף רק שבמקרה

של G נניח $x_i\in\mathbb{R}$ והמאפיינים נורמליים, ולכן

יש לנו תוחלת ושונות שונים בין הקלאסים בהתפלגות

נורמלית כלומר, נרצה למצוא ביטויים לפרמטרים

ל Naive bayes, תחת ההנחה ש $X|y=k\sim N(\mu_k,\sigma_k^2)$,

$y\sim Mult(\pi)$ עבור $\sum_{i=1}^{k}\pi_i=1$, $\pi\in[0,1]^k$.

נתחיל למצוא את פונ' הנראות בצורה דומה למה שראינו

בהרצאה:

$$L(\Theta|X,y)=f_{(X,y)|\Theta}\left(\left((x_i,y_i)\right)_{i=1}^{m}\right)\underset{\text{אי"ת}}{=}\prod_{i=1}^{m}f_{x,y|\Theta}(x_i,y_i)=$$

$$\underset{\text{בייס}}{=}\prod_{i=1}^{m}f_{x|y=y_i}(x_i)\cdot f_{y|\Theta}(y_i)=\prod_{i=1}^{m}N(x_i|\mu_{y_i},\sigma_{y_i}^2)Mult(y_i|\pi)$$

נפעיל log כדי למצוא את ה log-likelihood:

$$\ell(\Theta|(X,y))=\log\left(\prod_{i=1}^{m}N(x_i|\mu_{y_i},\sigma_{y_i}^2)Mult(y_i|\pi)\right)=$$

$$= \sum_{i=1}^{m} \log\left(\mathcal{N}\left(x_i \mid \mu_{y_i}, \sigma_{y_i}^2\right)\right) + \log\left(\mathrm{Mult}\left(y_i \mid \pi\right)\right) =$$

$$= \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi\sigma_{y_i}^2}} \exp\left(-\frac{(x_i-\mu_{y_i})^2}{2\sigma_{y_i}^2}\right)\right) + \log\left(\pi_{y_i}\right) =$$

$$= \sum_{i=1}^{m} -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma_{y_i}^2\right) - \frac{(x_i-\mu_{y_i})^2}{2\sigma_{y_i}^2} + \log\left(\pi_{y_i}\right) =$$

$$= -\frac{m}{2}\log(2\pi) + \left(\sum_{i=1}^{m} -\frac{1}{2}\log\left(\sigma_{y_i}^2\right) - \frac{(x_i-\mu_{y_i})^2}{2\sigma_{y_i}^2} + \log\left(\pi_{y_i}\right)\right) = \circledast$$

כעת, נסמן ב- n_k את מס' הדגימות כך

ש $y_i = k$, כלומר $n_k := \sum_{i=1}^{m} \mathbb{1}_{(y_i=k)}$. אם נשים לב לכך

אז הסכימה עוברת כך שעובר על מס' k השונים

ולא על הדגימות. נסמן ב K את מס' ה k השונים;

$$\circledast = \overbrace{-\frac{m}{2}\log(2\pi) + \left(\sum_{k=1}^{K} -\frac{n_k}{2}\log\left(\sigma_k^2\right) - \frac{1}{2\sigma_k^2}\left(\sum_{\substack{i \\ y_i=k}} (x_i-\mu_k)^2 + n_k \log\left(\pi_k\right)\right)\right)}^{\#}$$

כעת, נרצה למצוא את π_k, כאשר ההסתברויות לכך

ש $y_i=k$ לכל $k\in[K]$. כיוון ש $\sum_{k=1}^{K}\pi_k = 1$

נגדיר את g כך ש $g(\pi)=\sum \pi_k - 1$ ונשתמש על

#, כלומר, קיים ה log-likelihood. נגזור כעת לפי לגרנג'

$$\exists \lambda \in \mathbb{R} \quad s.t \quad \nabla l(\Theta|X,y) = \lambda \nabla g(\pi^0)$$

כי π^0 נק' קיצון. נגזור את $l(\Theta|X,y)$ לפי

מאחר ש π, תלוי רק בפרמטרי מלבד הקבועים נשאר

רק הביטוי $\sum_{k=1}^{K} n_k \log(\pi_k)$. נגזור לפי כל משתנה ונקבל:

$$\nabla l(\Theta|X,y) = \left(\frac{n_1}{\pi_1^0}, \frac{n_2}{\pi_2^0}, \ldots, \frac{n_K}{\pi_K^0}\right)^T.$$

נגזור את $g(\pi)$ ונקבל $\nabla g(\pi^0) = (1, 1, \ldots 1)^T$.

נציב את $\nabla l(\Theta|X,y) = \lambda \nabla g(\pi^0)$

$$\Updownarrow$$

$$\begin{pmatrix} \frac{n_1}{\pi_1^0} \\ \vdots \\ \frac{n_K}{\pi_K^0} \end{pmatrix} - \begin{pmatrix} \lambda \\ \vdots \\ \lambda \end{pmatrix} = 0$$

$$\Updownarrow$$

$$\forall k \in K \quad \lambda = \frac{n_k}{\pi_k^0}$$

$$\Updownarrow$$

$$\Updownarrow$$

$$\forall k\in K \quad \pi_k = \frac{n_k}{\lambda}$$

ידוע ש $\sum_{k=1}^{K} \pi_k = 1$ נקבל ש

$$\sum_{k=1}^{K} \pi_k = \sum_{k=1}^{K} \frac{n_k}{\lambda} = \frac{1}{\lambda}\sum_{k=1}^{K} n_k = 1 \Leftrightarrow \sum_{i=1}^{k} n_k = \lambda$$

ובנוסף שנציב את n_k $\left(\sum \mathbb{1}(y_i=k)\right)$ נקבל ש בסה"כ

שלכן ה- $\sum_{i=1}^{m} 1 = \lambda \Rightarrow \lambda = m$

לכן סה"כ קיבלנו כי $\underline{\pi_k^{MLE} = \frac{n_k}{m}}$

כעת נמצא אומדים עבור μ_k, σ_k^2 לכל $k\in K$,

נגזור את ⊕ לפי μ_k ונקבל:

$$\sum_{\{i|y_i=k\}} (x_i-\mu_k) = 0 \Rightarrow \left(\sum x_i\right) - n_k\cdot\mu_k = 0 \Rightarrow \sum x_i = n_k\mu_k \Rightarrow$$

$$\Rightarrow \underline{\mu_k^{MLE} = \frac{\sum_{\{i|y_i=k\}} x_i}{n_k}}$$

בהתאם, נגזור לפי σ_k^2 כדי למצוא את האומד:

$$-\frac{n_k}{2\sigma_k^2} - \frac{1}{2(\sigma_k^2)^2}\left(\sum_{\{i|y_i=k\}}(x_i-\mu_k)^2\right) \Rightarrow \underline{(\sigma_k^2)^{MLE} = \frac{1}{n_k}\sum_{\{i|y_i=k\}}(x_i-\mu_k)^2}$$

(b) כיוון שעבור כל גאוסיה x_i קיימים מס' מימדים,

ונחזור לרגע עם ההנחות לסעיף הקודם. נשים לב כי ה

מימדים הם בת"ל, ולכן הסתברות חיתוך המאורעות

שווה למכפלת הסתברויות המאורעות, וכך עבור כל

מימד בנפרד, וכיוון שכל מימד הוא סקלר ב- $\mathbb{R}$ נוכל להשתמש

בהתפלגות נורמלית. נמשיך את השוויון שהגענו בסעיף הקודם:

$$L(\Theta|X,y) = \prod_{i=1}^{m} f_{X|y=y_i}(x_i)\, f_{y|\Theta}(y_i) = \prod_{i=1}^{m} Mult(y_i|\pi)\cdot N(x_i|\mu_{y_i},\Sigma_{y_i}) =$$

$$= \prod_{i=1}^{m} Mult(y_i|\pi)\cdot\prod_{j=1}^{d} N(x_{i,j}|\mu_{y_{i,j}},\sigma_{y_{i,j}}) \Rightarrow$$

$$\Rightarrow l(\Theta|X,y) = \sum_{i=1}^{m}\left(\log(Mult(y_i|\pi)) + \sum_{j=1}^{d}\log(N(x_{i,j}|\mu_{y_{i,j}},\sigma_{y_{i,j}}))\right)$$

נקבל שוב למקסם לכל $k \in [K]$ עם ה- y_i השונים ונקבל:

$$\sum_{k=1}^{K}\left(n_k \log(\pi_k) + \sum_{i=1}^{m}\sum_{j=1}^{d} \mathbb{1}_{[y_i=k]} \cdot \log\left(N\left(x_{i,j} \mid \mu_{k,j}, \sigma_{k,j}^2\right)\right)\right)$$

כעת, כדי למצוא את $\pi_{k,j}^{MLE}$, μ_k^{MLE}, $\sigma_{k,j}^{MLE}$ נגזור ונשווה

לפי המשתנה, ונשים לב שהביטוי בכל מקרה רק שבנפרד

זה מקרה 6 כבר, לכן נקבל כמו:

$$\underline{\left(\sigma_{k,j}^2\right)^{MLE} = \frac{1}{n_k}\sum_{\{i \mid y_i=k\}}\left(x_i - \mu_k\right)^2}$$

$$\underline{\mu_{k,j}^{MLE} = \frac{\sum_{\{i \mid y_i=k\}} x_i}{n_k}}$$

$$\underline{\pi_k^{MLE} = \frac{n_k}{m}}$$

4 (a) נשים לב שהשאלה זהה לקודמת, מלבד סוג ההתפלגות

עבור ה פיזור, לכן נוכל להשתמש בחישובים ובהסברים

מהשאלה הקודמת עם שינוי סוג ההתפלגות:

$$L(\Theta \mid X, y) = \prod_{i=1}^{m} f_{X|y=y_i}(x_i)\, f_{y|\Theta}(y_i) = \prod_{i=1}^{m} mult(y_i \mid \pi)\cdot poi(x_i \mid \lambda_{y_i}) \Rightarrow$$

$$\xrightarrow[\log]{\text{הפעלת}} \sum_{i=1}^{m} \log\left(\frac{\lambda_{y_i}^{x_i} \exp(-\lambda_{y_i})}{x_i!}\right) + \log(\pi_{y_i}) = \sum_{i=1}^{m} x_i \log(\lambda_{y_i}) - \lambda_{y_i} - \log(x_i!) + \log(\pi_{y_i}) =$$

הגדרת poi

$$\xrightarrow{\substack{\text{נחלק}\\ \text{לפי הקבוצות } [K]}} \sum_{k=1}^{K}\left(\sum_{\substack{i=1\\ y_i=k}}^{m} \left(x_i \log(\lambda_k)\right) - n_k\lambda_k + n_k \log(\pi_k)\right) - \sum_{i=1}^{m} \log(x_i!)$$

נגזור את l לפי הפרמטר פואסון ונקבל כי הנגזרת היא $\left(\frac{1}{\lambda_k}\sum_{\substack{i=1\\ y_i=k}}^{m} x_i\right) - n_k$

נשווה לאפס ונקבל כי: $\lambda_k^{MLE} = \frac{1}{n_k}\sum_{\substack{i=1\\ y_i=k}}^{m} x_i = \frac{1}{n_k}\sum_{i=1}^{m} \mathbb{1}_{[y_i=k]} x_i$

את הפרמטר ההתפלגות המולטינומית נגזור כמו בשאלה

הקודמת ונקבל: $\pi_k^{MLE} = \frac{n_k}{m}$

(b) נשתמש בהגדרה G כמו סעיף (3)(b) רק עם התפלגות פואסון

נשתמש בסעיף הקודם:

$$L(\Theta|x,y)=\prod_{i=1}^{m}\left(\prod_{j=1}^{d} Poi(x_{i,j}|\lambda_{y_i,j})\cdot Mult(y_i|\pi)\right)\Rightarrow$$

$$\Rightarrow l(\Theta|x,y)=\sum_{i=1}^{m}\left(\sum_{j=1}^{d}\log(Poi(x_{i,j}|\lambda_{y_i,j}))+\log(Mult(y_i,\pi))\right)$$

מאחר ונרצה למקסם K, נפעיל ונפתח log:

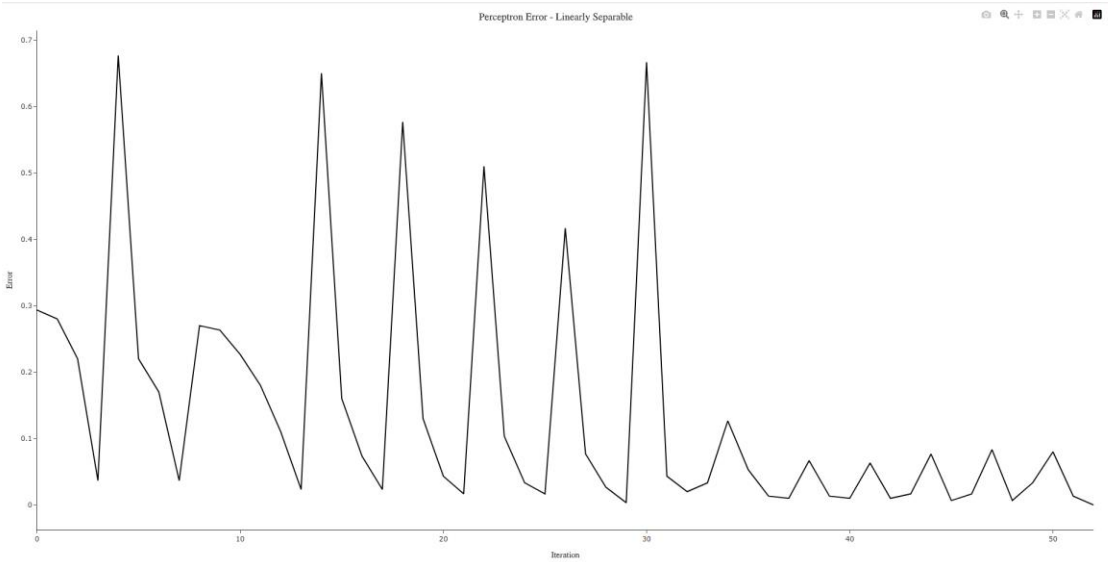
$$\sum_{k}^{K}\left(\sum_{j=1}^{d}\left(\sum_{\substack{i=1\\ y_i=k}}^{m}\left(x_{i,j}\log(\lambda_{k,j})\right)-n_k\lambda_{k,j}\right)+n_k\log(\pi_k)\right)+\sum_{i=1}^{m}\sum_{j=1}^{d}\log(x_{i,j}!)$$

נזכור כמו בסעיף קודם, אך פה פה נגזור בנפרד ונקבל:

$$\lambda_{k,j}^{MLE}=\frac{1}{n_k}\sum_{i=1}^{m}\mathbb{1}_{[y_i=k]}x_{i,j}\ ,\quad \pi_k^{MLE}=\frac{n_k}{m}$$

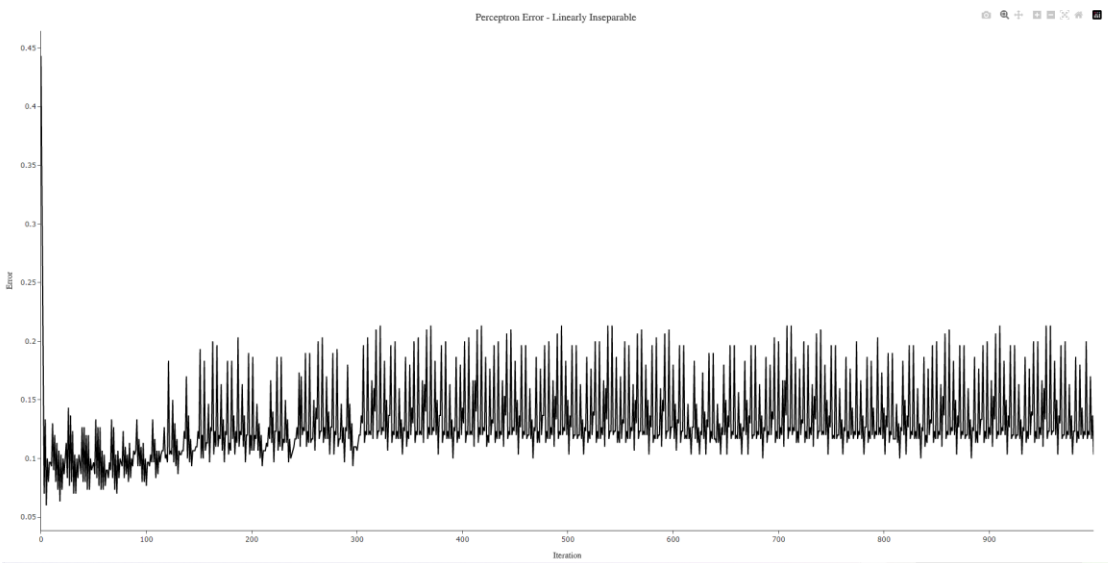
3.1)

1)



אנו יכולים ללמוד מהגרף כי ככל שכמות האיטרציות גדלה כך רמת הדיוק משתפרת, כיוון שבכל איטרציה האלגוריתם משפר את הW שלו, אנו יכולים לראות כי השגיאות החריגות (השפיצים) הולכות וקטנות עד שהשגיאה מתאפסת.

2)



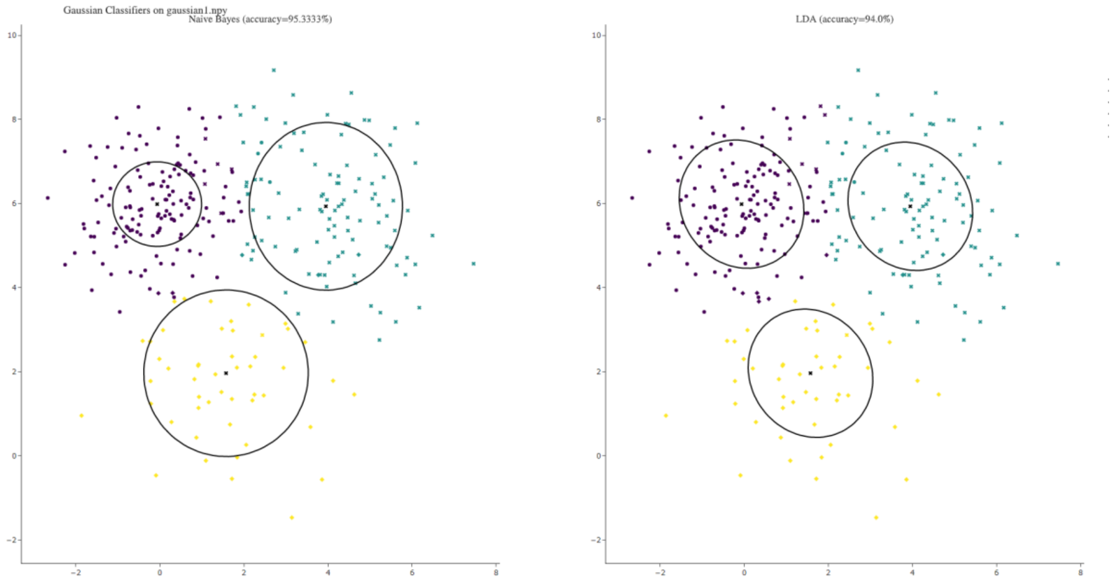
מכיוון שהפרספטרון מניח שקיימת הפרדה ליניארית לדגימות, הוא מתקשה בעיבוד דאטה שלא ניתן להפרדה ליניארית.

נוכל לשים לב כי האלגוריתם הגיע למספר האיטרציות המקסימלי שאפשרנו לו (1000 איטרציות) ועדיין לא הגיע לשגיאה שהיא 0, האלגוריתם ממשיך לחפש את ה- W הנכון, וכל שינוי שלו לדגימה

מסויימת גורר סטיה על דגימה אחרת, לכן יוכל להמשיך לרוץ לנצח, זאת בניגוד למידע בר ההפרדה משהשאלה הקודמת שלאחר 52 איטרציות כבר מצא על-מישור מפריד אופטימלי לדגימות.

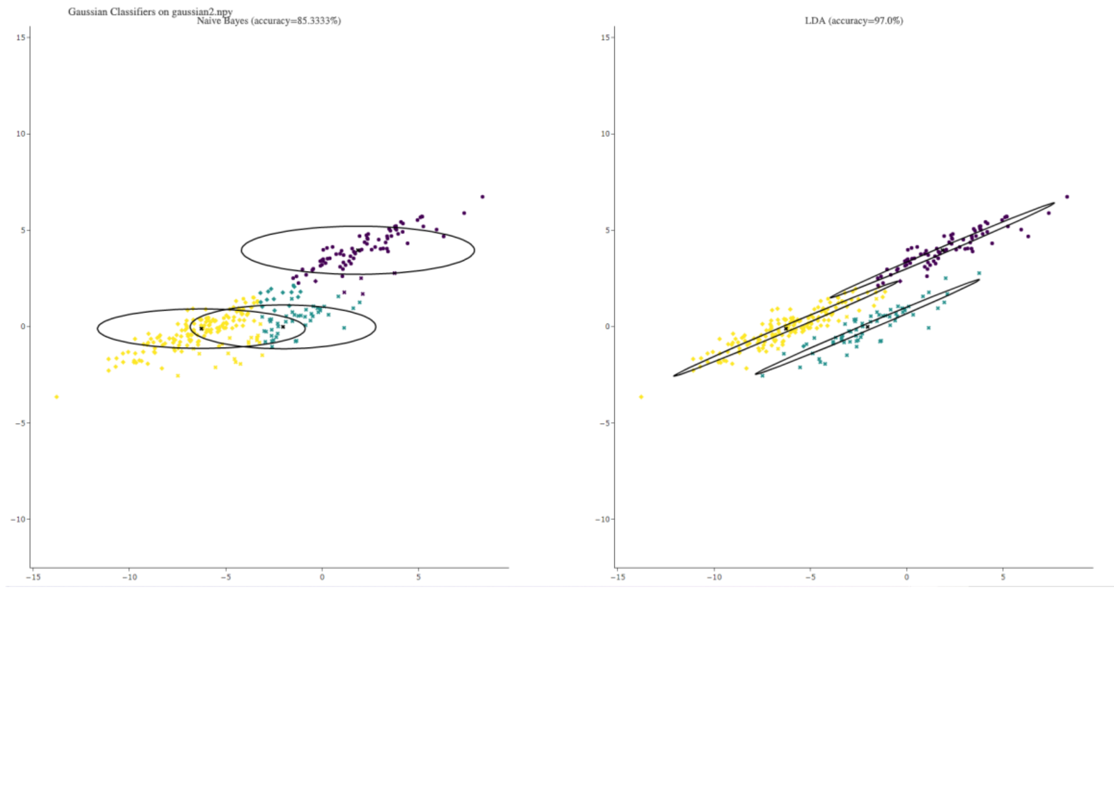
(3.2

(1



ניתן ללמוד מהגרף כי שני המודלים השיגו תוצאות טובות ולכן ניתן להניח כי הדגימות מתפלגות נורמלית עם שונות נמוכה

(2



מהגרף ניתן ללמוד כי לדגימות יש שונות יחסית גבוהה ושונות משותפת משמעותית. לכן עבור הדאטה הנ"ל ניתן לראות כי מודל ה LDA השיג תוצאות טובות באופן ניכר מה NAIVE BAYES ולכן נעדיף להשתמש בו.