Theoretical Part

2.1 Hard - & soft - SVM

1. נתון שאתקיים: $\|w\|^2 = w^T I w$ כש- I היא מטריצת היחידה.

מתקיים:

$$\forall i \quad y_i(\langle w, x_i\rangle + b) \geq 1 \iff \forall i \quad y_i(w^T x_i + b) \geq 1$$

$$\iff \forall i \quad -y_i(w^T x_i + b) \leq -1 \iff -y_i x_i^T w - y_i b \leq -1$$

ואז נקבל:

$$A = \begin{bmatrix} -y_1 x_{11} & -y_1 x_{12} & \dots & -y_1 x_{1n} & -y_1 \\ -y_2 x_{21} & -y_2 x_{22} & \dots & -y_2 x_{2n} & -y_2 \\ \vdots & & & & \\ -y_m x_{m1} & -y_m x_{m2} & \dots & -y_m x_{mn} & -y_m \end{bmatrix} = \begin{bmatrix} -y_1 x_1^T & \dots & -y_1 \\ -y_2 x_2^T & \dots & -y_2 \\ \vdots & & \vdots \\ -y_m x_m^T & & -y_m \end{bmatrix} \in M_{m \times (n+1)}$$

ואז נגדיר את וקטור $d \in \mathbb{R}^m$: $d = \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}$. בכך נקבל את המשוואה הרצויה $Av \leq d$.

נגדיר את v באופן הבא: $v = \begin{bmatrix} w \\ b \end{bmatrix}$ כלומר $v \in \mathbb{R}^{n+1}$

נגדיר את Q באופן הבא: $Q = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$ כאשר $Q \in M_{(n+1)\times(n+1)}$

באופן כזה כשנכפיל מכפלה $v^T Q v$ ה-b יוכפל באפס ובסוף נקבל $w^T I w = \|w\|^2$.

נגדיר את a כאוסף הבא: $a = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$.

ואז מתקיים

$$\underset{(w,b)}{\operatorname{argmin}} \|w\|^2 = \underset{(w,b)}{\operatorname{argmin}} w^T I w = \underset{(w,b)}{\operatorname{argmin}} v^T Q v$$

$$= \underset{(w,b)}{\operatorname{argmin}} \begin{bmatrix} w \\ b \end{bmatrix}^T Q \begin{bmatrix} w \\ b \end{bmatrix} = \underset{(w,b)}{\operatorname{argmin}} \frac{1}{2} \begin{bmatrix} w \\ b \end{bmatrix}^T Q \begin{bmatrix} w \\ b \end{bmatrix}$$

$$= \underset{(w,b)}{\operatorname{argmin}} \frac{1}{2} \begin{bmatrix} w \\ b \end{bmatrix}^T Q \begin{bmatrix} w \\ b \end{bmatrix} + 0^T \begin{bmatrix} w \\ b \end{bmatrix} = \underset{(w,b)}{\operatorname{argmin}} \frac{1}{2} v^T Q v + a^T v$$

Naive Bayes classifiers

2.a) בהינתן $X \in \mathbb{R}$ כך שלכל דגימה יש פיצ'ר אחד וכל x_i מתפלג בהתפלגות נורמלית עם k כך ש- $k \in \mathbb{N}$ הוא מספר התוויות, עם ממוצע μ_k ושונות σ_k^2 וההתפלגות של כל תווית y_i היא π_k.

ראינו בכיתה שפונקציית ה- likelihood היא:

$$L(\theta|X,y) \overset{i.i.d}{=} \prod_{i=1}^{m} f_{X,Y|\theta}(x_i, y_i) = \prod_{i=1}^{m} f_{X|Y=y}(x_i) \cdot f_{Y|\theta}(y_i)$$

$$= \prod_{i=1}^{m} \mathcal{N}(x_i|\mu_{y_i}, \Sigma) \cdot \mathrm{Mult}(y_i|\pi)$$

ורואינו שלמצוא את ה- MLE של ה likelihood שווה ל log-likelihood אני:

$$\ell(\theta|X,y) = \log\left(\prod_{i=1}^{m} \mathcal{N}(x_i|\mu_{y_i}, \Sigma) \cdot \mathrm{Mult}(y_i|\pi)\right)$$

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

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

כעת, כדי למצוא את argmax. נגזור לפי כל פרמטר בנפרד:

נמצא MLE לפי μ_k:

$$\frac{\partial}{\partial \mu_k} \ell(\theta | X,y) = \frac{\partial}{\partial \mu_k} \sum_{i=1}^{m} \left(-\frac{(x_i - \mu_{y_i})^2}{2\sigma_{y_i}^2} \right)$$

כעת, קבלל שאנחנו גוזרים לפי μ_k, כל הערכים מתאפסים חוץ מהבוטלנטים עם μ_k. כלומר, נקבל:

$$\frac{\partial}{\partial \mu_k} \left(-\frac{(x_i - \mu_{y_i})^2}{2\sigma_{y_i}^2} \right) = -\frac{1}{2\sigma_k^2} \cdot 2(x_i - \mu_k) \cdot -1 = \frac{x_i - \mu_k}{\sigma_k^2}$$

לכן, נקבל עבור $y_i = k$ לכל i

$$\frac{\partial}{\partial \mu_k} \sum_{i=1}^{m} \left(-\frac{(x_i - \mu_{y_i})^2}{2\sigma_{y_i}^2} \right) = \sum_{i=1}^{m} \frac{1}{\sigma_k^2} (x_i - \mu_k)$$

כלומר נקבל:

$$\boxed{\mu_k^{MLE} = \frac{\sum (x_i - \mu_k)\, \delta(y_i = k)}{\sigma_k^2}}$$

$\delta(y_i = k) = 1$ אינדיקטור מצ"ן

כעת, נגזור לפי $\sigma_{y_i}^2$:

$$\frac{\partial}{\partial \sigma_k^2} \ell(\theta | X,y) = \sum_{i=1}^{m} \frac{\partial}{\partial \sigma_k^2} \left(-\frac{1}{2}\log(2\pi\sigma_k^2) - \frac{(x_i - \mu_{y_i})^2}{2\sigma_k^2} \right)$$

גזירה לפי σ_k^2 תאפס את כל הערכים $y_i \neq k$ ותשאיר רק את $y_i = k$:

נגזור ונשווה ל-0 ונקבל:

$$\frac{\partial}{\partial \sigma_k^2} -\frac{1}{2}\log(2\pi\sigma_k^2) = -\frac{1}{2} \cdot \frac{1}{2\pi\sigma_k^2} \cdot 2\pi = -\frac{1}{4\pi\sigma_k^2} \cdot 2\pi = -\frac{1}{2\sigma_k^2}$$

$$\frac{\partial}{\partial \sigma_k^2} \frac{(x_i - \mu_{y_i})^2}{2\sigma_k^2} = \frac{(x_i - \mu_{y_k})^2}{2\sigma_k^4}$$

$$\Rightarrow \sum_{i=1}^{m} \left(-\frac{1}{2\sigma_k^2} + \frac{(x_i - \mu_k)^2}{2\sigma_k^4} \right) \delta(y_i = k) = 0$$

נכפיל ב- $2\sigma_k^4$ ונקבל:

$$\sum_{i=1}^{m}\left(-\sigma_k^2+(x_i-\mu_k)^2\right)\delta(y_i=k)=0$$

$$\rightarrow \sum_{i=1}^{m}-\sigma_k^2\,\delta(y_i=k)+\sum_{i=1}^{m}(x_i-\mu_k)^2\delta(y_i=k)=0$$

$$\rightarrow -\sigma_k^2\cdot n_k+\sum_{i=1}^{m}(x_i-\mu_k)^2\delta(y_i=k)=0$$

$$\rightarrow \boxed{\frac{\sum_{i=1}^{m}(x_i-\mu_k)^2\delta(y_i=k)}{n_k}=\sigma_k^2}$$

וכעת, נגזור לפי π_k:

$$\frac{\partial}{\partial\pi_k}\ell(\theta|X,y)=\frac{\partial}{\partial\pi_k}\log\pi_{y_i}$$

מכיוון שצריכים להתחשב באילוץ ש- $\sum_{i=1}^{k}\pi_i=1$ נשתמש בלגרנז'ן ונקבל:

$$\mathcal{L}=\ell(\theta|X,y)+\lambda\left(\sum_{i=1}^{k}\pi_i-1\right)$$

בכיתה ראינו

$$\frac{\partial\mathcal{L}}{\partial\pi_k}=\frac{n_k}{\pi_k}-\lambda=0 \iff \pi_k=\frac{n_k}{\lambda}$$

$\lambda=m$ כיון ש- $\sum_{i=1}^{k}\pi_i=1$ נקבל ש-

$$\boxed{\pi_k=\frac{n_k}{m}}$$

כאשר m הוא מספר כל הדגימות.

2.b עבור d פיצ'רים יתקיים:

$$\mathcal{L}(\theta \mid x,y) = \prod_{i=1}^{m} f_{X|Y=y}(x_i) \cdot f_{Y|\theta}(y_i)$$

$$= \prod_{j=1}^{d} \prod_{i=1}^{m} N(x_{ij} \mid \mu_{y_i j}, \sigma^2_{y_i j}) \cdot Mult(y_i \mid \pi)$$

נשים לב שלכל פיצ'ר הנצרת שנקבל עבור μ_k, π_k, σ_k^2 תואם אות כל האוטמנטים עם פיצ'רים אחרים. כלומר, הפיצ'רים הם בת"ל ולכן לכל $j \in [d]$ נקבל:

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

$$\sigma^2_{kj}{}^{MLE} = \frac{1}{n_k} \sum (x_{ij} - \mu_{kj})^2$$

$$\mu_{kj}^{MLE} = \frac{1}{n_k} \sum (x_{ij} - \mu_{kj})$$

3.a) נחשב את פונקציית ה-likelihood:

$$\mathcal{L}(\theta|x,y) = \prod_{i=1}^{m} f_{X,Y|\theta}(x_i, y_i) = \prod_{i=1}^{m} f_{X|Y=y}(x_i) \cdot f_{Y|\theta}(y_i)$$

$$= \prod_{i=1}^{m} \text{Pois}(x_i|\lambda_{y_i}) \cdot \text{Mult}(y_i)$$

נחשב את פונקציית log-likelihood:

$$\ell(\theta|x,y) = \log\left(\prod_{i=1}^{m} \text{Pois}(x_i|\lambda_{y_i}) \cdot \text{Mult}(y_i|\pi)\right)$$

$$= \sum_{i=1}^{m} \log(\text{Pois}(x_i|\lambda_{y_i}) + \log(\text{Mult}(y_i|\pi))$$

$$= \sum_{i=1}^{m} \log\left(\frac{e^{-\lambda_{y_i}} \cdot \lambda^{x_i}}{x_i!}\right) + \log(\pi_{y_i})$$

$$= \sum_{i=1}^{m} \log(e^{-\lambda_{y_i}} \cdot \lambda^{x_i}) - \log(x_i!) + \log(\pi_{y_i})$$

$$= \sum_{i=1}^{m} -\lambda_{y_i} + \log(\lambda^{x_i}) - \log(x_i!) + \log(\pi_{y_i})$$

$$= \sum_{i=1}^{m} -\lambda_{y_i} + x_i \log(\lambda) - \log(x_i!) + \log(\pi_{y_i})$$

$$= \sum_{y_i=k} \left[\log(\lambda_k) \cdot \sum_{i|y_i=k} x_i - n_k \cdot \lambda_k + n_k \cdot \log(\pi_k)\right] - \sum_{i=1}^{m} \log(x_i!)$$

כעת, נגזור לפי כל אחד מהפרמטרים ונשווה ל-0.

גזירה לפי λ_k:

$$\frac{\partial}{\partial \lambda_k} \ell(\theta|x,y) = \frac{\partial}{\partial \lambda_k} \sum_{y_i=k} \left[\log(\lambda_k) \cdot \sum_{i|y_i=k} x_i - n_k \cdot \lambda_k\right]$$

$$= \frac{1}{\lambda_k} \cdot \sum_{i|y_i=k} x_i - n_k$$

$$\frac{1}{\lambda_k} \cdot \sum_{i|y_i=k} x_i - n_k = 0 \iff \frac{1}{\lambda_k} \cdot \sum_{i|y_i=k} x_i = n_k$$

$$\iff \boxed{\lambda_k = \frac{1}{n_k} \cdot \sum_{i|y_i=k}^{m} x_i}$$

עבור π_k נקבל את אותה התוצאה מ- (a.2) ולכן

$$\boxed{\pi_k = \frac{n_k}{m}}$$

b.3) נחשב את פונקציית ה- likelihood

$$L(\theta|x,y) = \prod_{i=1}^{m} f_{X|Y=y}(x_i) \cdot f_{Y|\theta}(y_i)$$

$$= \prod_{1=i}^{m} \prod_{i=1}^{d} Pois(x_{ij}|\lambda_{y_i j}) \cdot Mult(y_i|\pi)$$

נחשב את פונקציית ה- log-likelihood :

$$\ell(\theta|x,y) = \log\left(\prod_{1=i}^{m}\prod_{i=1}^{d} Pois(x_{ij}|\lambda_{y_i j}) \cdot Mult(y_i|\pi)\right)$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{d} \log(Pois(x_{ij}|\lambda_{y_i j}) + \log(Mult(y_i|\pi))$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{d} \log(Pois(x_{ij}|\lambda_{y_i j}) + \sum_{i=1}^{m} \log(Mult(y_i|\pi))$$

$$= \sum_{k=1}^{K}\sum_{j=1}^{d} \log\left(\frac{e^{-\lambda_{kj}} \cdot \lambda^{x_{ij}}}{x_{ij}!}\right) + n_k \cdot \log(\pi_k)$$

$$= \sum_{k=1}^{K}\sum_{j=1}^{d}\left[n_k \cdot -\lambda_{kj} + \log(\lambda_{kj})\sum_{i:y_i=k} x_{ij} + n_k \cdot \log(\pi_k)\right] - \sum_{i=1}^{m}\log(x_{ij}!)$$

קבוצה δ(2.b) מתקיים ש-

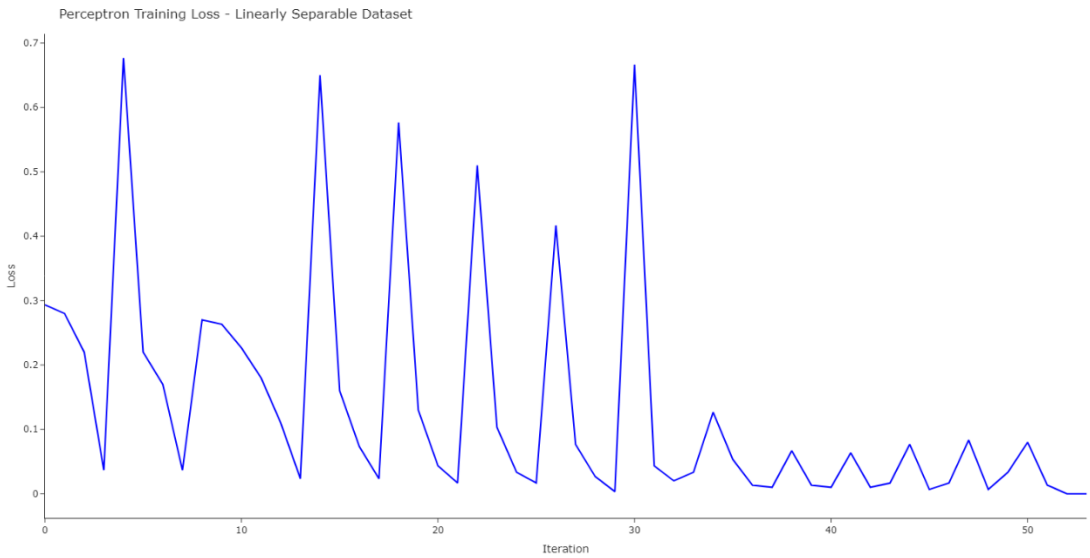
$$\boxed{\lambda_k^{MLE} = \frac{1}{n_k} \cdot \sum_{i|y_i=k}^{m} x_i}$$

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

Practical part – IML

3.1 Perceptron Classifier

1. Fitting and plotting over the linearly_seprable.npy dataset, what can we learn from the plot?



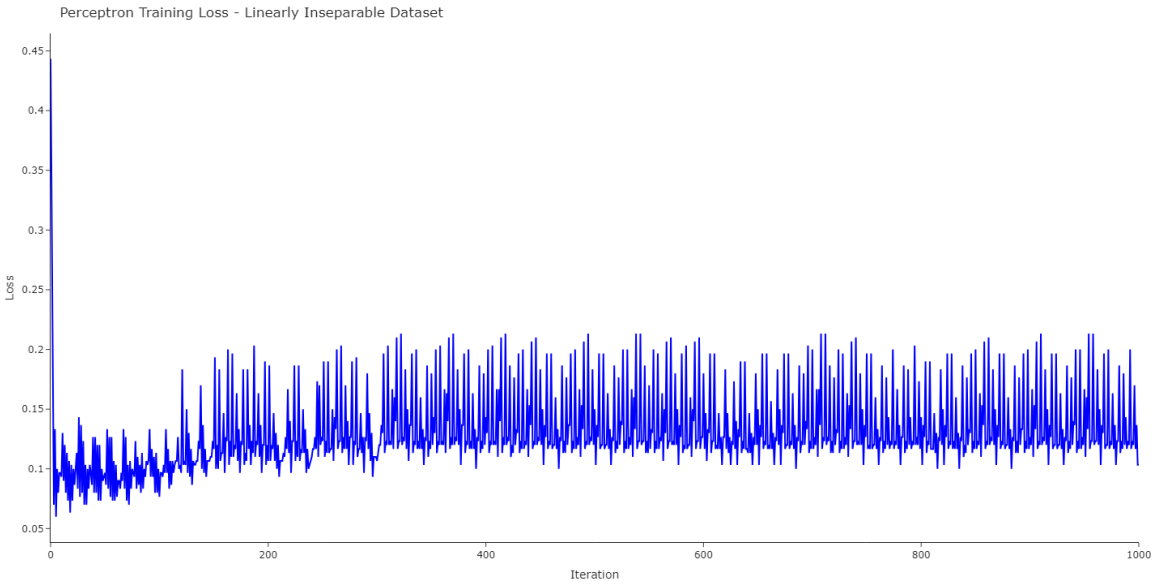
הגרף ממחיש את התהליך האיטרטיבי של אלגוריתם של Perceptron על מערך נתונים הניתן להפרדה ליניארית. זה מראה את יכולתו של האלגוריתם למצוא hyperplane מפריד בהצלחה, זאת ניתן לראות על ידי התייצבותו של הגרף בנקודה אפס לאחר 50+ איטרציות. התנהגות זו מדגימה את יעילותו ותכונת ההתכנסות של ה-Perceptron על מערכי נתונים הניתנים להפרדה ליניארית.

נשים לב שבתחילה, Perceptron מתחיל עם מספר גבוה של סיווגים שגויים, אך ככל מספר האיטרציות עולה, מתעדכן וקטור המשקולות ומספר הסיווגים השגויים הולך ופוחת משמעותית עד התייצבותו באפס.

התנודות החדות בגרף נובעות ככל הנראה מאופי עדכוני המשקל של ה-Perceptron , כאשר הוא מבצע תיקונים עבור נקודות שסווגו בצורה שגויה. אך החל מאיטרציה 30 הוא מצליח לסווג את הנקודות בצורה נכונה עד שלבסוף הוא מצליח להביא את ה MLE לאפס.

בנוסף, מכיוון שהמערך נתונים הוא מופרד ליניארית, מובטח שהמסווג יימצא hyperplane שיחלק אותו באופן לינארי ומוצלח ואכן הגרף מאשר זאת על ידי התייצבותו החל מאיטרציה 50+ סביב נקודת האפס.

2. <u>What is the difference between this plot and to the one in the previous question? How can we explain the difference in terms of the objective and parameter space?</u>



מהגרף ניתן לראות שעבור מערך נתונים שאינו ניתן להפרדה האלגוריתם אינו מתכנס ומתייצב סביב נקודת האפס כלל מה שממחיש את חוסר ההצלחה שלו בלמצוא hyperplane מפריד בהצלחה.

נשים לב לשונות בין שני הגרפים הבאה לידי ביטוי נקודות הבאות:

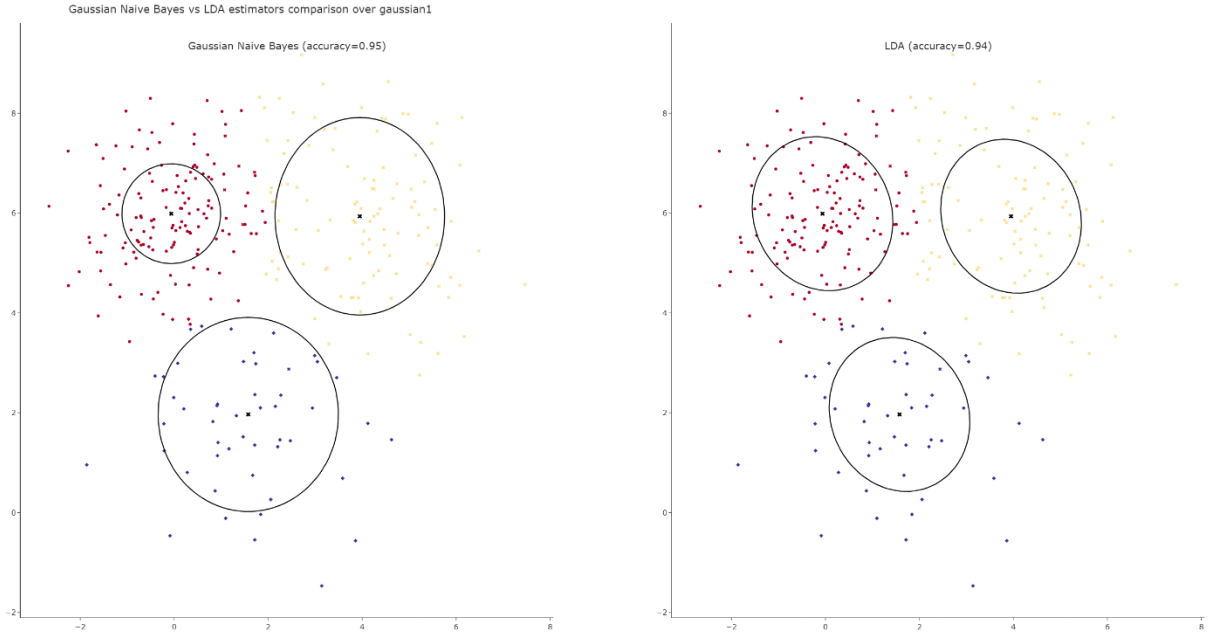
התנהגות הפונקציה: במקרה של מערך נתונים הניתן להפרדה, הפונקציה הולכת ומתייצבת על אפס בעוד שבגרף של מערך נתונים בלתי ניתן להפרדה הפונקציה עולה ויורדת ולא מתכנסת בשום שלב עד לעצירת האלגוריתם אחרי max_iter איטרציות.

ערכי loss נמוכים: במקרה של מערך נתונים הניתן להפרדה, ערך הloss הולך וקטן במהלך האיטרציות ואילו בגרף מערך נתונים בלתי ניתן להפרדה, ערכי הloss נשארים גבוהים ומתנודדים בערכי טווח מסוים, מה שמשקף סיווגים שגויים מתמשכים.

ההבדל בנתונים מדגיש את ההשפעה של הפרדת הנתונים על יכולתו של ה-Perceptron לייעל את המטרה שלו. עבור נתונים הניתנים להפרדה ליניארית, ה-Perceptron מוצא פתרון מושלם במרחב הפרמטרים שממזער את פונקציית המטרה לאפס. עבור נתונים בלתי ניתנים להפרדה באופן ליניארי, ה-Perceptron נאבק למצוא פתרון יציב, מה שמוביל לעלייה וירידה מתמשכת כאשר הוא מנסה להתייצב ולמזער ללא הצלחה סיווגים שונים.

3.2 Bayes Classifiers

1. Explain what can be learned from the plots above regarding the distribution used to sample the data?

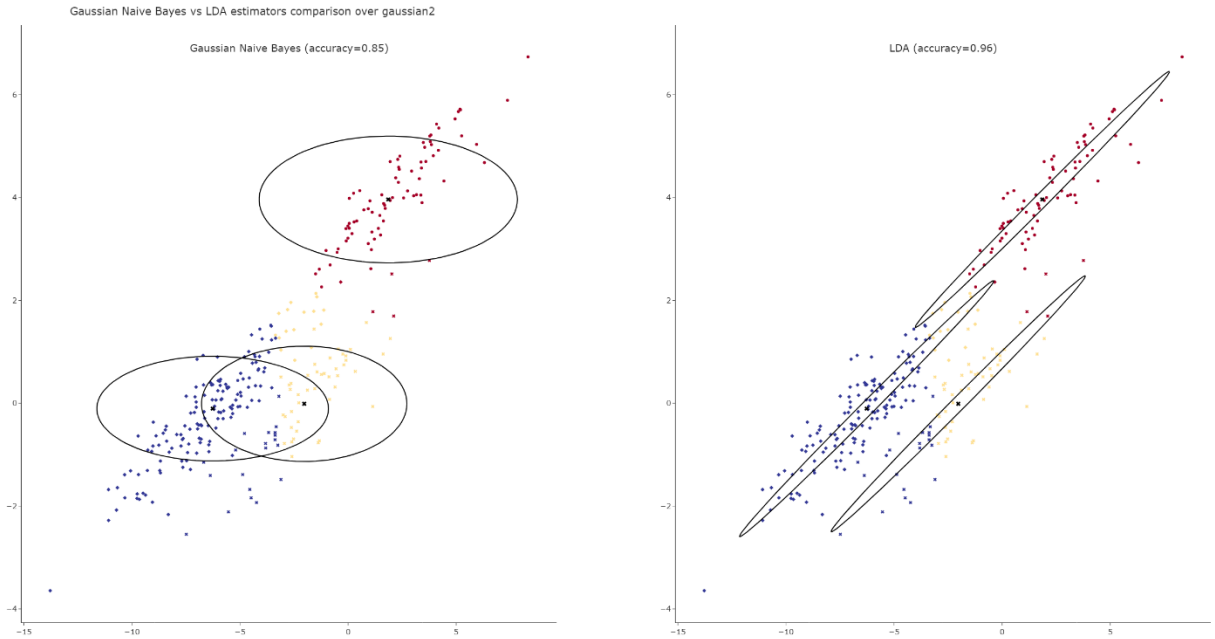


מהגרף ניתן להסיק כי ההתפלגות המשותפת שממנה נדגמו הדגימות היא נמוכה. נסתכל על תוצאות הגרף של ה LDA classifier, מודל זה משתמש בחישוביו בהנחה שיש תלות בין הפיצ'רים. כמו כן, עבור כל מחלקה הוא משתמש באותה מטריצת שונות משותפת בחישוב (לכן נקבל שהאליפסות הן זהות עבור כל מחלקה) בעוד שה GNB classifier אינו משתמש בהנחה שיש תלות בין הפיצ'רים ועבור כל מחלקה הוא משתמש בשונות שלה עצמה (לכן נקבל שכל אליפסה היא בגודל שונה).

לכן, מבחינת ביצועים שני הclassifiers משיגים תוצאות יחסית דומות וקרובות כיוון שהשונות המשותפת היא נמוכה מאוד ולכן בLDA אף שהוא משתמש בהנחה שיש תלות בין הפיצ'רים, זה לא משפיע על החישובים.

ניתן להסיק גם שההתפלגות המשותפת היא נמוכה על ידי פיזור הנקודות. נשים לב שעבור נקודה כלשהי, ערך ה x שלה לא משפיע על ערך ה y.

2. <u>What is the difference between the two scenarios? What can be learned regarding the distribution used to sample the data? Which of the two classifiers better matches this dataset and why?</u>



מהגרף ניתן להסיק שההתפלגות שממנה נדגמו הדגימות היא גבוהה משמעותית. ניתן לראות זאת מהגרף של LDA classifier שמניח שישנה תלות בין הפיצ'רים ואכן רואים שישנה התפלגות משותפת על ידי צורת האליפסה העקומה שנוצרה. בנוסף, ניתן לראות זאת על ידי פיזור הנקודות, שעבור נקודה כלשהי, ערך ה x משפיע על ערך הy (ככל שערך האיקס גבוה, ערך הy גבוה גם).

לעומת זאת, GNB classifier לא מצליח לתפוס את ההתפלגות, שכן הוא לא משתמש בהנחה שיש תלות בין הפיצ'רים, לכן, עבור הסט נתונים הזה, הביצועים שלו יהיו נמוכים לעומת ה LDA שמשיג תוצאה ממש גבוהה.

לכן נסיק שעבור סט הנתונים הזה, ההתפלגות המשותפת היא גבוהה ולכן המודל לחיזוי שיהיה מתאים יותר הוא LDA classifier.