IML 2 תשובות תרגיל
צבי מרמור 206949539

2. חלק תיאורטי

2.1. Hard & Soft SVM

1. ראשית נכתוב את הדרישה:

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

נכפיל את שני אגפי הא"ש ב -1.

$$\forall i \quad -y_i x_i^T w - y_i b \leq -1$$

נשים לב כי ניתן להגדיר כמו זה $A\in M_{m\times(d+1)}$ ווקטורי $v,d\in\mathbb{R}^{d+1}$

כך ש

$$\forall i \quad -y_i x_i^T w - y_i b \leq -1 \iff \forall i \quad Av\leq d$$

בצורה הבאה.

$$\forall i \quad [A]_i = -\begin{bmatrix} y_i x_i \\ \hline y_i \end{bmatrix} \qquad v_i = \begin{bmatrix} w \\ \hline b \end{bmatrix} \qquad d_i = -1$$

כך שמתקיים:

$$Av\leq d \iff \forall i \quad A_i v_i \leq -1$$

$$\iff -y_i x_i^T w - y_i b \leq -1 \iff y_i x_i^T w + y_i b \geq 1 \iff y_i(\langle w, x_i\rangle + b)\geq 1$$

נגדיר בנוסף $a=0$, ומטריצה $Q\in M_{(d+1)\times(d+1)}$ כך ש

$Q_{i,j} = \mathbb{1}_{i=j \cap i\leq d}$ ולכן מתקיים.

$$\arg\min \|w\|^2 = \arg\min \begin{bmatrix} w \\ \hline b \end{bmatrix}^T Q \begin{bmatrix} w \\ \hline b \end{bmatrix}$$

$$= \arg\min \frac{1}{2}\begin{bmatrix} w \\ \hline b \end{bmatrix}^T Q \begin{bmatrix} w \\ \hline b \end{bmatrix} + 0^T \begin{bmatrix} w \\ \hline b \end{bmatrix} = \arg\min \frac{1}{2} v^T Q v + a^T v$$

(לא משנה את ה argmin) (אפס)

כנדרש. ∎

② k. ראשית, נרשום את פונקציית ה likelihood.

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

$$=\prod_{i=1}^{m}\mathcal{N}(x_i|\mu_{y_i},\sigma^2_{y_i})\cdot \text{mult}(y_i|\pi)$$

מההפשטות $x_j|y$ ו y משפ' ב(4)

כעת, כמו שראינו בתרגול, נעביר את הפונקציה L ללוג ℓ הפונקציה על מנת לעבור לסכימת דברים (לא מכפלות):

$$\ell(\Theta|X,y)=\log\left(\prod_{i=1}^{m}\mathcal{N}(x_i|\mu_{y_i},\sigma^2_{y_i})\cdot \text{mult}(y_i|\pi)\right)$$

$$=\sum_{i=1}^{m}\log(\text{mult}(y_i|\pi))+\log(\mathcal{N}(x_i|\mu_{y_i},\sigma^2_{y_i}))$$

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

(ולפי ההפשטות הן)

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

$$=-\frac{m}{2}\log(2\pi)+\sum_k n_k\log(\pi_k)-\frac{1}{2}n_k\log(\sigma^2_k)-\frac{1}{2\sigma^2_k}\sum_{\{i|y_i=k\}}(x_i-\mu_k)^2$$

לכל $k\in[k]$ מגדירים $n_k=\sum_i \mathbb{1}_{[y_i=k]}$. נגדיר

$$\bar{X}_k=\frac{1}{n_k}\sum_{i|y_i=k}x_i$$

כעת נגזור את ה-ℓ ונשווה לאפס:

$$\frac{\partial\ell}{\partial\mu_k}=-\frac{1}{\sigma^2_k}\sum_{\{i|y_i=k\}}(x_i-\mu_k)=0$$

ולכן כי

$$\hat{\mu}_k^{MLE}=\bar{X}_k$$

(כאשר MLE = maximum Likelyhood estimators).

ולגבי ה-σ:

$$\frac{\partial\ell}{\partial\sigma^2_k}=-\frac{n_k}{2\sigma^2_k}-\frac{1}{2(\sigma^2_k)}\sum_{\{i|y_i=k\}}(x_i-\mu_k)^2$$

$$(\hat{\sigma}^2)_k^{MLE}=\frac{1}{n_k}\sum_{\{i|y_i=k\}}(x_i-\hat{\mu}_k^{MLE})^2$$

נמצא את ה maximizer של ההסתברויות $\{\pi_k\}$

$$L = l(\theta | x,y) - \lambda\left(\sum_k \pi_k - 1\right)$$ כאשר

ואז נגזור ונשווה ל0:

$$\frac{\partial L}{\partial \pi_k} = \frac{\partial l(\theta|x,y)}{\partial \pi_k} - \lambda = \frac{n_k}{\pi_k} - \lambda = 0$$

ולכן

$$\pi_k = \frac{n_k}{\lambda}$$

נמצא את λ: מאילוץ $\sum_k \pi_k = 1$ מהגדרה, ולכן

$$1 = \sum_k \pi_k \underset{\text{הגדרנו}}{=} \sum_k \frac{n_k}{\lambda}$$

ולכן מאילוץ $\lambda = m$. ולכן

$$\hat{\pi}_k^{MLE} = \frac{n_k}{m}$$

ב) דומה לסעיף א, רק עכשיו עם ה Likelihood func.

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

$$= \left(\prod_{j=1}^{d} \prod_{i=1}^{m} \mathcal{N}(x_{ij} | \mu_{y_i j}, \sigma^2_{y_i, j})\right) \cdot mult(y_i | \pi)$$

וכמו בסעיף א, נפעיל "תכונות ה log" על הפונקציה $l(\theta|x,y)$:

$$l(\theta|x,y) = \sum_{i=1}^{m}\left(\log(mult(y_i|\pi)) + \sum_j \log(\mathcal{N}(x_{ij} | \mu_{y_i j}, \sigma^2_{y_i j}))\right)$$

$$= \sum_k n_k \log(\pi_k) + \sum_k \sum_j \mathbb{1}_{[y_i = k]} \log(\mathcal{N}(x_{ij} | \mu_{kj}, \sigma^2_{kj}))$$

נשים לב שלכל $j \in [d]$ המשערך של כל אחד מהפרמטרים האופטימליים נגזר רק מ

ונראה שלא תלוי בפיצ'ר j ובשאר הפרמטרים. לכן, בהתאם לסעיף הקודם:

$$\hat{\pi}_k^{MLE} = \frac{n_k}{m}, \quad \hat{\mu}_{kj}^{MLE} = \frac{1}{n_k}\sum_{\{i | y_i = k\}} x_{ij}, \quad (\hat{\sigma}^2)_{kj}^{MLE} = \frac{1}{n_k}\sum_{\{i|y_i=k\}}\left(x_{ij} - \hat{\mu}_{kj}^{MLE}\right)^2$$

③ k. נגדיר (כמו בשאלה קודמת) רכיב/נפשו את Likelihood func

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

$$=\prod_{i=1}^{m} Poi(x_i|\lambda_{y_i})\cdot mult(y_i|\pi)$$

נעשה כך עם log likelihood:

$$l(\Theta|X,y)=\sum_{i=1}^{m} \log(Poi(x_i|\lambda_{y_i}))+\log(mult(y_i|\pi))$$

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

הצבת poi ופישוט

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

ניקח את הנגזרת בהתחשב ב $\{\lambda_k\}$ ונשווה ל0:

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

ונקבל ש

$$\hat{\lambda}_k^{MLE}=\frac{1}{n_k}\sum \mathbb{1}_{[y_i=k]}x_i$$

במקרה ומדובר ב $\{\pi_k\}$ הנגזרת זהה / לשאלה 2 ולכן:

$$\hat{\pi}_k^{MLE}=\frac{n_k}{m}.$$

ב. נגדיר כמו בסעיפים הקודמים ונגדיר את Likelihood func

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

$$=\prod_{i=1}^{m}\left(\prod_{j=1}^{d} Poi(x_{ij}|\lambda_{y_ij})\right)\cdot mult(y_i|\pi)$$

ולכן כך

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

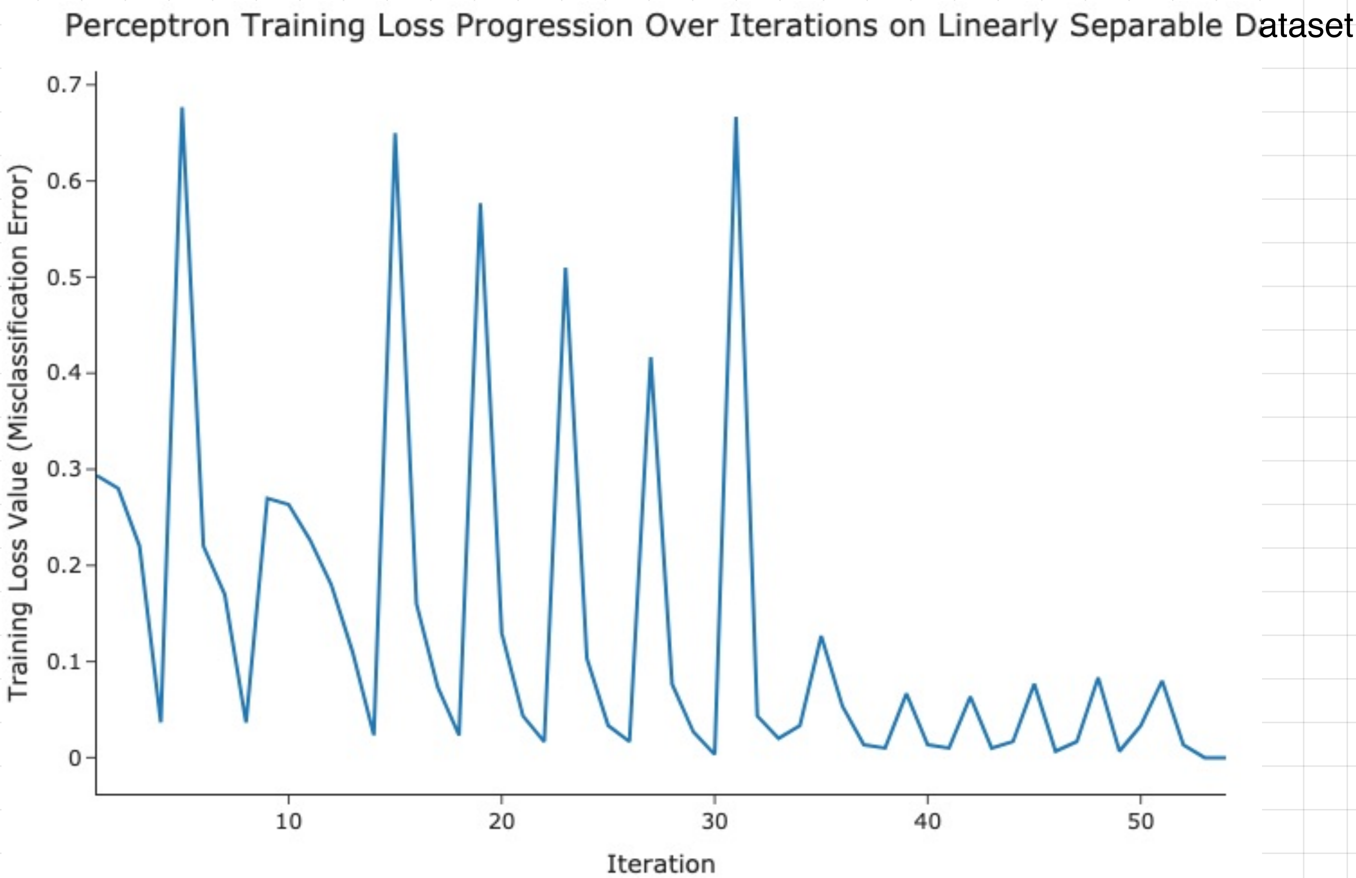
$$=\sum_k\sum_j\left(\log(\lambda_{kj})\sum_{i|y_i=k}x_{ij}-n_k\cdot\lambda_{kj}+n_k\cdot\log(\pi_k)\right)-\sum_{i,j}\log(x_{ij}!)$$

ולכן ה maximiers הם:

$$\hat{\lambda}_{kj}^{MLE}=\frac{1}{n_k}\sum_i \mathbb{1}_{[y_i=k]}x_{ij}\,,\quad \hat{\pi}_k^{MLE}=\frac{n_k}{m}$$

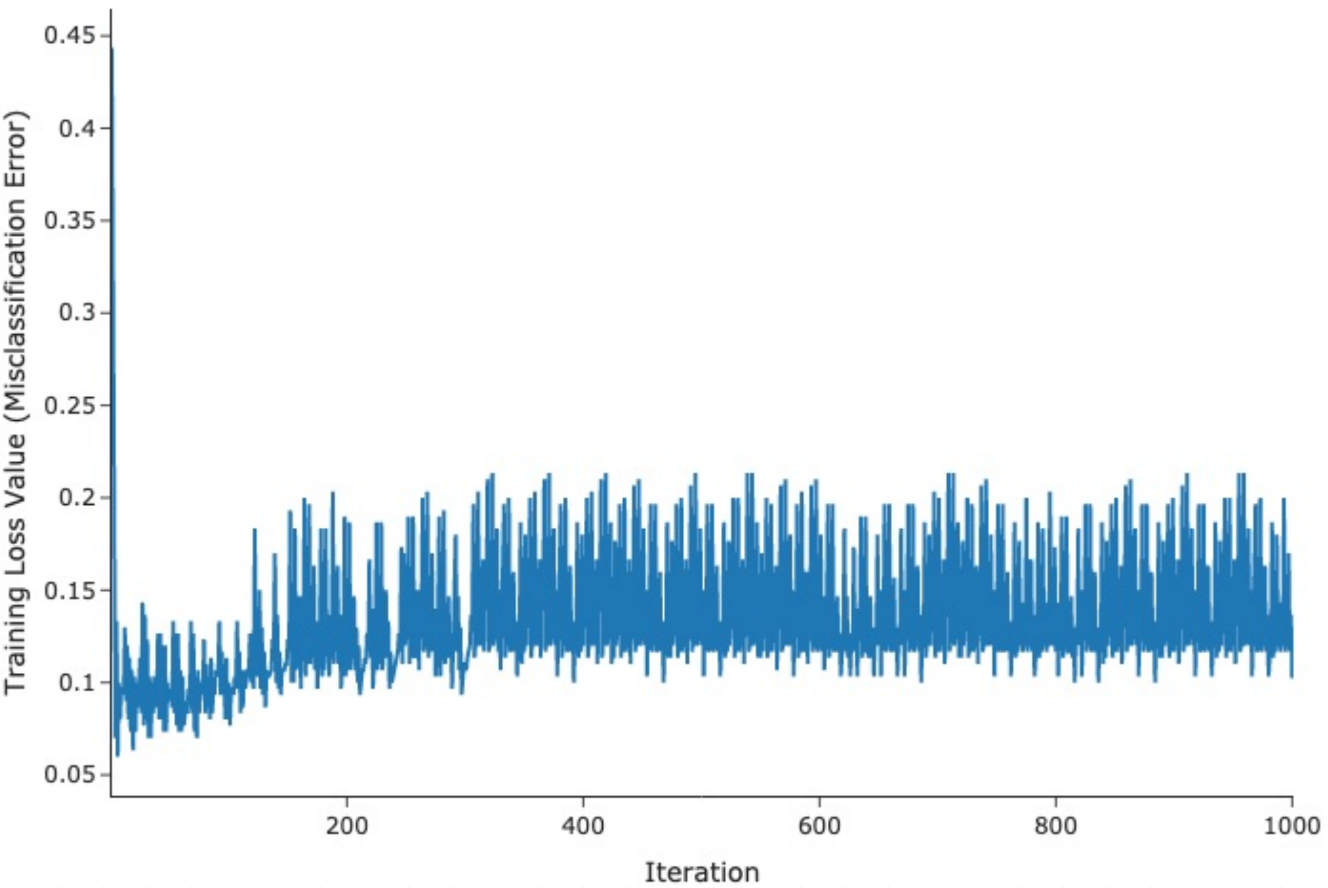
3 חלק פרקטי

3.1.1 מהגרף המוצג מטה אנו יכולים לראות את אופן פעולת האלגוריתם על dataset הניתן. ניתן לראות איך לאורך האיטרציות האלגוריתם ׳מצליח׳ להוריד את הmisclassification error בעזרת ההתאמות לcoeffs שמבוצעות על ידי האלגוריתם. האלגוריתם אכן ׳מצליח׳ לתקן את אופן החלוקה הלינארית שלו את הדגימות במרחב עד כדי הורדת הerror לס.
מכך שהאלגוריתם הצליח לעשות דבר זה, גם אם הדבר כלל מספר איטרציות אשר גרמו לאלגוריתם להתקשות עם הורדת הerror, מעידה על כך שככל הנראה הדאטה הניתן לאלגוריתם במקרה זה היה ניתן להפרדה לינארית ברורה, עובדה האיפשרה לאלגוריתם למצוא חלוקה זו לבסוף.

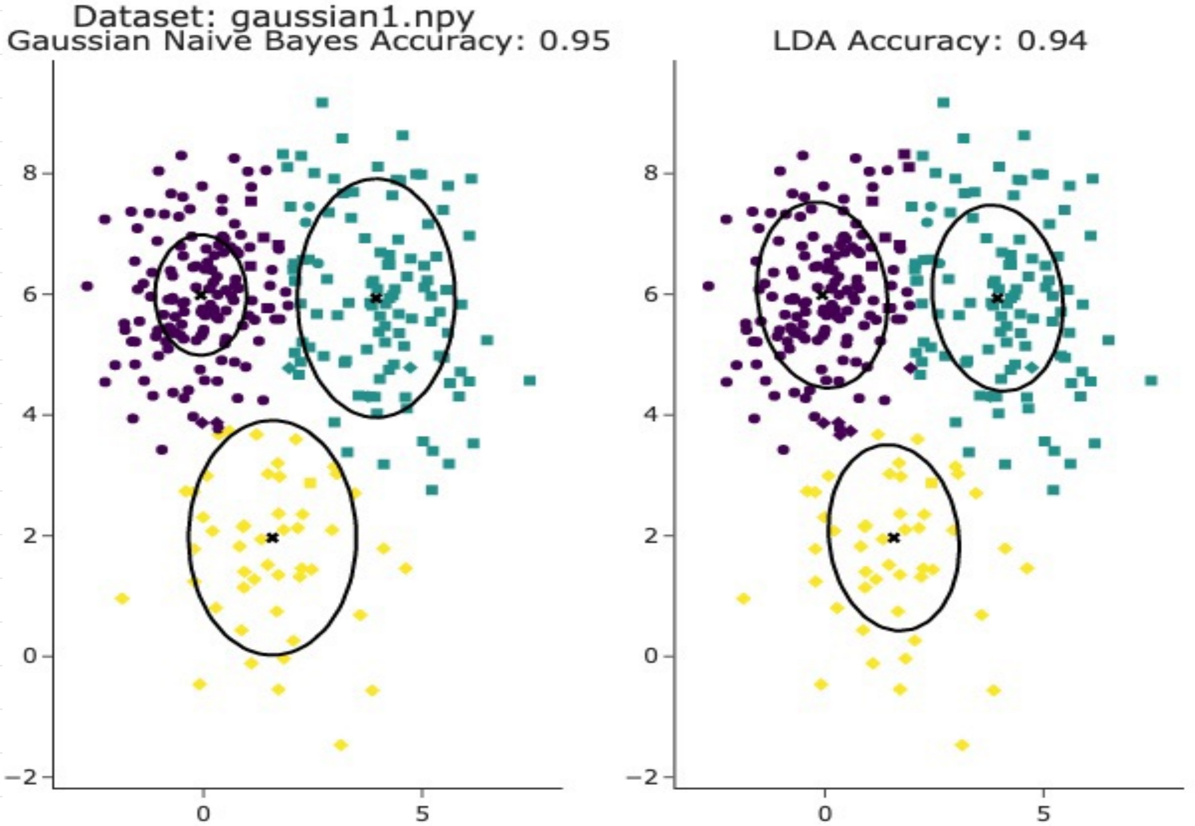


3.1.2 בגרף המוצג מטה אנו רואים את פעילות האלגוריתם על הdataset השני. בנקל ניתן לשים לב כי האלגוריתם לא 'מצליח' להוריד את הmissclassification error על ידי חלוקה יעילה של הדאטה ותיקון מיקום קו החלוקה על ידי פעילות האלגוריתם על הcoeffs. ניתן לראות כי האלגוריתם הגיע למספר האיטרציות המקסימלי- 1000 בעוד שהerror אינו מראה מגמה של ירידה או קרבה ל0.
ניתן להבין מכך על הדאטה הניתן כי ככל הנראה דאטה זה לא מאפשר חלוקה של כלל הsamples לשתי קבוצות ומעבר קו לינארי בדאטה, דבר המקשה על פעולת האלגוריתם ועל כך שבכל פעם שהוא 'מתקן' את הcoeffs הוא עדיין לא גורם להורדת הerror ועל כן הוא ממשיך לנסות על ידי תיקונים להוריד את הerror באיטרציות הבאות. דבר זה לא מתאפשר מכיוון שהדאטה הניתן ככל הנראה לא ניתן לחלוקה לינארית מבוקשת כזו.

Perceptron Training Loss Progression Over Iterations on Linearly Inseparable Dataset



3.2.1 בגרף המוצג מטה אנו יכולים לזהות שגם הlda classifier וגם הgaussian naive classifier מציגים את הsamples בצורה דומה ושרמת הדיוק שלהם כמעט זהה. על פי מה שלמדנו בתרגול ועל פי אופן הפעולה של שני הclassifiers אנו יכולים להבין כי ההתפלגות של הדגימות בדאטה סט זה היא גאוסיינית בעוד שרוב הדגימות הם בעלי covariance נמוך בין אחת לשנייה (וקרוב ל0), דבר אשר גרם לכך ששני המסווגים קיבלו תוצאות כמעט זהות.



3.2.2 בשונה מהגרף הקודם, פה אנו רואים כי הlda classifier השיג ביצועים גבוהים יותר מהgnb classifier. ניתן לראות זאת על ידי רמת הדיוק וההבדלים ביניהם (0.97 מול 0.85) וכן על פי האליפסות שנוצרו בשני המודלים וההתאמה שלהם אל הדגימות שבדאטה סט. מכיוון שהlda classifier ׳מתחשב׳ במקרים בהם קיימת covarience (שונות) גבוהה בין הדגימות בדאטה סט בעוד שהgnb classifier לא, אנו יכולים להבין כי הדגימות בדאטה סט אכן כללו שונות גבוהה ביניהן ועל כן נוצר פער גדול בין המסוגים, כאשר מסווג הlda הצליח ׳ללמוד׳ את השונות בין הדגימות ולקבל רמת דיוק גבוהה, בעוד שמסווג הgnb אשר מתבסס על הנחת שונות נמוכה עד אפסית בין הדגימות קיבל רמת דיוק נמוכה משמעותית.