מבוא למערכות לומדות | תרגיל 3

שם : אלון ויזנר | תז 318592052

23 באפריל 2022

2. חלק תיאורטי

2.1 מכונת וקטורים תומכים

1. הוכיחו שבעיית ה-Hard SVM הבאה היא בעיית תכנון ריבועי:

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

כלומר, מצאו מטריצות Q ו-A ווקטורים a, b כך שהבעיה לעיל תיכתב באופן הבא:

$$\underset{v\in\mathbb{R}^n}{\text{argmin}}\frac{1}{2}v^TQv + a^Tv\ \ s.t\ \ Av \leq d$$

ראשית, במקום לכתוב את $w \in \mathbb{R}^d$ ו-$b \in \mathbb{R}$, נייצג את b באמצעות קורדינטה נוספת ב-w כך ש-$w \in \mathbb{R}^{d+1}$, ובאופן זה גם הדגימות $x_i \in \mathbb{R}^{d+1}$. נקבל שהאילוצים לעיל שקולים ל-

$$\forall i: y_i\left\langle w, x_i\right\rangle \geq 1$$

אולם, פונקציית המטרה תשתנה, כיוון שאנו לא מעוניינים למזער את הקורדינטה שנוספה ל-w. הבעיה למעשה שקולה ל-

$$\underset{v\in\mathbb{R}^{d+1}}{\text{argmin}}\,||v - v_{d+1}||^2 \quad s.t\ \ \forall i: y_i\left\langle v, x_i\right\rangle \geq 1$$

כאשר v_{d+1} היא הקורדינטה האחרונה של הווקטור v. נשים לב ש-

$$||v - v_{d+1}||^2 = v^TQv$$

כאשר Q היא מטריצת היחידה ללא מופע של 1 בשורה האחרונה, כלומר -

$$Q = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{bmatrix}$$

ובסה"כ נקבל שהבעיה לעיל שקולה ל-

$$\underset{v\in\mathbb{R}^{d+1}}{\text{argmin}} v^T Q v \quad s.t \;\; \forall i : y_i \langle v, x_i \rangle \geq 1$$

את האילוצים $\forall i : y_i \langle v, x_i \rangle \geq 1$ נוכל לכתוב באופן הבא:

$$\begin{aligned} & y_i \langle v, x_i \rangle \geq 1 \iff \\ & -y_i \langle v, x_i \rangle \leq -1 \iff \\ & \langle v, x_i \rangle \leq y_i \iff \\ & Xv \leq y \end{aligned}$$

כאשר $X \in M_{n\times d+1}(\mathbb{R})$ היא המטריצה ש-$(x_1, ..., x_n)$ הן שורותיה, ו-$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$. בסה"כ קיבלנו את הבעיה השקולה-

$$\underset{v\in\mathbb{R}^{d+1}}{\text{argmin}} v^T Q v \quad s.t \;\; Xv \leq y$$

כנדרש, עבור $d = y$, $A = X$ ו-Q.

2. נביט בבעיית אופטימיזציה של Soft-SVM הבאה:

$$\underset{w,\{\zeta_i\}}{\text{argmin}}\frac{\lambda}{2}||w||^2+\frac{1}{m}\sum_i \zeta_i \quad s.t\ \forall i: y_i\langle w,x_i\rangle \geq 1-\zeta_i \wedge \zeta_i\geq 0$$

נסמן את פונקציית ה-hinge-loss ב-$l^{hinge}(a) := \max\{0,1-a\}$. הראו שבעיית אופטימיזציית ה-Soft-SVM שקולה לבעיית האופטימיזציה הבאה:

$$\underset{w,\{\zeta_i\}}{\text{argmin}}\frac{\lambda}{2}||w||^2+\frac{1}{m}\sum_i l^{hinge}(y_i\langle w,x_i\rangle)$$

כיוון שהאילוצים דורשים $\forall i: y_i\langle w,x_i\rangle \geq 1-\zeta_i \wedge \zeta_i\geq 0$, ההשמה האופטימלית עבור $\zeta_1,...,\zeta_m$ היא

$$\zeta_i=\begin{cases}0 & y_i\langle w,x_i\rangle\geq 1\\ 1-y_i\langle w,x_i\rangle & y_i\langle w,x_i\rangle\leq 1\end{cases}$$

או באופן שקול,

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

ולכן בעיית האופטימיזציה שקולה ל-

$$\underset{w,\{\zeta_i\}}{\text{argmin}}\frac{\lambda}{2}||w||^2+\frac{1}{m}\sum_i l^{hinge}(y_i\langle w,x_i\rangle) \quad s.t\ \forall i: y_i\langle w,x_i\rangle \geq 1-\zeta_i \wedge \zeta_i\geq 0$$

אך $\{\zeta_i\}_{i=1}^m$ שממזערים את פונקציית המטרה גם מקיימים את האילוצים, ולכן ניתן להשמיטם, ולמעשה הבעיה שקולה ל-

$$\underset{w}{\text{argmin}}\frac{\lambda}{2}||w||^2+\frac{1}{m}\sum_i l^{hinge}(y_i\langle w,x_i\rangle)$$

כפי שרצינו להראות.

2.2 סיווג בייסיאני נאיבי

3. מסווג בייסיאני נאיבי נורמלי מניח פריור מולטינומי, ו-likelihood בלתי-תלוי

$$\begin{aligned} & y \sim Multinomial\left(\pi\right) \\ & x_j | y = k \overset{ind}{\sim} \mathcal{N}\left(\mu_{kj}, \sigma_{kj}^2\right) \end{aligned}$$

עבור π **וקטור הסתברויות** $\pi \in [0,1]^K : \sum \pi_j = 1$**.**

(א) נניח ש-$x \in \mathbb{R}$**, כלומר לכל דגימה יש פיצ'ר יחיד. בהינתן דגימות** $\{(x_i, y_i)_{i=1}^m\}$**, התאימו מסווג בייסיאני נאיבי נורמלי שפותר את** (5) **תחת הנחות** (6)**.**

פונקציית הנראות ניתנת ע"י

$$\begin{aligned} \mathcal{L}\left(\Theta|X,y\right) = & \quad f_{X,y|\Theta}\left(\{x_i, y_i\}\right) = \\ iid & \quad \prod_{i=1}^m f_{X|Y=y_i}\left(x\right) f_Y\left(y_i\right) = \\ & \quad \prod_{i=1}^m \mathcal{N}\left(x_i|\mu_{y_i}, \sigma_{y_i}\right) \pi_{y_i} = \end{aligned}$$

הביטוי שממקסם את $\mathcal{L}\left(\Theta|X,y\right)$ ממקסם גם את $\log$ הביטוי, כלומר-

$$\begin{aligned} l\left(\Theta|X,y\right) = & \quad \sum_{i=1}^m \log\left(\mathcal{N}\left(x_i|\mu_{y_i}, \sigma_{y_i}\right) \pi_{y_i}\right) = \\ & \quad \sum_{i=1}^m \log \pi_{y_i} - \log \sigma_{y_i} - \frac{\left(x_i - \mu_{y_i}\right)^2}{2\sigma_{y_i}^2} = \\ & \quad \sum_k n_k \left(\log\left(\pi_k\right) - \log\left(\sigma_k\right)\right) - \frac{1}{2} \sum_{i:y_i=k} \frac{\left(x_i - \mu_k\right)^2}{\sigma_k^2} \end{aligned}$$

כאשר $n_k = \sum_i 1_{[y_i=k]}$. עלינו למצוא $\pi, \sigma \in \mathbb{R}^{[K]}$ הממקסמים את הביטוי לעיל תחת ההגבלה $\sum_k \pi_k = 1$. נשים לב שמקסום של הביטוי לעיל תחת ההגבלה שקול למקסום של

$$\mathcal{L} = l\left(\Theta|X,y\right) - \lambda\left(\sum_k \pi_k + 1\right)$$

נגזור את הביטוי לפי כל אחד מהמשתנים $\{(\pi_k), (\sigma_k), (\mu_k)\}$ ונשווה ל-0:

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

ועל מנת לקבל λ שמקיים את האילוצים,

$$1 = \sum_k \pi_k = \sum_k \frac{n_k}{\lambda} \implies \lambda = \sum_k n_k = m$$

על מנת למצוא את הפרמטרים $\{(\mu_k, \sigma_k)\}$ שממקסמים את הנראות: נשים לב שהאיברים בביטוי $l(\Theta|X, y)$ שתלויים ב-$\{(\mu_k, \sigma_k)\}$ זהים לאלו שהסקנו במקסום של פונקציית הנראות של משתנה גאוסייני יחיד, כפי שראינו בתרגול

$$\hat{\mu}_k^{MLE} = \frac{1}{n_k} \sum_i 1_{[y_i = k]} x_i, \quad \hat{\Sigma}^{MLE} = \frac{1}{m} \sum_i \left(x_i - \mu_{y_i}^{MLE}\right)^2$$

(ב) נניח שכל דגימה נמצאת ב- $x \in \mathbb{R}^d$. בהינתן דגימות $\{(x_i, y_i)_{i=1}^m\}$, התאימו מסווג בייסיאני נאיבי נורמלי שפותר את (5) תחת הנחות (6).

עלינו למצוא ביטוי שפותר את (5).
כיוון שאנו מניחים שהפיצ'רים בלתי תלויים, פונקציית הנראות היא מכפלת פונקציית הנראות במשנה יחיד.
כלומר, $x|Y = k$ מתפלג $\mathcal{N}(\mu_k, \Sigma_k)$ כאשר μ_k הוא וקטור עמודה $\mu_k = \begin{pmatrix} \mu_{k1} \\ \vdots \\ \mu_{kd} \end{pmatrix}$, ו-$\Sigma_k$ מטריצת השונות המשותפת המקיימת $\Sigma_k = diag(\sigma_{k1}^2, ..., \sigma_{kd}^2)$.
מההנחה שהפיצ'רים השונים ב"ת, הנראות ניתנת ע"י

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

מצאנו בסעיף הקודם ביטוי הממקסם את $\mathcal{L}(\Theta|x_i, y)$ עבור $1 \leq i \leq d$. מקסום המכפלה לעיל שקול למקסום כל אחד מהגורמים שלה, כיוון שכולם חיוביים. נובע ש–

$$\mu_{kj}^{MLE} = \frac{1}{n_k} \sum_i 1_{[y_i = k]} x_{ij} \quad \hat{\Sigma}^{MLE} = \frac{1}{m} \sum_i \left(x_i - \mu_{y_i}^{MLE}\right) \left(x_i - \mu_{y_i}^{MLE}\right)^T$$

כאשר כעת x_i הוא וקטור עם d עמודות.

4. מסווג בייסיאני נאיבי פואסוני מניח פריור מולטינומי, ו-likelihood בלתי-תלוי

$$y \sim Multinomial\left(\pi\right)$$
$$x_j|y=k \overset{ind}{\sim} Poi\left(\lambda_{kj}\right)$$

עבור π **וקטור הסתברויות** $\pi \in [0,1]^K : \sum \pi_j = 1$**.**

(א) נניח ש-$x \in \mathbb{R}$**, כלומר לכל דגימה יש פיצ'ר יחיד. בהינתן דגימות** $\{(x_i, y_i)_{i=1}^m\}$**, התאימו מסווג בייסיאני נאיבי פואסוני שפותר את** (5) **תחת הנחות** (7)**.**

באופן דומה לחישוב שביצענו בשאלה 3, נמצא את ממקסם לנראות עבור התפלגות פואסונית. ניזכר: $x \sim Poi\left(\lambda\right) \implies P\left(X=k\right) = e^{-\lambda}\cdot\frac{\lambda^k}{k!}$.
אם כן,

$$\mathcal{L}\left(\Theta|X,y\right) = f_{X,y|\Theta}\left(\{x_i,y_i\}\right) = \prod_{i=1}^{m} f_{X|Y=y_i}\left(x_i\right) f_Y\left(y_i\right) = \prod_{i=1}^{m} e^{-\lambda_{y_i}}\cdot\frac{\lambda_{y_i}^{x_i}}{x_i!}\cdot\pi_{y_i}$$

הפרמטרים שממקסמים את הביטוי לעיל ממקסמים גם את ה-log שלו,

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

ובהסרת ביטויים קבועים, ובסימון $n_k = \sum_i 1_{[y_i=k]}$ הביטוי לעיל שווה ל-

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

באופן דומה לשאלה 3, נגזור לפי כל רכיב בנפרד, ונקבל שעבור π מתקיים-

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

נגזור לפי λ_k עבור $k \in [K]$ ונשווה לאפס,

$$\frac{\partial \mathcal{L}}{\partial \lambda_k} = -n_k + \frac{1}{\lambda_k} \sum_{i:y_i=k} x_i = 0$$

$$\lambda_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

ו-$\{\lambda_k\}$ שמצאנו אכן ממקסמים את ה-likelihood.

(ב) נניח שכל דגימה נמצאת ב- $x \in \mathbb{R}^d$. בהינתן דגימות $\{(x_i, y_i)_{i=1}^m\}$, התאימו מסווג בייסיאני נאיבי פואסוני שפותר את (5) תחת הנחות (7).

אנו מניחים שהפיצ'רים ב"ת, לכן הנראות ניתנת ע"י

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

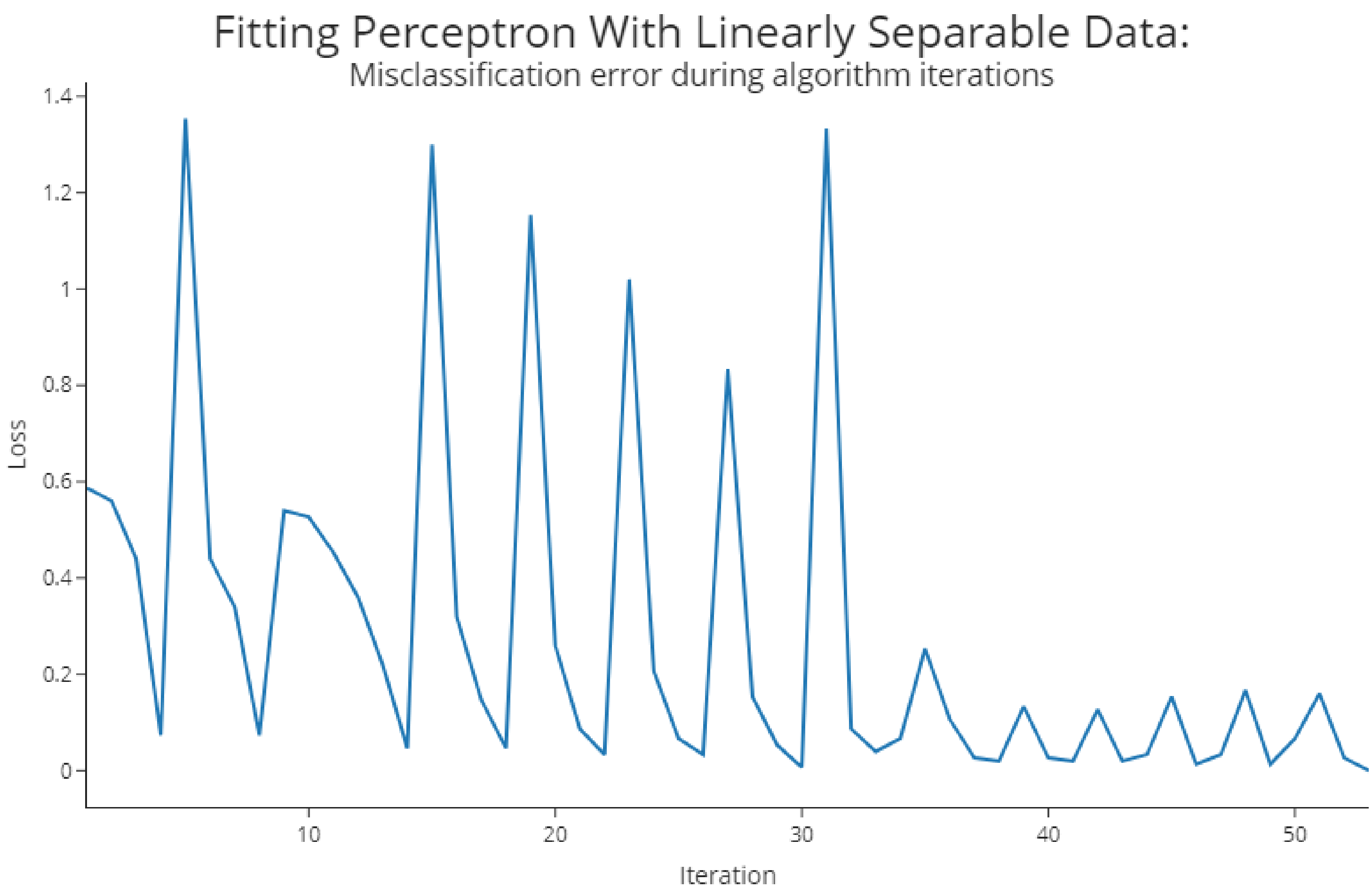
מצאנו בסעיף הקודם ביטוי הממקסם את $\mathcal{L}(\Theta|x_i, y)$ עבור $1 \leq i \leq d$. מקסום המכפלה לעיל שקול למקסום כל אחד מהגורמים שלה, כיוון שכולם חיוביים. נובע ש–

$$\lambda_{kj} = \frac{1}{n_k} \sum_{i:y_i=k} x_{ij}$$

3. חלק מעשי

3.1 פרספטרון

1. התאימו את הפרספטרון ל-Linearly Seperable Data. מה ניתן ללמוד מהגרף?

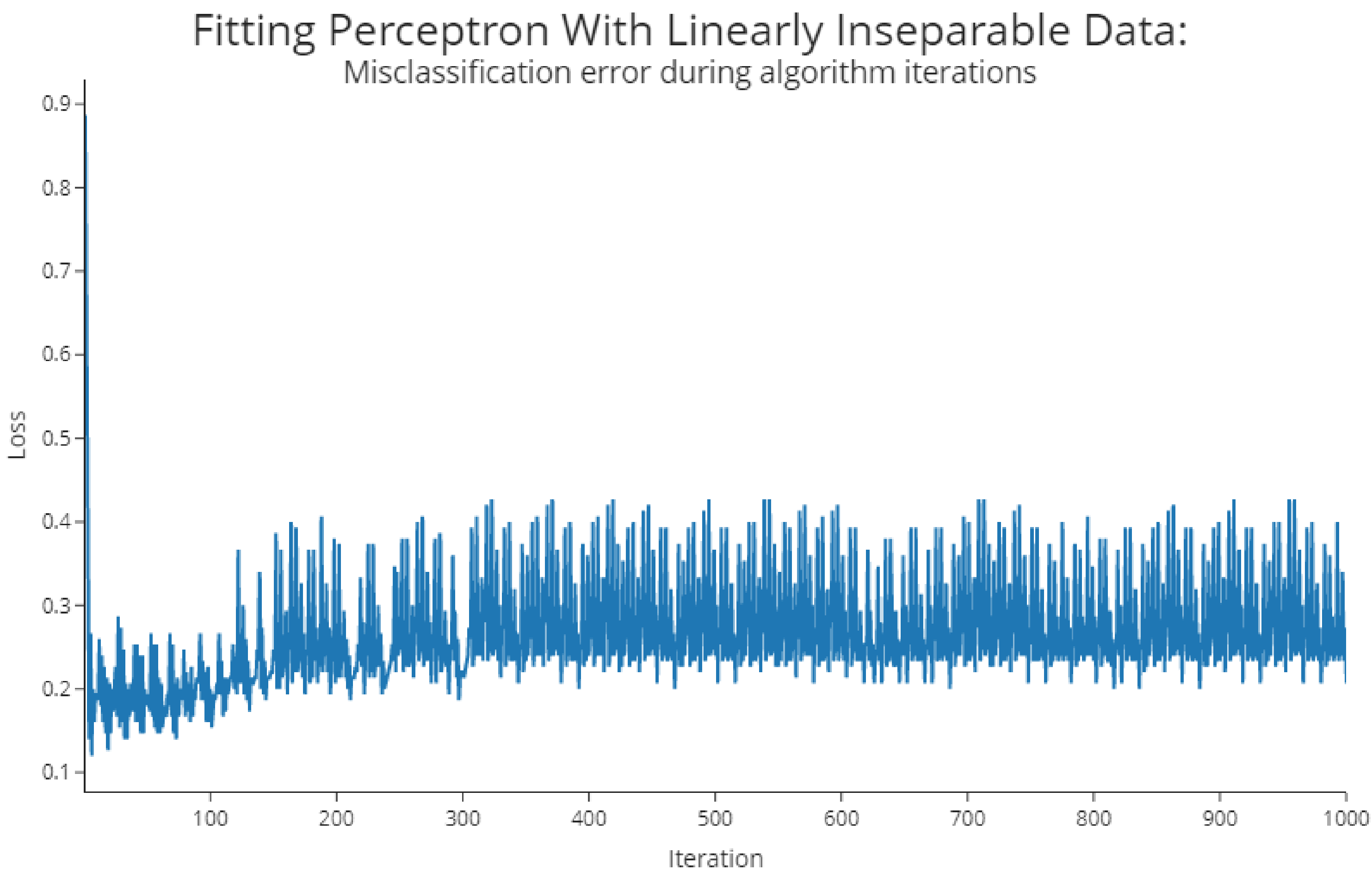


איור 1 : Fitting Perceptron over Linearly Seperable Data

מהגרף ניתן לראות שלאורך האיטרציות של הפרספטרון, המגמה הכללית של ה-loss נמצאת בירידה. עם זאת, הגרף אינו יורד לחלוטין, כלומר קיימים שינויים מסוימים, שלאחר שהאלגוריתם מבצע אותם, ה-loss גדל.
זאת אומרת ששינויים מסוימים שהאלגוריתם מבצע גורמים לכך שפחות דגימות ב-train set מסווגות נכון. לבסוף ה-loss מגיע לאפס, לאחר כ-50 איטרציות, כאשר האלגוריתם מצא על-מישור מפריד.

2. הריצו את הפרספטרון על ה-Linearly Inseperable Data ושרטטו את ה-loss כפונקציה של מספר האיטרציות. מה ההבדלים בין הגרף שהתקבל לזה בשאלה הקודמת?

כיצד ניתן להסביר זאת במונחי פונקציית המטרה ומרחב הפרמטרים?



איור 2: Fitting Perceptron over Linearly Inseperable Data

ראשית, במקרה של דאטה שהינו Linearly Inseperable, לא קיים על-מישור מפריד, וכן ניתן לראות שהאלגוריתם לא מגיע לפרמטרים שמאפסים את ה-loss.

יתר על כן, ה-misclassification loss על הפרמטרים המוחזרים אינו הנמוך ביותר שהושג לאורך האיטרציות.

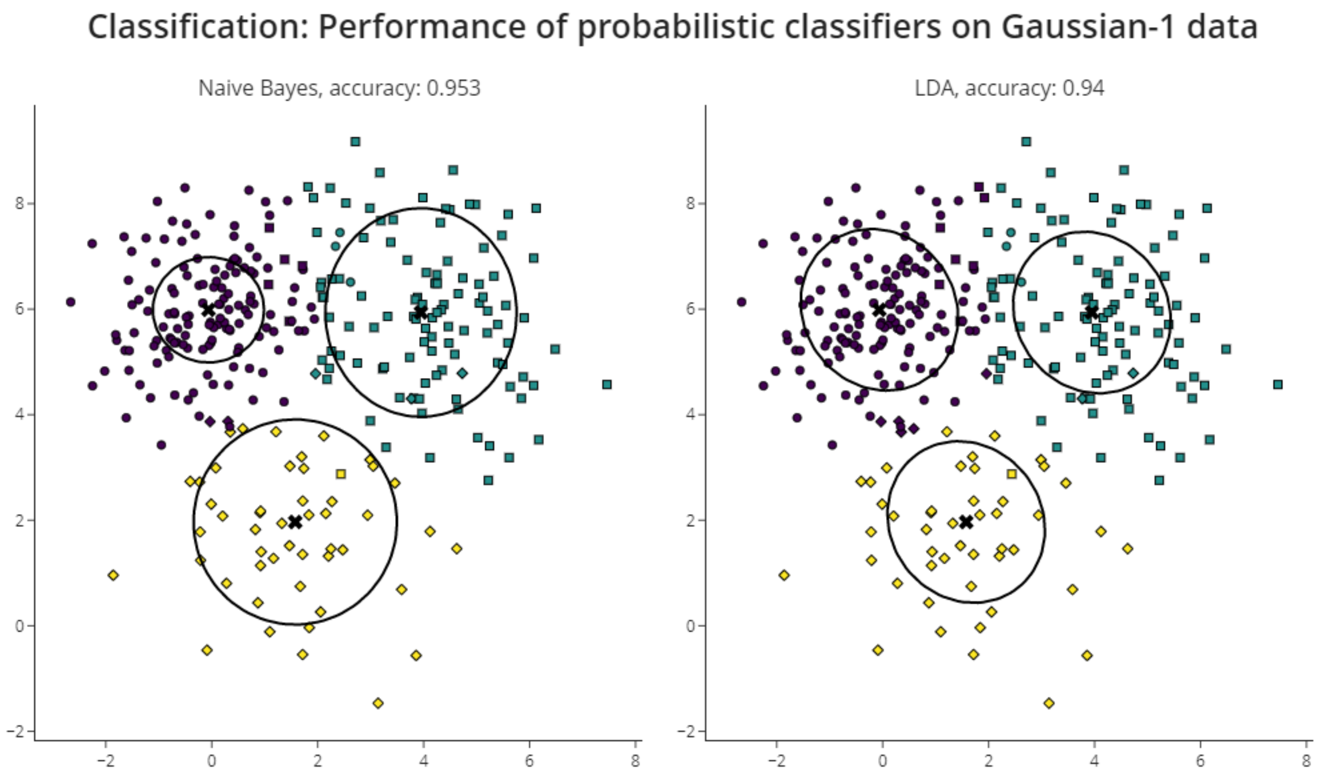
זאת כיוון שאלגוריתם הפרספטרון מתקן את העל-מישור המפריד לאורך האיטרציות, אך לא בהכרח באופן שממזער את ה-loss על כל ה-train set יחד, אלא בכל פעם על דגימה יחידה.

כך, תיקון הפרמטרים כך שיתאימו לדגימה יחידה, למעשה מגדיל את ה-loss בכך שהוא מביא לטעות בסיווג על דגימות אחרות.

ה-loss עולה ויורד לאורך כל ריצת האלגוריתם, כיוון שבכל איטרציה קיימות דגימות שלא מסווגות נכון, שאת סיווגן האלגוריתם מנסה לתקן.

3.2 מסווגים בייסיאנים

.1

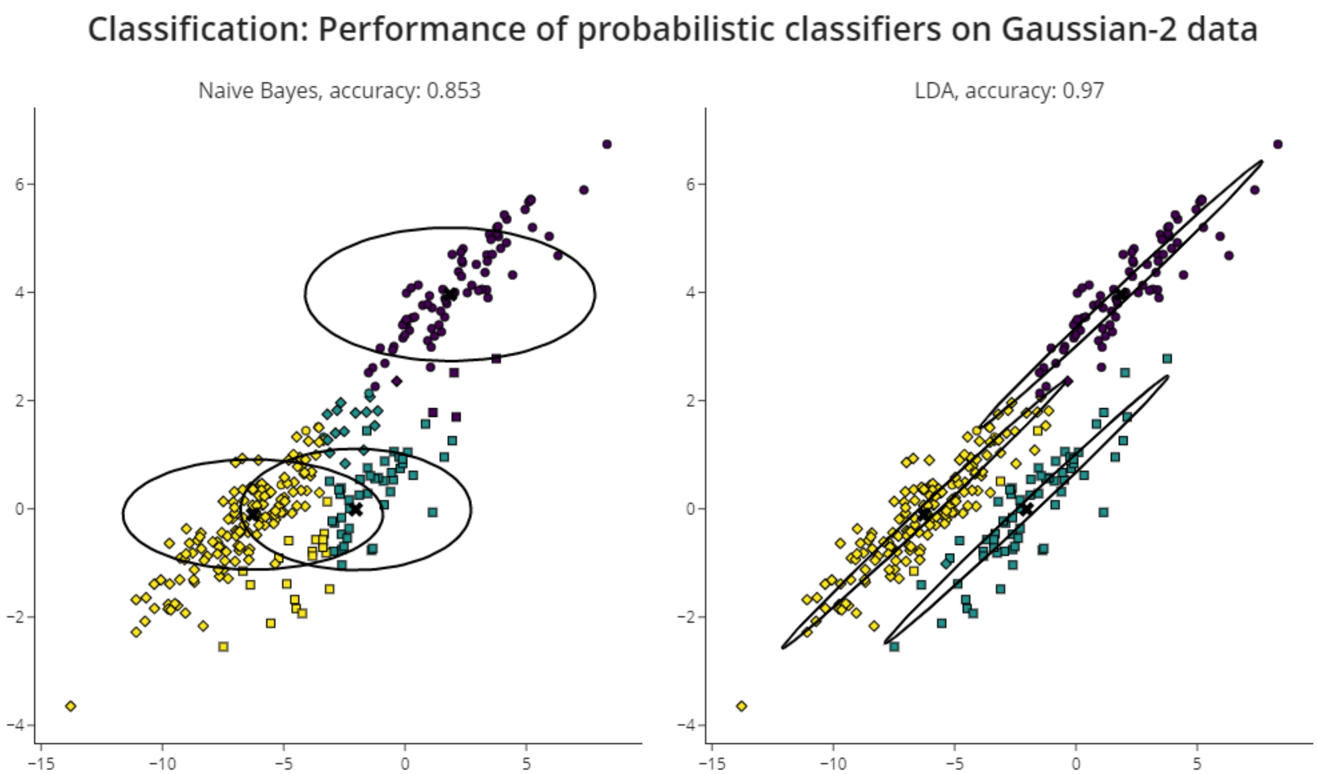


מהגרף ניתן לראות כי שני המסווגים, LDA ו-Naive Bayes השיגו תוצאות די דומות על הדאטה, אם כי Naive Bayes השיג דיוק טוב במקצת.

המסווג Naive Bayes מניח אי-תלות בין הפיצ'רים, וכייון שהשיג דיוק טוב יותר, ניתן להסיק שאכן הפיצ'רים בלתי תלויים. אכן, ניתן לראות שהאליפסות שמצוירות בגרף המתאים ל-LDA נוטות להיות כמעט ישרות, מה שמתאים למטריצת שונות משותפת אלכסונית, כלומר אי-תלות בין הפיצ'רים.

כמו כן, ה-Covariance של כל Class ככל הנראה שונה. ניתן לראות זאת בהבדלים המשמעותיים בגדלי האליפסות בגרף השמאלי, שניתן ע"י Naive Bayes, שלא מניח מטריצת Covariance זהה לכל הקלאסים, בניגוד ל-LDA.

.2



עבור הדאטה הנ"ל, מסווג ה-LDA השיג דיוק טוב יותר משמעותית. ניתן לראות על הגרף שהאליפסות שמודל זה התאים, מתאימות ל-classes באופן כמעט מושלם.

אכן ניתן לראות לפי פיזור הנקודות בגרף שקיים מתאם חיובי גבוה בין שני הפיצר'ים. כיוון ש-LDA אינו מניח אי-תלות בין הפיצ'רים, הוא הצליח להתאים Covariance באופן מדויק יותר מ-Naive Bayes.

כמו כן, ניתן לראות לפי הפיזור שהשונות של הפיצ'רים לא שונה משמעותית בין ה-classes השונים. LDA מניח גם תכונה זו, ולכן הצליח להתאים את ה-Covariance בצורה מדויקת יותר

לעומת Naive Bayes שלא מניח זאת, מה שפגע בהתאמה שלו.

לסיכום, ניתן ללמוד שהדאטה נשלף מהתפלגות עם שונות משותפת חיובית בין שני הפיצ'רים, כאשר השונות המשותפת זהה על פני כל הקלאסים, בדומה להנחות של LDA.