Exercise 3 IML

Eyal Perets 209541903

(ניתוח תיאורטי)

1. Hard SVM:

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

דרך: למצוא מטריצות → A, Q וקטורים d, v, a כך שניתן לבטא ← הבעיה כבעיה הנ"ל:

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

תחילה: נכתוב ← האילוצים במונחי $Av \leq d$

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

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

נגדיר $v = \begin{bmatrix} w \\ b \end{bmatrix}$ ו- $A = \begin{bmatrix} y_1 x_1^T & \cdots & y_1 x_1^T \cdots & y_1 \\ & & & \vdots \\ y_n x_n^T & \cdots & y_n x_n^T \cdots & y_n \end{bmatrix}$ ונקבל

$$-y_i x_i^T w - y_i b = A_i v \leq -1$$

עכשיו $\underset{(w,b)}{\arg\min} \|w\|^2$ נשים לב ש

$$\|w\|^2 = w^T I w$$

נגדיר את ← המטריצה Q כך:

$$Q = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

כאשר $I_{d \times d}$

ונקבל $a = 0^T$ ו-

$$\arg\min w^T I w = \arg\min \begin{bmatrix} w \\ b \end{bmatrix}^T \cdot Q \cdot \begin{bmatrix} w \\ b \end{bmatrix} + 0^T \begin{bmatrix} w \\ b \end{bmatrix} = \arg\min v^T \cdot Q \cdot v + a^T v$$

$$= \arg\min \frac{1}{2} v^T \cdot Q \cdot v + a^T v$$

ולכן הבעיה הנ"ל ניתנת לתיאור כבעיית תכנון ריבועי.

2. Soft SVM:

$$\arg\min \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_i \xi_i \quad s.t \ \forall_i \ y_i\langle w,x_i\rangle \geq 1-\xi_i \wedge \xi_i \geq 0$$

צריך להראות שבעיית Soft Svm זהה לבעיה הבאה:

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

שים לב ש l^{hinge} מבטיח לנו את האילוץ

: $y_i\langle w,x_i\rangle \geq 1-\xi_i \wedge \xi_i \geq 0$

וכן עבור $w \in \mathbb{R}^d$, $\forall_i$, אם $y_i\langle w,x_i\rangle \geq 1$,

אזי נשאף בבעיה את הערך y_i ו ξ_i , שיהיה הכי נמוך לכן

0.

אם $y_i\langle w,x_i\rangle < 1$ אזי הערך המינימלי

ביחד לכן $1-y_i\langle w,x_i\rangle$, ו 2 המקרים הללו

מתארים בדיוק את $l^{hinge}(y_i\langle w,x_i\rangle)$.

ולכן $\forall_i$ $\xi_i = l^{hinge}(y_i\langle w,x_i\rangle)$ והמשוואות מתקיימות

3. א. אנו נכתוב ← מקסימום likelihood של estimators:

נשים לב כי ה likelihood הינו:

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

$$=\prod_{i=1}^{m}\mathcal{N}(x_i|\mu_{y_i},\sigma^2_{y_i})\cdot Mult(y_i,\pi)$$ ← לקחנו מהגדרה:

ידוע ש-log פונקציה מונוטונית עולה, לכן המקסימום של ה likelihood, זהה למקסימום של log likelihood ולכן נגזור את ה log-likelihood:

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

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

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

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

$$=\sum_{k}\left[n_k\log(\pi_k)-\frac{1}{2}\sum_{i|y_i=k}(x_i-\mu_i)^T(\sigma^2)^{-1}(x_i-\mu_i)\right]-\frac{m}{2}\log(2\pi)-\frac{m}{2}\log|\sigma^2|$$

כאשר $n_k=\sum_i 1_{y_i=k}$, כעת נגזור את הביטוי, כדי למצוא את הפרמטרים האופטימליים: $\{\pi_k\},\{\mu_k\},\{\sigma^2\}$ ונשווה ל-0.

נשים לב ש $\pi \in [0,1]^k$ ו $\sum_k \pi_k = 1$

נשתמש בשיטת כופלי לגרנז', כלומר למצוא את המקסימום תחת ההנחה: $\sum_k \pi_k = 1$

$$L = l(\Theta|X,y) - \lambda g(\pi)$$

$$\frac{\partial L}{\partial \pi_k} = \frac{\partial}{\partial \pi_k} l(\Theta|X,y) - \lambda \frac{\partial}{\partial \pi_k} g(\pi) = \frac{n_k}{\pi_k} - \lambda = 0$$

$$\Downarrow$$

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

כעת נמצא את λ, ונציב את π_k במשוואת האילוץ:

$$\sum_{i=1}^{m} \frac{n_k}{\lambda} = 1 \iff m = \lambda$$

ולכן ה MLE הינו $\pi_k^{MLE} = \frac{n_k}{m}$

$$\mu_k^{MLE} = \frac{1}{n_k} \sum_i 1_{y_i=k} X_i \qquad \sigma^{2\,MLE} = \frac{1}{m} \sum_i (X_i - \mu_{y_i}^{MLE})(X_i - \mu_{y_i}^{MLE})^T$$

ג. נשים לב כי ההבדל כעת היחיד הוא בפונקציה של הפרמטרים, כלומר, כעת ה likelihood יהיה כדלקמן:

$$\prod_{j=1}^{k} \prod_{i=1}^{m} N(X_i | \mu_{y_i}, \sigma^2_{y_i}) \cdot Mult(y_i, \pi)$$

ולכן ה log likelihood יהיה:

$$= \sum_k \left[n_k \log(\pi_k) - \frac{1}{2} \sum_{i|y_i=k} (x_i - \mu_i)^T (\sigma^2)^{-1} (x_i - \mu_i) \right] - \frac{md}{2} \log(2\pi) - \frac{m}{2} \log|\sigma^2|$$

(הפיתוחים היו במחברת Matlab)

נשים לב כי לכל $j \in [d]$ המינוחים

ולכן בכלל, המקסימום לפרמטר j, ולכל k הפרמטרים:

$$\pi_k^{MLE} = \frac{n_k}{m} \qquad \mu_{kj}^{MLE} = \frac{1}{n_k} \sum_{i|y_i=k} x_{ij} \qquad (\sigma^2)_{kj}^{MLE} = \frac{1}{n_k} \sum_{i|y_i=k} (x_{ij} - \mu_{kj}^{MLE})^2$$

4. א. באותה הדרך בדיוק, כמו בסעיף הקודם, ה-likelihood היינו:

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

לכן ה-log-likelihood היינו: (הסימון לשימוש ה מופיע בדף בסוף.)

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

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

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

כעת נגזור לפי λ ונשווה ל-0:

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

$$\Downarrow$$

$$\lambda_k^{MLE} = \frac{1}{n_k} \sum_{i=1}^{m} \mathbb{1}_{y_i=k} x_i$$

π_k לכל π_k בשאלה 3.א. (הגזירה זהה)

ולכן $\pi_k^{MLE} = \frac{n_k}{m}$

ב.

כמו בשאלה 3.ב. אפשר להפריד כגק יהיה:

בנוסף והמתבקש, נקבל ה-likelihood יהיה:

$$\prod_{j=1}^{d} \prod_{i=1}^{m} Pois(X_i \mid \mu_{y_i}, \lambda_{y_i}) \cdot Mult(y_i, \pi)$$

ובאותם עיונים בדיון, ה- log-likelihood יהיה:

$$\sum_{i=1}^{m} \sum_{j=1}^{d} \log(\pi_k) + X_{ij}\log(\lambda_{kj}) - \lambda_{kj} - \log(X_{ij}!)$$

(החישובים היו באמצעות Matlab)

זה מקסימום מתקיים עבור

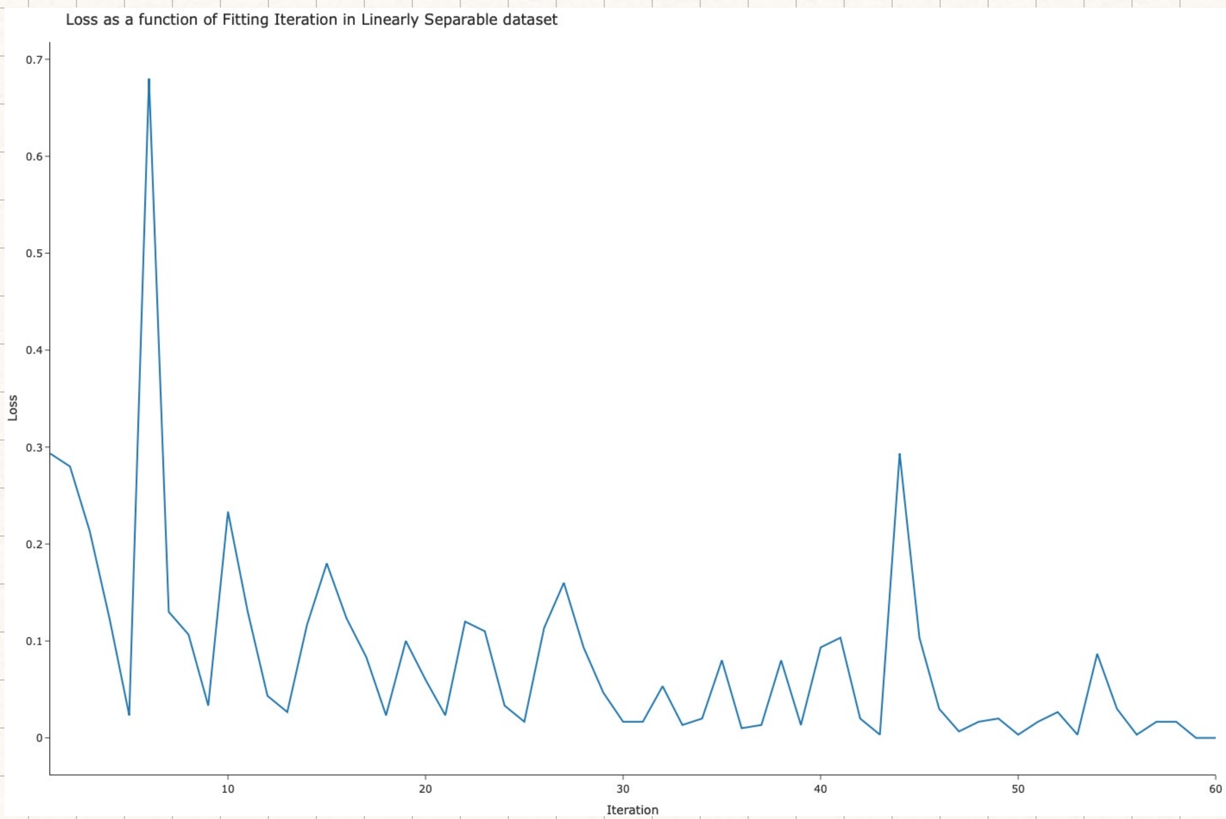
יש השמטה

ה-j כפי שמצאנו ← $\lambda_{kj}^{MLE} = \frac{1}{n_k} \sum_{i=1}^{k} \mathbb{1}_{y_i = k} X_{ij}$

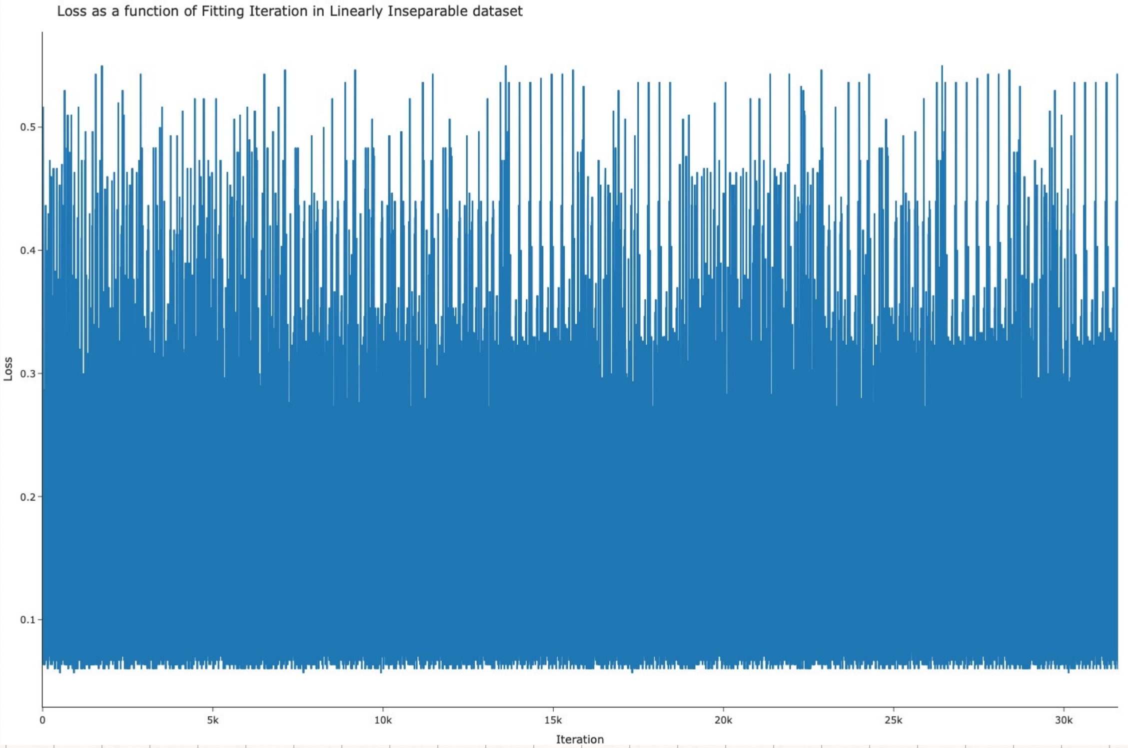
וגם - $\pi_k = \frac{n_k}{m}$

ויחוק הפרוק׳

ס1



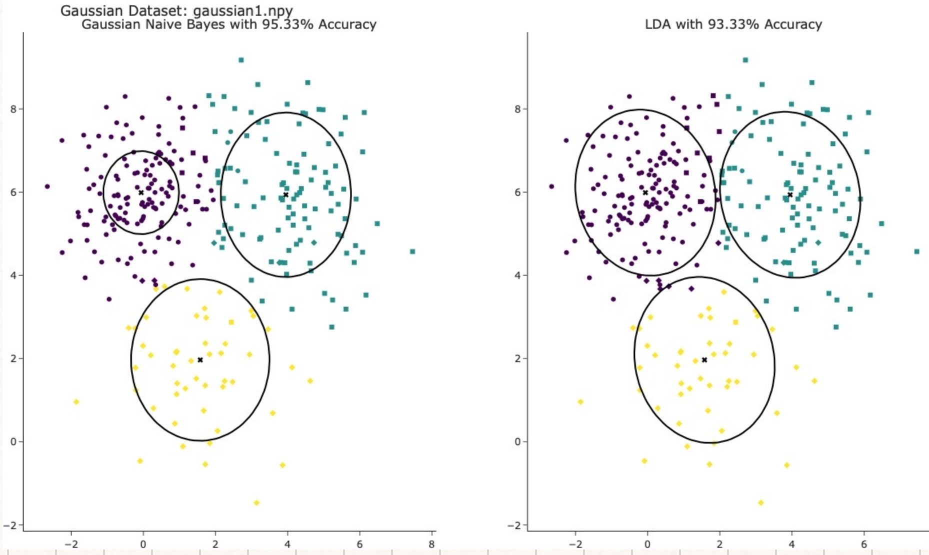
ניתן ללמוד מגרף זה שככל שכמות האיטרציות של Perceptron
עולה, ה MSE יורד, כלומר האלגוריתם מצליח לבצע קלסיפיקציה
טובה עד כדי שגיאה אפסית. ניתן לראות שבחלק מהאיטרציות
ישנה "קפיצה" ב MSE, שככל הנראה נובעת מהוספה
בגלל מידי, שלעיתים ישנו תיוגן דגם השני.



2.

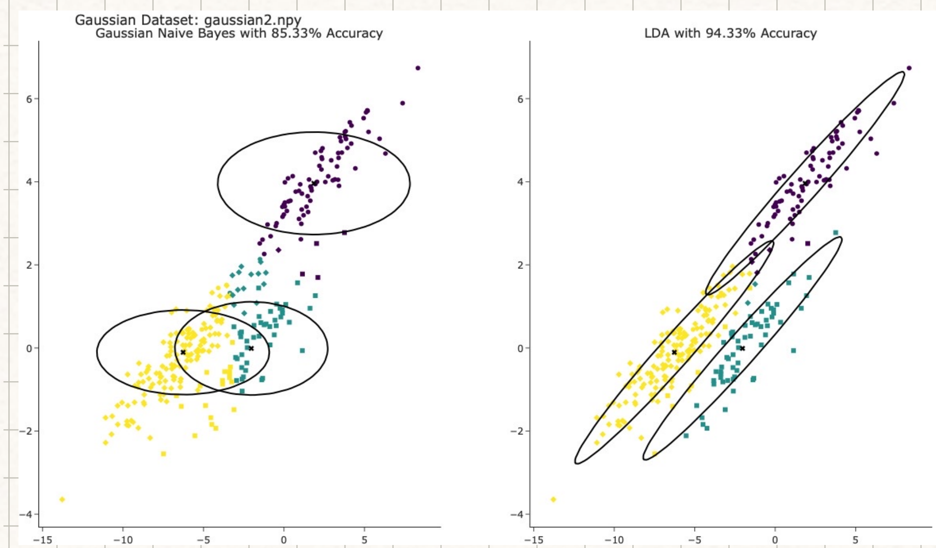
בגרף הזה לעומת הגרף הקודם, ה perceptron לא מצליח להגיע לפתרון אופטימלי, ללא קשר במספר האיטרציות.
זה נובע מכך שאין פתרון אופטימלי עבור דאטה שאינו מופרד לינארית, ולכן ללא קשר בכמות האיטרציות, ה perceptron לעולם לא ימצא את הנקודה הנכונה, ולכן ה MSE לעולם לא יתאפס. במונחי y,x,b,w - לא קיים (w,b), כך ש

$$\forall i \quad y_i \langle x_i, w \rangle \geq 0 .$$



3.

ניתן ללמוד מהגרפים הנל כי השונות המשותפת של הדגימות
היא אפס, ועל כן אין הבדל בין ה LDA וה
Gaussian Naive Bayes המוצגים.



4.

ההבדל בין 2 המקרים הנל שכאן מדובר על
שונות משותפת שאינה אפס ולכן ההנחה של המשתנים
הגאוסיאנים שהשונות המשותפת אפס היא אינה נכונה,
ועל כן ה LDA מסווג תוצאות טובות יותר.