2.1 Hard- & Soft-SVM

Based on Lecture 2 and Recitation 4

In class we saw the Hard-SVM classification model.

1. Prove that the following Hard-SVM optimization problem is a Quadratic Programming problem:

$$\underset{(\mathbf{w},b)}{\operatorname{argmin}} \|\mathbf{w}\|^2 \quad \text{s.t.} \quad \forall i\, y_i(\langle \mathbf{w},\mathbf{x}_i\rangle + b) \geq 1 \tag{1}$$

That is, find matrices Q and A and vectors $\mathbf{a}$ and $\mathbf{d}$ such that the above problem can be written in the following format

$$\underset{\mathbf{v}\in\mathbb{R}^n}{\operatorname{argmin}} \tfrac{1}{2}\mathbf{v}^\top Q\mathbf{v} + \mathbf{a}^\top\mathbf{v} \quad \text{s.t.} \quad A\mathbf{v} \leq \mathbf{d} \tag{2}$$

Hint: Observe that $\|\mathbf{w}\|^2 = \mathbf{w}^\top I\mathbf{w}$

$$\underset{(w,b)}{\operatorname{argmin}} |w|^2 \iff \operatorname{argmin} \left\|\begin{pmatrix} w \\ b\end{pmatrix}\right\|^2$$

$$\Rightarrow \underset{(w,b)}{\operatorname{argmin}} \left(\begin{pmatrix} w \\ b\end{pmatrix}^T I \begin{pmatrix} w \\ b\end{pmatrix}\right)$$

נפיק את הריבוע

$$\text{s.t } y_i(\langle w, x_i\rangle + b) \geq 1$$

$$\iff$$

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

לכל i לכן אפשר להפוך למטריצה

$$\Rightarrow \underbrace{\begin{pmatrix} -y_1 x_1^T & -y_1 \\ & \vdots \\ -y_n x_n^T & -y_n \end{pmatrix}}_{A\in\mathbb{R}^{m\times(d+1)}} \circ \underbrace{\begin{pmatrix} w \\ b \end{pmatrix}}_{v} \leq \underbrace{-1}_{d}$$

לכן $v = \begin{pmatrix} w \\ b\end{pmatrix}$

(2) צריך ש
$$a^T\begin{pmatrix} w \\ b\end{pmatrix} + \frac{1}{2}\begin{pmatrix} w \\ b\end{pmatrix}^T Q \begin{pmatrix} w \\ b\end{pmatrix} = w^T I w$$

לכן נבחר $a=\bar{0}$ ואז Q מטריצה שתוגדר באופן הבא

$$Q_{ij} = \begin{cases} 1/2 & i=j \wedge i\leq d \\ 0 & \text{else}\end{cases}$$

כלומר מטריצה מהצורה

$$\begin{pmatrix} 1/2 & 0 & \cdots & \cdots & 0 \\ 0 & 1/2 & & & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & 1/2 & \vdots \\ 0 & \cdots & 0 & 0 & 0 \end{pmatrix}$$

עבורה מתקיים

$$a^T \begin{pmatrix} w \\ b \end{pmatrix} + \frac{1}{2} \begin{pmatrix} w \\ b \end{pmatrix}^T Q \begin{pmatrix} w \\ b \end{pmatrix} = \frac{1}{2} w^T 2I w = w^T I w = |w|^2$$

כנדרש.

$$Q \in R^{m \times (d+1)}$$

שאלה 2

a. 1)

$$\arg\max_{k\in[K]} \frac{f_{X|y=k}(x)\, f_Y(k)}{f_X(x)}$$

$$= \arg\max_{k\in[K]} f_{X|y=k}(x)\, f_Y(k)$$

$$= \arg\max_{k\in[K]} \pi_k \cdot N(x \mid \mu_k, \sigma_k^2)$$

$$\arg\max_{k\in[K]} \pi_k \cdot \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)$$

נחשב את הלייקליהוד

$$L(\theta \mid X, y) = P_{X,y|\theta}\left(\{X_i, y_i\}_{i=1}^m\right)$$

$$= \prod_i^m f_{X,y|\theta}((x_i, y_i))$$

$$= \prod_i^n f_{X,y=y_i|\theta}(x)\, f_{y|\theta}(y_i)$$

$$= \prod_i^m N(x \mid \mu_{y_i}, \sigma_{y_i}^2)\, M(y_i \mid \pi)$$

$$= \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma_{y_i}^2}} \exp\left(-\frac{(x-\mu_{y_i})^2}{2\sigma_{y_i}^2}\right) \cdot \pi_{y_i}$$

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

נרצה למצוא את המקסימום ובגלל ש- log הוא מונוטוני עולה המקסימום של log הלייקליהוד שקול למקסימום של הלייקליהוד.

$$\log(L(\Theta \mid x,y)) = \log\left(\prod_i^m \pi_{y_i}\right) - \frac{m}{2}\log\left(2\pi\sigma^2_{y_i}\right) + \sum_{i=1}^m -\frac{(x-\mu_{y_i})^2}{2\sigma^2_{y_i}}$$

$$= \sum_{i=1}^m \left(-\frac{1}{2}\log(2\pi\sigma_{y_i}) - \frac{(x-\mu_i)^2}{2\sigma^2_{y_i}} + \log(\pi_{y_i}) \right)$$

נגזור לפי $\sigma_{y_i}, \pi_{y_i}, \sigma_{y_i}$ ← אפשר לעשות לפי כופלי לגרנג'

סה"כ נקבל לפי התרגול ש

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i:y_i=k} x_i \qquad \hat{\pi}_k^{MLE} = \frac{n_k}{m} \Rightarrow n_k = \hat{\pi}_k m$$

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

ב. עבור $x \in \mathbb{R}^d$ נקבל כמו ש k שלנו מתפלג

בצורה הבאה $x \sim N\left(\mu_{kj}, \sigma^2_{kj}\right)$

מכיוון שהמשתנים הם ב"ת נקבל שהתפלגות

המשותפת היא המכפלה של התפלגויות

לכן

$$\hat{y} = \underset{k \in K}{\operatorname{argmax}} \left(\pi_k \cdot \prod_{j=1}^{d} N\left(x_j, \mu_{kj}, \sigma_{kj}^2\right)\right)$$

$$= \prod_{j=1}^{d} \underset{k \in K}{\operatorname{argmax}} \left(\pi_k \cdot N\left(x_j, \mu_{kj}, \sigma_{kj}^2\right)\right)$$

באותו אופן נקבל ש

$$L(\Theta|x,y) = \prod_{j=1}^{d} \left(\prod_{i=1}^{m} f_{x,y=y_i|\Theta}(x_{ji}) \cdot \pi_{y_i} \right)$$

$$= \prod_{j=1}^{d} \left(\left(\prod_{i=1}^{m} \pi_{y_i} \right) \left(1/(2\pi\sigma_{jy_i}^2)^{\frac{1}{2}} \right)^m \exp\left(\sum_{i=1}^{m} \frac{(x_{ji} - \mu_{jy_i})^2}{2\sigma_{jy_{jy_i}}^2} \right) \right)$$

$\Rightarrow$

לוג הנראות יתן לנו

$$\log(L(\Theta|x,y)) = \sum_{j=1}^{d} \left(\sum_{i=1} \pi_{y_i} + \sum_{i=1} f_{x_{ij}}(x_{ij}) \right)$$

$\Leftarrow$ כעת נגזור כדי למצוא את $\hat{\sigma}_k, \mu_k$. סה"כ נקבל

על סמך סעיף קודם

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

$$\hat{\mu}_{kj} = \frac{\sum_{i=1}^{m} \prod_{y_i=k} x_{ji}}{n_k}$$

$$\hat{\sigma}_k = \frac{1}{n_k} \sum_{i=1}^{m} \prod_{(y_i=k)} (x_{ij} - \mu_{kj})^2$$

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

2

$$\arg\max_{k\in[K]} \frac{f_{X|y=k}(x)\, f_Y(k)}{f_X(x)}$$

$$= \arg\max_{k\in[K]} f_{X|y=k}(x)\, f_Y(k)$$

$$= \arg\max_{k\in[K]} \pi_k \cdot \text{Poi}(\lambda_k)$$

$$\arg\max_{k\in[K]} \pi_k \cdot \left(\frac{\lambda^k}{k!} e^{-\lambda}\right)$$

נחשב את הנראות

$$L(\theta | x, y) = P_{x,y|\theta}\left(\{x_i, y_i\}_{i=1}^m\right)$$
$$= \prod_{i=1}^m f_{x,y|\theta}((x_i, y_i))$$
$$= \prod_{i=1}^m f_{x|y=y_i|\theta}(x) \cdot f_{y|\theta}(y_i)$$
$$= \prod_{i=1}^m \pi_{y_i} \cdot \frac{\lambda_{y_i}^{x_i} e^{-\lambda_{y_i}}}{x_i!}$$

נחשב את log הנראות

קירוב סטרלינג.

$$\Rightarrow \log(L(\theta | x, y)) \overset{\sim}{=} \sum_{i=1}^m \left((\log(\pi_{y_i})) + x_i \log(\lambda_{y_i}) - \lambda_{y_i} - (i\ln(i) - i) \right)$$

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

נמצא מקסימום בעזרת כופלי לגרנז'.

$$\frac{\partial}{\partial \pi_k} = 0 = \frac{n_k}{\pi_k} - n \rightarrow \hat{\pi}_k = \frac{n_k}{n} = \frac{n_k}{m}$$

$$\Rightarrow 1 = \sum_{k=1}^{m} \pi_k = \sum \frac{n_k}{n} \rightarrow \boxed{n = m}$$

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

$$\frac{\partial}{\partial n_k} = 0 = \sum_{i=1}^{m} \frac{x_i}{n_{y_i}} - 1 \rightarrow$$ (אנחנו רוצים $y_i = k$ לכן)

$$\Rightarrow \left(\sum_{i=1}^{m} x_i\right) / n_k - n_k = 0$$

$$\Rightarrow \hat{n}_k = \frac{\sum_{i=1}^{m} \mathbb{I}_{y_i = k}(x_i)}{n_k}$$

ב. כמו בסעיף קודם נסיק שמדובר בכפל של התפלגויות

לכן נקבל ש $\log$ הנראות הינו מהצורה:

$$\sum_{j=1}^{d} \sum_{i=1}^{m} \left((\log(\pi_{y_i})) + x_{ij} \log\left(\frac{n_{y_i}}{x_{ij}}\right) - n_{y_i} + x_{ij} \right)$$

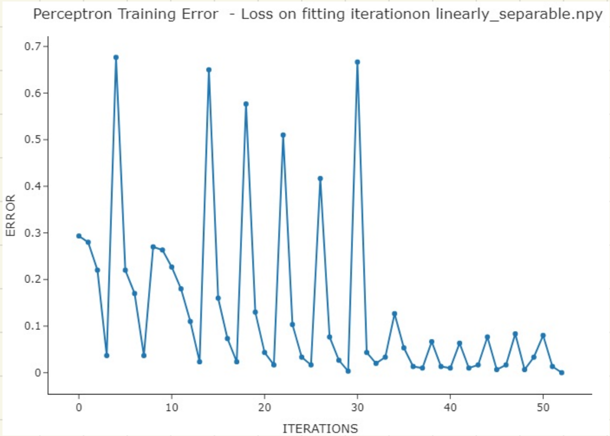
סה"כ נקבל גם פה ש

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

$$\hat{\pi}^{MLE}_{kj} = \frac{\sum_{i=1}^{m} \mathbb{I}_{y_i = k}(x_{ij})}{n_k}$$

3.1 Perceptron Classifier

Based on Lecture 2 and Reci



הציור מייצג את האלגורים

ה Peceptron .

לפי הגרף אפשר לראות שעבור סט נתונים שניתן לבדל

אותו לינארית, יש מגמת ירידה ב Error (Misclassification loss)

שזה אומר שככל שיש לנו יותר איטרציות המגמה תרד

והחל ממספר 52 זה אפס.

בעקבות זאת אפשר להגיד שאלגוריתם עובד

בצורה טובה על data שהוא לינארי.

2)

• אפשר לראות בברור זה שמדובר בסט נתונים שלא ניתן להפרדה לינארית. בברור החל מערך מסויים האלגוריתם מדפיס בדפוס קבוע כלומר מגמת השגיאה נשארת ולא משתנה (בפרט היא לא אפס).

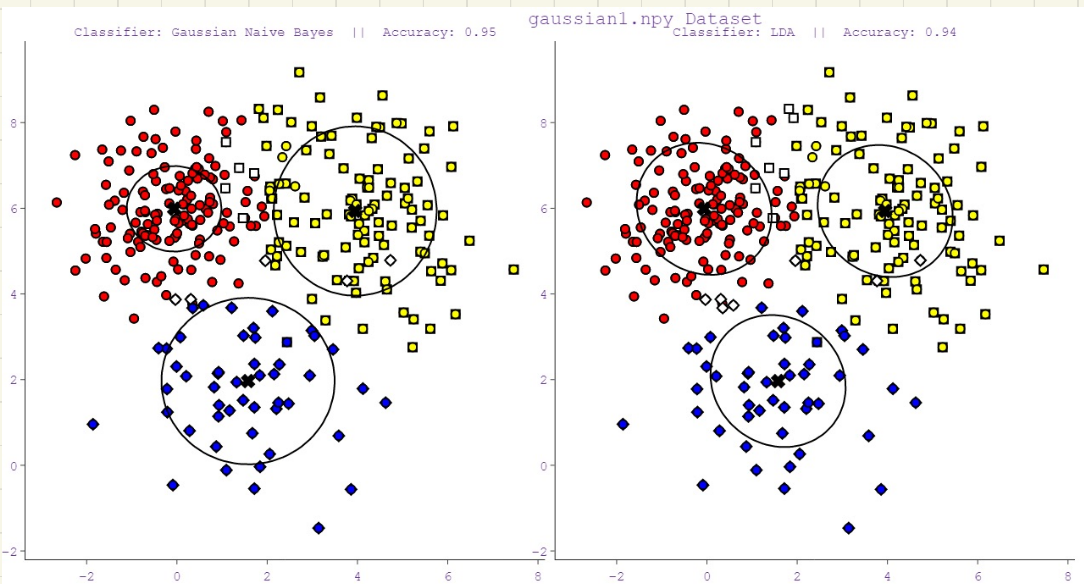
← מכאן אפשר להסיק שיש הבדל בין data שניתן להבדיל לינארי ל data שלא ניתן להפריד לינארי. שכן אם ה data לא ניתן להפרדה לינארית, תמיד נחזור אל טווח מגמת שגיאה. (האלגוריתם ימשיך לחפש פיתרון, אבל אין כזה מה שגורר את הדפוס שגיאה שחוזר על עצמו.)

3.2 Bayes Classifiers

בסרטון ניתן לראות GNB ו LDA על ה dataset הראשון.

=> אפשר לראות ש ה - GNB משיג 95.3% accuracy וה LDA 94%.

לכן נסיק שהנקודות מפוזרות יחסית בצורה אחידה סביב התוחלת של כל אחת => ששתי המחלקות כנראה מתפלגות נורמלי.



מהגרפים הנ״ל אפשר לראות:

ש- GNB משיג 85.33% accuracy

ש- LDA משיג 97% accuracy

אפשר לראות שב data יש סוג של תלות לינארית, כי ה data מתפזר בערך לינארית (אפשר לצייר קו), ולכן אפשר להגיד שמודל ה LDA יותר מדוייק מהמודל GNB

בגלל ש- GNB מניח אי תלות בין הפיצ'רים ← שזה מה שפוגע ביכולת החיזוי שלו. ו- LDA מניח תלות לינארית, כלומר שה data ניתן להפרדה לינארית.