מבוא למערכות לומדות (67577) | תרגיל 3

שם: נמרוד בר גיורא | ת"ז: 207090622

חלק תאורטי

Hard and Soft SVM

שאלה 1

1. Prove that 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

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

הוכחה: נשים לב ש־hyperplane ־ (w^*, b^*), עם $w^* \in \mathbb{R}^d = \begin{bmatrix} w_1^* \\ \vdots \\ w_d^* \end{bmatrix}$ ו־ $b^* \in \mathbb{R}$ הוא פתרון אופטימלי של הבעיה:

$$\underset{(w,b)}{\operatorname{argmin}} \|w\|^2 \quad \text{s.t} \quad \forall\, i \in [m] :\ y_i\left(\langle w \mid x_i\rangle + b\right) \geqslant 1$$

<u>אם ורק אם</u> הוא פתרון אופטימלי של הבעיה:

$$\underset{(w,b)}{\operatorname{argmin}} \left\| \begin{bmatrix} w_1 \\ \vdots \\ w_d \\ b \end{bmatrix} \right\|^2 \quad \text{s.t} \quad \forall\, i \in [m] :\ y_i\left\langle \begin{bmatrix} w_1 \\ \vdots \\ w_d \\ b \end{bmatrix} \mid x_i \right\rangle \geqslant 1$$

כאשר מוסיפים למטריצה X עמודת 1־דים. כלומר $[x_i]_{d+1} = 1$ (זו הכניסה ה־$d+1$ של השורה ה־i של X) לכל $i \in [m]$.
הסיבה לכך היא שלכל $i \in [m]$ מתקיים ש־

$$y_i\left(\langle w^* \mid x_i\rangle + b\right) \geqslant 1 \iff y_i\left(\sum_{j=1}^{d} w_j^* [x_i]_j + b\right) \geqslant 1 \iff y_i\left(\sum_{j=1}^{d} w_j^* [x_i]_j + b\,[x_i]_{d+1}\right) \geqslant 1 \iff y_i\left\langle \begin{bmatrix} w_1 \\ \vdots \\ w_d \\ b \end{bmatrix} \mid x_i \right\rangle \geqslant 1$$

ובנוסף לכל $w \in \mathbb{R}^d$ מתקיים ש־

$$\|w^*\|^2 \leqslant \|w\|^2 \iff \sum_{i=1}^{d} (w_i^*)^2 \leqslant \sum_{i=1}^{d} (w_i)^2 \iff \sum_{i=1}^{d} (w_i^*)^2 + b^2 \leqslant \sum_{i=1}^{d} (w_i)^2 + b^2 \iff \left\| \begin{bmatrix} w_1^* \\ \vdots \\ w_d^* \\ b \end{bmatrix} \right\|^2 \leqslant \left\| \begin{bmatrix} w_1 \\ \vdots \\ w_d \\ b \end{bmatrix} \right\|^2$$

ולכן אם נוסיף עמודת 1־דים כנ"ל ל־X, נוכל פשוט לעבוד עם הבעיה:

$$\underset{w\in\mathbb{R}^{d+1}}{\operatorname{argmin}} \|w\|^2 \quad \text{s.t} \quad \forall\, i \in [m] :\ y_i\,\langle w \mid x_i\rangle \geqslant 1$$

נגדיר $Q = 2I_{d+1}$, $\mathbf{a} = 0_{\mathbb{R}^{d+1}}$ ו־

$$\mathbb{R}^m \ni \mathbf{d} = \begin{bmatrix} -1 \\ \vdots \\ -1 \end{bmatrix} \qquad , \qquad \mathbb{R}^{m\times(d+1)} \ni A = -1 \cdot \begin{bmatrix} - & y_1 x_1^\top & - \\ & \vdots & \\ - & y_m x_m^\top & - \end{bmatrix}$$

ונקבל ש־$w^* \in \mathbb{R}^{d+1}$ הוא פתרון של הבעיה הנ"ל <u>אם ורק אם</u> לכל $w \in \mathbb{R}^{d+1}$ מתקיים ש־

$$\|w^*\|^2 \leqslant \|w\|^2 \iff w^{*\top} w^* \leqslant w^\top w \iff w^{*\top} I_d w^* \leqslant w^\top I_d w \iff$$
$$\iff \frac{1}{2} w^{*\top} Q w^* + \overbrace{\mathrm{a}w^*}^{=0} \leqslant \frac{1}{2} w^\top Q w + \overbrace{\mathrm{a}w}^{=0} \iff w^* = \operatorname*{argmin}_{w\in\mathbb{R}^{d+1}} \left(\frac{1}{2} w^\top Q w + \mathbf{a}^\top w\right)$$

ובנוסף:

$$\forall i \in [m] :\ y_i \langle w^* \mid x_i \rangle \geqslant 1 \iff \forall i \in [m] :\ -y_i \langle w^* \mid x_i \rangle \leqslant -1 \iff$$
$$\iff \forall i \in [m] :\ -y_i x_i^\top w^* \leqslant -1 \iff A w^* \leqslant \mathbf{d}$$

כלומר <u>אם ורק אם</u> הוא פתרון אופטימלי של הבעיה:

$$\operatorname*{argmin}_{w\in\mathbb{R}^{d+1}} \left(\frac{1}{2} w^\top Q w + \mathbf{a}^\top w\right) \quad \text{s.t.} \quad Aw \leqslant \mathbf{d}$$

■

שאלה 2

2. Consider the Soft-SVM optimization problem:

$$\operatorname*{argmin}_{\mathbf{w},\{\xi_i\}} \frac{\lambda}{2} ||\mathbf{w}||^2 + \frac{1}{m} \sum_i \xi_i \quad \text{s.t.} \quad \forall i\, y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1 - \xi_i \wedge \xi_i \geq 0 \tag{3}$$

Denote the hinge-loss function as $\ell^{hinge}(a) := \max\{0, 1-a\}$. Show that the Soft-SVM optimization problem is equivalent to the following unconstraint optimization problem:

$$\operatorname*{argmin}_{\mathbf{w},\{\xi_i\}} \frac{\lambda}{2} ||\mathbf{w}||^2 + \frac{1}{m} \sum_i \ell^{hinge}(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) \tag{4}$$

הוכחה: יהי $w^* \in \mathbb{R}^d$, $\{\xi_i^*\}_{i=1}^m \subseteq \mathbb{R}$ - פתרון אופטימלי לבעיה (3). כלומר לכל $w \in \mathbb{R}^d$ ולכל $\{\xi_i\}_{i=1}^m$ מתקיים ש־

$$\frac{\lambda}{2} \|w^*\|^2 + \frac{1}{m} \sum_{i=1}^m \xi_i^* \leqslant \frac{\lambda}{2} \|w\|^2 + \frac{1}{m} \sum_{i=1}^m \xi_i$$

נשים לב שלכל $i \in [m]$ מתקיים ש־

$$\xi_i^* = \ell^{\text{hinge}}(y_i \langle w \mid x_i \rangle) = \max\{0, 1 - y_i \langle w^* \mid x_i \rangle\} = \begin{cases} 0 & 1 < y_i \langle w^* \mid x_i \rangle \\ 1 - y_i \langle w \mid x_i \rangle & 1 \geqslant y_i \langle w^* \mid x_i \rangle \end{cases}$$

כי אם נניח בשלילה שקיים j כך ש־ $\xi_j^* \neq \ell^{\text{hinge}}(y_j \langle w^* \mid x_j \rangle)$ אז:

- אם $1 < y_j \langle w^* \mid x_j \rangle$ אז $\ell^{\text{hinge}}(y_j \langle w^* \mid x_j \rangle) = 0$, ולכן מכך ש־ $\xi_j^* \geqslant 0$ נובע ש־ $\ell^{\text{hinge}}(y_j \langle w^* \mid x_j \rangle) < \xi_j^*$.
- אם $1 \geqslant y_j \langle w^* \mid x_j \rangle$ אז $\ell^{\text{hinge}}(y_j \langle w^* \mid x_j \rangle) = 1 - y_j \langle w^* \mid x_j \rangle$ ולכן מכך ש־ $1 - \xi_j^* \leqslant y_j \langle w^* \mid x_j \rangle$ נובע ש־

$$1 - y_j \langle w^* \mid x_j \rangle = \ell^{\text{hinge}}(y_j \langle w^* \mid x_j \rangle) < \xi_j^*$$

ובכל מקרה נקבל שהפתרון $w^*, \{\xi_i'\}_{i=1}^m$ כאשר $\xi_i' = \xi_i^*$ לכל $i \neq j$ ו־ $\xi_j' = \ell^{\text{hinge}}\left(y_j \langle w^* \mid x_j \rangle\right)$ **קטן** יותר מהפתרון האופטימלי, וזו סתירה.
לכן w^* הנ"ל הוא פתרון אופטימלי לבעיה (4).
בכיוון השני, אם w^* הוא פתרון אופטימלי ל־(4) אז לכל $w \in \mathbb{R}^d$ מתקיים ש־

$$\frac{\lambda}{2}\|w^*\|^2 + \frac{1}{m}\sum_{i=1}^n \ell^{\text{hinge}}\left(y_i \langle w^* \mid x_i \rangle\right) \leqslant \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^n \ell^{\text{hinge}}\left(y_i \langle w \mid x_i \rangle\right)$$

נסמן $\xi_i^* = \ell^{\text{hinge}}\left(y_i \langle w^* \mid x_i \rangle\right)$ לכל $i \in [m]$ ונקבל שהאילוצים של בעיה (3) מתקיימים, ובנוסף לכי שהראינו לעיל לכל $\{\xi_i\}_{i=1}^m \subseteq \mathbb{R}$ מתקיים ש־

$$\forall i \in [m]: \quad \xi_i^* = \ell^{\text{hinge}}\left(y_i \langle w^* \mid x_i \rangle\right) \leqslant \xi_i \quad \Longrightarrow \quad \frac{1}{m}\sum_{i=1}^m \xi_i^* \leqslant \frac{1}{m}\sum_{i=1}^m \xi_i$$

ולכן $w^*, \{\xi_i^*\}_{i=1}^m$ הוא פתרון אופטימלי לבעיה (3). ■

Naive Bayes Classifiers

שאלה 3

3. The ***Gaussian* Naive Bayes** classifier assumes a multinomial prior and independent feature-wise Gaussian likelihoods:

$$\begin{gathered} y \sim \text{Multinomial}(\boldsymbol{\pi}) \\ x_j | y = k \overset{ind.}{\sim} \mathcal{N}\left(\mu_{kj}, \sigma_{kj}^2\right) \end{gathered} \tag{6}$$

for $\boldsymbol{\pi}$ a probability vector: $\boldsymbol{\pi} \in [0,1]^K, \sum \boldsymbol{\pi}_j = 1$.
   (a) Suppose $x \in \mathbb{R}$ (i.e each sample has a single feature). Given a trainset $\{(x_i, y_i)\}_{i=1}^m$ fit a Gaussian Naive Bayes classifier solving (5) under assumptions (6).
   (b) Suppose $\mathbf{x} \in \mathbb{R}^d$ (i.e each sample has d feature). Given a trainset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ fit a Gaussian Naive Bayes classifier solving (5) under assumptions (6). You are encouraged to use the results from (3.a).

(a) כאשר הדגימות בעלות פיצ'ר בודד (חד־מימדיות) ההנחות של המודל Gaussian Naive Bayes הן שלכל $i \in [m]$:

$$\begin{gathered} y_i \sim \text{Mult}(\boldsymbol{\pi}) \\ x_i \mid y_i \sim \mathcal{N}\left(\mu_{y_i}, \sigma_{y_i}^2\right) \end{gathered}$$

כאשר $\boldsymbol{\pi} \in [0,1]^K$ ומתקיים $\sum_{j=1}^{K} \boldsymbol{\pi}_j = 1$, ובנוסף - $\sigma_k^2, \mu_k \in \mathbb{R}$ לכל $k \in [K]$.
לכן כדי לאמן את המודל צריך לחשב אומד של $\{\boldsymbol{\pi}_k\}_{k\in[K]}$ - ההסתברויות של המחלקות, אומד של $\{\mu_k\}_{k\in[K]}$ - התוחלות של המחלקות, ואומד של $\left\{\sigma_k^2\right\}_{k\in[k]}$ - השונויות של המחלקות. נחשב את ערך ה־likelihood בהינתן הדגימות:

$$\begin{aligned} \mathcal{L}(\Theta \mid \mathbf{x}, \mathbf{y}) = f_{X,Y|\Theta}\left(\{(x_i, y_i)\}_{i=1}^m\right) &= \prod_{i=1}^m f_{X,Y|\Theta}(x_i, y_i) = \\ &= \prod_{i=1}^m f_{X|Y=y_i}(x_i) \cdot f_{Y|\Theta}(y_i) = \prod_{i=1}^m \mathcal{N}\left(x_i \mid \mu_{y_i}, \sigma_{y_i}^2\right) \cdot \text{Mult}(y_i \mid \boldsymbol{\pi}) = \\ &= \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma_{y_i}^2}} \cdot \exp\left(-\frac{(x_i - \mu_{y_i})^2}{2\sigma_{y_i}^2}\right) \cdot \boldsymbol{\pi}_{y_i} \end{aligned}$$

כדי למצוא argmax ל־likelihood נוכל להשתמש ב־log-likelihood:

$$\begin{aligned}
\ell\left(\Theta\mid\mathbf{x},\mathbf{y}\right) &= \ln\left(\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\sigma_{y_i}^2}}\cdot\exp\left(-\frac{\left(x_i-\mu_{y_i}\right)^2}{2\sigma_{y_i}^2}\right)\cdot\boldsymbol{\pi}_{y_i}\right)=\\
&=\sum_{i=1}^{m}\ln\left(\frac{1}{\sqrt{2\pi\sigma_{y_i}^2}}\cdot\exp\left(-\frac{\left(x_i-\mu_{y_i}\right)^2}{2\sigma_{y_i}^2}\right)\cdot\boldsymbol{\pi}_{y_i}\right)=\\
&=\sum_{i=1}^{m}\left[\ln\left(\left(2\pi\sigma_{y_i}^2\right)^{-\frac{1}{2}}\right)-\frac{\left(x_i-\mu_{y_i}\right)^2}{2\sigma_{y_i}^2}+\ln(\boldsymbol{\pi}_{y_i})\right]=\\
&=\sum_{i=1}^{m}\left[\ln(\boldsymbol{\pi}_{y_i})-\frac{1}{2}\ln\left(2\pi\right)-\frac{1}{2}\ln\left(\sigma_{y_i}^2\right)-\frac{\left(x_i-\mu_{y_i}\right)^2}{2\sigma_{y_i}^2}\right]=\\
&=\sum_{i=1}^{m}\left[\ln(\boldsymbol{\pi}_{y_i})-\frac{1}{2}\ln\left(\sigma_{y_i}^2\right)-\frac{\left(x_i-\mu_{y_i}\right)^2}{2\sigma_{y_i}^2}\right]-\frac{m}{2}\ln\left(2\pi\right)=\\
&=\sum_{k\in[K]}\left[n_k\cdot\ln(\boldsymbol{\pi}_k)-\frac{n_k}{2}\ln\left(\sigma_k^2\right)-\sum_{i:\,y_i=k}\frac{\left(x_i-\mu_k\right)^2}{2\sigma_k^2}\right]-\frac{m}{2}\ln\left(2\pi\right)
\end{aligned}$$

כאשר $n_k=\sum_{i=1}^{m}\mathbb{1}_{[y_i=k]}$ לכל $k\in[K]$. כדי למצוא את ה־argmax נגזור ביחס לפרמטרים השונים ־ $\boldsymbol{\pi}_k,\mu_k,\sigma_k^2$ ־ ונשווה ל־0:

כדי לגזור ביחס ל־$\boldsymbol{\pi}_k$ נשתמש בשיטת כופלי לגרנז' ־ מכיוון שיש לנו את האילוץ $\sum_{k\in[K]}\boldsymbol{\pi}_k=1 \iff \sum_{k\in[K]}\boldsymbol{\pi}_k-1=0$ נגדיר:

$$g\left(\boldsymbol{\pi}\right)=\sum_{k\in[K]}\boldsymbol{\pi}_k-1$$

ונגזור את ־ $\mathcal{L}=\ell\left(\Theta\mid\mathbf{x},\mathbf{y}\right)-\lambda g\left(\boldsymbol{\pi}\right)$ כאשר λ הוא משתנה חדש:

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\pi}_k}=\frac{\partial}{\partial\boldsymbol{\pi}_k}\ell\left(\Theta\mid\mathbf{x},\mathbf{y}\right)-\lambda\frac{\partial}{\partial\boldsymbol{\pi}_k}g\left(\boldsymbol{\pi}\right)=\frac{n_k}{\boldsymbol{\pi}_k}-\lambda$$

נשווה ל־0 ונקבל ש־

$$\frac{n_k}{\boldsymbol{\pi}_k}-\lambda=0 \quad\iff\quad \boldsymbol{\pi}_k=\frac{n_k}{\lambda}$$

וכדי למצוא את λ נציב באילוץ את הערך של $\boldsymbol{\pi}_k$ שמצאנו:

$$1=\sum_{k\in[K]}\boldsymbol{\pi}_k=\sum_{k\in[K]}\frac{n_k}{\lambda} \quad\iff\quad \lambda=m$$

לכן אומד להסתברויות של המחלקות הוא:

$$\forall k\in[K]:\quad \boxed{\hat{\boldsymbol{\pi}}_k^{\mathrm{MLE}}=\frac{n_k}{m}}$$

נגזור ביחס ל־μ_k:

$$\frac{\partial}{\partial\mu_k}\ell\left(\Theta\mid\mathbf{x},\mathbf{y}\right)=\frac{\partial}{\partial\mu_k}\Big(-\sum_{i:\,y_i=k}\frac{\left(x_i-\mu_k\right)^2}{2\sigma_k^2}\Big)=-\sum_{i:\,y_i=k}\frac{\partial}{\partial\mu_k}\Big(\frac{\left(x_i-\mu_k\right)^2}{2\sigma_k^2}\Big)=\sum_{i:\,y_i=k}\frac{x_i-\mu_k}{\sigma_k^2}$$

ונשווה ל־0:

$$\sum_{i:\,y_i=k}\frac{x_i-\mu_k}{\sigma_k^2}=0 \quad\iff\quad \sum_{i:\,y_i=k}\left(x_i-\mu_k\right)=0 \quad\iff\quad n_k\mu_k=\sum_{i:\,y_i=k}x_i \quad\iff\quad \mu_k=\frac{1}{n_k}\sum_{i=1}^{m}\mathbb{1}_{[y_i=k]}x_i$$

לכן אומד לתוחלות של המחלקות הוא:

$$\forall k \in [K]: \quad \boxed{\hat{\mu}_k^{\mathrm{MLE}} = \frac{1}{n_k} \sum_{i=1}^{m} \mathbb{1}_{[y_i = k]} x_i}$$

נגזור ביחס ל־σ_k^2:

$$\begin{aligned}\frac{\partial}{\partial \sigma_k^2} \ell(\Theta \mid \mathbf{x}, \mathbf{y}) &= \frac{\partial}{\partial \sigma_k^2}\left(-\frac{n_k}{2}\ln\left(\sigma_k^2\right)\right) - \sum_{i:\, y_i = k} \frac{\partial}{\partial \sigma_k^2}\left(\frac{(x-\mu_k)^2}{2\sigma_k^2}\right) = \\ &= -\frac{n_k}{2\sigma_k^2} + \sum_{i:\, y_i = k} \frac{(x-\mu_k)^2}{2\left(\sigma_k^2\right)^2}\end{aligned}$$

ונשווה ל־0:

$$\begin{aligned}-\frac{n_k}{2\sigma_k^2} + \sum_{i:\, y_i = k} \frac{(x-\mu_k)^2}{2\left(\sigma_k^2\right)^2} = 0 \quad &\iff \quad \frac{\sum_{i:\, y_i=k} (x-\mu_k)^2 - \sigma_k^2 n_k}{2\left(\sigma_k^2\right)^2} = 0 \quad \iff \\ &\iff \quad \sum_{i:\, y_i=k} (x-\mu_k)^2 - \sigma_k^2 n_k = 0 \quad \iff \\ &\iff \quad \sigma_k^2 = \frac{1}{n_k} \sum_{i:\, y_i=k} (x-\mu_k)^2\end{aligned}$$

לכן אומד לשונויות של המחלקות הוא:

$$\forall k \in [K]: \quad \boxed{\hat{\sigma^2}_k^{\mathrm{MLE}} = \frac{1}{n_k} \sum_{i:\, y_i=k} \left(x - \hat{\mu}_k^{\mathrm{MLE}}\right)^2}$$

(b) כאשר הדגימות הן בעלות d פיצ'רים ההנחות של המודל Naive Bayes Classifier הן שלכל $i \in [m]$:

$$\begin{aligned} y_i &\sim \mathrm{Mult}(\boldsymbol{\pi}) \\ [\mathbf{x}_i]_j \mid y_i &\overset{\text{ind}}{\sim} \mathcal{N}\left(\mu_{y_i,j}, \sigma_{y_i,j}^2\right)\end{aligned}$$

כלומר הלייבל ה־i מתפלג מולטינומי עם $\boldsymbol{\pi} \in [0,1]^K$, $\sum_{k\in[K]} \boldsymbol{\pi}_k = 1$, והפיצ'ר ה־$j$ של כל דגימה $\mathbf{x}_i \in \mathbb{R}^d$ בהינתן y_i מתפלג נורמלי עם תוחלת $\mu_{y_i,j} \in \mathbb{R}$ ועם שונות $\sigma_{y_i,j}^2 \in \mathbb{R}$.
לכן כדי לאמן את המודל צריך למצוא אומד ל־$\{\boldsymbol{\pi}_k\}_{k\in[K]}$, $\{\mu_{k,j}\}_{k\in[K]}^{j\in[d]}$, $\{\sigma_{k,j}^2\}_{k\in[K]}^{j\in[d]}$. פונקציית ה־likelihood היא:

$$\begin{aligned}\mathcal{L}(\Theta \mid \mathbf{X}, \mathbf{y}) &= f_{X,Y|\Theta}\left(\{(\mathbf{x}_i, y_i)\}_{i=1}^m\right) = \prod_{i=1}^m f_{X,Y|\Theta}(\mathbf{x}_i, y_i) = \\ &= \prod_{i=1}^m f_{X|Y=y_i}(\mathbf{x}_i) \cdot f_{Y|\Theta}(y_i) = \prod_{i=1}^m f_{X_1|Y=y_i,\ldots,X_d|Y=y_i}\left([\mathbf{x}_i]_1, \ldots, [\mathbf{x}_i]_d\right) \cdot f_{Y|\Theta}(y_i) = \\ &\overset{☼}{=} \prod_{i=1}^m \left[f_{Y|\Theta}(y_i) \cdot \prod_{j=1}^d f_{X_j|Y=y_i}\left([\mathbf{x}_i]_j\right)\right] = \\ &= \prod_{i=1}^m \left[\mathrm{Mult}(y_i \mid \boldsymbol{\pi}) \cdot \prod_{j=1}^d \mathcal{N}\left([\mathbf{x}_i]_j \mid \mu_{y_i,j}, \sigma_{y_i,j}^2\right)\right] = \\ &= \prod_{i=1}^m \left[\boldsymbol{\pi}_{y_i} \cdot \prod_{j=1}^d \frac{1}{\sqrt{2\pi\sigma_{y_i,j}^2}} \exp\left(-\frac{\left([\mathbf{x}_i]_j - \mu_{y_i,j}\right)^2}{2\sigma_{y_i,j}^2}\right)\right]\end{aligned}$$

☼: לפי ההנחה של המודל ש־$\left[\mathbf{x}_i\right]_j \mid y_i \overset{\text{ind}}{\sim} \mathcal{N}\left(\mu_{y_i,j}, \sigma^2_{y_i,j}\right)$.
כדי למצוא argmax ל־likelihood נשתמש ב־log-likelihood:

$$\begin{aligned}
\ell\left(\Theta \mid \mathbf{X},\mathbf{y}\right) = \ln\left(\prod_{i=1}^{m}\left[\boldsymbol{\pi}_{y_i}\cdot \prod_{j=1}^{d}\frac{1}{\sqrt{2\pi\sigma^2_{y_i,j}}}\exp\left(-\frac{\left(\left[\mathbf{x}_i\right]_j-\mu_{y_i,j}\right)^2}{2\sigma^2_{y_i,j}}\right)\right]\right) = \\
=\sum_{i=1}^{m}\ln\left(\boldsymbol{\pi}_{y_i}\cdot \prod_{j=1}^{d}\frac{1}{\sqrt{2\pi\sigma^2_{y_i,j}}}\exp\left(-\frac{\left(\left[\mathbf{x}_i\right]_j-\mu_{y_i,j}\right)^2}{2\sigma^2_{y_i,j}}\right)\right) = \\
=\sum_{i=1}^{m}\left[\ln(\boldsymbol{\pi}_{y_i}) - \sum_{j=1}^{d}\left[\frac{1}{2}\ln\left(2\pi\right)+\frac{1}{2}\ln\left(\sigma^2_{y_i,j}\right)+\frac{\left(\left[\mathbf{x}_i\right]_j-\mu_{y_i,j}\right)^2}{2\sigma^2_{y_i,j}}\right]\right] = \\
=\sum_{i=1}^{m}\left[\ln(\boldsymbol{\pi}_{y_i}) - \frac{1}{2}\sum_{j=1}^{d}\ln\left(\sigma^2_{y_i,j}\right) - \sum_{j=1}^{d}\frac{\left(\left[\mathbf{x}_i\right]_j-\mu_{y_i,j}\right)^2}{2\sigma^2_{y_i,j}}\right] - \frac{m\cdot d}{2}\ln\left(2\pi\right) = \\
=\sum_{k\in[K]}\left[n_k\ln(\boldsymbol{\pi}_k) - \frac{n_k}{2}\sum_{j=1}^{d}\ln\left(\sigma^2_{k,j}\right) - \sum_{i:\,y_i=k}\sum_{j=1}^{d}\frac{\left(\left[\mathbf{x}_i\right]_j-\mu_{k,j}\right)^2}{2\sigma^2_{k,j}}\right] - \frac{m\cdot d}{2}\ln\left(2\pi\right)
\end{aligned}$$

כאשר $n_k = \sum_{i=1}^{m}\mathbb{1}_{[y_i=k]}$ לכל $k\in[K]$. כדי למצוא את ה־argmax נגזור ביחס לפרמטרים – $\boldsymbol{\pi}_k, \mu_{k,j}, \sigma^2_{k,j}$ – ונשווה ל־0:
כדי לגזור ביחס ל־$\boldsymbol{\pi}_k$ נשתמש בכופלי לגרנז'. נגדיר –

$$g\left(\boldsymbol{\pi}\right) = \sum_{k\in[K]}\boldsymbol{\pi}_k - 1$$

ונגזור את $\mathcal{L} = \ell\left(\Theta \mid \mathbf{X},\mathbf{y}\right) - \lambda g\left(\boldsymbol{\pi}\right)$ כאשר λ הוא משתנה חדש:

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\pi}_k} = \frac{\partial}{\partial\boldsymbol{\pi}_k}\ell\left(\Theta \mid \mathbf{X},\mathbf{y}\right) - \frac{\partial}{\partial\boldsymbol{\pi}_k}\lambda g\left(\boldsymbol{\pi}\right) = \qquad \frac{n_k}{\boldsymbol{\pi}_k} - \lambda$$

נשווה ל־0 ונקבל ש־ $\boldsymbol{\pi}_k = \frac{n_k}{\lambda}$, וע"י הצבה של הערך הזה באילוץ נקבל ש־ $\lambda = m \iff \sum_{k\in[K]}\frac{n_k}{\lambda} = 1$.
לכן אומד להסתברויות של המחלקות הוא:

$$\forall k\in[K]: \quad \boxed{\hat{\boldsymbol{\pi}}_k^{\text{MLE}} = \frac{n_k}{m}}$$

נגזור את $\ell\left(\Theta \mid \mathbf{X},\mathbf{y}\right)$ ביחס ל־$\mu_{k,j}$:

$$\begin{aligned}
\frac{\partial}{\partial\mu_{k.j}}\ell\left(\Theta \mid \mathbf{X},\mathbf{y}\right) = \frac{\partial}{\partial\mu_{k,j}}\Big(-\sum_{i:\,y_i=k}\frac{\left(\left[\mathbf{x}_i\right]_j-\mu_{k,j}\right)^2}{2\sigma^2_{k,j}}\Big) = \\
= -\sum_{i:\,y_i=k}\frac{\partial}{\partial\mu_{k,j}}\Big(\frac{\left(\left[\mathbf{x}_i\right]_j-\mu_{k,j}\right)^2}{2\sigma^2_{k,j}}\Big) = \sum_{i:\,y_i=k}\frac{\left[\mathbf{x}_i\right]_j-\mu_{k,j}}{\sigma^2_{k,j}}
\end{aligned}$$

ונשווה ל־0:

$$\sum_{i:\,y_i=k}\frac{\left[\mathbf{x}_i\right]_j-\mu_{k,j}}{\sigma^2_{k,j}} = 0 \quad\iff\quad \sum_{i:\,y_i=k}\left(\left[\mathbf{x}_i\right]_j-\mu_{k,j}\right) = 0 \quad\iff\quad \mu_{k,j} = \frac{1}{n_k}\sum_{i:\,y_i=k}\left[\mathbf{x}_i\right]_j$$

לכן אומד לתוחלות הוא:

$$\forall k\in[K]\ ,\ j\in[d]: \quad \boxed{\hat{\mu}_{k,j}^{\text{MLE}} = \frac{1}{n_k}\sum_{i:\,y_i=k}\left[\mathbf{x}_i\right]_j}$$

נגזור ביחס ל־$\sigma_{k,j}^2$:

$$\frac{\partial}{\partial\sigma_{k,j}^2}\ell\left(\Theta \mid \mathbf{X},\mathbf{y}\right)=\frac{\partial}{\partial\sigma_{k,j}^2}\left(-\frac{n_k}{2}\sum_{j=1}^{d}\ln\left(\sigma_{k,j}^2\right)-\sum_{i:\,y_i=k}\sum_{j=1}^{d}\frac{\left(\left[\mathbf{x}_i\right]_j-\mu_{k,j}\right)^2}{2\sigma_{k,j}^2}\right)=$$

$$=-\frac{n_k}{2}\cdot\frac{\partial}{\partial\sigma_{k,j}^2}\ln\left(\sigma_{k,j}^2\right)-\sum_{i:\,y_i=k}\frac{\partial}{\partial\sigma_{k,j}^2}\left(\frac{\left(\left[\mathbf{x}_i\right]_j-\mu_{k,j}\right)^2}{2\sigma_{k,j}^2}\right)=$$

$$=-\frac{n_k}{2\sigma_{k,j}^2}+\sum_{i:\,y_i=k}\frac{\left(\left[\mathbf{x}_i\right]_j-\mu_{k,j}\right)^2}{2\left(\sigma_{k,j}^2\right)^2}$$

ונשווה ל־0:

$$-\frac{n_k}{2\sigma_{k,j}^2}+\sum_{i:\,y_i=k}\frac{\left(\left[\mathbf{x}_i\right]_j-\mu_{k,j}\right)^2}{2\left(\sigma_{k,j}^2\right)^2}=0 \iff \frac{\sum_{i:\,y_i=k}\left(\left[\mathbf{x}_i\right]_j-\mu_{k,j}\right)^2-\sigma_{k,j}^2 n_k}{2\left(\sigma_{k,j}^2\right)^2}=0 \iff$$

$$\iff \sigma_{k,j}^2=\frac{1}{n_k}\sum_{i:\,y_i=k}\left(\left[\mathbf{x}_i\right]_j-\mu_{k,j}\right)^2$$

לכן אומד לשונויות הוא:

$$\forall k\in[K]\ ,\ j\in[d]:\quad \boxed{\hat{\sigma^2}_{k,j}^{\text{MLE}}=\frac{1}{n_k}\sum_{i:\,y_i=k}\left(\left[\mathbf{x}_i\right]_j-\hat{\mu}_{k,j}^{\text{MLE}}\right)^2}$$

שאלה 4

4. The ***Poisson*** **Naive Bayes** classifier assumes a multinomial prior and independent feature-wise Poisson likelihoods:

$$\begin{aligned} y &\sim \text{Multinomial}\left(\boldsymbol{\pi}\right) \\ x_j|y=k &\overset{ind.}{\sim} \text{Poi}\left(\lambda_{kj}\right) \end{aligned} \tag{7}$$

for $\boldsymbol{\pi}$ a probability vector: $\boldsymbol{\pi}\in[0,1]^K, \sum\boldsymbol{\pi}_j=1$.

(a) Suppose $x\in\mathbb{R}$ (i.e each sample has a single feature). Given a trainset $\{(x_i,y_i)\}_{i=1}^m$ fit a Gaussian Naive Bayes classifier solving (5) under assumptions (7).

(b) Suppose $\mathbf{x}\in\mathbb{R}^d$ (i.e each sample has d feature). Given a trainset $S=\{(\mathbf{x}_i,y_i)\}_{i=1}^m$ fit a Poisson Naive Bayes classifier solving (5) under assumptions (7). You are encouraged to use the results from (4.a).

(a) כאשר הדגימות הן בעלות פיצ'ר בודד ההנחות של המודל Poisson Naive Bayes הן שלכל $i\in[m]$:

$$\begin{aligned} y_i &\sim \text{Mult}\left(\boldsymbol{\pi}\right) \\ x_i \mid y_i &\overset{\text{ind}}{\sim} \text{Poi}(\lambda_{y_i}) \end{aligned}$$

כאשר $\boldsymbol{\pi}\in[0,1]^K$, $\sum_{k\in[K]}\boldsymbol{\pi}_k=1$. כדי לאמן את המודל צריך לחשב אומד של $\{\boldsymbol{\pi}_k\}_{k\in[K]}$ ושל $\{\lambda_k\}_{k\in[K]}$.
נתחיל בחישוב פונקציית ה־likelihood:

$$\mathcal{L}\left(\Theta \mid \mathbf{x},\mathbf{y}\right)=p_{X,Y|\Theta}\left(\{(x_i,y_i)\}_{i=1}^m\right)=\prod_{i=1}^m p_{X,Y|\Theta}\left(x_i,y_i\right)=\prod_{i=1}^m p_{X|Y=y_i}\left(x_i\right)\cdot f_{Y|\Theta}\left(y_i\right)=$$

$$=\prod_{i=1}^m \text{Poi}\left(x_i \mid \lambda_{y_i}\right)\cdot\text{Mult}\left(y_i \mid \boldsymbol{\pi}\right)=\prod_{i=1}^m\frac{e^{-\lambda_{y_i}}\cdot\left(\lambda_{y_i}\right)^{x_i}}{x_i!}\cdot\boldsymbol{\pi}_{y_i}$$

כדי למקסם את ה־likelihood אפשר למקסם את ה־log-likelihood:

$$\begin{aligned}
\ell\left(\Theta \mid \mathbf{x},\mathbf{y}\right) = \ln\left(\prod_{i=1}^{m}\frac{e^{-\lambda_{y_i}}\cdot\left(\lambda_{y_i}\right)^{x_i}}{x_i!}\cdot\boldsymbol{\pi}_{y_i}\right) &= \sum_{i=1}^{m}\left[\ln\left(\frac{e^{-\lambda_{y_i}}\cdot\left(\lambda_{y_i}\right)^{x_i}}{x_i!}\right)+\ln(\boldsymbol{\pi}_{y_i})\right] = \\
&= \sum_{i=1}^{m}\left[\ln\left(e^{-\lambda_{y_i}}\right)+\ln\left(\left(\lambda_{y_i}\right)^{x_i}\right)-\ln\left(x_i!\right)+\ln(\boldsymbol{\pi}_{y_i})\right] = \\
&= \sum_{i=1}^{m}\left[-\lambda_{y_i}+x_i\cdot\ln(\lambda_{y_i})-\ln\left(x_i!\right)+\ln\left(\boldsymbol{\pi}_{y_i}\right)\right] = \\
&= \sum_{k\in[K]}\left[-n_k\lambda_k+\ln(\lambda_k)\cdot\sum_{i:\,y_i=k}x_i-\sum_{i:\,y_i=k}\ln\left(x_i!\right)+n_k\ln(\boldsymbol{\pi}_k)\right]
\end{aligned}$$

כאשר $n_k = \sum_{i=1}^{m}\mathbb{1}_{[y_i=k]}$ לכל $k\in[K]$.

נגזור ביחס לפרמטר $\boldsymbol{\pi}_k$ בעזרת כופלי לגרנז. נגדיר $g\left(\boldsymbol{\pi}\right)=\sum_{k\in[K]}\boldsymbol{\pi}_k-1$ ונגזור את $\ell\left(\Theta\mid\mathbf{x},\mathbf{y}\right)-\gamma g\left(\boldsymbol{\pi}\right)$ כאשר γ הוא משתנה חדש:

$$\frac{\partial}{\partial\boldsymbol{\pi}_k}\ell\left(\Theta\mid\mathbf{x},\mathbf{y}\right)-\gamma\frac{\partial}{\partial\boldsymbol{\pi}_k}g\left(\boldsymbol{\pi}\right)=\frac{n_k}{\boldsymbol{\pi}_k}-\gamma$$

נשווה ל־0 ונקבל ש־ $\boldsymbol{\pi}_k=\frac{n_k}{\gamma}$, וע"י הצבה של הערך הזה באילוץ נקבל ש־ $\gamma=\sum_{k\in[K]}n_k=m$. לכן אומד להסתברויות של המחלקות הוא:

$$\forall k\in[K]:\quad \boxed{\hat{\boldsymbol{\pi}}_k^{\text{MLE}}=\frac{n_k}{m}}$$

נגזור את $\ell\left(\Theta\mid\mathbf{x},\mathbf{y}\right)$ ביחס ל־λ_k:

$$\frac{\partial}{\partial\lambda_k}\ell\left(\Theta\mid\mathbf{x},\mathbf{y}\right)=-n_k+\frac{1}{\lambda_k}\cdot\sum_{i:\,y_i=k}x_i-0+0$$

נשווה ל־0 ונקבל:

$$-n_k+\frac{1}{\lambda_k}\cdot\sum_{i:\,y_i=k}x_i=0\quad\iff\quad\lambda_k=\frac{1}{n_k}\sum_{i:\,y_i=k}x_i=\frac{1}{n_k}\sum_{i=1}^{m}\mathbb{1}_{[y_i=k]}\cdot x_i$$

לכן אומד מתאים לפרמטרים של ההתפלגות הוא:

$$\forall k\in[K]:\quad \boxed{\hat{\lambda}_k^{\text{MLE}}=\frac{1}{n_k}\sum_{i=1}^{m}\mathbb{1}_{[y_i=k]}\cdot x_i}$$

(b) כאשר הדגימות $\left\{(\mathbf{x}_i,y_i)\right\}_{i=1}^{m}$ הן בעלות d פיצ'רים ההנחות של המודל Poisson Naive Bayes הן שלכל $i\in[m]$:

$$y_i\sim\text{Mult}\left(\boldsymbol{\pi}\right)$$

$$\left[\mathbf{x}_i\right]_j\mid y_i\overset{\text{ind}}{\sim}\text{Poi}(\lambda_{y_i,j})$$

כדי לאמן את המודל צריך לחשב אומדים להסתברויות של המחלקות $\left\{\boldsymbol{\pi}_k\right\}_{k\in[K]}$ ולפרמטרים של ההתפלגות פואסון $\left\{\lambda_{k,j}\right\}_{k\in[K]}^{j\in[d]}$.
נחשב את פונקציית ה־likelihood:

$$\begin{aligned}
\mathcal{L}\left(\Theta\mid\mathbf{X},\mathbf{y}\right)=p_{X,Y\mid\Theta}\left(\left\{(\mathbf{x}_i,y_i)\right\}_{i=1}^{m}\right) &= \prod_{i=1}^{m}p_{X,Y\mid\Theta}\left(\mathbf{x}_i,y_i\right)=\prod_{i=1}^{m}p_{X\mid Y=y_i}\left(\mathbf{x}_i\right)\cdot p_{Y\mid\Theta}\left(y_i\right)= \\
&= \prod_{i=1}^{m}p_{X_1\mid Y=y_i,\ldots,X_d\mid Y=y_i}\left(\left[\mathbf{x}_i\right]_1,\ldots,\left[\mathbf{x}_i\right]_d\right)\cdot p_{Y\mid\Theta}\left(y_i\right)=\prod_{i=1}^{m}\left[p_{Y\mid\Theta}\left(y_i\right)\cdot\prod_{j=1}^{d}p_{X_j\mid Y=y_i}\left(\left[\mathbf{x}_i\right]_j\right)\right] \\
&= \prod_{i=1}^{m}\left[\text{Mult}\left(y_i\mid\boldsymbol{\pi}\right)\cdot\prod_{j=1}^{d}\text{Pois}\left(\left[\mathbf{x}_i\right]_j\mid\lambda_{y_i,j}\right)\right]=\prod_{i=1}^{m}\left[\boldsymbol{\pi}_{y_i}\cdot\prod_{j=1}^{d}\frac{e^{-\lambda_{y_i,j}}\cdot\left(\lambda_{y_i,j}\right)^{\left[\mathbf{x}_i\right]_j}}{\left[\mathbf{x}_i\right]_j!}\right]
\end{aligned}$$

☼: לפי ההנחה של המודל שהפיצ'רים בלתי תלויים.

כדי למקסם את ה־likelihood אפשר למקסם את ה־log-likelihood:

$$\begin{aligned}\ell\left(\Theta \mid \mathbf{X},\mathbf{y}\right) &= \ln\left(\prod_{i=1}^{m}\left[\boldsymbol{\pi}_{y_i}\cdot\prod_{j=1}^{d}\frac{e^{-\lambda_{y_i,j}}\cdot\left(\lambda_{y_i,j}\right)^{[\mathbf{x}_i]_j}}{[\mathbf{x}_i]_j!}\right]\right) = \sum_{i=1}^{m}\ln\left(\boldsymbol{\pi}_{y_i}\cdot\prod_{j=1}^{d}\frac{e^{-\lambda_{y_i,j}}\cdot\left(\lambda_{y_i,j}\right)^{[\mathbf{x}_i]_j}}{[\mathbf{x}_i]_j!}\right) = \\ &= \sum_{i=1}^{m}\left[\ln(\boldsymbol{\pi}_{y_i}) + \ln\left(\prod_{j=1}^{d}\frac{e^{-\lambda_{y_i,j}}\cdot\left(\lambda_{y_i,j}\right)^{[\mathbf{x}_i]_j}}{[\mathbf{x}_i]_j!}\right)\right] = \\ &= \sum_{i=1}^{m}\left[\ln(\boldsymbol{\pi}_{y_i}) + \sum_{j=1}^{d}\left[-\lambda_{y_i,j} + [\mathbf{x}_i]_j\cdot\ln(\lambda_{y_i,j}) - \ln([\mathbf{x}_i]_j)\right]\right] = \\ &= \sum_{i=1}^{m}\left[\ln(\boldsymbol{\pi}_{y_i}) - \sum_{j=1}^{d}\lambda_{y_i,j} + \sum_{j=1}^{d}[\mathbf{x}_i]_j\cdot\ln(\lambda_{y_i,j}) - \sum_{j=1}^{d}\ln([\mathbf{x}_i]_j)\right] = \\ &= \sum_{k\in[K]}\left[n_k\ln(\boldsymbol{\pi}_k) - n_k\sum_{j=1}^{d}\lambda_{k,j} + \sum_{j=1}^{d}\left[\ln(\lambda_{k,j})\cdot\sum_{i:\,y_i=k}[\mathbf{x}_i]_j\right] - \sum_{i:\,y_i=k}\sum_{j=1}^{d}\ln([\mathbf{x}_i]_j)\right]\end{aligned}$$

כאשר $n_k = \sum_{i=1}^{m}\mathbb{1}_{[y_i=k]}$ לכל $k\in[K]$.

נגזור ביחס לפרמטר $\boldsymbol{\pi}_k$ בעזרת כופלי לגרנז. נגדיר $g\left(\boldsymbol{\pi}\right) = \sum_{k\in[K]}\boldsymbol{\pi}_k - 1$ ונגזור את $\ell\left(\Theta \mid \mathbf{x},\mathbf{y}\right) - \gamma g\left(\boldsymbol{\pi}\right)$ כאשר γ הוא משתנה חדש:

$$\frac{\partial}{\partial\boldsymbol{\pi}_k}\ell\left(\Theta \mid \mathbf{X},\mathbf{y}\right) - \gamma\frac{\partial}{\partial\boldsymbol{\pi}_k}g\left(\boldsymbol{\pi}\right) = \frac{n_k}{\boldsymbol{\pi}_k} - \gamma$$

נשווה ל־0 ונקבל ש־ $\boldsymbol{\pi}_k = \frac{n_k}{\gamma}$, וע"י הצבה של הערך הזה באילוץ נקבל ש־ $\gamma = \sum_{k\in[K]}n_k = m$. לכן אומד להסתברויות של המחלקות הוא:

$$\forall k\in[K]:\quad \boxed{\hat{\boldsymbol{\pi}}_k^{\text{MLE}} = \frac{n_k}{m}}$$

נגזור את $\ell\left(\Theta \mid \mathbf{X},\mathbf{y}\right)$ ביחס ל־$\lambda_{k,j}$:

$$\frac{\partial}{\partial\lambda_{k,j}}\ell\left(\Theta \mid \mathbf{X},\mathbf{y}\right) = 0 - n_k + \frac{1}{\lambda_{k,j}}\cdot\sum_{i:\,y_i=k}[\mathbf{x}_i]_j - 0$$

נשווה ל־0 ונקבל:

$$-n_k + \frac{1}{\lambda_{k,j}}\cdot\sum_{i:\,y_i=k}[\mathbf{x}_i]_j = 0 \iff -n_k\lambda_{k,j} + \sum_{i:\,y_i=k}[\mathbf{x}_i]_j = 0 \iff \lambda_{k,j} = \frac{1}{n_k}\cdot\sum_{i:\,y_i=k}[\mathbf{x}_i]_j = \frac{1}{n_k}\sum_{i=1}^{m}\mathbb{1}_{[y_i=k]}\cdot[\mathbf{x}_i]_j$$

לכן אומד מתאים לפרמטרים של ההתפלגות הוא:

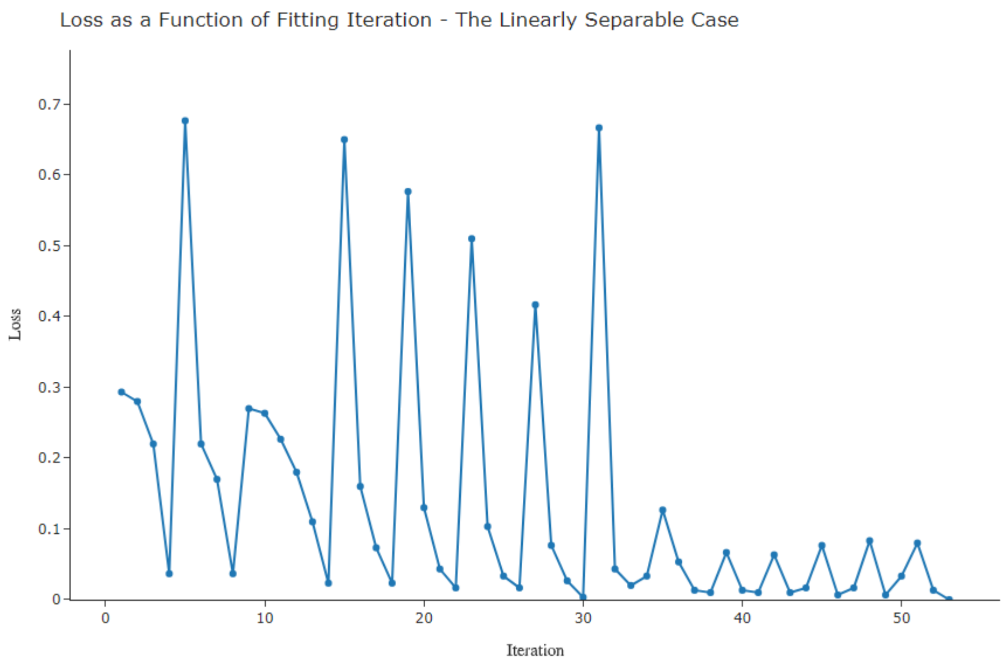
$$\forall k\in[K]\ ,\ j\in[d]:\quad \boxed{\hat{\lambda}_{k,j}^{\text{MLE}} = \frac{1}{n_k}\sum_{i=1}^{m}\mathbb{1}_{[y_i=k]}\cdot[\mathbf{x}_i]_j}$$

חלק פרקטי

Perceptron Classifier

שאלה 1

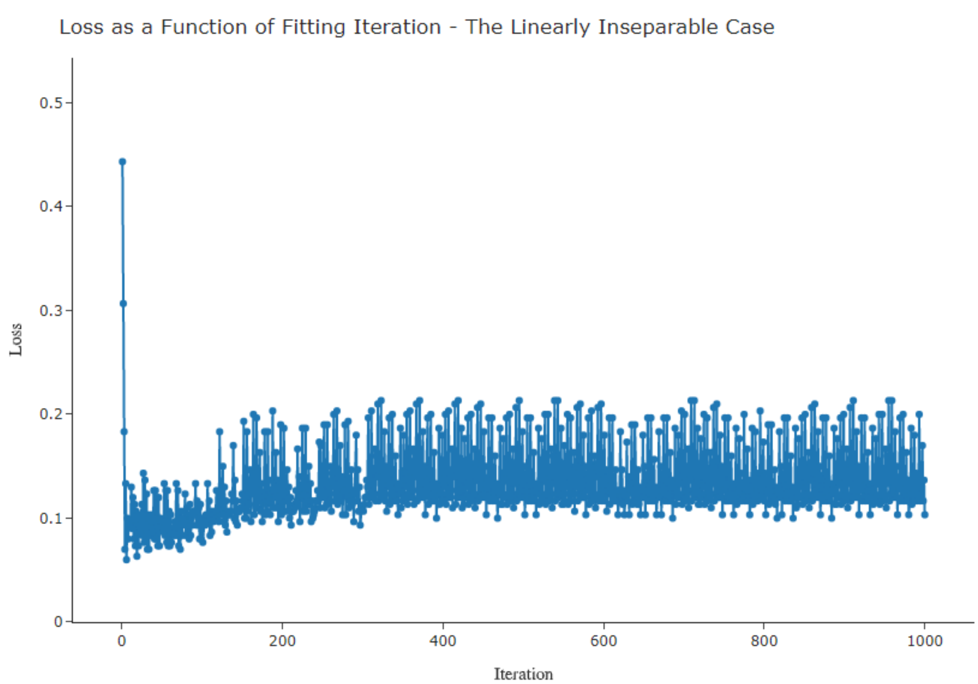
גרף המציג את ההתקדמות של האלגוריתם על הדאטא שניתן להפרדה ליניארית - מציג את ערך ה-Loss על ה-training data כפונקציה של מספר האיטרציות של האלגוריתם:



ניתן לראות שלאחר 50 איטרציות של האלגוריתם הוא מתכנס - ערך ה-Loss שווה ל-0 ממש. כלומר האלגוריתם מצליח להפריד ליניארית את הדאטא.

שאלה 2

גרף המציג את ההתקדמות של האלגוריתם על הדאטא **שלא** ניתן להפרדה ליניארית:



ההבדל בין שני הגרפים ניכר ־ רואים בגרף השני שהאלגוריתם **לא מתכנס** על הדאטא שלא ניתן להפרדה ליניארית ־ גם אחרי 1000 איטרציות של האלגוריתם ערך ה־Loss נשאר גדול מ־0, כלומר האלגוריתם לא מצליח להפריד ליניארית את הדאטא.
במונחים פורמליים ־ רק עבור הדאטא שניתן להפרדה ליניארית האלגוריתם מצא וקטור $w \in \mathbb{R}^{d+1}$ כך שלכל $i \in [m]$ מתקיים ש־

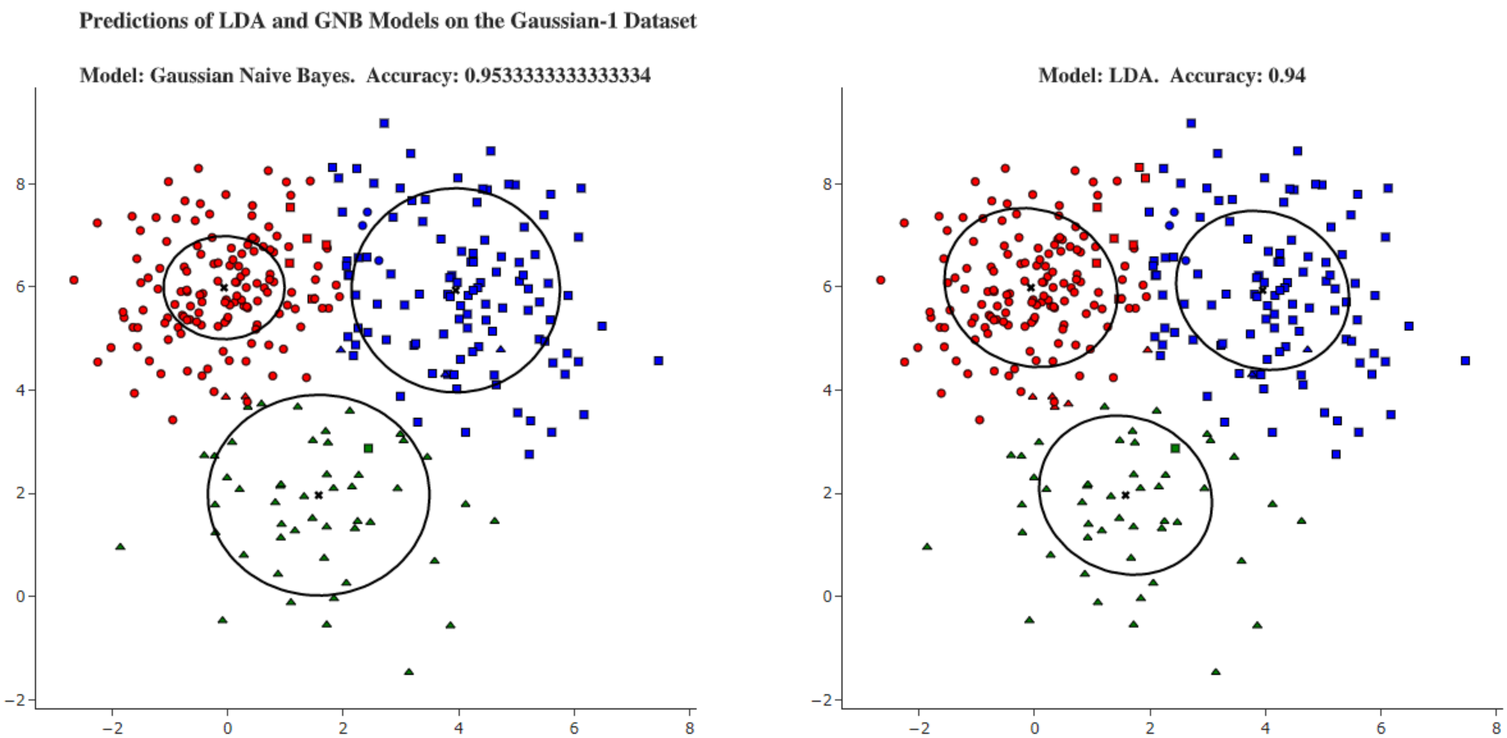
$$y_i \cdot \langle w \mid x_i \rangle > 0$$

(כשמוסיפים ל־x את ה־intercept כדי שהעל־מישור יוכל לא לעבור בראשית הצירים). עבור הדאטא שאינו ניתן להפרדה ליניארית לא קיים וקטור w כזה, ולכן האלגוריתם לא התכנס.

Bayes Classifiers

שאלה 1

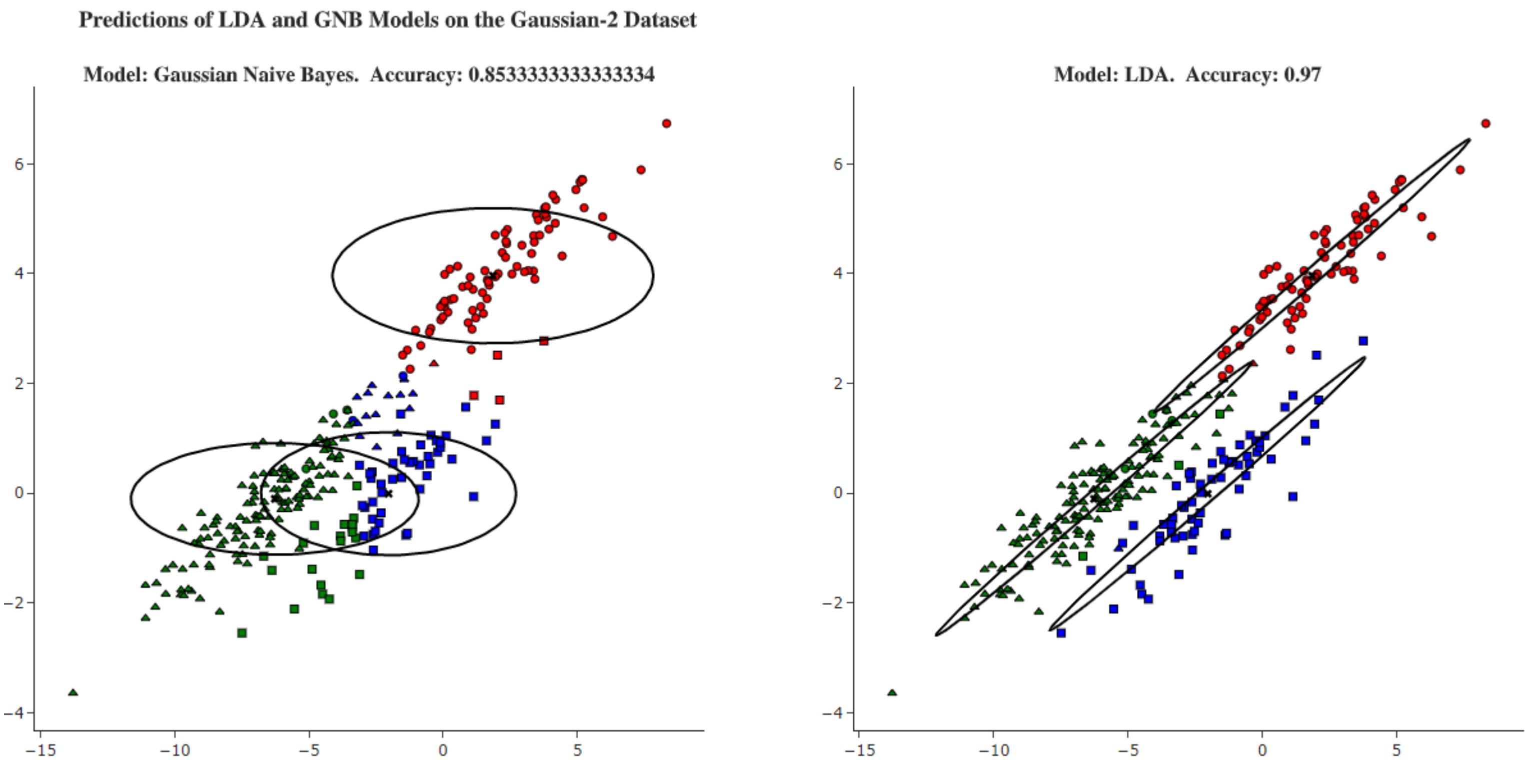
הגרף הנדרש:



לפי הגרפים הללו אפשר להסיק שההתפלגות שממנה נדגם הדאטא דומה יותר להתפלגות שהמודל Gaussian Naive Bayes מניח, כי הוא הצליח להפריד את הדאטא קצת יותר טוב מה־LDA. כלומר ־ הפיצ'רים בדאטא הם בלתי תלויים, ו־3 המחלקות נדגמו עם התפלגות נורמלית סביב 3 הנקודות (שמסומנות ב־$\times$) שניתן לראות בגרף.
להס

שאלה 2

הגרף:



בתרחיש הזה, דווקא המודל LDA הפריד את הדאטא יותר טוב מאשר ה־GNB.

באמת ניתן לראות שבדאטא הזה דווקא **יש** תלות בין הפיצ'רים. דרך אחת לראות זאת היא ע"י האליפסות שבגרפים המציגות את מטריצת השונות המשותפת של הפיצ'רים - ה־LDA לא מניח אי־תלות בין הפיצ'רים והוא באמת בונה מודל שבו השונות המשותפת שלהם שונה מ־0. לעומת זאת ההנחה של ה־GNB שהשונות המשותפת היא 0 גורמת לכך שהוא חוזה את הפיזור של הנקודות באופן פחות מדוייק.

בנוסף ניתן לראות גם שכאשר הערך בציר ה־x גדל - כלומר הערך של הפיצ'ר הראשון גדל - גם הערך בציר ה־y גדל - כלומר הערך של הפיצ'ר השני גדל.

מעבר לכך מכיוון שה־LDA זיהה ברמת דיוק גבוהה את הנקודות ניתן להסיק שההתפלגות שממנה נדגם הדאטא היא מהתפלגות נורמלית עם 3 מוקדים שקרובים לנקודות (שמסומנות ב־×) שניתן לראות בגרף.