מערכות לומדות תרגיל 3

28 באפריל 2022

חלק תיאורטי

1. הוכיחו כי בעיית האופטימיזציה Hard-SVM היא בעיית תכנות ריבועית:

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

כלומר, מצאו מטריצות Q ו־ A ווקטורים a ו־d כך שהבעיה הנ"ל יכולה להיכתב כך:

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

רמז: שימו לב כי $\|w\|^2=w^TIw$

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

$$\text{argmin}_{(w,b)}\begin{pmatrix} w \\ b \end{pmatrix}^T I \begin{pmatrix} w \\ b \end{pmatrix} \quad \text{s.t} \quad \begin{pmatrix} (y_1x_1) & y_1 \\ \vdots & \vdots \\ (y_mx_m) & y_m \end{pmatrix}\begin{pmatrix} w \\ b \end{pmatrix}\geq\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$\text{argmin}_{(w,b)}\begin{pmatrix} w \\ b \end{pmatrix}^T I \begin{pmatrix} w \\ b \end{pmatrix} \quad \text{s.t} \quad \begin{pmatrix} (y_1x_1) & y_1 \\ \vdots & \vdots \\ (y_mx_m) & y_m \end{pmatrix}\begin{pmatrix} w \\ b \end{pmatrix}\geq\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

אם נכפול ונחלק את הביטוי הראשון ב־2 (אין שינוי), נגדיר $a=\vec{0}$ ונכפיל את הביטוי השני בשתי צדדיו ב־-1 נקבל:

$$\text{argmin}_{(w,b)}\frac{1}{2}\begin{pmatrix} w \\ b \end{pmatrix}^T 2\cdot I\begin{pmatrix} w \\ b \end{pmatrix}+0^T\begin{pmatrix} w \\ b \end{pmatrix} \quad \text{s.t} \quad -\begin{pmatrix} (y_1x_1) & y_1 \\ \vdots & \vdots \\ (y_mx_m) & y_m \end{pmatrix}\begin{pmatrix} w \\ b \end{pmatrix}\leq\begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix}$$

לכן קיבלנו: $v=\begin{pmatrix} w \\ b \end{pmatrix}$, $Q=2\cdot I$, $A=-\begin{pmatrix} (y_1x_1) & y_1 \\ \vdots & \vdots \\ (y_mx_m) & y_m \end{pmatrix}$ ו־ $d=\begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix}$.

2. בעיית Soft-SVM אופטימיזציה:

$$\mathbf{argmin}_{w,\{\varepsilon_i\}}\frac{\lambda}{2}\|w\|^2+\frac{1}{m}\sum\varepsilon_i\quad \mathbf{s.t}\quad \forall i\ y_i\left\langle w,x_i\right\rangle\geq 1-\varepsilon_i\wedge\varepsilon_i\geq 0$$

סמנו את פונקציית $hinge-loss$ כ־ $\ell^{hinge}(a):=\mathbf{max}\{0,1-a\}$. הראו כי בעיית האופטימיזציה Soft-SVM היא שקולה לבעיית האופטימיזציה ללא אילוצים הבאה:

$$\mathbf{argmin}_{w,\{\varepsilon_i\}}\frac{\lambda}{2}\|w\|^2+\frac{1}{m}\sum_i\ell^{hinge}\left(y_i\left\langle w,x_i\right\rangle\right)$$

נשים לב כי התנאי $1-\varepsilon_i\wedge\varepsilon_i\geq 0$ הוא בדיוק התנאי בפונקציה ℓ^{hinge}. אם $0\geq\varepsilon_i$ ו ומתקיים $y_i\left\langle w,x_i\right\rangle\geq 1-\varepsilon_i$ אזי $\ell^{hinge}\left(y_i\left\langle w,x_i\right\rangle\right)=1-y_i\left\langle w,x_i\right\rangle=\varepsilon_i$ ואחרת יהיה 0.

3. א. הניחו $x\in\mathbb{R}^d$. בהינתן trainset $S=\{(x_i,y_i)\}_{i=1}^m$, התאימו מסווג Gaussian Naive Bayes הפותר את (5) תחת הנחות (6).

נרצה למצוא θ הממקסמת את הנראות.

$$\begin{aligned}L(\theta\mid X,y)&=f_{X,y|\theta}\left(\{(x_i,y_i)\}_{i=1}^m\right)\overset{iid}{=}\prod_{i=1}^m f_{X,y|\theta}(X_i,y_i)\\&=\prod_{i=1}^m f(y_i\mid\theta)\cdot f(x_i\mid y=y_i)=\prod_{i=1}^m\pi_{y_i}\cdot N\left(X_i\mid\mu_{y_i},\sigma_{y_i}^2\right)\\&=\prod_{i=1}^m\pi_{y_i}\cdot\frac{1}{\sqrt{2\pi\left(\sigma_{y_i}\right)^2}}\cdot\exp\left(-\frac{\left(X_i-\mu_{y_i}\right)^2}{2\left(\sigma_{y_i}\right)^2}\right)\end{aligned}$$

נפעיל log על הביטוי כדי לפשט אותו.

$$\begin{aligned}&=\sum_{i=1}^m log\left(\pi_{y_i}\right)+log\left(\frac{1}{\sqrt{2\pi\sigma_{y_i}^2}}\right)-\frac{\left(X_i-\mu_{y_i}\right)^2}{2\sigma_{y_i}^2}\\&=\sum_{i=1}^m log\left(\pi_{y_i}\right)-\frac{1}{2}log\left(2\pi\sigma_{y_i}^2\right)-\frac{\left(X_i-\mu_{y_i}\right)^2}{2\sigma_{y_i}^2}\\&=\sum_{i=1}^m log\left(\pi_{y_i}\right)-\frac{1}{2}log\left(2\pi\right)-\frac{1}{2}log\left(\sigma_{y_i}^2\right)-\frac{\left(X_i-\mu_{y_i}\right)^2}{2\sigma_{y_i}^2}\end{aligned}$$

נרצה לחשב את θ הממקסמת את הנראות, לכן

$$\underset{\theta}{\operatorname{argmax}} \ell(\theta, S) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{m} log\left(\pi_{y_i}\right) - \frac{1}{2} log\left(\sigma_{y_i}^2\right) - \frac{\left(X_i - \mu_{y_i}\right)^2}{2\sigma_{y_i}^2}$$
$$= \underset{\theta}{\operatorname{argmax}} \sum_{k} \left(n_k \cdot log\left(\pi_k\right) - \frac{1}{2} n_k \cdot log\left(\sigma_k^2\right) - \sum_{i:y_i=k} \frac{\left(X_i - \mu_k\right)^2}{2\sigma_k^2} \right)$$

נגזור את הביטוי וכל פעם עבור משתנה אחר: $\{\pi_k\}, \{\mu_k\}, \{\sigma_k\}$

π_k : נקבל מהגזירה אותו דבר כמו החישוב בתרגול (כי יש את אותו ביטוי עם π_k) לכן $\hat{\pi}_k = \frac{n_k}{m}$.

μ_k :

$$\frac{\partial \ell}{\partial \mu_k} = \sum_{i:y_i=k} \frac{-2\left(X_i - \mu_k\right)}{2\sigma_k^2} = 0$$
$$\sum_{i:y_i=k} -2\left(X_i - \mu_k\right) = 0$$
$$\sum_{i:y_i=k} X_i - n_k \cdot \mu_k = 0$$
$$\mu_k = \frac{1}{n_k} \cdot \sum_{i:y_i=k} X_i$$

σ_k^2 :

$$\frac{\partial \ell}{\partial \mu_n} = -\frac{1}{2}\frac{n_k}{\sigma_k^2} - \sum_{i:y_i=k} \frac{\left(X_i - \mu_k\right)^2}{2} \cdot \left(-\frac{1}{\sigma_k^4}\right) = 0$$
$$\frac{n_k}{\sigma_k^2} = \sum_{i:y_i=k} \left(X_i - \mu_k\right)^2 \cdot \frac{1}{\sigma_k^4}$$
$$n_k = \sum_{i:y_i=k} \left(X_i - \mu_k\right)^2 \cdot \frac{1}{\sigma_k^2}$$
$$\hat{\sigma}_k^2 = \sum_{i:y_i=k} \frac{\left(X_i - \mu_k\right)^2}{n_k}$$

ב. הניחו $x \in \mathbb{R}^d$ (לכל דגימה יש d פיצ'רים). בהניתן trainset $S = \{(X_i, y_i)\}_{i=1}^m$, התאימו מסווג Gaussian Naive Bayes הפותר את (5) תחת הנחות (6).

נשנה את ההנחה ל- $X \in \mathbb{R}^d$. כלומר, כעת כל דגימה היא בעלת d פיצ'רים, כאשר כל פיצ'ר X_{ij} מתפלג $X_{ij} \mid y_i = k \sim N\left(\mu_{n_j}, \sigma_{k_j}^2\right)$. בנוסף $y \sim Mult\left(\pi\right)$, $\sum \pi_i = 1$, $\pi = \left(\pi_1, .., \pi_n\right)$.

נכתוב שוב את פונקציית הנראות עבור $\theta = \left(\{\pi_k\}, \{\mu_k\}, \left\{\sigma_{k_j}^2\right\}\right)$.

$$\begin{aligned}\ell\left(\theta \mid \{X_i, y_i\}\right) &= f_{X,y|\theta}\left(\{X_i, y_i\}_{i=1}^m\right) \overset{iid}{=} \prod_{i=1}^m f_{X,y|\theta}\left(X_i, , y_i\right) \\ &= \prod_{i=1}^m \prod_{j=1}^d f_{X_i,y|\theta}\left(X_{ij}, y_i\right) = \prod_{i=1}^m \prod_{j=1}^d f_y\left(y_i, \theta\right) \cdot f\left(X_{ij} \mid y = y_i\right) \\ &= \prod_{i\in[m], j\in[d]} f_y\left(y_i \mid \theta\right) \cdot f\left(X_{ij} \mid y = y_i\right)\end{aligned}$$

נגדיר $S' = \{(X_{ij}, y_i) : i \in [m], j \in [d]\}$

$: \pi_k$

אם נסמן ב־n_k את כמות הדגימות בסט S המקורי ששייכות למחלקה k אז בקבוצה S' קיימות $n_k \cdot d$ דגימות כאלה, שכן כל פיצ'ר של כל אחת מהדגימות הנ"ל קיבל כעת את התווית y_i.

לכן ראינו $\hat{\pi}_k = \frac{|\{y_i = k\}|}{\text{num of samples}}$ לכן עבור הסט S' נקבל $\hat{\pi}_k = \frac{n_k \cdot d}{m \cdot d} = \frac{n_k}{m}$.

$: \mu_{k_j}$

$$\hat{\mu}_{k,j} = \frac{1}{n_{k,j}} \sum_{i:y_i=k} X_{ij} = \frac{1}{n_k} \sum_{i:y_i=k} X_{ij}$$

$: \sigma_{k_j}^2$

$$\sigma_{k_j}^2 = \frac{1}{n_{k,j}} \cdot \sum_{i:y_i=k} \left(X_{ij} - \mu_{k,j}\right)^2 = \frac{1}{n_k} \cdot \sum_{i:y_i=k} \left(X_{ij} - \mu_{k_j}\right)^2$$

4. א. הניחו $x \in \mathbb{R}^d$. בהינתן trainset $S = \{(x_i, y_i)\}_{i=1}^m$, התאימו מסווג Gaussian Naive Bayes הפותר את (5) תחת הנחות (7).

נרצה למצוא θ הממקסמת את הנראות.

$$\begin{aligned}L\left(\theta \mid X, y\right) &= f_{X,y|\theta}\left(\{(x_i, y_i)\}_{i=1}^m\right) \overset{iid}{=} \prod_{i=1}^m f_{X,y|\theta}\left(X_i, y_i\right) \\ &= \prod_{i=1}^m f\left(y_i \mid \theta\right) \cdot f\left(x_i \mid y = y_i\right) = \prod_{i=1}^m \pi_{y_i} \cdot Poi\left(X_i \mid \lambda_{k,i}\right) \\ &= \prod_{i=1}^m \pi_{y_i} \cdot \frac{e^{-\lambda_{k,i}} \lambda_{k,i}^{y_i}}{y_i!}\end{aligned}$$

נפעיל log כדי לפשט את הביטוי

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

נרצה לחשב את θ הממקסמת את הנראות, לכן

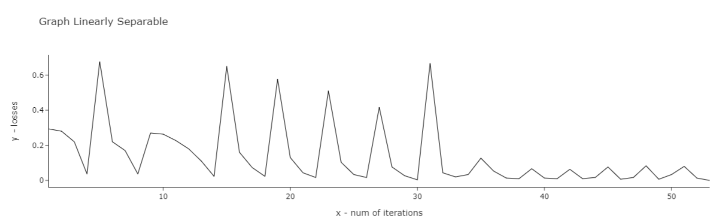
$$\underset{\theta}{\operatorname{argmax}}\ell\left(\theta,S\right)=\underset{\theta}{\operatorname{argmax}}\sum_{i=1}^{m}log\left(\pi_{y_i}\right)-\lambda_{k,i}+y_i\cdot log\left(\lambda_{k,i}\right)-log\left(y_i!\right)$$
$$=\underset{\theta}{\operatorname{argmax}}\sum_{k=1}^{d}n_k\cdot log\left(\pi_k\right)+\sum_{i:y_i=k}-\lambda_{k,i}+y_i\cdot log\left(\lambda_{k,i}\right)-log\left(y_i!\right)$$

$\underline{\pi_k :}$ נקבל מהגזירה אותו דבר כמו החישוב בתרגול (כי יש את אותו ביטוי עם π_k) לכן $\hat{\pi}_k=\frac{n_k}{m}$.
$\underline{\lambda_{k,i} :}$

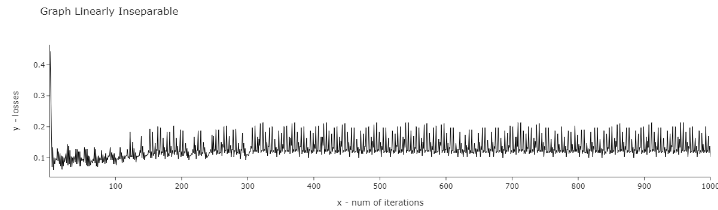
$$\frac{\partial\ell}{\partial\lambda_{k,i}}=\sum_{i:y_i=k}-1+y_i\cdot\frac{1}{\lambda_{k,i}}=0$$
$$-n_k+n_k\cdot y_i\cdot\frac{1}{\lambda_{k,i}}=0$$
$$\lambda_{k,i}=y_i$$

חלק פרקטי

Perceptron Classifier

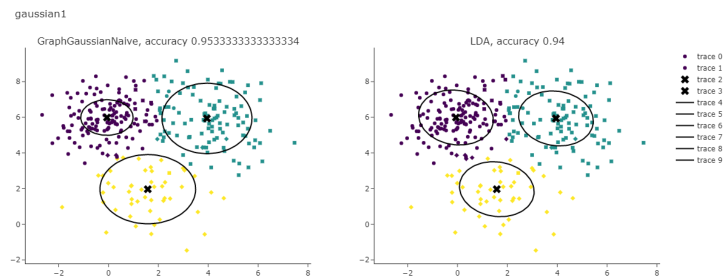


אנו יכולים להסיק מהגרף כי עבור המקרה בו הנתונים יכולים להיחלק לינארית, האלגוריתם יכול להגיע לתוצאה טובה (להפריד את הנתונים בצורה מושלמת) אחרי מספר איטרציות הקטן ממספר האיטרציות המקסימלי.

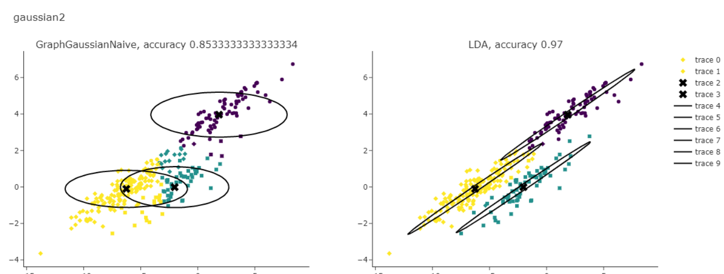


אנו יכולים להסביר את ההבדל בין שני הגרפים בכך שהגרף הזה עובד על נתונים שאינם יכולים להיפרד ע"י חלוקה לינארית. לכן האלגוריתם perceptron אינו יכול להגיע לתשובה טובה אף פעם. לכן הוא פעל עד מספר האיטרציות הסופי - שבמקרה שלנו 1000.

Bayes Classifiers



אנו יכולים להסיק שעבור הדאטא הראשון, לשני המסווגים תוצאות דומות.



ניתן להסיק מהגרף הזה כי לdataset הנוכחי כדאי להשתמש בLDA משום שהוא סיווג יותר טוב, עם $accuracy = 0.97$.

ניתן להבין מכך שהsampels בדאטא היו תלויים לינארית. לכן החיזוי של Gaussian Naive היה פחות טוב כי הוא עובד עם מידע בלתי תלוי לינארית.