מבוא למערכות לומדות - תרגיל 3

גיא קורנבליט, ת.ז 308224948

אומד בייס אופטימלי ו-LDA

יהיו מרחב מדגם $\mathcal{X} = \mathbb{R}^d$ ו-$\mathcal{Y} = \{\pm 1\}$. יהא מדגם S , נניח כי הדגימות ב-S נדגמו באופן בלתי תלוי מההתפלגות המשותפת מעל $\mathcal{X} \times \mathcal{Y}$.

שאלה 1

טענה: נניח שההתפלגות $\mathcal{D}$ ידועה, יהי $h_{\mathcal{D}}$ אומד בייס אופטימלי מוגדר על ידי

$$\forall x \in \mathcal{X} \quad h_{\mathcal{D}}(x) = \begin{cases} 1 & \mathbb{P}(y=1|x) \geq \frac{1}{2} \\ -1 & o.w \end{cases}$$

אזי מתקיים

$$h_{\mathcal{D}} = \underset{y \in \{\pm 1\}}{\operatorname{argmax}} \mathbb{P}(\boldsymbol{x}|y)\,\mathbb{P}(y)$$

הוכחה: נבחין כי הפונקציה מחזירה 1 אם"ם

$$\mathbb{P}(y=-1|\boldsymbol{x}) < \frac{1}{2} \iff \mathbb{P}(y=1|x) \geq \mathbb{P}(y=-1|x)$$

לפיכך, נוכל להתייחס ל-h כפונקציה של y, כך שהערך שמחזירה h הוא הערך שממקסם את ההסתברות, כלומר

$$\begin{aligned} \forall \boldsymbol{x} \in \mathcal{X} \quad h_{\mathcal{D}}(x) &= \underset{y \in \{\pm 1\}}{\operatorname{argmax}} \{\mathbb{P}(Y=y|\boldsymbol{x})\} \\ &\underset{\text{Bayes rule}}{=} \underset{y \in \{\pm 1\}}{\operatorname{argmax}} \{\mathbb{P}(\boldsymbol{x}|Y=y)\,\mathbb{P}(Y=y)\} \end{aligned}$$

כנדרש. ■

שאלה 2

נניח כי לכל $\boldsymbol{x} \in \mathcal{X} = \mathbb{R}^d$ מתקיים $\boldsymbol{x}|y \sim \mathcal{N}(\mu_y, \Sigma)$, כאשר $\mu_y \in \mathbb{R}^d$ ו-$\Sigma \in \mathbb{R}^{d\times d}$. נניח Σ ידועה, ו-μ_y ידוע לכל $y \in \{\pm 1\}$, אזי

$$h_{\mathcal{D}}(\boldsymbol{x}) = \underset{y\in\{\pm1\}}{\text{argmax}}\delta_y(\boldsymbol{x})$$

כאשר $\delta_y = \boldsymbol{x}^\top \Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^\top \Sigma^{-1}\mu_y + \ln \mathbb{P}(y)$.

הוכחה: בטענה הקודמת הראינו כי אומד בייס אופטימלי הינו מהצורה

$$\forall \boldsymbol{x} \in \mathcal{X} \quad h_{\mathcal{D}}(x) = \underset{y\in\{\pm1\}}{\text{argmax}}\left\{\mathbb{P}_{\mathcal{D}}(\boldsymbol{x}|y)\,\mathbb{P}_Y(y)\right\}$$

כאשר הביטוי $\mathbb{P}_{\mathcal{D}}(\boldsymbol{x}|y)$ מתייחס להתפלגות הנקודתית של הוקטור $\boldsymbol{x} \in \mathbb{R}^d$ והביטוי $\mathbb{P}_Y(y)$ מתייחס להתפלגות השולית של המשתנה Y. כעת, מפני שההתפלגות המשותפת $\mathcal{D} = \mathcal{N}$ הינה רציפה, ההתפלגות הנקודתית שקולה לפונקציית הצפיפות של המשתנה המקרי X. בנוסף, ממונוטוניות הלוגריתם, מתקיים לכל $y \in \{\pm 1\}$

$$\begin{aligned} f_{\mathcal{D}}(\boldsymbol{x}|y)\,\mathbb{P}_Y(y) &= \ln\left(f_{\mathcal{D}}(\boldsymbol{x}|y)\,\mathbb{P}_Y(y)\right) \\ &= \ln\left(f_{\mathcal{D}}(\boldsymbol{x}|y)\right) + \ln\left(\mathbb{P}_Y(y)\right) \end{aligned}$$

נבחין כי $\ln\left(f_{\mathcal{D}}(\boldsymbol{x}|y)\right)$ היא בדיוק לוג פונקציית הנראות ומתקיים

$$\begin{aligned} \ln\left(f_{\mathcal{D}}(\boldsymbol{x}|y)\right) &= \ln\left(\frac{1}{\sqrt{(2\pi)^d det(\Sigma)}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}-\mu_y\right)^\top \Sigma^{-1}\left(\boldsymbol{x}-\mu_y\right)\right\}\right) \\ &= \underbrace{\ln\left(\frac{1}{\sqrt{(2\pi)^d det(\Sigma)}}\right)}_{:=c} + -\frac{1}{2}\left(\boldsymbol{x}-\mu_y\right)^\top \Sigma^{-1}\left(\boldsymbol{x}-\mu_y\right) \\ &= c - \frac{1}{2}\left(\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x} - \boldsymbol{x}^T\Sigma^{-1}\mu_y - \mu_y^T\Sigma^{-1}\boldsymbol{x} + \mu_y^T\Sigma^{-1}\mu_y\right) \\ &= c - \frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x} + \boldsymbol{x}^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y \end{aligned}$$

נבחין כי הביטוי $c = \frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}$ לא תלוי ב-y ולכן לא רלוונטי למציאת y הממקסם את ערך הפונקציה. בסה"כ נקבל

$$\begin{aligned} h_{\mathcal{D}}(\boldsymbol{x}) &= \underset{y\in\{\pm1\}}{\operatorname{argmax}} \left\{ f_{\mathcal{D}}(\boldsymbol{x}|y)\,\mathbb{P}_Y(y) \right\} \\ &= \underset{y\in\{\pm1\}}{\operatorname{argmax}} \left\{ \ln\left(f_{\mathcal{D}}(\boldsymbol{x}|y)\,\mathbb{P}_Y(y) \right) \right\} \\ &= \underset{y\in\{\pm1\}}{\operatorname{argmax}} \left\{ \underbrace{\boldsymbol{x}^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + \ln\mathbb{P}(y)}_{\delta_y(\boldsymbol{x})} \right\} \end{aligned}$$

כנדרש. ■

שאלה 3

כדי להמיר את אומד בייס, במקרה הריאלי בו ההתפלגות אינה ידועה, נדרש לאמוד את μ_{+1}, μ_{-1} ו-Σ. בהינתן מדגם S מגודל m, נסמן $S_- = \{(\boldsymbol{x}, -1) \in S\}$ ו-$S_+ = \{(\boldsymbol{x}, 1) \in S\}$, וניעזר באומד בלתי מוטה שראינו עבור התוחלת, כלומר $\hat{\mu}_1 = \frac{1}{|S_+|}\sum_{\boldsymbol{x}\in S_+}\boldsymbol{x}$ ו-$\hat{\mu}_{-1} = \frac{1}{|S_+|}\sum_{\boldsymbol{x}\in S_+}\boldsymbol{x}$. נאמוד את $\mathbb{P}(Y)$ באמצעות פונקציית השכיחות.

כדי לאמוד את מטריצת השונות המשותפת Σ של המשתנה המקרי $X|Y$, נשתמש באומדים שהראינו למטריצת השונות המשותפת בהינתן כל ערך של y, ונשתמש בממוצע משוקלל (Pooled Covariance). כלומר

$$\hat{\Sigma} = \frac{1}{m-2}\sum_{y=\pm1}\sum_{i:y_i=y}(x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^T$$

שאלה 4

במקרה של מסווג דואר זבל, יתכנו הטעויות הבאות:

(1) סיווג דואר זבל כדואר רגיל.

(2) סיווג דואר רגיל כזבל - זו הטעות בעלת נזק פוטנציאלי גדול יותר מהטעות השניה, ולכן נעדיף להימנע ממנה לחלוטין.

נסווג דואר רגיל בתווית -1 (שלילי) ודואר זבל בתווית 1 (חיובי), כך שטעות false-positive תהיה כאשר $y_i = -1$ אבל המסווג יחזיר $\hat{y}_i = 1$.

שאלה 5

נתבונן בבעיית Hard-SVM שפיתחנו בתרגול,

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

נבחין כי נוכל לקודד את הסטיה $b \in \mathbb{R}$ בקוארדינטה ה-$n+1$ ל-w, כאשר נידרש להוסיף פיצ'ר intercept לכל דגימה $\boldsymbol{x}_i$. נגדיר

$$w' = \begin{bmatrix} b \\ w \end{bmatrix} \in \mathbb{R}^{n+1},\ \boldsymbol{x}_i' = \begin{bmatrix} 1 \\ \boldsymbol{x}_i \end{bmatrix},\ Q = 2 \begin{bmatrix} 0 & \\ & I_n \end{bmatrix}$$

$$\vec{a} = \vec{0} \in \mathbb{R}^{n+1},\ d = -\vec{1}_m$$

ואת המטריצה $A \in \mathbb{R}^{m \times n+1}$ כך ששורותיה מוגדרות לכל $i \in [m]$ כך ש- $A_i = -y_i \left(\boldsymbol{x}_i'\right)^T$. נראה כי הבעיה שקולה לבעיית QP מהצורה הקנונית:

$$\underset{w \in \mathbb{R}^{n+1}}{\text{argmin}} \left(\tfrac{1}{2} w'^\top Q w' + a^\top w'\right)$$
$$\text{s.t. } Aw' \leq d$$

(הסימון $w \in \mathbb{R}^{n+1}$ הינו עבור $w = w'$ שהגדרנו, לטובת נוחות הסימון). ראשית, נטען כי פונקציית המטרה זהה -

$$\frac{1}{2} w'^\top Q w' + a^\top w' = w'^\top \begin{bmatrix} 0 & \\ & I_n \end{bmatrix} w' = b \cdot 0 \cdot b + w^\top I_n w$$
$$= \langle w, w \rangle = \|w\|^2$$

כמו כן, נראה כי האילוצים שקולים -

$$Aw' \leq d \iff \forall i \in [m] \quad A_i w' \leq d_i \iff -y_i \left(\boldsymbol{x}_i'\right)^T w' \leq -1$$

$$\iff y_i \begin{bmatrix} 1 \\ \boldsymbol{x}_i \end{bmatrix}^T \begin{bmatrix} b \\ w \end{bmatrix} \geq 1 \iff y_i \left(b + \boldsymbol{x}_i^\top w\right) \geq 1$$

$$\iff y_i \left(w^\top \boldsymbol{x}_i + b\right) \geq 1 \iff y_i \left(\langle w, \boldsymbol{x}_i \rangle + b\right) \geq 1$$

כלומר, האילוצים ופונקציית המטרה שקולים ולכן הבעיות שקולות. כנדרש. ■

שאלה 6

נתבונן בבעיית Soft-SVM בנוסח הבא

$$\arg\min_{\mathbf{w},\{\xi_i\}} \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} \xi_i \text{ s.t. } \forall_i, y_i \langle w, \boldsymbol{x}_i \rangle \geqslant 1 - \xi_i \text{ and } \xi_i \geqslant 0$$

ונבחין כי לכל $i \in [m]$ מתקיים

$$y_i \langle w, \boldsymbol{x}_i \rangle \geqslant 1 - \xi_i \text{ and } \xi_i \geqslant 0 \iff \xi_i \geq 1 - y_i \langle w, \boldsymbol{x}_i \rangle \text{ and } \xi_i \geqslant 0$$

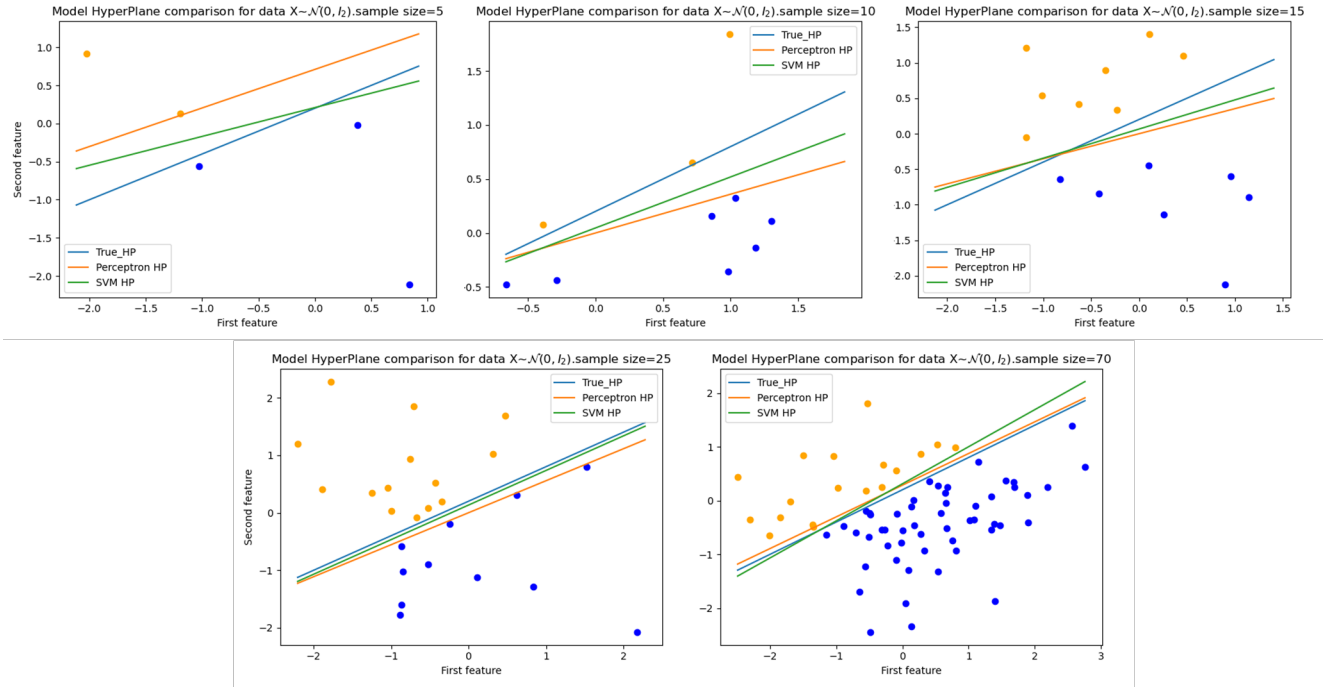
כלומר, במידה והדגימה נמצאת מעבר ל-margin של מחלקת התיוג ו-$y_i \langle w, \boldsymbol{x}_i \rangle < 1$, הערך שממזער את ξ_i ביחס לפונקציית המטרה הינו $1 - y_i \langle w, \boldsymbol{x}_i \rangle$. אחרת, אם אין הפרה של הדגימה את ה-margin אז נוכל להימנע מהעלות של החריגה, והערך שימזער את ξ_i ויעמוד בתנאים (כי $1 - y_i \langle w, \boldsymbol{x}_i \rangle < 0$) הינו $\xi_i = 0$. לפיכך, הבחירה של ξ_i שתמזער את פונקציית המטרה ותעמוד באילוצים הינה

$$\ell^{\text{hinge}} 1 - y_i \langle w, \boldsymbol{x}_i \rangle = \xi_i = \max\{0, 1 - y_i \langle w, \boldsymbol{x}_i \rangle\}$$

ולכן נוכל לנסח את הבעיה באופן שקול

$$\arg\min_{\mathbf{w}} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \ell^{hinge}\left(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle\right)$$

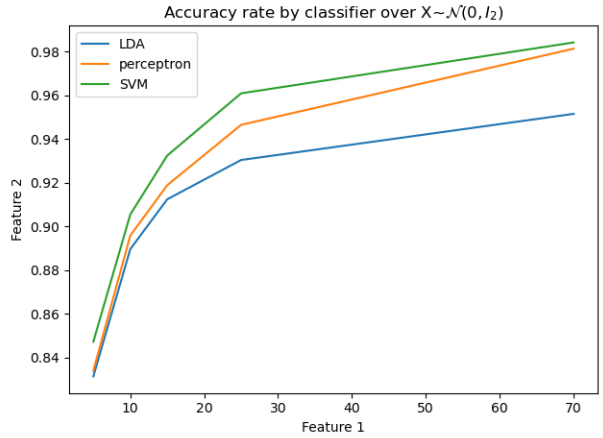
שאלה 9



איור 0.1: השוואת היפר-מישורים של כל מודל, עבור train-set בגודל שונה

בשאלה זו, כאשר ההתפלגות ופונקציית התיוג ידועה, נבדקו המודלים ביחס ל-ground truth, נעיר כי SVM ממומש עם קבוע רגולריזציה גדול מאוד, ולכן המסווג שקול למסווג Hard-SVM שמצליח במשימתו מפני שבמקרה זה אנו במקרה הרלייזבילי. ראשית, נבחין כי עם עליית גודל המדגם (סט האימון) המסווגים הולכים ומתקרבים ל-ground-truth, לכן שגיאת סיווג-לא-נכון תתקרב ואף תתאפס, בסט האימון, ועבור נקודות חדשות.

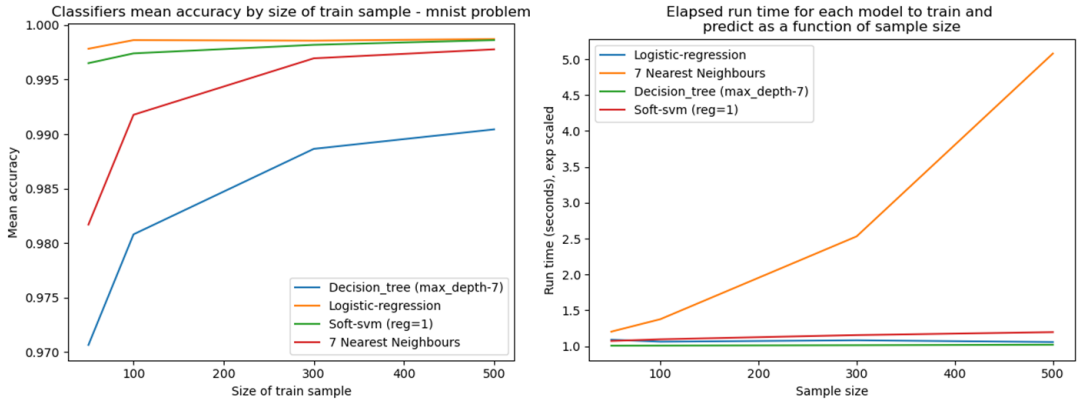
שאלה 11



איור 0.2: Mean accuracy rate for each model on the Normal-bivariate distribtion

בגרף זה מתואר מדד הדיוק ($\frac{\text{True positives+True negative}}{\text{Positive + Negative}}$) של כל מסווג, מול גודל המדגם עליו התאמן המסווג, כאשר הדאטה בצורה סינטטית מההתפלגות הדו-נורמלית למשתנים בלתי מתואמים עם תוחלת 0 ושונות 1. ניכר, כי SVM שולט בתוצאות מדד הדיוק לכל גודל מדגם. להבנתי, ההבדל טמון בפונקציית התיוג האמיתית, שהוגדרה להיות היפר המישור $f(\boldsymbol{x}) = \text{sign}\left\langle \begin{pmatrix} 0.3 \\ -0.5 \end{pmatrix}, \boldsymbol{x} \right\rangle + 0.1$. כלומר, הדאטה ניתן להפרדה ליניארית, ולכן סביר להניח שאלגוריתמים עם מחלקת היפותזות של מפרידים ליניארים ינצלו את המבנה של הבעיה כדי ללמוד קירוב טוב יותר של ההיפותזה. אבחנה זו מסבירה את הפער המתמשך בין הביצועים של מודל LDA ההסתברותי, לבין הביצועים של SVM ו-Perceptron. בהתייחס לשני האחרונים, ניכר כי ההבדל המרכזי באופן פעולתם הוא אופן הלמידה. מודל SVM מתחשב בשתי המחלקות כדי למצוא את היפר-המישור (כלומר, השוליים של היפר-המישור רחוקים במידה שווה משני הוקטורים התומכים), כאשר ה-Perceptron מחפש את כל וקטור שעומד בכל התנאים, ואיטרטיבית לומד על כל נקודה בסט האימון בנפרד, וכך מתחשב בכל צעד בפחות אילוצים. כלומר, perceptron פועל בתנאי רלקסציה ביחס ל-SVM, ולכן סביר שאם הדאטה ניתן להפרדה ליניארית אז Hard-SVM יצליח להשיג ציון טוב גבוה יותר.

שאלה 14



איור 0.3: Mean accuracy as a function of sample size for each algorithm

גרף זה מציג את ההבדל במדד הדיוק (accuracy) בין המסווגים השונים, כפונקציה של גודל המדגם עליו התאמנו בבעיית סיווג תמונות של הספרות 0 ו-1 למחלקה הנכונה. כמו כן, מתואר גרף של זמני הריצה של אימון מודל לכל גודל מדגם, וחיזוי על test-set קבוע, לכל מודל. נבחין במגמות הבאות:

(1) מודל KNN משיג זמני ריצה גרועים ביחס לשאר המודלים. זמן הריצה של המודל נשלט על ידי זמן החיזוי של דגימה מה-test, מפני שהאלגוריתם מחשב את המרחקים של כל דגימה ב-train (ששמר בשלב האימון), ממיין ובוחר את השכנים הקרובים ביותר. נשים לב כי מדד הדיוק של האלגוריתם משיג תוצאות קרובות למודלים הטובים ביותר ככל שסט האימון גדל (במחיר זמן ריצה משמעותי).

(2) רגרסיה לוגיסטית ו-Soft-svm יחסית בביצועים גם במדד הדיוק וגם בזמני הריצה. להבנתי, זמני הריצה הנמוכים נובעים על רקע פתרונות יעילים שיש לבעיות אופטימיזציה קמורה (למדנו על מימושים ספציפיים עבור רגרסיה, ותכנון ריבועי עבור SVM).

(3) עץ ההחלטה משיג זמני ריצה מהירים מאוד, אך במדד הדיוק המודל משיג את התוצאות הנמוכות ביותר לכל גודל מדגם (באופן יחסי, עדיין מתייצב מעל 90 אחוזי דיוק ככל שגודל סט האימון עולה). יצוין, כי גם עבור מודל עם עומק עץ גדול יותר (15 במקום 7) השיפור לא שינה מגמה זו. סביר להניח, כמו עבור 7-NN, כי הביצועים הנמוכים עבור סט אימון קטן יחסית נובעים מ-under-fit ביחס לכמות החלוקות שהעץ מבצע.