מבוא למערכות לומדות
תרגיל 4

רן שחם - 203781000 - ransha

2 ביוני 2017

1 שקילות של הגדרות Soft-SVM

נראה את שקילות ההגדרות:

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

$$(2)\min_{\mathbf{w},\{\xi_i\}}\frac{\lambda}{2}\left\|\mathbf{w}\right\|^{2}+\frac{1}{m}\sum_{i=1}^{m}\xi_{i}\text{ such that }\forall i,y_{i}\left\langle \mathbf{w},\mathbf{x}_{i}\right\rangle \geq1-\xi_{i}\text{ and }\xi_{i}\geq0$$

נניח ש־$\mathbf{w}$ הוא פתרון ל־(1). נזכור כי $\ell^{hinge}\left(y_{i}\left\langle \mathbf{w},\mathbf{x}_{i}\right\rangle \right)=\max\left\{ 0,1-y_{i}\left\langle \mathbf{w},\mathbf{x}_{i}\right\rangle \right\}$. יהיו $\{\xi_i\}$ כך שלכל i מתקיים $y_{i}\left\langle \mathbf{w},\mathbf{x}_{i}\right\rangle \geq1-\xi_{i}$ וגם $\xi_i\geq0$, כלומר $\mathbf{w},\{\xi_i\}$ הוא פתרון כלשהו ל־(2). אם כך:

$$\begin{aligned} & y_{i}\left\langle \mathbf{w},\mathbf{x}_{i}\right\rangle \geq1-\xi_{i}\\ \iff & \xi_{i}\geq1-y_{i}\left\langle \mathbf{w},\mathbf{x}_{i}\right\rangle \end{aligned}$$

כמוכן, $\xi_i\geq0$ לכן: $\xi_{i}\geq\max\left\{ 0,1-y_{i}\left\langle \mathbf{w},\mathbf{x}_{i}\right\rangle \right\} =\ell^{hinge}\left(y_{i}\left\langle \mathbf{w},\mathbf{x}_{i}\right\rangle \right)$. זה נכון לכל i וכל הביטויים הנדונים הם אי־שליליים. לכן:

$$\frac{1}{m}\sum_{i=1}^{m}\xi_{i}\geq\frac{1}{m}\sum_{i=1}^{m}\ell^{hinge}\left(y_{i}\left\langle \mathbf{w},\mathbf{x}_{i}\right\rangle \right)$$

אם כך, $\{\xi_i\}$ הטובים ביותר שאפשר לבחור (ולפיכך מהווים פתרון אופטימלי ל־(2)) הם $\xi_{i}'=\ell^{hinge}\left(y_{i}\left\langle \mathbf{w},\mathbf{x}_{i}\right\rangle \right)$, כלומר (1) ו־(2) נותנים את אותו פתרון.

2 גרעין חוקי (Valid Kernel)

נגדיר מיפוי $\psi:\left\{ M,\ldots,N\right\} \to\left\{ 0,1\right\} ^{N}$ באופן הבא:

$$x\mapsto(\overbrace{11\cdots1}^{x}\overbrace{0\cdots0}^{N-x})$$

לכל $x\in\left\{ M,\ldots,N\right\}$. נראה שלכל $x,x'\in\left\{ M,\ldots,N\right\}$ מתקיים $K\left(x,x'\right)=\left\langle \psi\left(x\right),\psi\left(x'\right)\right\rangle$. יהיו x,x' כנ"ל. נניח בה"כ ש־$x\leq x'$. לכן $M\leq\min\left\{ x,x'\right\} =x\leq x'\leq N$. מתקיים:

$$\begin{aligned} & \psi\left(x\right)=(\overbrace{11\cdots1}^{x}\overbrace{0\cdots0}^{N-x})\\ \implies & \forall i\in\left[N\right]\ \left(\psi\left(x\right)\right)_{i}\overset{*}{=}\begin{cases} 1, & i\leq x\left(\leq x'\right)\\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

ובאופן דומה:

$$\begin{aligned}
\left(\psi\left(x'\right)\right)_i &\overset{**}{=} \begin{cases} 1, & i \le x' \\ 0, & \text{otherwise} \end{cases} \\
\implies \left\langle \psi\left(x\right), \psi\left(x'\right) \right\rangle &\overset{1}{=} \sum_{i=1}^{N} \left(\psi\left(x\right)\right)_i \cdot \left(\psi\left(x'\right)\right)_i \\
&\overset{2}{=} \sum_{i=1}^{x} \overbrace{1 \cdot 1}^{i \le x \le x'} + \sum_{i=x+1}^{x'} \overbrace{1 \cdot 0}^{x < i \le x'} + \sum_{i=x'+1}^{N} \overbrace{0 \cdot 0}^{i > x'} \\
&= x \cdot 1 + \left(x' - x\right) \cdot 0 + \left(N - x'\right) \cdot 0 = x \\
&= \min\left\{x, x'\right\} = K\left(x, x'\right)
\end{aligned}$$

כאשר:

1. לפי הגדרת המכפלה הפנימית
2. נובע מ־$(*)$ ו־$(**)$. נשים לב שהסכום האמצעי יכול להיות ריק (אם $x = x'$)

לכן ψ מקיימת: $K\left(x, x'\right) = \left\langle \psi\left(x\right), \psi\left(x'\right) \right\rangle$ לכל $x, x' \in \{M, \dots, N\}$; כלומר K היא פונקציית kernel חוקית.

3 בחירת מודל

3.1

לכל היפותזה $h \in \mathcal{H}_k$ מתקיים:

$$L_{\mathcal{D}}\left(h\right) = \underset{S_{all} \sim \mathcal{D}^m}{\mathbb{E}} \left[L_{S_{all}}\left(h\right)\right]$$

לכן מאי־שוויון הופדינג מתקיים שלכל $\delta' \in (0,1)$:

$$\begin{aligned}
&\mathbb{P}\left[\left|L_{S_{all}}\left(h\right) - L_{\mathcal{D}}\left(h\right)\right| \le \sqrt{\frac{\ln\left(2/\delta'\right)}{2m}}\right] \ge 1 - \delta' \\
\implies &\mathbb{P}\left[\left|L_{S_{all}}\left(h\right) - L_{\mathcal{D}}\left(h\right)\right| \ge \sqrt{\frac{\ln\left(2/\delta'\right)}{2m}}\right] \le \delta' \\
\overset{\delta' = \frac{\delta}{|\mathcal{H}_k|}}{\implies} &\mathbb{P}\left[\left|L_{S_{all}}\left(h\right) - L_{\mathcal{D}}\left(h\right)\right| \ge \sqrt{\frac{\ln\left(2 \cdot |\mathcal{H}_k| / \delta\right)}{2m}}\right] \le \frac{\delta}{|\mathcal{H}_k|}
\end{aligned}$$

אם כך, מחסם האיחוד מתקיים:

$$\mathbb{P}\left[\exists h \in \mathcal{H}_k : \left|L_{S_{all}}\left(h\right) - L_{\mathcal{D}}\left(h\right)\right| \ge \sqrt{\frac{\ln\left(2|\mathcal{H}_k|/\delta\right)}{2m}}\right] \le |\mathcal{H}_k| \frac{\delta}{|\mathcal{H}_k|} = \delta$$

נקבל שאם $h^* \in \mathrm{ERM}_{\mathcal{H}_k}\left(S_{all}\right)$, בסיכוי לפחות $1 - \delta$ מתקיים שלכל $h \in \mathcal{H}_k$:

$$\begin{aligned}
L_{\mathcal{D}}\left(h^*\right) &\le L_{S_{all}}\left(h^*\right) + \sqrt{\frac{\ln\left(2|\mathcal{H}_k|/\delta\right)}{2m}} \le L_{S_{all}}\left(h\right) + \sqrt{\frac{\ln\left(2|\mathcal{H}_k|/\delta\right)}{2m}} \\
&\le L_{\mathcal{D}}\left(h\right) + \sqrt{\frac{\ln\left(2|\mathcal{H}_k|/\delta\right)}{2m}} + \sqrt{\frac{\ln\left(2|\mathcal{H}_k|/\delta\right)}{2m}} = L_{\mathcal{D}}\left(h\right) + 2\sqrt{\frac{\ln\left(2|\mathcal{H}_k|/\delta\right)}{2m}} \\
&= L_{\mathcal{D}}\left(h\right) + \sqrt{\frac{2\ln\left(2|\mathcal{H}_k|/\delta\right)}{m}}
\end{aligned}$$

ובפרט, הנ"ל מתקיים גם עבור $h \in \arg\min_{h \in \mathcal{H}_k} L_{\mathcal{D}}(h)$. לכן:

$$L_{\mathcal{D}}(h^*) \leq \min_{h \in \mathcal{H}_k} L_{\mathcal{D}}(h) + \sqrt{\frac{2\ln\left(2|\mathcal{H}_k|/\delta\right)}{m}}$$

3.2

נניח ש־$\arg\min_{h \in \mathcal{H}_k} L_{\mathcal{D}}(h) \notin \mathcal{H}_{j-1}$ ‏$\ni \mathcal{H}_j$ (כלומר, ההיפותזה הממזערת את שגיאת ההכללה "מגיעה" מהמחלקה $\mathcal{H}_j$). נשים לב ש־$|S| = (1-\alpha)m, |V| = \alpha m, |\mathcal{H}| = k$. מהסעיף הקודם, עבור $h^* \in \mathrm{ERM}_{\mathcal{H}}(V)$ מתקיים בסיכוי לפחות $1 - \frac{\delta}{2}$:

$$(1)\ : L_{\mathcal{D}}(h^*) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \sqrt{\frac{2}{\alpha m}\ln\left(\frac{4k}{\delta}\right)}$$

ומכיוון ש־$h_i \in \mathrm{ERM}_{\mathcal{H}_i}(S)$, בסיכוי לפחות $1 - \frac{\delta}{2}$ מתקיים גם:

$$(2)\ : L_{\mathcal{D}}(h_i) \leq \min_{h \in \mathcal{H}_i} L_{\mathcal{D}}(h) + \sqrt{\frac{2}{(1-\alpha)m}\ln\left(\frac{4|\mathcal{H}_i|}{\delta}\right)}$$

לכל $i \in [k]$. לכן, בסיכוי לפחות $1 - \delta$ מתקיים:

$$\begin{aligned}
L_{\mathcal{D}}(h^*) &\overset{(1)}{\leq} \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \sqrt{\frac{2}{\alpha m}\ln\left(\frac{4k}{\delta}\right)} \leq L_{\mathcal{D}}(h_j) + \sqrt{\frac{2}{\alpha m}\ln\left(\frac{4k}{\delta}\right)} \\
&\overset{(2)}{\leq} \min_{h \in \mathcal{H}_j} L_{\mathcal{D}}(h) + \sqrt{\frac{2}{\alpha m}\ln\left(\frac{4k}{\delta}\right)} + \sqrt{\frac{2}{(1-\alpha)m}\ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)} \\
&= \min_{h \in \mathcal{H}_k} L_{\mathcal{D}}(h) + \sqrt{\frac{2}{\alpha m}\ln\left(\frac{4k}{\delta}\right)} + \sqrt{\frac{2}{(1-\alpha)m}\ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)}
\end{aligned}$$

3.3

$$\begin{aligned}
\epsilon_{est}^{MS} &= L_{\mathcal{D}}(h^*) - \min_{h \in \mathcal{H}_k} L_{\mathcal{D}}(h) \overset{2}{=} \sqrt{\frac{2}{\alpha m}\ln\left(\frac{4k}{\delta}\right)} + \sqrt{\frac{2}{(1-\alpha)m}\ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)} \\
\epsilon_{est}^{S} &\overset{1}{=} \sqrt{\frac{2\ln\left(2|\mathcal{H}_k|/\delta\right)}{m}} \\
\implies \frac{\epsilon_{est}^{MS}}{\epsilon_{est}^{S}} &= \frac{\sqrt{\frac{2}{\alpha m}\ln\left(\frac{4k}{\delta}\right)} + \sqrt{\frac{2}{(1-\alpha)m}\ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)}}{\sqrt{\frac{2\ln\left(2|\mathcal{H}_k|/\delta\right)}{m}}} = \sqrt{\frac{\ln\left(\frac{4k}{\delta}\right)}{\alpha\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}} + \sqrt{\frac{\ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)}{(1-\alpha)\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}}
\end{aligned}$$

כאשר השוויון המסומן ב־1 נובע מסעיף 1 וכנ"ל לגבי 2. שאר המעברים הם אלגבריים. נשים לב ש־$\mathcal{H}_j \subset \mathcal{H}_k$ לכן שני הביטויים שהתקבלו (בשורש) חסומים, כלומר שגיאת ה־estimation בתהליך ה־MS לא יכולה להיות הרבה יותר גרועה מזו בשיטה הרגילה. אפשר גם לראות שאם הופכים את היחס $\left(\frac{\epsilon_{est}^{S}}{\epsilon_{est}^{MS}}\right)$ מקבלים ביטוי שגדל באופן פרופורציוני ליחס בין הגודל של $\mathcal{H}_k$ לגודל של $\mathcal{H}_j$. כלומר, ככל שניקח $\mathcal{H}_k$ גדולה מ־$\mathcal{H}_j$, נקבל ביטוי גדול יותר ל־$\frac{\epsilon^{S}}{\epsilon^{MS}}$.

באופן מפורש, אם ניקח את $j = k$, כלומר ההיפותזה ה"טובה ביותר" מגיעה מ־$\mathcal{H}_k$, נקבל שהחסם בסעיף 1 הדוק יותר מבסעיף 2 ־ כלומר שיטה זו "טובה יותר".

בכיוון השני, ניקח את המחלקה $\mathcal{H}_i$ להיות בגודל $2^{2^{t\cdot i}}$ לכל $i\in[k]$ וקבוע $t\geq 1$. נבחן את החסמים שהתקבלו בסעיפים הקודמים:

$$(S):L_{\mathcal{D}}\left(h^{*}\right)\leq\min_{h\in\mathcal{H}_{k}}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2\ln\left(2\cdot2^{2^{tk}}/\delta\right)}{m}}=\min_{h\in\mathcal{H}_{k}}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2\ln\left(2^{1+2^{tk}}/\delta\right)}{m}}$$

$$(MS):L_{D}\left(h^{*}\right)\leq\min_{h\in\mathcal{H}_{k}}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2}{\alpha m}\ln\left(\frac{4k}{\delta}\right)}+\sqrt{\frac{2}{\left(1-\alpha\right)m}\ln\left(\frac{2^{2+2^{tj}}}{\delta}\right)}$$

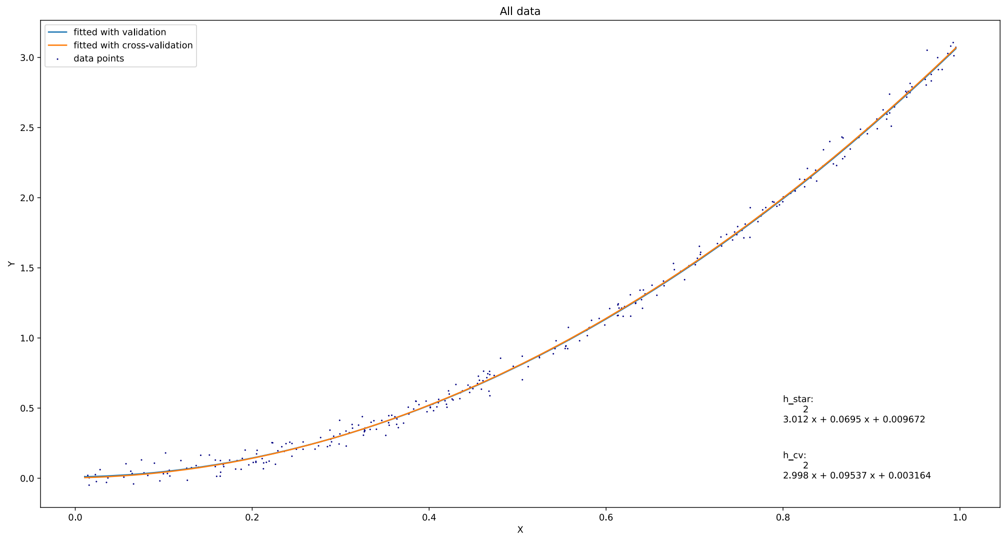
$$\Longrightarrow\frac{\epsilon_{est}^{S}}{\epsilon_{est}^{MS}}=\frac{\sqrt{\frac{2\ln\left(2^{1+2^{tk}}/\delta\right)}{m}}}{\sqrt{\frac{2}{\alpha m}\ln\left(\frac{4k}{\delta}\right)}+\sqrt{\frac{2}{(1-\alpha)m}\ln\left(\frac{2^{2+2^{tj}}}{\delta}\right)}}\xrightarrow[t\to\infty]{}\infty$$

כלומר במקרה זה MS טוב בהרבה מהשיטה הרגילה.

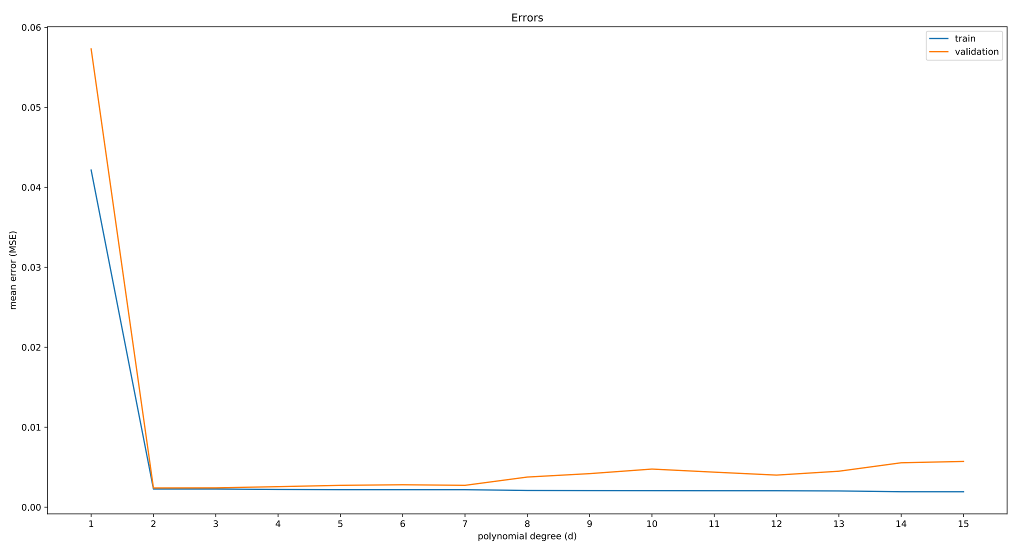
4 חלק תכנותי - Validation

הערה 4.1 נא להריץ את הקובץ `ex4.py` באמצעות פייתון 3 (נבדק ועובד באקווריום)

- שגיאת ההיפותזה h^* שהתקבלה על נתוני ה־test בתהליך הולידציה היא: ≈ 0.00226
- שגיאת ההיפותזה h_{cv} שהתקבלה על נתוני ה־test בתהליך הקרוס־ולידציה היא: ≈ 0.00225 (כמעט זהה)
- תהליך ה־5-fold cross-validation החזיר פולינום מדרגה 2 (דרגה זהה לדרגה שהתקבלה בולידציה רגילה) עם מקדמים דומים מאוד, כפי שאפשר לראות באיור 1



איור 1: אפשר לראות שהפולינום h^* שהותאם בתהליך ה־validation כמעט זהה ל־h_{cv} שהותאם ב־cross-validation



איור 2: ככל שדרגת הפולינום גדלה, שגיאת האימון קטנה אך שגיאת הולדיציה עולה קצת (בגלל overfitting). עבור פולינום מדרגה 1 ברור שיש underfitting ולכן גם שגיאת האימון וגם הולידציה גדולות.