

מבוא למערכות לומדות

תרגיל 4

2 ביוני 2017

1 שקילות של הגדרות Soft-SVM

נראה את שקילות ההגדרות:

$$(1) \min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^m \ell^{hinge}(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle)$$
$$(2) \min_{\mathbf{w}, \{\xi_i\}} \frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^m \xi_i \text{ such that } \forall i, y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

נניח ש- \mathbf{w} הוא פתרון ל-(1). נזכור כי $\ell^{hinge}(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) = \max\{0, 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle\}$. יהיו $\{\xi_i\}$ כך שלכל i מתקיים $y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1 - \xi_i$ וגם $\xi_i \geq 0$, כלומר $\{\xi_i\}$ הוא פתרון כלשהו ל-(2). אם כך:

$$y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1 - \xi_i$$
$$\iff \xi_i \geq 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle$$

כמוכן, $\xi_i \geq 0$: לכן: $\xi_i \geq \max\{0, 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle\} = \ell^{hinge}(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle)$. זה נכון לכל i וכל הביטויים הנדונים הם אי-שליים. לכן:

$$\frac{1}{m} \sum_{i=1}^m \xi_i \geq \frac{1}{m} \sum_{i=1}^m \ell^{hinge}(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle)$$

אם כך, $\{\xi_i\}$ הטובים ביותר שאפשר לבחור (ולפיכך מהווים פתרון אופטימלי ל-(2)) הם $\xi'_i = \ell^{hinge}(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle)$, כלומר (1) ו-(2) נותנים את אותו פתרון.

2 גרעין חוקי (Valid Kernel)

נגדיר מיפוי $\psi : \{M, \dots, N\} \rightarrow \{0, 1\}^N$ באופן הבא:

$$x \mapsto (\overbrace{11 \dots 1}^x \overbrace{0 \dots 0}^{N-x})$$

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

$$\psi(x) = (\overbrace{11 \dots 1}^x \overbrace{0 \dots 0}^{N-x})$$
$$\implies \forall i \in [N] (\psi(x))_i = \begin{cases} 1, & i \leq x (\leq x') \\ 0, & \text{otherwise} \end{cases}$$

$$\begin{aligned}
(\psi(x'))_i &\stackrel{**}{=} \begin{cases} 1, & i \leq x' \\ 0, & \text{otherwise} \end{cases} \\
\Rightarrow \langle \psi(x), \psi(x') \rangle &\stackrel{1}{=} \sum_{i=1}^N (\psi(x))_i \cdot (\psi(x'))_i \\
&\stackrel{2}{=} \sum_{i=1}^x \widehat{1 \cdot 1} + \sum_{i=x+1}^{x'} \widehat{1 \cdot 0} + \sum_{i=x'+1}^N \widehat{0 \cdot 0} \\
&= x \cdot 1 + (x' - x) \cdot 0 + (N - x') \cdot 0 = x \\
&= \min\{x, x'\} = K(x, x')
\end{aligned}$$

כאשר:

1. לפי הגדרת המכפלה הפנימית

2. נובע מ- $(*)$ ו- $(**)$. נשים לב שהסכום האמצעי יכול להיות ריק (אם $x = x'$)לכן ψ מקיימת: $K(x, x') = \langle \psi(x), \psi(x') \rangle$ לכל $x, x' \in \{1, \dots, N\}$; כלומר K היא פונקציית kernel חוקית.

3 בחירת מודל

3.1

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

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

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

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

אם כן, מחסם האיחוד מתקיים:

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

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

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

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

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

3.2

נניח ש- $\mathcal{H}_j \ni \arg \min_{h \in \mathcal{H}_k} L_{\mathcal{D}}(h) \notin \mathcal{H}_{j-1}$. נשים לב ש- $|S| = (1 - \alpha)m$, $|V| = \alpha m$, $|\mathcal{H}| = k$:
 $h^* \in \text{ERM}_{\mathcal{H}}(V)$ מתקיים בסיכוי לפחות $1 - \frac{\delta}{2}$:

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

ומכיון ש- $h_i \in \text{ERM}_{\mathcal{H}_i}(S)$, 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:

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

3.3

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

כאשר השוויון המסומן ב-1 נובע מסעיף 1 וכן"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- $\frac{\epsilon_{est}^S}{\epsilon_{est}^{MS}}$.

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 $2^{2^{t-i}}$ לכל $i \in [k]$ וקבוע $t \geq 1$. נבחן את החסמים שהתקבלו בסעיפים הקודמים:

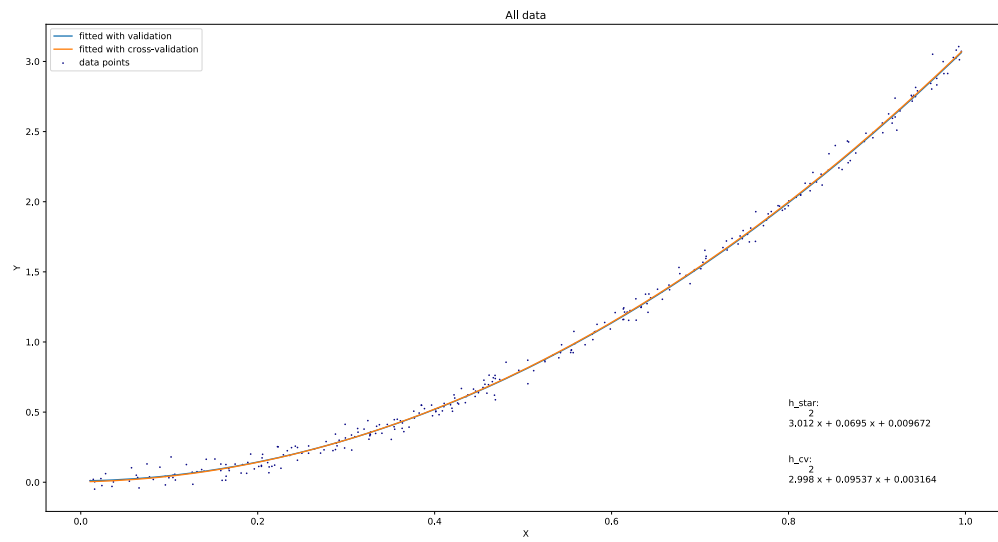
$$\begin{aligned} (S) : L_{\mathcal{D}}(h^*) &\leq \min_{h \in \mathcal{H}_k} L_{\mathcal{D}}(h) + \sqrt{\frac{2 \ln(2 \cdot 2^{2^k} / \delta)}{m}} = \min_{h \in \mathcal{H}_k} L_{\mathcal{D}}(h) + \sqrt{\frac{2 \ln(2^{1+2^k} / \delta)}{m}} \\ (MS) : L_{\mathcal{D}}(h^*) &\leq \min_{h \in \mathcal{H}_k} L_{\mathcal{D}}(h) + \sqrt{\frac{2}{\alpha m} \ln\left(\frac{4k}{\delta}\right)} + \sqrt{\frac{2}{(1-\alpha)m} \ln\left(\frac{2^{2+2^k} / \delta}{\delta}\right)} \\ \Rightarrow \frac{\epsilon_{est}^S}{\epsilon_{est}^{MS}} &= \frac{\sqrt{\frac{2 \ln(2^{1+2^k} / \delta)}{m}}}{\sqrt{\frac{2}{\alpha m} \ln\left(\frac{4k}{\delta}\right)} + \sqrt{\frac{2}{(1-\alpha)m} \ln\left(\frac{2^{2+2^k} / \delta}{\delta}\right)}} \xrightarrow{t \rightarrow \infty} \infty \end{aligned}$$

כלומר במקרה זה MS טוב בהרבה מהשיטה הרגילה.

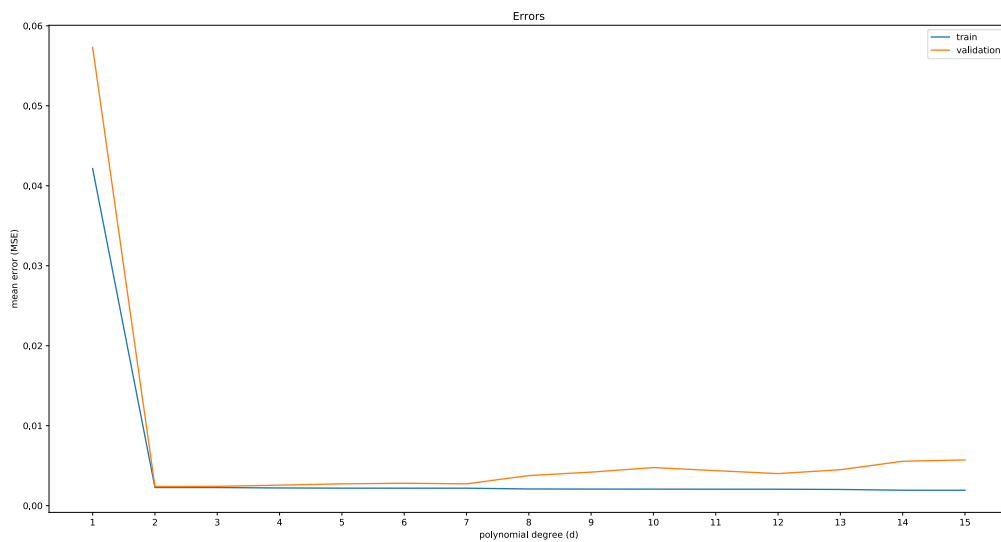
4 חלק תכנותי - Validation

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 הולדציה עולה קצת (בגלל overfitting). עבור פולינום מדרגה 1 ברור שיש underfitting ולכן גם שגיאת האימון וגם הולדציה גדולות.