תרגול 3

14/4/20 - ג'

אחת Supervised learning, נדבר על רגרסיה לינארית.

רעיון המידה:

הגדרת הבעיה
- Domain Set: $\mathcal{X}\subseteq\mathbb{R}^d$
- Response / Label Set: $\mathcal{Y}\subseteq\mathbb{R}$
- Rule/ hypothesis: $f:\mathcal{X}\to\mathcal{Y}$
- Hypothesis class: $\mathcal{H}\subseteq\mathcal{Y}^{\mathcal{X}}$

הגדרת הפתרון
- Training / Sample Set: $S=\{(\overset{\in\mathcal{X}}{x_1},\overset{\in\mathcal{Y}}{y_1}),\dots,(x_m,y_m)\}$
- Prediction rule: $\hat{f}\in\mathcal{H}$
- Loss function: $L(\hat{f},(x_i,y_i))\in\mathbb{R}$
- Empirical risk: $L_S(\hat{f})=\frac{1}{m}\sum_{i=1}^{m}L(\hat{f},(x_i,y_i))$
- The ERM rule: $\underset{f\in\mathcal{H}}{\arg\min}\ L_S(f)$

רגרסיה לינארית

בבעיות רגרסיה לינארית, מניחים שה-labels מוגדרות באופן לינארי,
כלומר עבור (x_i,y_i), $x_i\in\mathbb{R}^d$ ולכן נסמנו: $\begin{pmatrix}x_{i,1}\\ \vdots\\ x_{i,d}\end{pmatrix}$, אז מתקיים:
$\boxed{y_i=\sum_{j=1}^{d}w_j x_{i,j}=\langle w,x_i\rangle}$, עבור w_j משקולות.
בעית הלמידה היא מציאת $w\in\mathbb{R}^d$.

פונק' כזו היא פונק' הומוגנית. בעולם הלא הומוגני נוסיף עוד תנאי:
bias $\to b\in\mathbb{R}$, $\boxed{y_i=\langle w,x_i\rangle+b}$, זה מאפשרים הסטה מראשית הצירים.

- נבחין ש-b בעית "זהה": נגדיר: $x_i'=\begin{pmatrix}1\\ \top\\ x_i\\ \bot\end{pmatrix}$, $w'=\begin{pmatrix}b\\ \top\\ w\\ \bot\end{pmatrix}$. אז $y_i=\langle w',x_i'\rangle$
היא תרגום של הבעיה הלא-הומוגנית כבעיה הומוגנית, במחיר של הוספת מימד.
- ההיפותזה הלינארית: נקראת f_w, והיא ממירה בעיה לא-הומוגנית לבעיה הומוגנית.
כלומר: $f_{w'}(x)=\langle w',\begin{pmatrix}1\\ \top\\ x\\ \bot\end{pmatrix}\rangle \doteq y$, $w'\in\mathbb{R}^{d+1}$, $x\in\mathbb{R}^d$, $y\in\mathbb{R}$
- מחלקת ההיפותזות של הרגרסיה הלינארית:

$$\mathcal{H}_{lin}=\{f_w \mid w\in\mathbb{R}^{d+1}\}=\{x\mapsto\langle w',\begin{pmatrix}1\\ \top\\ x\\ \bot\end{pmatrix}\rangle \mid w'\in\mathbb{R}^{d+1}\}$$

נתעלם מהמקרה הלא-הומוגני (הבנו שהוא ניתן בקלות לפשוט למקרה הומוגני).

- השגיאה: (The squared loss): ל-$w\in\mathbb{R}^d$, (x_i,y_i): $L(f_w,(x_i,y_i))=\ell_2=(y_i-\langle w,x_i\rangle)^2$.
- חוק ה-ERM: $\underset{w\in\mathbb{R}^d}{\arg\min}\ L_S(f_w)=\underset{w\in\mathbb{R}^d}{\arg\min}\ \frac{1}{m}\sum_{i=1}^{d}(y_i-\langle w,x_i\rangle)^2$

או בהצגה המטריציונית: $X = \begin{bmatrix} x_1^T & \cdots & x_m^T \end{bmatrix} \in \mathbb{R}^{d\times m}$, $y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^m$

מתקיים: $\underset{w\in\mathbb{R}^d}{\arg\min}\, L_S(f_w) = \underset{w\in\mathbb{R}^d}{\arg\min}\, \frac{1}{m}\|y - X^T w\|^2$

[מכיוון שהוקטור שימצא את ה-loss, ימצא גם בביטוי בלי $\frac{1}{m}$] $= \underset{w\in\mathbb{R}^d}{\arg\min}\, \|y - X^T w\|^2$

* לא בטוח שנמצא w עבורו y_i יהיה בדיוק ה-x_i שניתן לכל i.
נרצה w שיתן <u>קירוב</u> טוב יותר.

* כשיש יותר משוואות מפרמטרים (כלומר כאשר $X\in\mathbb{R}^{d\times m}$ ו- $d<m$)
לא נמצא פתרון (בסיכוי גבוה) או שנמצא לכל היותר פתרון אחד.
מצב זה מכונה המצב הסינגולרי

* נסתכל על $XX^T \in \mathbb{R}^{d\times d}$. מניחים ש-$d \geq m$, כלומר קיבלנו שיש d וקטורים בת"ל ב-XX^T, ולכן היא מטריצה מדרגה d בסיכוי גבוה.

מצב לא סינגולרי

→ • <u>אם XX^T הפיכה</u>: אז: נסמן: $\hat{w} = \arg\min_w \|y - X^T w\|^2$

$\hat{w}$ מתקיים: $\hat{w} = (XX^T)^{-1}Xy$, והפתרון יחיד.

<u>הוכחה</u>: נסמן $f(w) = y - X^T w$. אז:

$$\frac{df(w)}{dw} = Xy - XX^T w \xRightarrow{\text{נשווה ל-0}} Xy - XX^T w = 0 \Leftrightarrow$$

$$Xy = XX^T w \Leftrightarrow (XX^T)^{-1}Xy = w$$

ולכן $\hat{w} = (XX^T)^{-1}Xy$. ∎

☆ הוקטור $\hat{w}$ בעצם אומר לנו משהו על השפעת הפרמטרים שבו x שנפסל.
(עבור - שהתפוגן הוא צפוי bias)..

מצב סינגולרי

→ • <u>אם XX^T לא הפיכה</u>: נשתמש ב-pseudo-inverse של X:

עבור $D = \mathrm{diag}(a_{11}, \ldots, a_{dd})$ יהיה: $D^{\dagger} = \begin{cases} a_{ii} = 0: & 0 \\ a_{ii} \neq 0: & \frac{1}{a_{ii}} \end{cases}$.

ולכן עבור פרוק ה-SVD: $X = U\Sigma V^T$ מתקיים: $X^{\dagger} = V\Sigma^{\dagger}U^T$

משפט: $\hat{w} = X^{\dagger}y$ ו- מתקיים: $y = X^T w$, ולכן השגיאה: $\|y - X^T w\|^2 = 0$

* יש במצב כזה ∞ פתרונות אפשריים ל-$\hat{w}$, זה אחד מהפתרונות, ובו נוח.

רגרסיה פולינומיאלית

הכללה של בעיית הרגרסיה הלינארית.

$$P(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_k x^k = \sum_{j=0}^{k} a_j x^j = \langle a, \Psi(x) \rangle$$

כאשר:

$$a = \begin{pmatrix} a_0 \\ \vdots \\ a_k \end{pmatrix}, \quad \Psi(x) = \begin{pmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^k \end{pmatrix}$$

ולכן במקום להסתכל על מטריצה X, נרצה להסתכל על מטריצה Ψ_X.

$$\begin{bmatrix} | & & | \\ \Psi(x_1) & \dots & \Psi(x_m) \\ | & & | \end{bmatrix} \qquad \begin{bmatrix} | & & | \\ x_1 & \dots & x_m \\ | & & | \end{bmatrix}$$

ולמעשה הפכנו את הבעיה הפולינומיאלית לבעיה לינארית.

זהירות - צריך להיות עדינים: ניתן לקחת פולינום מדרגה מאוד גבוהה, שעובר בכל הנקודות, אבל לא ידע לחזות בהצלחה דגימות חדשות, כי עושה התאמת יתר לבעיה - overfitting. מנגד אם נקח פולינום מדרגה מאוד נמוכה, הוא כנראה לא יתאים, וה-loss יהיה גבוה.

The bias-variance Tradeoff:

בהינתן x ו-$\hat{f}$ שאנו נרצה למצוא את y^* (קבוע ב-R) - התווית האמיתית ל-y. נשערך, ונחזיר את: $\hat{f}(x) = \hat{y}$.

מעניינת אותנו השגיאה הממוצעת על הדגימות: $E[(y^* - \hat{y})^2]$ - תוחלת השגיאה לפי המשערך $\hat{f}$.

נגדיר מ"מ: $\bar{y} = E[\hat{y}]$ → קבוע, ממוצע המשערך על נסויים נפרדים.

נרצה ש- $E[\bar{y}] = y^*$, אך זה לא תמיד נכון.

$$E[(y^* - \hat{y})^2] = E[((y^* - \bar{y}) + (\bar{y} - \hat{y}))^2] = E[(y^* - \bar{y})^2] +$$

$$+ E[(\bar{y} - \hat{y})^2] + 2E[(y^* - \bar{y})(\bar{y} - \hat{y})] \overset{\text{נוציא קבועים}}{=}$$

$$= E[(y^* - \bar{y})^2] + E[(\bar{y} - \hat{y})^2] + 2(y^* - \bar{y})E[\bar{y} - \hat{y}] \overset{\text{לינאריות}}{=}$$

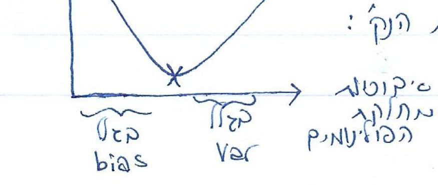
$$= \text{―\,"\,―} + \text{―\,"\,―} + 2(y^* - \bar{y})\underbrace{(E[\bar{y}] - E[\hat{y}])}_{=0} =$$

תוחלת של מספר ולכן $= \bar{y}$; מהגדרה $= \bar{y}$

$$= \underbrace{E[(y^* - \bar{y})^2]}_{bias^2} + \underbrace{E[(\bar{y} - \hat{y})^2]}_{Var(\hat{y})}$$

ההטיה של המדגם מהממוצע - $bias^2$ (למה אם שווים רק מצב ע"י טעויות)

$Var(\hat{y})$ (כמה המדגם שוני יחסית לממוצע, דגימות שונות)



אז: הדרגה למעלה נמוכה מייצרת bias גבוה.

דרגת פולינום גבוהה מייצרת שונות גבוהה.

אנחנו רוצים שהשגיאה תהיה קטנה, ולכן צריך למצוא את הנק': בדרך יעילה וטובה.

רגרסיה אקספוננציאלית

ננסה להתאים את הבעיה האקספו' לבעיה לינארית, הפעם ע״י שינוי y.

הבעיה מתנהגת כמו $y \approx e^{\langle w,x \rangle} = f(x)$ ו-w אשר נרצה ללמוד.

אם נקח $\log(y) \approx \langle w,x \rangle$ זה יראה לנו כבעיה לינארית!