תרגיל 5 - IML

מגיש: רועם כרולן

1. a) נגדיר $\varepsilon = \sqrt{\frac{1}{2m}\ln\left(\frac{2}{\delta}\right)}$, ומתוך שוויון ההאפדינג נקבל

$$\mathbb{P}\left(|L_{S_{all}}(h) - L_D(h)| \le \sqrt{\frac{\ln(2/\delta)}{2m}}\right) \ge 1-\delta$$

לכל סדרה δ הנתונה $h \in \mathcal{H}_k$. ולכן מתקיים

$$\underbrace{\mathbb{P}\left(|L_{S_{all}}(h_i) - L_D(h_i)| \ge \sqrt{\frac{\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{2m}}\right)}_{\text{מאחר שקבוצה ע"י הסכום המאיחוד מתקיים לכם}} \le \frac{\delta}{|\mathcal{H}_k|}$$

נסמן ב-T

$$T \le \frac{|\mathcal{H}_k|\,\delta}{|\mathcal{H}_k|} = \delta$$

כלומר

$$|L_{S_{all}}(h) - L_D(h)| \le \sqrt{\frac{\ln\left(\frac{2|H_k|}{\delta}\right)}{2m}}$$

לכל $h \in \mathcal{H}_k$.

$$\Rightarrow \mathbb{P}\left(L_D(h^*) \le \min_{h\in\mathcal{H}_k} L_D(h) + \sqrt{\frac{2\ln\left(\frac{2|H_k|}{\delta}\right)}{m}}\right) \ge 1-\delta$$

b) מאותו הקו נקבל שמתקיים

$$\mathbb{P}\left(L_D(h^*) \le \min_{h\in\mathcal{H}} L_D(h) + \sqrt{\frac{2\ln(4k/\delta)}{\alpha m}}\right) \ge 1-\frac{\delta}{2}$$

ובדומה סדרה δ שלקחנו הנתונה $\mathcal{H}_i$ כך ש $i\in[k]$

מתקיים

$$\mathbb{P}\left(L_D(h) \le \min_{h\in\mathcal{H}_i} L_D(h) + \sqrt{\frac{2\ln\left(\frac{4|H_i|}{\delta}\right)}{(1-\alpha)m}}\right) \ge 1-\frac{\delta}{2}$$

ולכן בהסתברות $1-\delta$ לפחות, ובסדרה δ $j\in[k]$ כך ש

$L_D(h) \in \mathcal{H}_j$ מתקיים

$$L_D(h^*) \le L_D(h_j) + \sqrt{\frac{2\ln(4k/\delta)}{\alpha m}} \le \min_{h\in\mathcal{H}_j} L_D(h) + \sqrt{\frac{2\ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)}{(1-\alpha)m}} + \sqrt{\frac{2\ln(4k/\delta)}{\alpha m}}$$

$$= \min_{h \in \mathcal{H}_k} L_b(h) + \sqrt{\frac{2\ln\left(4|\mathcal{H}_k|/\delta\right)}{(1-\alpha)m}} + \sqrt{\frac{2\ln(4k/\delta)}{\alpha m}}$$

ג)

אם נשתמש בהנחה $|\mathcal{H}_i| = 2^{2^n}$. נקבל את הסיבוכיות לבחירת

$$L_b(h^*) \le \min_{h \in \mathcal{H}_k} L_b(h) + \sqrt{\frac{2\ln\left(\frac{2^{2^n}}{\delta}\right)}{m}}$$

נסמן ב B

בשביל זה בחירת מודל לבחירת

$$L_b(h^*) \le \min_{h \in \mathcal{H}_k} L_b(h) + \sqrt{\frac{2\ln\left(\frac{2^{2^n}}{\delta}\right)}{(1-\alpha)m}} + \sqrt{\frac{2\ln\left(\frac{4k}{\delta}\right)}{\alpha m}}$$

נסמן ב V

כאשר אם נשתמש ב B ב V לבחירת מסווג יש

$$\lim_{n \to \infty} \frac{B}{V} = \infty$$

ובעצם לא כדאי לשנות בשביל בחירת מודל אלא
לנוח בתיבה בחירת תיבה נשימוש של סיבוכיות
בגדלים הקטנים הסבירים.

2. a) הרינו בכתבה את הפתרון לשגיאת ridge שהיא

$$\hat{w}_\lambda^{ridge} = (X X^T + \lambda I)^{-1} X y$$

במקרה של orthogonal design נקבל

$$\hat{w}_\lambda^{ridge} = Xy(I + \lambda I)^{-1} = Xy((1+\lambda)I)^{-1} = \frac{\hat{w}^{LS}}{1+\lambda}$$

b) במקרה של orthogonal design הרינו כי $\hat{w}^{LS} = X^T y$.

אני מבדיל לעיר $v \in \mathbb{R}^d$ משמעות שה Loss שלו לא הוסר

הוא:

$$\|y - Xv\|^2 = \sum_{i=1}^{n} (\hat{w}_i^{LS} - v_i)^2$$

לאחר מכן הוא מחפש אלו מקסמל את הפונקציה הבאה:

$$\underset{v \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} (\hat{w}_i^{LS} - v_i)^2 + \lambda \|v\|_0$$

נבחין בין שני מקרים:

- אם $(\hat{w}_i^{LS})^2 \le \lambda$ אזי $|\hat{w}_i^{LS}| \le \sqrt{\lambda}$ ולכן $v_i = 0$
- אם $(\hat{w}_i^{LS})^2 > \lambda$ אז $v_i = \hat{w}_i^{LS}$

לכן מקבלים מתמטית הנוסחה ל $\hat{w}_\lambda^{subset} = \eta^{hard}_{\sqrt{\lambda}}(\hat{w}^{LS})$

3. a)

$$A_\lambda \hat{w} = (X^T X + \lambda I)^{-1} (X^T X)(X^T X)^{-1} X^T y$$

$$= (X^T X + \lambda I) X^T y = \hat{w}(\lambda)$$

b)

$$E[\hat{w}(\lambda)] = A_\lambda E[\hat{w}] = (X^T X + \lambda I)^{-1} (X^T X) w$$

לכן לכל $\lambda > 0$ מתקיים $E[\hat{w}(\lambda)] \neq w$

c)

$$Var(\hat{w}(\lambda)) = A_\lambda Var(\hat{w}) A_\lambda^T = \sigma^2 A_\lambda (X^T X)^{-1} A_\lambda^T$$

d) אפשר לכתוב את הביטויים הבאים:

- $bias^2(\lambda) = \|(A_\lambda - I)w\|^2 = w^T (A_\lambda - I)^T (A_\lambda - I) w$
- $Var(\lambda) = \sigma^2 \, Tr(A_\lambda (X^T X)^{-1} A_\lambda^T)$

$MSE(\lambda) = bias^2(\lambda) + Var(\lambda)$

≥ 7 נק' הבנה התקדם עקרון

כעת נחפש את שני הביטויים לפי λ:

$$\frac{d}{d\lambda} bias^2(\lambda)\Big|_{\lambda=0} = \frac{d}{d\lambda}\left(\sum_i \left(\sum_j (A_\lambda - I)_{ij} w_j\right)^2\right) = 0$$

$$\frac{d}{d\lambda} Var(\lambda)\Big|_{\lambda=0} = \sum_{ij} \frac{\partial Var(\lambda)}{\partial (X^T X + \lambda I)_{ij}} \cdot \frac{d(X^T X + \lambda I)_{ij}}{d\lambda} =$$

$$Tr\left(\frac{d Var(\lambda)}{d(X^T X + \lambda I)} \cdot \frac{d(X^T X + \lambda I)}{d\lambda}\right) =$$

$$\overset{*}{=} -2\sigma^2 Tr\left((X^T X)^{-1}(X^T X)^{-1}\right) < 0$$

כאשר השתמשנו ב* נגזרת מרכזית ב

$$\frac{d}{d\lambda} Var(\lambda) = -2\sigma^2\left((X^T X)^{-1}(X^T X)^{-1}\right)$$

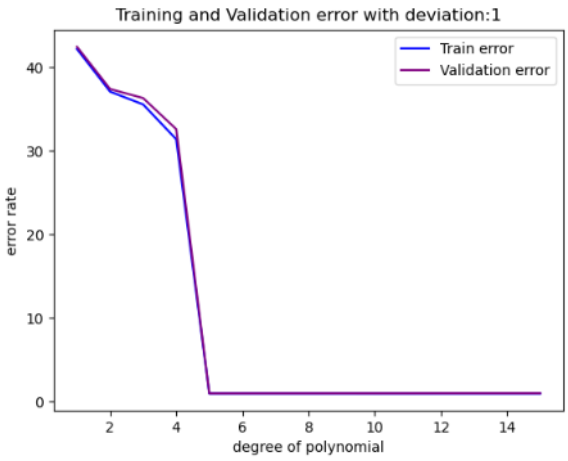
לכן נוכל לבטא את הנגזרת עם ל $\frac{d}{d\lambda} MSE(\lambda)\Big|_{\lambda=0} < 0$

e) נמצאנו בדרך הביטוי שקיים $\lambda > 0$ שמקטינה את הטעות הריבועית אלו MSE ולכן קיים גם כזה עם $MSE(\lambda) < MSE(0)$.

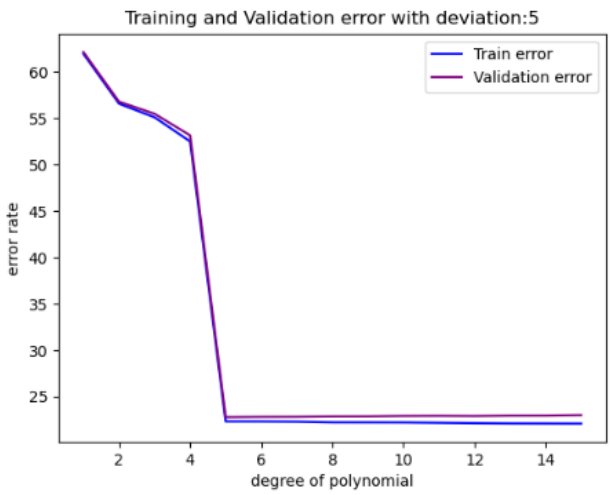
4.

e)

The degree of the polynomial with the lowest validation error for deviation 1 is 5



The degree of the polynomial with the lowest validation error for deviation 5 is 5



g)

Both polynomials (h*) that performed the best by the ERM principal were of degree 5, and the following is their test error:

test error for deviation 1: 0.5244495414916898

test error for deviation 5: 13.764926054734158

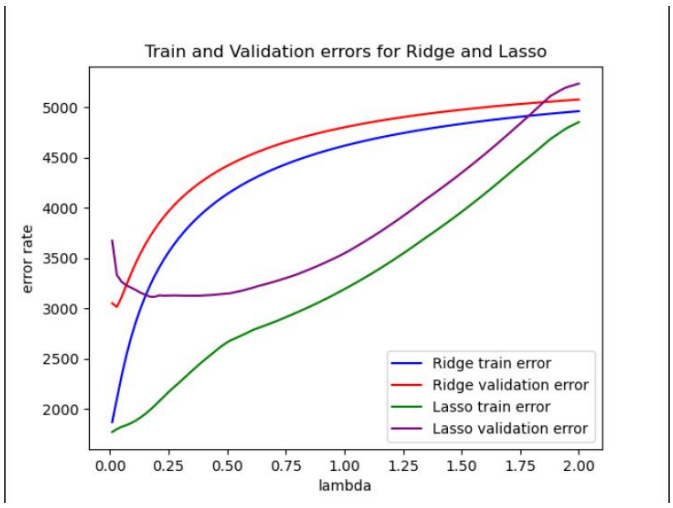
The validation error in both cases was similar to the test error

5.

cii)

The values of lambda that I chose to check were between 0.01(close to zero) and 2. I started with such a low number as to see if there is a difference between using a very small regularization term that is close to zero, which is very close to regular Linear regression as we have learned. By checking enough examples of lambda between these two numbers, we can clearly tell by the graph that there is a ceil to how much we can improve our classification if our regularization term(lambda) is higher.

d)



e)

Best lambda for Ridge is 0.03010101010101010101

Best lambda for Lasso is 0.19090909090909092

f)

Test error for best Ridge hypothesis 3216.0093244988275

Test error for best Lasso hypothesis 3392.0050650611543

Test error for best Linear Regression 3612.249688324898

g)

We can clearly tell from section f) that the regularization has indeed helped in getting better results, compared to regular Linear Regrression.