תרגיל 5 | IML (67329)

שם: רונאל חרדים | ת"ז:208917641

חלק I

$Theoretical$:

שאלה 1

(א)

נראה כי $\hat{\mathbf{w}}_\lambda = A_\lambda \hat{\mathbf{w}}$:

מתקיים השויון הבא $A_\lambda \hat{w} = A_\lambda \hat{w}(\lambda = 0)$ ואנו יודעים כי מהפתרון ללא רגולציה מתקיים

$$\hat{w}(\lambda = 0) \iff argmin_w \left(||y - Ww||_2^2\right) = \left(X^T X\right)^{-1} X^T y$$

ולכן:

$$A_\lambda \hat{w} = \left(X^T X + \lambda I_d\right)^{-1} \left(X^T X\right) \cdot \left(X^T X\right)^{-1} X^T y = \left(X^T X + \lambda I_d\right)^{-1} X^T y$$

ראינו בכיתה כי זה שווה ל$\hat{w}(\lambda)$ כנדרש.

(ב)

נראה כי $\lambda > 0 \Rightarrow \mathbb{E}[\hat{w}(\lambda)] \neq w$:

$$E[\hat{w}(\lambda)] \overset{(3.A)}{=} E\left[A_\lambda \hat{w}\right] = E\left[\left(X^T X + \lambda I_d\right)^{-1} \left(X^T X\right) \cdot \hat{w}\right]$$

מכיוון שהמטריצה X קבועה אנו יודעים כי $E\left[A_\lambda\right] = A_\lambda$. ולכן

$$= \left(X^T X + \lambda I_d\right)^{-1} \left(X^T X\right) E[\widehat{w}] \overset{E[\widehat{w}]=w}{=} \left(X^T X + \lambda I_d\right)^{-1} \left(X^T X\right) \cdot w$$

כעת, אם $\lambda = 0$ אזי $A_\lambda = I$, אך מכיוון ש $\lambda > 0$ מתקיים:

$$\left(X^T X + \lambda I_d\right)^{-1} \left(X^T X\right) \neq I_d, \text{ therefore } \left(X^T X + \lambda I_d\right)^{-1} \left(X^T X\right) w \neq w$$

כנדרש.

(ג)

נראה כי $\operatorname{Var}(\hat{w}(\lambda)) = \sigma^2 A_\lambda \left(X^T X\right)^{-1} A_\lambda^T$:
מהרמז אנו יודעים כי

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

כנדרש.

(ד)

(ה)

באופן דומה למה שראינו בכיתה, נניח כי y^* ההיפוזה האמיתית, $\bar{y}$הנצפה, ו $\hat{y}$האומדן. תחת ההגדרות הללו מתקיים:

$$E\left[||\hat{y} - y^*\|^2\right] = \operatorname{Var}[\hat{y}] + \operatorname{bias}^2[\hat{y}].$$

מכיוון ש

$$\operatorname{Var}[\hat{y}] = E\left[\|\hat{y} - \bar{y}\|^2\right] \ \wedge \ bias[\hat{y}] = \|\bar{y} - y^*\|^2.$$

ומתקיים כי $y^* = \hat{w}X$ ללא תנאי רגולריזציה. לכן $y^* = E\left[\hat{w}\right]$, בנוסף האסטימטור שלנו הוא $\hat{y} = \hat{w}(\lambda)$ ולכן גם $\bar{y} = E\left[\hat{w}(\lambda)\right]$.
נחשב:

$$\operatorname{Var}(\lambda) = \operatorname{Tr}\left(\operatorname{Var}(\hat{w}(\lambda)) = \sigma^2 \operatorname{Tr}\left(A_\lambda \left(X^T X\right)^{-1} A_\lambda^T\right)\right.$$

נגדיר $\left(X^T X + \lambda I_d\right) = X'$ ולכן $A_\lambda = X'^{-1}\left(X^T X\right)$, וגם

$$\operatorname{Var}(\lambda) = \sigma^2 \operatorname{Tr}\left(X'^{-1}\left(X^T X\right)\left(X^T X\right)^{-1}\left(X'^{-1}\left(X^T X\right)\right)^T\right) = \sigma^2 \operatorname{Tr}\left(X'^{-1}\left(X^T X\right) X'^{-1}\right)$$

כעת, נגדור את השונות לפי X':

$$\left.\frac{d \operatorname{Var}(\lambda)}{dX'}\right|_{\lambda=0} = \sigma^2 \operatorname{Tr}\left(\frac{d}{dX'}\left(X'^{-1}\left(X^T X\right) X'^{-1}\right)\right)_{\lambda=0} = -2\sigma^2 \left(X^T X\right)^{-1}\left(X^T X\right)^{-1}$$

ולפי λ:

$$\left.\frac{d \operatorname{Var}(\lambda)}{d\lambda}\right|_{\lambda=0} = \operatorname{Tr}\left(\frac{d \operatorname{Var}(\lambda)}{dX'} \frac{dX'}{d\lambda}\right) = -2\sigma^2 \operatorname{Tr}\left(\left(X^T X\right)^{-1}\left(X^T X\right)^{-1}\right)$$

הערך הוא אי שלילי.

עבור ה $bias$:

$$\begin{gathered}\operatorname{bias}(\lambda)=\|\bar{y}-y^*\|=\|E[\hat{w}(\lambda)]-E[\hat{w}]\|=\|E\left[A_\lambda w\right]-E[\hat{w}]\|=\|A_\lambda E[w]-w\| \\ =\|A_\lambda w-w\|=\|\left(A_\lambda-I\right)w\| \\ \operatorname{bias}^2(\lambda)=\|\left(A_\lambda-I\right)w\|^2=w^T\left(A_\lambda-I\right)^T\left(A_\lambda-I\right)w \\ \text{in terms of } X'\end{gathered}$$

עבור X':

$$bias^2(\lambda)=w^T\left(X'^{-1}\left(X^TX\right)-I\right)^T\left(X'^{-1}\left(X^TX\right)-I\right)w$$

נגזור :

$$\begin{gathered}\left.\frac{d\text{ bias }^2}{d\lambda}\right|_{\lambda=0}=\left.\frac{d}{d\lambda}\left(\sum_i\left(\sum_j X'_{ij}w_j\right)^2\right)\right|_{\lambda=0} \\ =2\sum_i\left(\sum_j X'_{ij}w_j\right)_{\lambda=0}\cdot\frac{d}{d\lambda}\left(\sum_j X_{ij}w_j\right)_{\lambda=0}=0\end{gathered}$$

השגיאה נתונה ע"י:

$$MSE=\text{ bias }^2+\text{ var }=w^T\left(A_\lambda-I\right)^T\left(A_\lambda-I\right)w+\sigma^2\operatorname{Tr}\left(A_\lambda\left(X^TX\right)^{-1}A_\lambda^T\right.$$

הנגזרת היא סכום הנגזרות וקטנה מ 0 . מסכיוון שהנגזרת של ה $bias^2=0$ והנגזרת של השונות שלילית.

(ו)

אנו יודעים כי מודש לינארי בלי רגולריזציה מתקיים כאשר נשווה $\lambda=0$. כעת ממה שראינו בסעיף קודם שעבור עברכים מסויימים מתקיים $MSE(\lambda)<0 \Rightarrow \lambda>0$ ניתן להסיק כי λ כזה מספק $MSE(\lambda)<MSE(0)$ כלומר הרגולריזציה מקטינה את הטעות.

שאלה 3

בהינתן קרנל $k(x,x')$ נגדיר את הקרנל המנורמל להיות עבור כל x ־ $\widetilde{k}(x,x)=1$.
נגדיר:

$$\tilde{k}\left(x,x'\right)=\frac{k\left(x,x'\right)}{\sqrt{k(x,x)\cdot k\left(x',x'\right)}}$$

ראישת נשים לב כי הקרנד המנורמל ולידי, והוא מייצג מ"פ $\psi/||\psi||$:

$$\tilde{k}\left(x,x'\right)=\frac{\left\langle \psi(x),\psi\left(x'\right)\right\rangle }{\sqrt{\left\langle \psi(x),\psi(x)\right\rangle \left\langle \psi\left(x'\right),\psi\left(x'\right)\right\rangle }}=\frac{\left\langle \psi(x),\psi\left(x'\right)\right\rangle }{\sqrt{||\psi(x)\|^{2}||\psi\left(x'\right)|^{2}}}=\frac{\left\langle \psi(x),\psi\left(x'\right)\right\rangle }{||\psi(x)\,\|\cdot||\psi\left(x'\right)\|}$$

$$=\left|\frac{\psi(x)}{||\psi(x)\|},\frac{\psi\left(x'\right)}{||\psi\left(x'\right)\|}\right|$$

ומתקיים

$$\tilde{k}(x,x)=\frac{\langle\psi(x),\psi(x)\rangle}{\|\psi(x)\|\cdot\|\psi(x)\|}=\frac{\|\psi(x)\|^{2}}{\|\psi(x)\|^{2}}=1$$

כנדרש.

שאלה 4

נביא דוגמה כנדרש:

עבור $d=2,S=S_1\cup S_2$ עבור $S_i=\left\{(x,y):r_i\leq\sqrt{x^2+y^2}\leq r_{i+1}\right\}$ המייצג סט של נקודות בתוך טבעת עם רדיוס פנימי r_i, ועם רדיוס חיצוני r_{i+1}כך ש $r_1<r_2<r_3$.

נגדיר את פונקציית המיפוי באופן הבא: עבור נקודות x,y

$$(x,y)\in S,\psi(x,y)=\left(x,y,\sqrt{x^2+y^2}\right)$$

הפונקציה ψ ממפה את הנקודות לעצמן, עם אבלואציה $z=\sqrt{x^2+y^2}$.

מכיוון ש $r_1<r_2<r_3$ הדאטה יהיה מופרד לינארית ב R^3 כנדרש.

חלק II

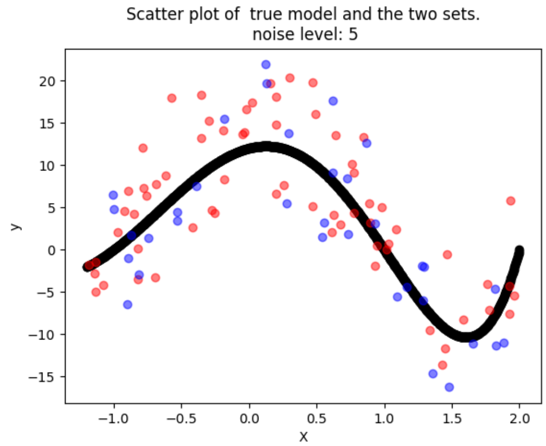
:$Practical$

שאלה 1

גרף הדגימות עם רעש 5:

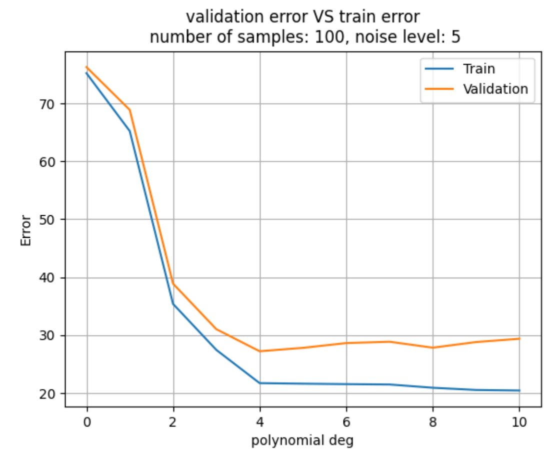
$train=red$

$test=blue$



שאלה 2

הגרף עבור $poly\ fit$ **עם** $k = 0...10$**:**



הסבר: אנו רואים כי בטעות על סט האימון יורדת ככל ש k עולה, כי האלגוריתם התאמן עליו כבר ומתאים את עצמו לסט טוב יותר ככל שדרגת הפולינום עולה. בעוד שעבור סט המבחן הוא חוזה טוב יותר עבור $k = 4$. כי ה var גבוה על סט האימון ככל ש k עולה.

שאלה 3

הערך k הטוב ביותר היה $k = 4$.

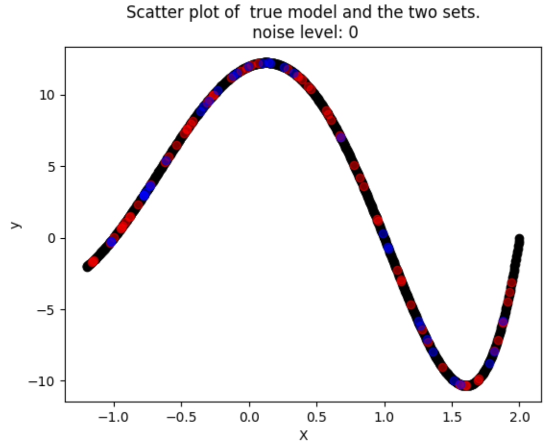
הטעות שהתקבלה על סט הבחינה הייתה $test\ error = 27.16$.

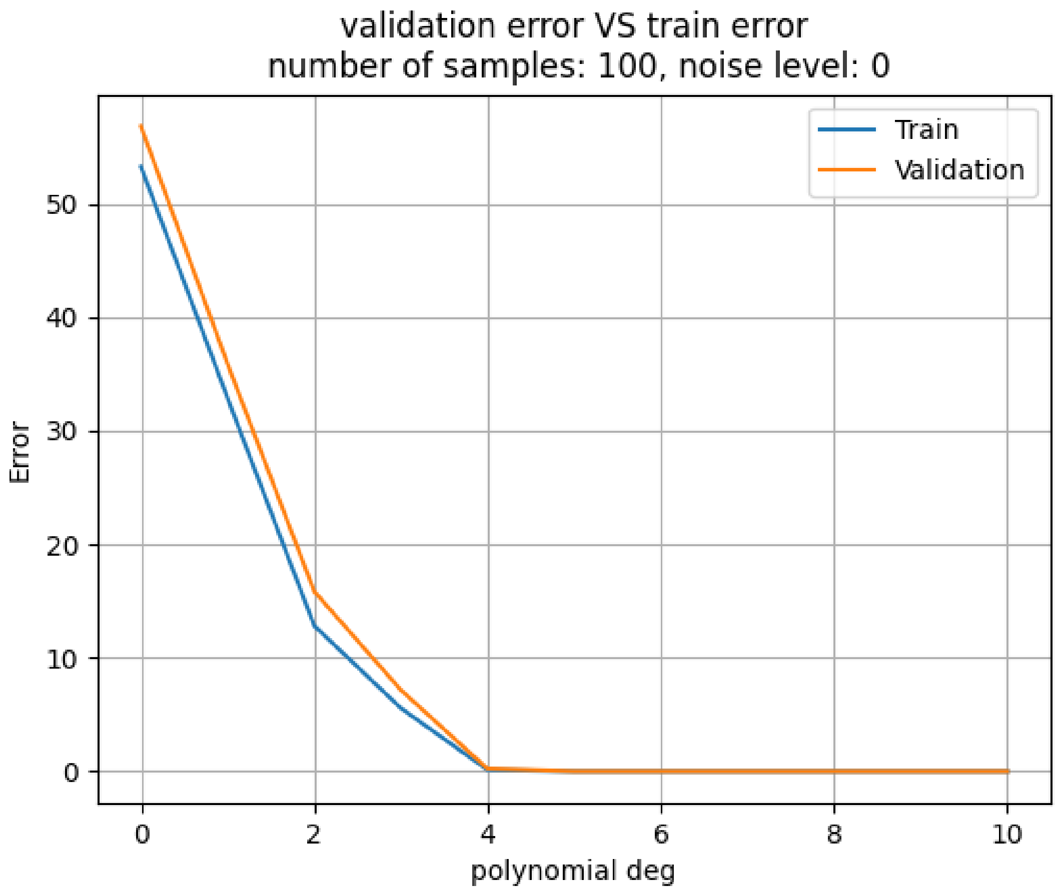
$validation\ error = 25.19$, לכן לא התרקיים שוויון בניהם, אך הם קרובים יחסית.

שאלה 4

$train = red$

$test = blue$



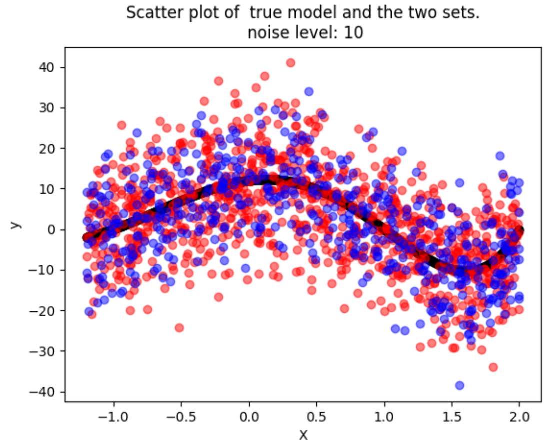


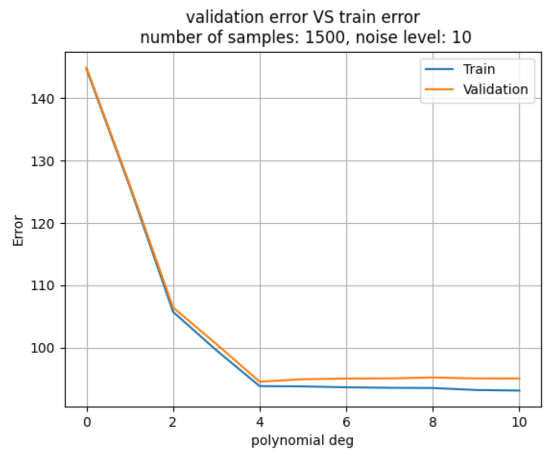
הסבר: עבור $k = 4$ מתקיים מפגש בין הטעויות של הטסט סט ו $validation$ משום שאין רעז אזי הטעות קרובה והאגוריתם חוזה טוב בשני המקרים. עבור $train\ set$ הטעות שווה ל 2.15 - שונה אך לא בהרבה.

שאלה 5

$$train = red$$

$$test = blue$$



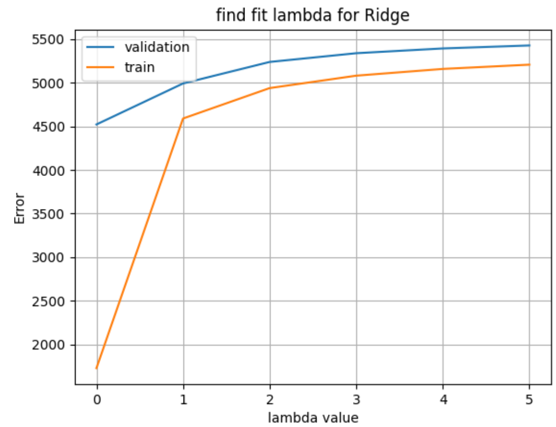


הסבר: נשים לב כי ככל שמספר הדגימות עולה טווח הטעות בין סט האימון ל $validation$ מצטצמם והם מתנהגים כאילו אין רעש כלל. ככל שיש לאלגוריתם יותר דגימות הוא מאומן טוב יותר ולכן האומדן שלו קרוב יותר למציאות. זה נתמך גם ע"י חוק המספרים הגדולים - התכנסות לתוחלת.

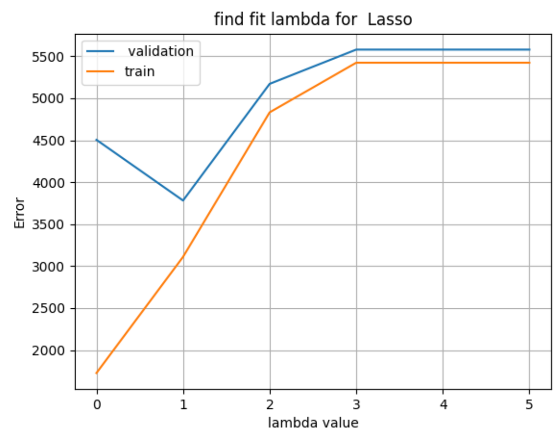
שאלה 7

הטווח הטוב ביותר עבור האלגוריתמים הוא: $0-6$

$Ridge\ plot$**:**



:*Lasso plot*



נסביר את ההבדלים בין הגרפים ובין הסטים:

אנו רואים כי יש הבדלים בין האלגוריתים, משום שהם ממשקלים את טעויות האלגוריתם עם נורמה שונה. אך שניהם מקבלים ערך נמוך כאשר למדה קטנה.

שאלה 8

ההפסד המינימלי מתקבל עבור פרמטרי רגולריזציה הבאים:

1 :*Lasso*

0 :*Ridge*

נחשב את הטעות לטסט סט:

4040.64 :*Lasso*

3612.24 :*Ridge*

3612.24 :*Linear Regression*