מערכות לומדות תרגיל 5

שיר שבח 322407701

31 במאי 2022

1 חלק תיאורטי

1.1 Regularization

1. בשאלה הבאה תראו כי למרות שאומדן Ridge מוטה הוא יכול להשיג MSE נמוך בהשוואה לאומדן LS.

יהי $X \in \mathbb{R}^{m\times d}$ מטריצה קבועה, $y \in \mathbb{R}^d$ וקטור response, והניחו כי X^TX הפיך. נסמן $\hat{w}$ פתרון LS ו־ $\hat{w}_\lambda$ פתרון ridge עבור פרמטר רגולריזציה $\lambda \geq 0$ (כאשר $\hat{w}_0 = \hat{w}$).

הניחו כי המודל הלינארי נכון, כלומר $y = Xw + \varepsilon$ כאשר $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$.

היזכרו כי במקרה הזה $\mathbb{E}\left(\hat{w}\right) = w$.

א. הראו כי $\hat{w}_\lambda = A_\lambda \hat{w}$ כאשר $A_\lambda = \left(X^TX + \lambda I_d\right)^{-1}\left(X^TX\right)$.

נסמן (לפי פירוק SVD) $X = U\Sigma V^T$.

הוכחנו בתרגול כי פתרון ridge הוא $\hat{w}_\lambda = V\Sigma_\lambda U^T y$.

ופתרון LS הוא: $\hat{w} = X^\dagger y = V\Sigma^\dagger U^T y$.

$$\begin{aligned}
A_\lambda \hat{w} &= \left(X^TX + \lambda I_d\right)^{-1}\left(X^TX\right)V\Sigma^\dagger U^T y \\
&= \left(V\Sigma^T U^T U\Sigma V^T + \lambda I_d\right)^{-1}\left(V\Sigma^T U^T U\Sigma V^T\right)V\Sigma^\dagger U^T y \\
&= \left(V\Sigma^T\Sigma V^T + \lambda I_d\right)^{-1} V\Sigma^T\Sigma\Sigma^\dagger U^T y \\
&= V\left(\Sigma^T\Sigma + \lambda I_d\right)^{-1} V^T V\Sigma^T\Sigma\Sigma^\dagger U^T y \\
&= V\underbrace{\left(\Sigma^T\Sigma + \lambda I_d\right)^{-1}\Sigma^T}_{\Sigma_\lambda}\Sigma\Sigma^\dagger U^T y \\
\circledast &= V\Sigma_\lambda\Sigma\Sigma^\dagger U^T y \\
\circledast\circledast &= V\Sigma_\lambda U^T y \\
&= \hat{w}_\lambda
\end{aligned}$$

$\circledast$ הוכחנו בתרגול $\Sigma_\lambda = \left(\Sigma^T\Sigma + \lambda I_d\right)^{-1}\Sigma^T$.

$\circledast\circledast$ $\Sigma\Sigma^\dagger = I$

ב. מלמעלה, הסיקו כי לכל $\lambda > 0$ האומדן ridge הוא אומדן מוטה של w. כלומר, הראו כי לכל $\lambda > 0$, $\mathbb{E}\left(\hat{w}_\lambda\right) \neq w$.

$$\mathbb{E}\left(\hat{w}_\lambda\right) = \mathbb{E}\left(A_\lambda \hat{w}\right) \underset{\circledast}{=} A_\lambda \mathbb{E}\left(\hat{w}\right) \underset{\circledast\circledast}{=} A_\lambda w$$

$\circledast$ לינאריות התוחלת.
$\circledast\circledast$ במקרה בו המודל הלינארי נכון, מתקיים $\mathbb{E}\left(\hat{w}\right) = w$.
נרצה להוכיח כי w הוא לא ווקטור עצמי של A_λ עם ע"ע 1.
נניח בשלילה שכן, כלומר, $A_\lambda w = w$. אזי:

$$\begin{aligned}\left(X^TX + \lambda I_d\right)^{-1}\left(X^TX\right)w &= w\\ \left(X^TX\right)w &= \left(X^TX + \lambda I_d\right)w\\ X^TXw &= X^TXw + \lambda w\\ 0 &= \lambda w\end{aligned}$$

אולם $\lambda > 0$ וגם $w \neq 0$. סתירה.
לכן קיבלנו:

$$A_\lambda w \neq w$$

ג. הראו כי: $\mathbf{Var}\left(\hat{w}_\lambda\right) = \sigma^2 A_\lambda \left(X^TX\right)^{-1} A_\lambda^T$, עבור σ^2 השונות של רעש משוער. רמז: היזכרו כי עבור מטריצה קבועה B ווקטור מקרי z מתקיים כי $\mathbf{Var}\left(Bz\right) = B \cdot \mathbf{Var}\left(z\right) \cdot B^T$ וכי $\mathbf{Var}\left(\hat{w}\right) = \sigma^2\left(X^TX\right)^{-1}$.

$$\begin{aligned}\mathrm{Var}\left(\hat{w}_\lambda\right) &= \mathrm{Var}\left(A_\lambda \hat{w}\right) = A_\lambda \cdot \mathrm{Var}\left(\hat{w}\right) \cdot A_\lambda^T\\ &= A_\lambda \cdot \sigma^2\left(X^TX\right)^{-1} \cdot A_\lambda^T\\ &= \sigma^2 A_\lambda \cdot \left(X^TX\right)^{-1} \cdot A_\lambda^T\end{aligned}$$

ד. הפיקו ביטויים מפורשים עבור ה$bias$ (מרובעת) וה$variace$ של $\hat{w}_\lambda$ כפונקציה של λ, כלומר כתבו פירוק $bias - variace$ עבור הMSE (mean square error) של $\hat{w}_\lambda$.

רמז: היזכרו כי עבור המקרה הרב־משתני הMSE מוגדר להיות:

$$\mathbf{MSE}\left(\hat{y}\right) = \mathbb{E}\left(\|\hat{y} - y\|^2\right) = \mathbb{E}\left(\left(\hat{y} - y\right)^T\left(\hat{y} - y\right)\right)$$

כאשר y הוא הערך האמיתי ו־$\hat{y}$ זה האומדן.

ראינו שמתקיים כי עבור MSE יש פירוק ל$bias - variace$:

$$\begin{aligned}\mathbb{E}\left(\|\hat{y} - y\|^2\right) &= \mathbb{E}\left(\|\hat{y} - \mathbb{E}\left(\hat{y}\right)\|^2\right) + \|\mathbb{E}\left(\hat{y}\right) - y\|^2\\ &= Var\left(\hat{y}\right) + bias^2\left(\hat{y}\right)\end{aligned}$$

ולכן:

$$\mathbb{E}\left(\|\hat{w}_\lambda-\hat{w}\|^2\right)=\underbrace{\mathbb{E}\left(\|\hat{w}_\lambda-\mathbb{E}\left(\hat{w}_\lambda\right)\|^2\right)}_{Var(\hat{w}_\lambda)}+\underbrace{\|\mathbb{E}\left(\hat{w}_\lambda\right)-\hat{w}\|^2}_{bias^2(\hat{w}_\lambda)}$$

<u>שונות:</u>

$$\begin{aligned}
Var\left(\hat{w}_\lambda\right) &= \mathbb{E}\left(\|\hat{w}_\lambda-\mathbb{E}\left(\hat{w}_\lambda\right)\|^2\right)\\
&= \mathbb{E}\left(\|A_\lambda\hat{w}-\mathbb{E}\left(A_\lambda\hat{w}\right)\|^2\right)\\
&= \mathbb{E}\left(\|A_\lambda\hat{w}-A_\lambda w\|^2\right)\\
&= \mathbb{E}\left(\|A_\lambda\left(\hat{w}-w\right)\|^2\right)\\
&= \mathbb{E}\left(\left(\left(\hat{w}-w\right)A_\lambda\right)^T\left(\hat{w}-w\right)A_\lambda\right)\\
&= \mathbb{E}\left(A_\lambda^T\left(\hat{w}-w\right)^T\left(\hat{w}-w\right)A_\lambda\right)\\
&= A_\lambda^T\mathbb{E}\left(\left(\hat{w}-w\right)^2\right)A_\lambda\\
&= A_\lambda^T Var\left(\hat{w}\right)A_\lambda\\
&= \sigma^2 A_\lambda\left(X^TX\right)^{-1}A_\lambda^T
\end{aligned}$$

קיבלנו בדיוק לפי הסעיף הקודם.

<u>הטיה בריבוע:</u>

$$\begin{aligned}
bias^2\left(\hat{w}_\lambda\right) &= \|\mathbb{E}\left(\hat{w}_\lambda\right)-\mathbb{E}\left(\hat{w}\right)\|^2\\
&= \|A_\lambda w-w\|^2\\
&= \|\left(A_\lambda-I\right)w\|^2\\
&= \left(\left(A_\lambda-I\right)w\right)^T\left(A_\lambda-I\right)w\\
&= w^T\left(A_\lambda-I\right)^T\left(A_\lambda-I\right)w
\end{aligned}$$

ה. הראו ע"י גזירה כי:

$$\frac{\partial}{\partial\lambda}\mathbf{MSE}\left(\hat{w}_\lambda\right)|_{\lambda=0}=\frac{\partial}{\partial\lambda}bias^2\left(\hat{w}_\lambda\right)|_{\lambda=0}+\frac{\partial}{\partial\lambda}Var\left(\hat{w}_\lambda\right)|_{\lambda=0}<0$$

כלומר, חשבו את הנגזרת של הפונקציה לעיל ביחס לλ בנקודה $\lambda = 0$.

$$\begin{aligned}\frac{\partial}{\partial\lambda}\mathrm{MSE}\left(\hat{w}_\lambda\right) &= \frac{\partial}{\partial\lambda}\left(Var\left(\hat{w}_\lambda\right) + bias^2\left(\hat{w}_\lambda\right)\right)\\ \circledast &= \frac{\partial}{\partial\lambda}\left(Var\left(\hat{w}_\lambda\right)\right) + \frac{\partial}{\partial\lambda}\left(bias^2\left(\hat{w}_\lambda\right)\right)\end{aligned}$$

$\circledast$ לינאריות הנגזרת.

<u>נגזור את ה$bais^2$:</u>

$$\begin{aligned}\frac{\partial}{\partial\lambda}\left(bias^2\left(\hat{w}_\lambda\right)\right) &= \frac{\partial}{\partial\lambda}\left(w^T\left(A_\lambda - I\right)^T\left(A_\lambda - I\right)w\right)\\ &= \frac{\partial}{\partial\lambda}\left(\|\left(A_\lambda - I\right)w\|^2\right)\\ &= \frac{\partial}{\partial\lambda}\left(\sum_{i=1}^{n}\left(\left(A_\lambda - I\right)_i w\right)^2\right)\\ &= \frac{\partial}{\partial\lambda}\left(\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\left(A_\lambda - I\right)_{ij}w_j\right)^2\right)\\ &= 2\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\left(A_\lambda - I\right)_{ij}w_j\right)\cdot\left(\sum_{j=1}^{n}\left(A_\lambda - I\right)_{ij}w_j\right)'\end{aligned}$$

נשים לב כי מתקיים: $A_\lambda - I = 0$ עבור $\lambda = 0$. לכן כל הביטוי הנ"ל שווה ל-0. ז"א $\frac{\partial}{\partial\lambda}\left(bias^2\left(\hat{w}_\lambda\right)\right) = 0$.

<u>נגזור את הvar:</u>

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

לכן:

$$\begin{aligned}\sigma^2 tr\left(\left(X^TX + \lambda I\right)^{-1}\left(X^TX\right)\left(X^TX\right)^{-1}A_\lambda^T\right) &= \sigma^2 tr\left(\left(X^TX + \lambda I\right)^{-1}\left(\left(X^TX + \lambda I\right)^{-1}\left(X^TX\right)\right)^T\right)\\ &= \sigma^2 tr\left(\left(X^TX + \lambda I\right)^{-1}\left(X^TX\right)\left(\left(X^TX + \lambda I\right)^{-1}\right)^T\right)\end{aligned}$$

X^TX מטריצה סימטרית לכן הפירוק SVD הוא: $X^TX = UDU^T$. וגם: $\left(X^TX\right)^{-1} = \left(UDU^T\right)^{-1} = UD^{-1}U^T$

בנוסף, $X^TX+\lambda I$ מטריצה סימטרית לכן הפירוק SVD הוא: $X^TX=VRV^T$. וגם: $\left(X^TX+\lambda I\right)^{-1}=\left(VRV^T\right)^{-1}=$ $VR^{-1}V^T$

$$\begin{aligned}
&=\sigma^2 tr\left(\left(VR^{-1}V^T\right)\left(UDU^T\right)\left(VR^{-1}V^T\right)^T\right)\\
&=\sigma^2 tr\left(\left(VR^{-1}V^T\right)\left(VR^{-1}V^T\right)\left(UDU^T\right)\right)\\
&=\sigma^2 tr\left(VR^{-1}R^{-1}V^T\left(VRV^T-\lambda I\right)\right)\\
&=\sigma^2 tr\left(VR^{-1}R^{-1}V^TVRV^T-\lambda VR^{-1}R^{-1}V^T\right)\\
&=\sigma^2 tr\left(VR^{-1}V^T-\lambda V\left(R^{-1}\right)^2V^T\right)\\
&=\sigma^2\left(tr\left(VR^{-1}V^T\right)-\lambda tr\left(V\left(R^{-1}\right)^2V^T\right)\right)\\
&=\sigma^2\left(tr\left(R^{-1}\right)-\lambda tr\left(\left(R^{-1}\right)^2\right)\right)
\end{aligned}$$

משום שמתקיים:

$$R=\begin{bmatrix}\alpha_1 & & \\ & \ddots & \\ & & \alpha_d\end{bmatrix}, D=\begin{bmatrix}b_1 & & \\ & \ddots & \\ & & b_d\end{bmatrix}$$

אזי מתקיים עבור ווקטור עצמי u_i של X^TX:

$$\begin{gathered}
X^TXu_i=b_iu_i\\
\left(X^TX+\lambda I\right)u_i=b_iu_i+\lambda u=\left(b_i+\lambda\right)u_i
\end{gathered}$$

לכן קיבלנו כי $\alpha_i=b_i+\lambda$ ואז:

$$R=D+\lambda I\rightarrow R^{-1}=\begin{bmatrix}\frac{1}{b_1+\lambda} & & \\ & \ddots & \\ & & \frac{1}{b_d+\lambda}\end{bmatrix}$$

לכן $tr\left(\left(R^{-1}\right)\right)=\sum_{i=1}^{d}\left(\frac{1}{b_i+\lambda}\right)$

ולכן הביטוי שהגענו אליו קודם שווה ל:

$$=\sigma^2\left(tr\left(\sum_{i=1}^{d}\left(\frac{1}{b_i+\lambda}\right)\right)-\lambda tr\left(\sum_{i=1}^{d}\left(\frac{1}{b_i+\lambda}\right)^2\right)\right)$$

נגזור את הביטוי הזה שפיתחנו:

$$\frac{\partial var\left(\hat{w}_{\lambda}\right)}{\partial\lambda}=\frac{\partial}{\partial\lambda}\left(\sigma^{2}\left(tr\left(\sum_{i=1}^{d}\left(\frac{1}{b_{i}+\lambda}\right)\right)-\lambda tr\left(\sum_{i=1}^{d}\left(\frac{1}{b_{i}+\lambda}\right)^{2}\right)\right)\right)$$
$$=-\sigma^{2}\left(\sum_{i=1}^{d}\left(\frac{1}{b_{i}+\lambda}\right)^{2}+\lambda\cdot\sum_{i=1}^{d}-2\frac{1}{\left(b_{i}+\lambda\right)^{3}}\right)$$
$$=-\sigma^{2}\sum_{i=1}^{d}\left(\frac{1}{\left(b_{i}+\lambda\right)^{2}}-\frac{2\lambda}{\left(b_{i}+\lambda\right)^{3}}\right)$$

נציב $\lambda=0$ ונבדוק את ערך הנגזרת:

$$-\sigma^{2}\cdot\sum_{i=1}^{d}\frac{1}{b_{i}^{2}}<0$$

לכן קיבלנו

$$\frac{\partial}{\partial\lambda}MSE\left(\hat{w}_{\lambda}\right)|_{\lambda=0}<0$$

ו. הסיקו מכך, כי אם המודל הלינארי נכון רגולריזציה קטנה של $ridge$ עוזרת להפחית את הMSE.

לפי הסעיף הקודם ראינו שבסביבת $\lambda=0$, כאשר מגדילים את λ הMSE יורד ־ משום שהנגזרת שלילית, וזה אומר כי עבור רגולריזציה קטנה ־ שמתקרבת ל$\lambda=0$, טעות הMSE פוחתת.

PCA 1.2

2. יהי $X:\Omega\to\mathbb{R}^{d}$ משתנה מקרי עם תוחלת אפס וקובריאנס $\Sigma\in\mathbb{R}^{d\times d}$. הראו כי לכל $v\in\mathbb{R}^{d}$, כאשר $\|v\|_{2}=1$, השונות של $\langle v,X\rangle$ זה לא גדול יותר מהשונות המתקבלת על ידי הטמעת PCA של X על תת מרחב חד מימדי (הניחו כי PCA משתמש בΣ בפועל).

נפתח את הביטוי $Var\left(v^{T}X\right)$ ונרצה להגיע לכך ש $Var\left(v^{T}X\right)\le Var\left(u_{1}^{T}X\right)$.

$$Var\left(v^{T}X\right)=\mathbb{E}\left(\left(v^{T}X-\mathbb{E}\left(v^{T}X\right)\right)^{2}\right)$$
$$=\mathbb{E}\left(\left(v^{T}X-\mathbb{E}\left(\sum_{i=1}^{d}v_{i}X_{i}\right)\right)^{2}\right)$$
$$=\mathbb{E}\left(\left(v^{T}X-\sum_{i=1}^{d}v_{i}\mathbb{E}\left(X_{i}\right)\right)^{2}\right)$$
$$\circledast=\mathbb{E}\left(\left(v^{T}X\right)^{2}\right)$$

⊛ נשים לב כי אם $\mathbb{E}(X)=0$ אזי $\begin{pmatrix}0\\ \vdots\\ 0\end{pmatrix}=\mathbb{E}\begin{pmatrix}X_1\\ \vdots\\ X_d\end{pmatrix}$ לכן גם לכל $i\in[d]$: $\mathbb{E}(X_i)=0$. ואז $\sum_{i=1}^{d}v_i\mathbb{E}(X_i)=0$.
נמשיך לפתח את הביטוי:

$$\begin{aligned}Var\left(v^TX\right)&=\mathbb{E}\left(\left(v^TX\right)^2\right)=\mathbb{E}\left(v^TX\left(v^TX\right)^T\right)\\&=\mathbb{E}\left(v^TXX^Tv\right)=v^T\mathbb{E}\left(XX^T\right)v\end{aligned}$$

נשים לב כי XX^T זוהי מכפלה חיצונית. ובה מקבלים:

$$XX^T=\begin{bmatrix}X_1\\ \vdots\\ X_d\end{bmatrix}\begin{bmatrix}X_1&\cdots&X_d\end{bmatrix}=\begin{bmatrix}X_1X_1&\cdots&X_1X_d\\ \vdots&\ddots&\vdots\\ X_dX_1&\cdots&X_dX_d\end{bmatrix}$$

ולכן

$$\mathbb{E}\left(XX^T\right)=\mathbb{E}\left(\begin{bmatrix}X_1X_1&\cdots&X_1X_d\\ \vdots&\ddots&\vdots\\ X_dX_1&\cdots&X_dX_d\end{bmatrix}\right)=\begin{bmatrix}\mathbb{E}(X_1X_1)&\cdots&\mathbb{E}(X_1X_d)\\ \vdots&\ddots&\vdots\\ \mathbb{E}(X_dX_1)&\cdots&\mathbb{E}(X_dX_d)\end{bmatrix}$$

ונשים לב כי מתקיים שלכל $i,j\in[d]$:

$$\mathbb{E}(X_iX_j)=cov(X_i,X_j)$$

משום ש:

$$\begin{aligned}cov(X_i,X_j)&=\mathbb{E}\left(\left(X_i-\mathbb{E}(X_i)\right)\left(X_j-\mathbb{E}(X_j)\right)\right)\\ [\forall k\in[d]\ \ \mathbb{E}(X_k)=0]&=\mathbb{E}(X_iX_j)\end{aligned}$$

לכן קיבלנו כי $\mathbb{E}\left(XX^T\right)=\Sigma$.

$$Var\left(v^TX\right)=v^T\Sigma v$$

נכתוב את v כצירוף לינארי של וקטורי הבסיס העצמיים $(u_1,..,u_d)$: $v=\sum_{i=1}^{d}\alpha_iu_i$ עבור $(\alpha_1,..,\alpha_d)$ סקלרים לא כולם אפס.
נשים לב כי מתקיים:

$$
\begin{aligned}
1 = \|v\|^2 = \|\sum_{i=1}^{d}\alpha_i u_i\|^2 &= \left\langle \sum_{i=1}^{d}\alpha_i u_i, \sum_{i=1}^{d}\alpha_i u_i \right\rangle \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d}\alpha_i\alpha_j \left\langle u_i, u_j\right\rangle \\
[\text{orthogonal basis}] &= \sum_{i=1}^{d}\alpha_i^2\|u_i\|^2 = \sum_{i=1}^{d}\alpha_i^2
\end{aligned}
$$

נמשיך לפתח. נרשום $\vec{\alpha} = \begin{pmatrix}\alpha_1\\ \vdots \\ \alpha_d\end{pmatrix}$ ונקבל:

$$
\begin{aligned}
Var\left(v^T X\right) = v^T\Sigma v &= \left(\sum_{i=1}^{d}\alpha_i u_i\right)^T \Sigma \left(\sum_{i=1}^{d}\alpha_i u_i\right) \\
&= \left(\vec{\alpha}^T U\right)^T \Sigma\vec{\alpha}U \\
&= (U\vec{\alpha})^T \left(UDU^T\right)U\vec{\alpha} \\
&= \vec{\alpha}^T U^T U D\vec{\alpha} \\
&= \vec{\alpha}^T D\vec{\alpha}
\end{aligned}
$$

מתקיים

$$
\begin{aligned}
\alpha^T D\alpha &= \begin{bmatrix}\alpha_1 & \cdots & \alpha_d\end{bmatrix}\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_d\end{bmatrix}\begin{bmatrix}\alpha_1\\ \vdots \\ \alpha_d\end{bmatrix} = \begin{bmatrix}\alpha_1 & \cdots & \alpha_d\end{bmatrix}\begin{bmatrix}\alpha_1\lambda_1\\ \vdots \\ \alpha_d\lambda_d\end{bmatrix} \\
&= \alpha_1^2\lambda_1 + \alpha_2^2\lambda_2 + ... + \alpha_d^2\lambda_d \le \lambda_1 \cdot \sum_{i=1}^{d}\alpha_i^2 = \lambda_1
\end{aligned}
$$

לכן אם נציב $\alpha_1 = a, \alpha_2 = 0, ..., \alpha_d = 0$ אזי

$$
v = \sum_{i=1}^{d}\alpha_i u_i = u_1
$$

לכן

$$
var\left(u_1^T X\right) = \alpha_1^2\lambda_1 = \lambda_1
$$

וקיבלנו כי

$$var\left(v^T X\right) \leq var\left(u_1^T X\right)$$

מה שהיה להוכיח.

Kernels

3. יהי $k\left(x,x'\right)$ קרנל PSD חוקי. ספקו קרנל PSD חוקי $\widetilde{k}\left(x,x'\right)$, הנבנה מ-$k$, כאשר הוא מנורמל. כלומר, לכל x מתקיים $\widetilde{k}\left(x,x\right)=1$. הוכיחו את תשובתכם.

משום שהקרנל k PSD חוקי, אזי גם לכל f פונקציה מתקיים כי גם

$$\widetilde{k}\left(x,y\right)=f\left(x\right)k\left(x,y\right)f\left(y\right)$$

חוקית.
לכן אם נגדיר $f\left(x\right)=\frac{1}{\sqrt{k\left(x,x\right)}}$ אזי נקבל:

$$\widetilde{k}\left(x,x'\right)=\frac{1}{\sqrt{k\left(x,x\right)}}k\left(x,x'\right)\frac{1}{\sqrt{k\left(x',x'\right)}}$$

וגם מתקיים כי:

$$\widetilde{k}\left(x,x\right)=\frac{1}{\sqrt{k\left(x,x\right)}}k\left(x,x\right)\frac{1}{\sqrt{k\left(x,x\right)}}=\frac{k\left(x,x\right)}{k\left(x,x\right)}=1$$

כנדרש.

4. נתון ה-data set $S=\left\{\left(x_i,y_i\right)\right\}_{i=1}^m$ כאשר $x_i\in\mathbb{R}^d$ ו-$y_i\in\left\{\pm1\right\}$, וממפה פיצ'רים $\psi:\mathbb{R}^d\to\mathcal{F}$ כאשר $\mathcal{F}$ זה מרחב פיצ'רים כלשהו. תנו דוגמה של data set S ומפת פיצ'רים כך ש-S לא ניתן לחלוקה לינארית ב-$\mathbb{R}^d$ (עבור $d\geq2$) אבל הדאטא שעבר שינוי $S_\psi=\left\{\left(\psi\left(x_i\right),y_i\right)\right\}_{i=1}^m$ ניתן לחלוקה לינארית ב-$\mathcal{F}$.

$$S=\left\{\left(\left(-2,-2\right),1\right),\left(\left(1,1\right),-1\right),\left(\left(2,2\right),-1\right)\right\}$$

נשים לב כי הדאטא הזה אינו ניתן לחלוקה לינארית.
נגדיר $\psi\begin{pmatrix}x\\y\end{pmatrix}=\left(x^2,y^2,x^2+y^2\right)$, ואז מתקיים:

$$\psi\begin{pmatrix}2\\2\end{pmatrix}=\left(4,4,8\right)$$
$$\psi\begin{pmatrix}1\\1\end{pmatrix}=\left(1,1,2\right)$$
$$\psi\begin{pmatrix}-2\\-2\end{pmatrix}=\left(4,4,8\right)$$

והנקודות האלו ניתנות לחלוקה לינארית במימד 3.

5. עבור כל אחת מהפונקציות הבאות, הוכיחו אם היא קרנל PSD חוקי או הביאו דוגמה נגדית.

א. $k(x,y)=\exp\left(\|x-y\|^2\right)$.

$$x=\begin{pmatrix}1\\0\end{pmatrix},y=\begin{pmatrix}0\\1\end{pmatrix}$$

$$k(e_1,e_2)=\exp\left(\|\begin{pmatrix}1\\0\end{pmatrix}-\begin{pmatrix}0\\1\end{pmatrix}\|^2\right)=e^2$$

$$G=\begin{bmatrix}1&e^2\\e^2&1\end{bmatrix}$$

נחשב ערכים עצמיים:

$$\begin{aligned}det\left(\begin{bmatrix}1&e^2\\e^2&1\end{bmatrix}-\lambda I\right)&=det\left(\begin{bmatrix}1-\lambda&e^2\\e^2&1-\lambda\end{bmatrix}\right)\\&=(1-\lambda)^2-\left(e^2\right)^2\end{aligned}$$

$$\begin{aligned}(1-\lambda)^2-e^4&=0\\e^4&=(1-\lambda)^2\\e^2&=1-\lambda\\\lambda_1&=1-e^2\end{aligned}$$

או

$$\begin{aligned}-e^2&=1-\lambda\\\lambda_2&=1+e^2\end{aligned}$$

קיבלנו כי ערך עצמי אחד הוא שלילי - $\lambda_1=1-e^2$, לכן זאת לא מטריצת PSD.

ב. $k(x,y)=k_1(x,y)-k_2(x,y)$ **עבור** k_1 **ו**k_2 **קרנלים חוקיים.**

נגדיר $k_1(x,y)=5$ ו־ $k_2(x,y)=10$. הוכחנו בתרגול כי קרנל קבוע (שנותן לכל x,y אותו מספר) הוא קרנל חוקי. אולם $k(x,y)=k_1(x,y)-k_2(x,y)=-5$. ואז מתקיים שהמטריצת G של k 2×2 היא:

$$G=\begin{bmatrix}-5&-5\\-5&-5\end{bmatrix}$$

ואז עבור $x=\begin{pmatrix}1\\0\end{pmatrix}$ מתקיים:

$$xGx^T=\begin{pmatrix}1\\0\end{pmatrix}\begin{bmatrix}-5&-5\\-5&-5\end{bmatrix}\begin{pmatrix}1&0\end{pmatrix}=-5<0$$

לכן לפי התנאים של מטריצת PSD, G לא מקיימת את אחד התנאים. לכן k לא PSD חוקי.

ג. $k\left(x,y\right)=k_a\left(x_a,y_a\right)+k_b\left(x_b,y_b\right)$ **עבור** k_1 **ו**k_2 **קרנלים חוקיים,** $x=\left[\begin{smallmatrix}x_a\\x_b\end{smallmatrix}\right],y=\left[\begin{smallmatrix}y_a\\y_b\end{smallmatrix}\right]$**.**
מתקיים:

$$\begin{aligned}G&=\begin{bmatrix}k_a\left(x_a,x_a\right)+k_b\left(x_b,x_b\right)&k_a\left(x_a,y_a\right)+k_b\left(x_b,y_b\right)\\k_a\left(y_a,x_a\right)+k_b\left(y_b,x_b\right)&k_a\left(y_a,y_a\right)+k_b\left(y_b,y_b\right)\end{bmatrix}\\&=\begin{bmatrix}k_a\left(x_a,x_a\right)&k_a\left(x_a,y_a\right)\\k_a\left(y_a,x_a\right)&k_a\left(y_a,y_a\right)\end{bmatrix}+\begin{bmatrix}k_b\left(x_b,x_b\right)&k_b\left(x_b,y_b\right)\\k_b\left(y_b,x_b\right)&k_b\left(y_b,y_b\right)\end{bmatrix}\\&=G_a+G_b\end{aligned}$$

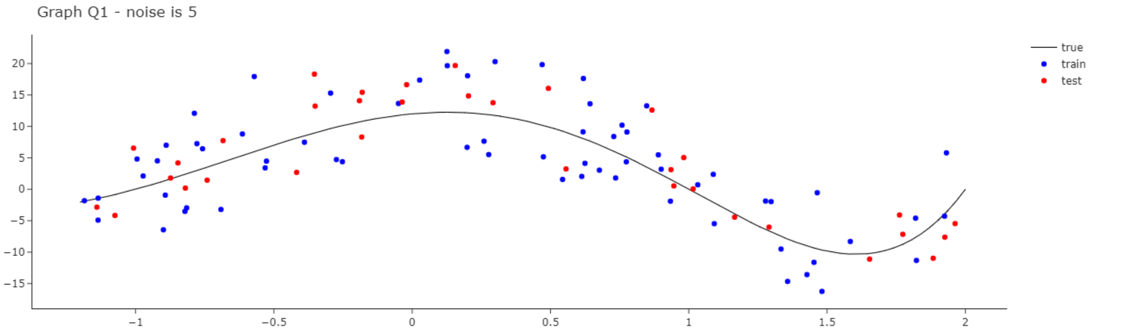
לכן:

$$x^TGx=x^T\left(G_a+G_b\right)x=x^TG_ax+x^TG_bx\geq0$$

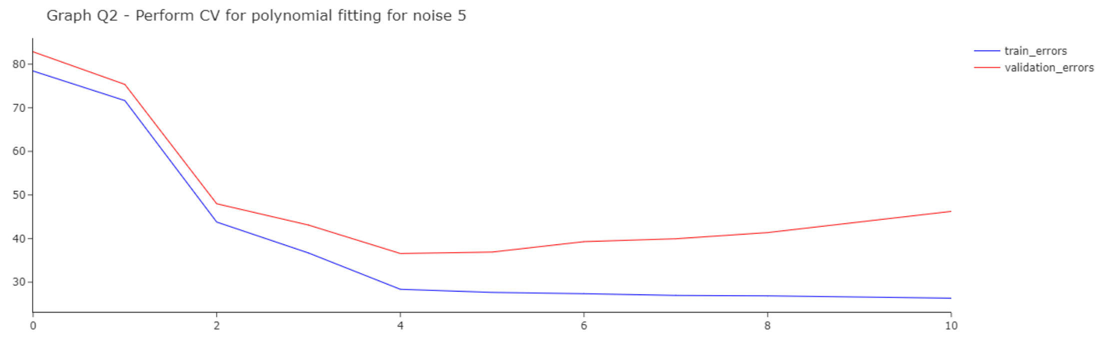
לכן G מטריצת PSD ולכן k חוקי.

חלק פרקטי

1.



2.

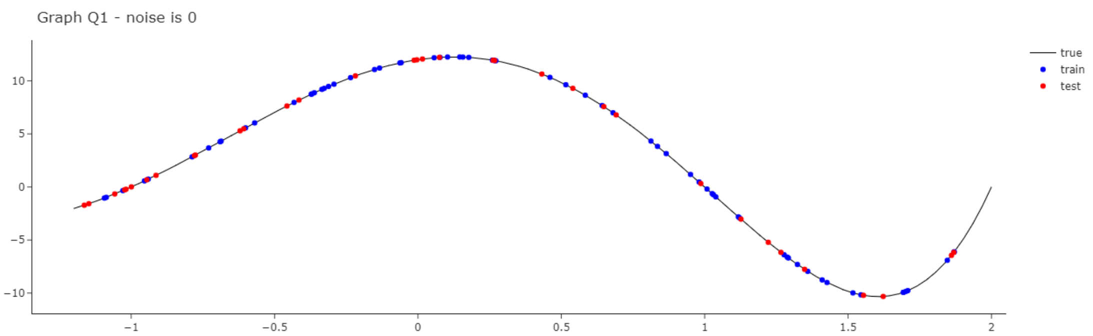


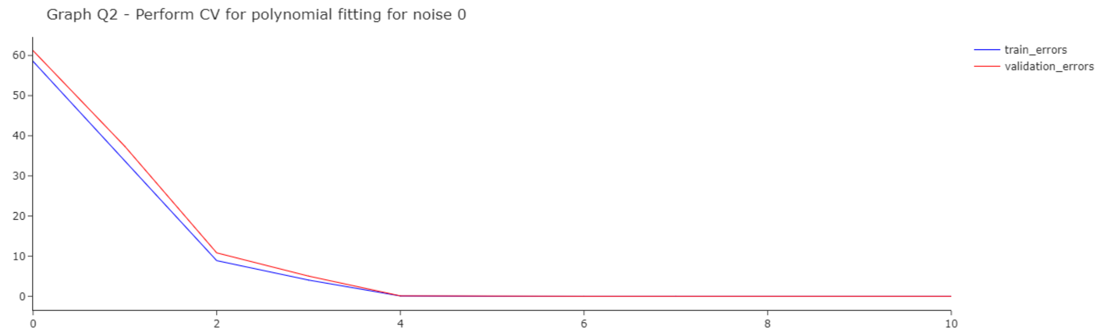
נשים לב כי הגרף של הvalidation error לא מבצע overfit לכן באיזשהו שלב הטעות עולה. ול-train error יש overfitting לכן הטעות יורדת ויורדת (אולם השונות עולה).

3.

הk בו הvalidation error הכי נמוך זה k=4. בו הtest error הוא -16.967. הvalidation error הוא 36.572 ולכן שני הerror-ים לא דומים.

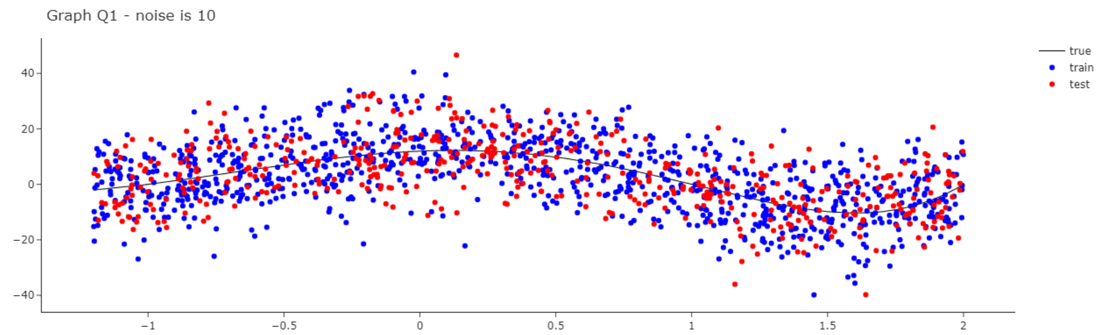
4.

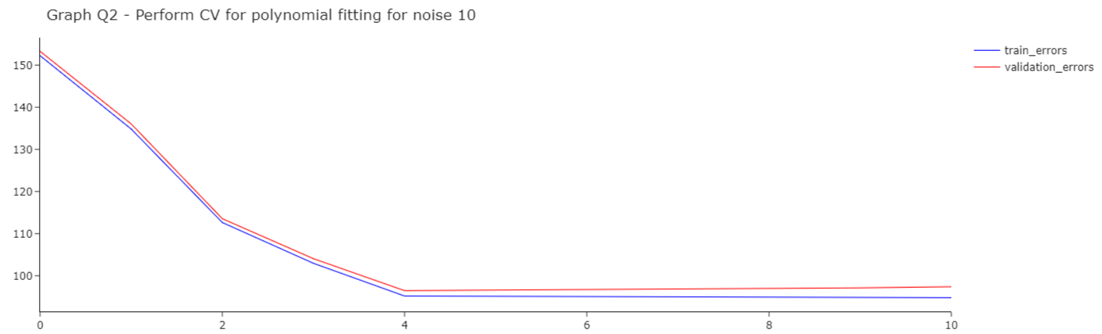




הk בו הvalidation error הכי נמוך זה k=5. בו הtest error הוא (מעוגל) 0. הvalidation error הוא גם מעוגל ל0 ולכן שני הerror-ים דומים.

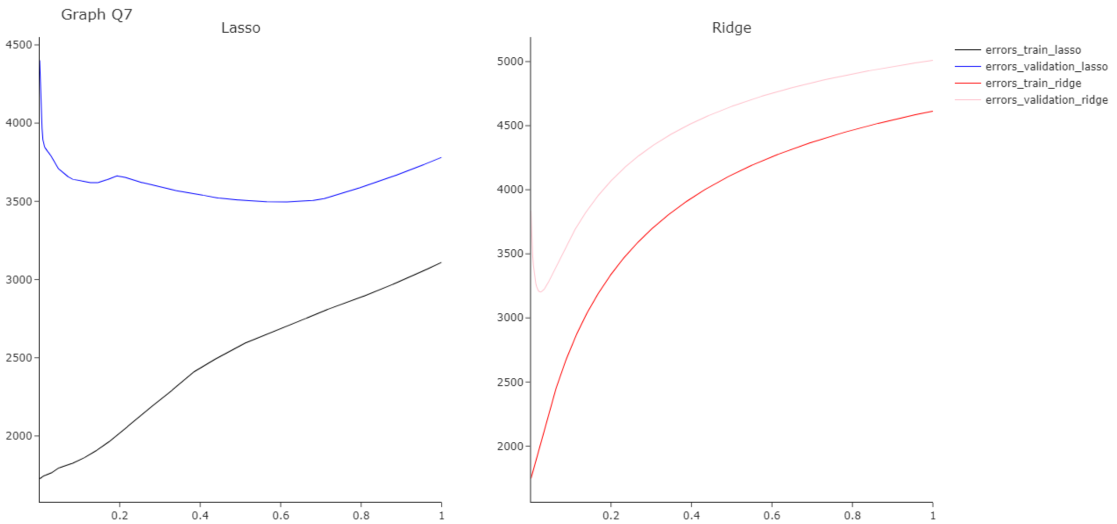
5.





הk בו הvalidation error הכי נמוך זה k=4. בו הtest error הוא 95.050. הvalidation error הוא 96.458 ולכן שני הerror-ים דומים. למרות שהעלנו את רמת הרעש, הוספנו בהרבה את מספר הדגימות. לכן הטעות בין שני הerrorים דומה – בהתאם לחוק המספרים הגדולים (אולם עדיין הטעות גדולה מאוד, פשוט דומה לvalidation).

7.



עבור טווח של לאמדות (0.001,1).

אנו שמים לב כי לlasso, בגרף של הטעות של הvalidation, יש לאמדות דומות, ולאחר מכן ככל שעולים בערך הלאמדה, הטעות עולה עד התייצבות.

.8

הלאמדה המינימלית בlasso: 0.597

הלאמדה המינימלית בridge: 0.025

הטעות בLinearRegression: 3612.249

הטעות בlasso: 3641.447

הטעות בridge: 3243.425

המודל שמקבל את הטעות הקטנה ביותר הוא ridge.