(67577) מבוא למערכות לומדות | תרגיל 5

שם: נמרוד בר גיורא | ת"ז: 207090622

חלק תאורטי

רגולריזציה

Let $\mathbf{X} \in \mathbb{R}^{m\times d}$ be a **constant** design matrix, $\mathbf{y} \in \mathbb{R}^d$ a response vector, and assume that $\mathbf{X}^\top\mathbf{X}$ is invertible. Denote $\hat{\mathbf{w}}$ the LS solution and $\hat{\mathbf{w}}_\lambda$ the ridge solution for the regularization parameter $\lambda \geq 0$ (where $\hat{\mathbf{w}}_0 \equiv \hat{\mathbf{w}}$)

- Assume the linear model is correct, namely $\mathbf{y} = X\mathbf{w} + \varepsilon$ where $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)$.
- Recall that in this case: $\mathbb{E}[\hat{\mathbf{w}}] = \mathbf{w}$.

שאלה 1

(a) Show that $\hat{\mathbf{w}}_\lambda = A_\lambda \hat{\mathbf{w}}$ where $A_\lambda := (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}(\mathbf{X}^\top\mathbf{X})$

הוכחה: יהי $X = U\Sigma V^\top$ פירוק ה־SVD של X. ראינו בתרגול שהפתרון של בעיית הרגרסיה הכולל רגולריזציה בשיטת ridge הוא:

$$\hat{w}_\lambda = V\Sigma_\lambda U^\top \cdot y \quad , \quad [\Sigma_\lambda]_{ii} = \frac{\sigma_i}{\sigma_i^2 + \lambda}$$

והפתרון שראינו עבור בעיית הרגרסיה שאינו כולל רגולריזציה הוא:

$$\hat{w} = V\Sigma^\dagger U^\top \cdot y \quad , \quad \left[\Sigma^\dagger\right]_{ii} = \begin{cases} 1/\sigma_i & \sigma_i \neq 0 \\ 0 & \sigma_i = 0 \end{cases}$$

נשים לב ש־

$$\begin{aligned} A_\lambda &= \left(X^\top X + \lambda I_d\right)^{-1}\left(X^\top X\right) = \left(\left(V\Sigma^\top U^\top\right)\left(U\Sigma V^\top\right) + \lambda I_d\right)^{-1}\left(V\Sigma^\top U^\top\right)\left(U\Sigma V^\top\right) = \\ &\overset{(1)}{=} \left(V\Sigma^\top\Sigma V^\top + \lambda I_d\right)^{-1} V\Sigma^\top\Sigma V^\top \overset{(2)}{=} \left(V\left(\Sigma^\top\Sigma + \lambda I_d\right)V^\top\right)^{-1} V\Sigma^\top\Sigma V^\top = \\ &\overset{(3)}{=} V\left(\Sigma^\top\Sigma + \lambda I_d\right)^{-1} V^\top V\Sigma^\top\Sigma V^\top \overset{(1)}{=} V\left(\Sigma^\top\Sigma + \lambda I_d\right)^{-1}\Sigma^\top\Sigma V^\top \overset{(4)}{=} V\Sigma_\lambda \cdot \Sigma V^\top \end{aligned}$$

(1): V ו־U מטריצות אורתוגונליות ולכן $U^\top U = I_m$ ו־ $V^\top V = I_d$.
(2): לפי למה 2.1 שראינו בתרגול.
(3): נובע מתכונה של מטריצות אורתוגונליות שראינו בתרגול (ו־V אורתוגונלית).
(4): ראינו בתרגול שמהגדרת Σ, Σ_λ מתקבל השיוויון $\left(\Sigma^\top\Sigma + \lambda I_d\right)^{-1}\Sigma^\top = \Sigma_\lambda$.
ולכן:

$$A_\lambda \cdot \hat{w} = \left(V\Sigma_\lambda\Sigma V^\top\right)\left(V\Sigma^\dagger U^\top\right) \overset{(1)}{=} V\Sigma_\lambda\Sigma\Sigma^\dagger U^\top \overset{(5)}{=} V\Sigma_\lambda I U^\top = V\Sigma_\lambda U^\top = \hat{w}_\lambda$$

(5): נתון ש־ $X^\top X$ **הפיכה** ולכן כל הערכים הסינגולריים שלה **שונים** מ־0. לכן מהגדרת $\Sigma^\dagger$ נובע ש־ $\left[\Sigma\Sigma^\dagger\right]_{ii} = \sigma_i \cdot \frac{1}{\sigma_i} = 1$ לכל $i \in [d]$. ■

(b) From the above, conclude that for any $\lambda > 0$ the ridge estimator is a biased estimator of $\mathbf{w}$. That is, show that for any $\lambda > 0$ $\mathbb{E}[\hat{\mathbf{w}}_\lambda] \neq \mathbf{w}$.

הוכחה: יהי $\lambda > 0$. מכך ש־$X^\top X$ הפיכה נובע שפתרון לבעיית הרגרסיה הליניארית הוא:

$$\hat{w} = \left(X^\top X\right)^{-1} X^\top y$$

נציב את הזהות שקיבלנו בסעיף הקודם עבור $\hat{w}_\lambda$:

$$\mathbb{E}\left(\hat{w}_\lambda\right) = \mathbb{E}\left(A_\lambda \hat{w}\right) \overset{☼}{=} A_\lambda \mathbb{E}\left(\hat{w}\right) = A_\lambda \mathbf{w}$$

☼: X מטריצה קבועה ולכן A_λ מטריצה קבועה, ולכל $i \in [d]$ הכניסה ה־i של $\mathbb{E}\left(A_\lambda \hat{w}\right)$ היא $\mathbb{E}\left(a^i \hat{w}\right)$, כאשר a^i היא השורה ה־i של A_λ. לכן:

$$\mathbb{E}\left(a^i \hat{w}\right) = \mathbb{E}\left(\sum_{j=1}^{d} a^i_j \hat{w}_j\right) \overset{\text{ליניאריות בתוחלת}}{=} \sum_{j=1}^{d} a^i_j \mathbb{E}\left(\hat{w}_j\right) = a^i \cdot \mathbb{E}\left(\hat{w}\right)$$

ולכן $\mathbb{E}\left(A_\lambda \hat{w}\right) = A_\lambda \mathbb{E}\left(\hat{w}\right)$.

נשים לב ש־$A_\lambda \neq I_d$ כי אם נניח בשלילה ש־$A_\lambda = I_d$ נקבל ש־

$$I_d = A_\lambda = \left(X^\top X + \lambda I_d\right)^{-1} \left(X^\top X\right)$$

ולכן

$$\left(X^\top X + \lambda I_d\right)^{-1} = \left(X^\top X\right)^{-1}$$

מכיוון ש־$\lambda > 0$ קיבלנו שקיימות ל־$X^\top X$ **שתי** מטריצות הפוכות **שונות**, וזו סתירה.

לכן $\mathbb{E}\left(\hat{w}_\lambda\right) = A_\lambda \mathbf{w} \neq \mathbf{w}$. ■

(c) Show that: $\mathrm{Var}(\hat{\mathbf{w}}_\lambda) = \sigma^2 A_\lambda (\mathbf{X}^\top \mathbf{X})^{-1} A_\lambda^\top$, for σ^2 the variance of the assumed noise.
Hint:: Recall that for a constant matrix B and a random vector $\mathbf{z}$ it holds that$\mathrm{Var}(B\mathbf{z}) = B \cdot \mathrm{Var}(\mathbf{z}) \cdot B^\top$ and that $\mathrm{Var}(\hat{\mathbf{w}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

הוכחה:

$$\mathrm{Var}\left(\hat{w}_\lambda\right) = \mathrm{Var}\left(A_\lambda \hat{w}\right) = A_\lambda \mathrm{Var}\left(\hat{w}\right) A_\lambda^\top \overset{\text{מהרמז}}{=} A_\lambda \sigma^2 \left(X^\top X\right)^{-1} A_\lambda^\top$$

■

(d) Derive explicit expressions for the (squared) bias and variance of $\hat{\mathbf{w}}_\lambda$ as a function of λ, i.e. write a bias-variance decomposition for the mean square error of $\hat{\mathbf{w}}_\lambda$.

הוכחה: לפי ההגדרה של ה־$\mathrm{Bias}, \mathrm{Var}$ ומה שחישבנו בסעיפים הקודמים:

$$\mathrm{Var}\left(\hat{w}_\lambda\right) = \mathbb{E}\left(\left(\hat{w}_\lambda - \mathbb{E}(\hat{w}_\lambda)\right)\left(\hat{w}_\lambda - \mathbb{E}(\hat{w}_\lambda)\right)^\top\right) \overset{\text{מהסעיף הקודם}}{=} \sigma^2 A_\lambda \left(X^\top X\right)^{-1} A_\lambda^\top := \mathrm{Var}\left(\lambda\right)$$

$$\left\|\mathrm{Bias}\left(\hat{w}_\lambda\right)\right\|^2 = \left\|\mathbb{E}\left(\hat{w}_\lambda\right) - \mathbf{w}\right\|^2 \underset{\text{מסעיף ב'}}{=} \left\|A_\lambda \mathbf{w} - \mathbf{w}\right\|^2 := \mathrm{Bias}^2\left(\lambda\right)$$

לפי ההגדרה, ה־MSE של $\hat{w}_\lambda$ הוא:

$$\mathrm{MSE}\left(\hat{w}_\lambda\right)=\mathbb{E}\left(\left\|\hat{w}_\lambda-\mathbf{w}\right\|^2\right)$$

נסמן $\overline{w}=\mathbb{E}\left(\hat{w}_\lambda\right)$. נשים לב שלפי ההגדרה של מכפלה חיצונית האלכסון של מטריצת ה־covariance $\left(\hat{w}_\lambda-\overline{w}\right)\left(\hat{w}_\lambda-\overline{w}\right)^\top$ מכיל בדיוק את הערכים:

$$\left[\left(\hat{w}_\lambda-\overline{w}\right)\right]_i\left[\left(\hat{w}_\lambda-\overline{w}\right)^\top\right]_i=\left[\left(\hat{w}_\lambda-\overline{w}\right)\right]_i\cdot\left[\left(\hat{w}_\lambda-\overline{w}\right)\right]_i=\left[\left(\hat{w}_\lambda-\overline{w}\right)\right]_i^2$$

כלומר את הריבועים של הקואורדינטות של הוקטור $\left(\hat{w}_\lambda-\overline{w}\right)$. לכן, מכך שתוחלת של מטריצת משתנים מקריים היא מטריצת התוחלות שלהם ומליניאריות התוחלת נובע ש־

$$\mathrm{Tr}\left(\mathrm{Var}\left(\lambda\right)\right)=\mathrm{Tr}\left(\mathbb{E}\left(\left(\hat{w}_\lambda-\overline{w}\right)\left(\hat{w}_\lambda-\overline{w}\right)^\top\right)\right)=\sum_{i=1}^{d}\mathbb{E}\left(\left[\left(\hat{w}_\lambda-\overline{w}\right)\right]_i^2\right)=\mathbb{E}\left(\sum_{i=1}^{d}\left[\left(\hat{w}_\lambda-\overline{w}\right)\right]_i^2\right)=\mathbb{E}\left(\left\|\left(\hat{w}_\lambda-\overline{w}\right)\right\|^2\right)$$

נפתח את הביטויים $\mathrm{Bias}^2\left(\lambda\right),\mathrm{Tr}\left(\mathrm{Var}\left(\lambda\right)\right),\mathrm{MSE}\left(\hat{w}_\lambda\right)$ בנפרד:

$$\begin{aligned}\mathrm{Tr}\left(\mathrm{Var}\left(\lambda\right)\right)&=\mathbb{E}\left(\left\|\left(\hat{w}_\lambda-\overline{w}\right)\right\|^2\right)=\mathbb{E}\left(\left\|\hat{w}_\lambda\right\|^2-2\overline{w}^\top\hat{w}_\lambda+\left\|\overline{w}\right\|^2\right)=\mathbb{E}\left(\left\|\hat{w}_\lambda\right\|^2\right)-2\mathbb{E}\left(\overline{w}^\top\hat{w}_\lambda\right)+\mathbb{E}\left(\left\|\overline{w}\right\|^2\right)=\\&=\mathbb{E}\left(\left\|\hat{w}_\lambda\right\|^2\right)-2\overline{w}^\top\mathbb{E}\left(\hat{w}_\lambda\right)+\left\|\overline{w}\right\|^2=\mathbb{E}\left(\left\|\hat{w}_\lambda\right\|^2\right)-2\left\|\overline{w}\right\|^2+\left\|\overline{w}\right\|^2=\mathbb{E}\left(\left\|\hat{w}_\lambda\right\|^2\right)-\left\|\overline{w}\right\|^2\end{aligned}$$

$$\mathrm{Bias}^2\left(\lambda\right)=\left\|\overline{w}-\mathbf{w}\right\|^2=\left\|\overline{w}\right\|^2-2\mathbf{w}^\top\overline{w}+\left\|\mathbf{w}\right\|^2$$

$$\begin{aligned}\mathrm{MSE}\left(\hat{w}_\lambda\right)&=\mathbb{E}\left(\left\|\hat{w}_\lambda-\mathbf{w}\right\|^2\right)=\mathbb{E}\left(\left\|\hat{w}_\lambda\right\|^2-2\mathbf{w}^\top\hat{w}_\lambda+\left\|\mathbf{w}\right\|^2\right)=\mathbb{E}\left(\left\|\hat{w}_\lambda\right\|^2\right)-2\mathbb{E}\left(\mathbf{w}^\top\hat{w}_\lambda\right)+\mathbb{E}\left(\left\|\mathbf{w}\right\|^2\right)=\\&=\mathbb{E}\left(\left\|\hat{w}_\lambda\right\|^2\right)-2\mathbf{w}^\top\mathbb{E}\left(\hat{w}_\lambda\right)+\left\|\mathbf{w}\right\|^2=\mathbb{E}\left(\left\|\hat{w}_\lambda\right\|^2\right)-2\mathbf{w}^\top\overline{w}+\left\|\mathbf{w}\right\|^2\end{aligned}$$

ונקבל ש־

$$\mathrm{Tr}\left(\mathrm{Var}\left(\lambda\right)\right)+\mathrm{Bias}^2\left(\lambda\right)=\mathbb{E}\left(\left\|\hat{w}_\lambda\right\|^2\right)-\left\|\overline{w}\right\|^2+\left\|\overline{w}\right\|^2-2\mathbf{w}^\top\overline{w}+\left\|\mathbf{w}\right\|^2=\mathbb{E}\left(\left\|\hat{w}_\lambda\right\|^2\right)-2\mathbf{w}^\top\overline{w}+\left\|\mathbf{w}\right\|^2=\mathrm{MSE}\left(\hat{w}_\lambda\right)$$

וזה פירוק ה־bias-variance של ה־MSE, כלומר:

$$\mathrm{MSE}\left(\hat{w}_\lambda\right)=\mathrm{Tr}\left(\sigma^2A_\lambda\left(X^\top X\right)^{-1}A_\lambda^\top\right)+\left\|A_\lambda\mathbf{w}-\mathbf{w}\right\|^2=\sigma^2\mathrm{Tr}\left(A_\lambda\left(X^\top X\right)^{-1}A_\lambda^\top\right)+\left\|A_\lambda\mathbf{w}-\mathbf{w}\right\|^2$$

■

(e) Show by differentiation that

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

That is, calculate the derivative of the functions above with respect to λ at point $\lambda=0$.

הוכחה: נשתמש במהלך ההוכחה בזהויות הבאות (שפורסמו בפורום התרגיל):

$$\frac{\partial}{\partial\lambda}\mathrm{Tr}\left(A_\lambda\right)=\mathrm{Tr}\left(\frac{\partial}{\partial\lambda}\left[A_\lambda\right]\right)$$

$$\frac{\partial}{\partial\lambda}\left[\left(A+\lambda B\right)^{-1}\right]=-\left(A+\lambda B\right)^{-1}B\left(A+\lambda B\right)^{-1}$$

$$\frac{\partial}{\partial\lambda}\left[A_\lambda\cdot B_\lambda\right]=\frac{\partial}{\partial\lambda}\left[A_\lambda\right]\cdot B_\lambda+A_\lambda\cdot\frac{\partial}{\partial\lambda}\left[B_\lambda\right]$$

נגזור תחילה את A_λ כפונקציה של λ (כלומר פונקציה מ־$\mathbb{R}$ ל־$\mathbb{R}^{d\times d}$):

$$\begin{aligned}\frac{\partial}{\partial\lambda}\left[A_\lambda\right] &= \frac{\partial}{\partial\lambda}\left[\left(X^\top X+\lambda I_d\right)^{-1}\left(X^\top X\right)\right] = \frac{\partial}{\partial\lambda}\left[\left(X^\top X+\lambda I_d\right)^{-1}\right]\cdot\left(X^\top X\right) = \\ &= -\left(X^\top X+\lambda I_d\right)^{-1}\cdot I_d\cdot\left(X^\top X+\lambda I_d\right)^{-1}\left(X^\top X\right) = -\left(X^\top X+\lambda I_d\right)^{-1}A_\lambda\end{aligned}$$

נגזור את $\mathrm{Tr}\left(\mathrm{Var}\left(\lambda\right)\right)$ כפונקציה של λ (מ־$\mathbb{R}$ ל־$\mathbb{R}$):

$$\begin{aligned}\frac{\partial}{\partial\lambda}\mathrm{Tr}\left(\mathrm{Var}\left(\lambda\right)\right) &= \mathrm{Tr}\left(\frac{\partial}{\partial\lambda}\left[\mathrm{Var}\left(\hat{w}_\lambda\right)\right]\right) = \mathrm{Tr}\left(\frac{\partial}{\partial\lambda}\left[\sigma^2A_\lambda\left(X^\top X\right)^{-1}A_\lambda^\top\right]\right) = \sigma^2\mathrm{Tr}\left(\frac{\partial}{\partial\lambda}\left[A_\lambda\left(X^\top X\right)^{-1}A_\lambda^\top\right]\right) = \\ &= \sigma^2\mathrm{Tr}\left(\frac{\partial}{\partial\lambda}\left[A_\lambda\right]\cdot\left(X^\top X\right)^{-1}A_\lambda^\top + A_\lambda\cdot\frac{\partial}{\partial\lambda}\left[\left(X^\top X\right)^{-1}A_\lambda^\top\right]\right) = \\ &= \sigma^2\mathrm{Tr}\left(\frac{\partial}{\partial\lambda}\left[A_\lambda\right]\left(X^\top X\right)^{-1}A_\lambda^\top + A_\lambda\left(X^\top X\right)^{-1}\frac{\partial}{\partial\lambda}\left[A_\lambda^\top\right]\right) = \\ &= \sigma^2\mathrm{Tr}\left(\frac{\partial}{\partial\lambda}\left[A_\lambda\right]\left(X^\top X\right)^{-1}A_\lambda^\top + A_\lambda\left(X^\top X\right)^{-1}\left(\frac{\partial}{\partial\lambda}\left[A_\lambda\right]\right)^\top\right) = \\ &= \sigma^2\mathrm{Tr}\left(-\left(X^\top X+\lambda I_d\right)^{-1}A_\lambda\left(X^\top X\right)^{-1}A_\lambda^\top + A_\lambda\left(X^\top X\right)^{-1}\left(-\left(X^\top X+\lambda I_d\right)^{-1}A_\lambda\right)^\top\right) = \\ &= \sigma^2\mathrm{Tr}\left(-\left(X^\top X+\lambda I_d\right)^{-1}A_\lambda\left(X^\top X\right)^{-1}A_\lambda^\top - A_\lambda\left(X^\top X\right)^{-1}A_\lambda^\top\left(\left(X^\top X+\lambda I_d\right)^{-1}\right)^\top\right) = \\ &\overset{☼}{=} -\sigma^2\mathrm{Tr}\left(\left(X^\top X+\lambda I_d\right)^{-1}A_\lambda\left(X^\top X\right)^{-1}A_\lambda^\top + A_\lambda\left(X^\top X\right)^{-1}A_\lambda^\top\left(X^\top X+\lambda I_d\right)^{-1}\right) = \\ &= -\sigma^2\mathrm{Tr}\left(\left(X^\top X+\lambda I_d\right)^{-1}A_\lambda\left(X^\top X\right)^{-1}A_\lambda^\top\right) - \sigma^2\mathrm{Tr}\left(A_\lambda\left(X^\top X\right)^{-1}A_\lambda^\top\left(X^\top X+\lambda I_d\right)^{-1}\right) = \\ &= -2\sigma^2\mathrm{Tr}\left(\left(X^\top X+\lambda I_d\right)^{-1}A_\lambda\left(X^\top X\right)^{-1}A_\lambda^\top\right)\end{aligned}$$

☼:כי $X^\top X+\lambda I_d$ מטריצה סימטרית והפיכה, ולכן גם ההופכית שלה - $\left(X^\top X+\lambda I_d\right)^{-1}$ היא מטריצה סימטרית.
נשים לב ש־ $A_\lambda = I_d$ כאשר $\lambda = 0$, ולכן כשנציב $\lambda = 0$ בנגזרת נקבל ש־

$$\frac{\partial}{\partial\lambda}\mathrm{Tr}\left(\mathrm{Var}\left(\hat{w}_\lambda\right)\right)\Big|_{\lambda=0} = -2\sigma^2\mathrm{Tr}\left(\left(X^\top X\right)^{-1}\left(X^\top X\right)^{-1}\right) = -2\sigma^2\mathrm{Tr}\left(\left(X^\top X\right)^{-1}\right)^2 < 0$$

כי $\sigma^2 > 0$ וכי כל הערכים על האלכסון של $\left(\left(X^\top X\right)^{-1}\right)^2$ **גדולים ממש** מ־0, כי אחרת נקבל שקיימת עמודה $x\in\mathbb{R}^d$ של $\left(X^\top X\right)^{-1}$ המקיימת ש־

$$x^\top x = 0 \quad\Longrightarrow\quad x = 0$$

בסתירה לכך ש־$\left(X^\top X\right)^{-1}$ **הפיכה**. לכן ה־trace של המטריצה הזו **גדול ממש** מ־0.
נגזור עכשיו את $\mathrm{Bias}^2\left(\lambda\right)$:

$$\begin{aligned}\frac{\partial}{\partial\lambda}\mathrm{Bias}^2\left(\lambda\right) &= \frac{\partial}{\partial\lambda}\left\|A_\lambda\mathbf{w}-\mathbf{w}\right\|^2 \overset{\text{כלל השרשרת}}{=} 2\left(A_\lambda\mathbf{w}-\mathbf{w}\right)^\top\frac{\partial}{\partial\lambda}\left[A_\lambda\mathbf{w}-\mathbf{w}\right] = \\ &= 2\left(\left(A_\lambda-I_d\right)\mathbf{w}\right)^\top\left(\frac{\partial}{\partial\lambda}\left[A_\lambda\right]\mathbf{w}-\frac{\partial}{\partial\lambda}\left[\mathbf{w}\right]\right) = 2\mathbf{w}^\top\left(A_\lambda-I_d\right)^\top\frac{\partial}{\partial\lambda}\left[A_\lambda\right]\mathbf{w} = \\ &= -2\mathbf{w}^\top\left(A_\lambda-I_d\right)^\top\left(X^\top X+\lambda I_d\right)^{-1}A_\lambda\mathbf{w}\end{aligned}$$

נציב $\lambda = 0$ ונקבל ש־

$$\frac{\partial}{\partial\lambda}\mathrm{Bias}^2\left(\lambda\right)\Big|_{\lambda=0} = -2\mathbf{w}^\top\left(I_d-I_d\right)^\top\left(X^\top X\right)^{-1}I_d\mathbf{w} = -2\mathbf{w}^\top 0\left(X^\top X\right)^{-1}\mathbf{w} = 0$$

ולכן מהפירוק של ה־MSE מהסעיף הקודם ומכך שהנגזרת שלה לפי λ היא אדטיבית נובע ש־

$$\frac{\partial}{\partial\lambda}\text{MSE}\left(\hat{w}_\lambda\right)\big|_{\lambda=0}=\frac{\partial}{\partial\lambda}\text{Tr}\left(\text{Var}\left(\hat{w}_\lambda\right)\right)\big|_{\lambda=0}+\frac{\partial}{\partial\lambda}\text{Bias}^2\left(\lambda\right)\big|_{\lambda=0}<0$$

כנדרש. ∎

> (f) Conclude that, if the linear model is correct, a little Ridge regularization helps to reduce the MSE.

הוכחה: נחשוב על ה־MSE כעל פונקציה של $\lambda\in\mathbb{R}_+$ שמחזירה ערך ב־$\mathbb{R}$. לפי הסעיף הקודם - הנגזרת שלה בנקודה $\lambda=0$ מקיימת ש־

$$\frac{\partial}{\partial\lambda}\text{MSE}\left(\hat{w}_\lambda\right)\big|_{\lambda=0}=\frac{\partial}{\partial\lambda}\text{MSE}\left(\hat{w}_0\right)<0$$

ולכן קיימת סביבה ימנית של 0 שבה הפונקציה $\text{MSE}\left(\lambda\right)$ היא מונוטונית יורדת. לכן עבור כל $\lambda>0$ בסביבה הזו מתקיים ש־

$$\text{MSE}\left(\hat{w}_0\right)>\text{MSE}\left(\hat{w}_\lambda\right)$$

כלומר עבור ערכי λ גדולים מ־0 אבל קטנים מספיק כדי להישאר בסביבה הזו מתקיים שה־MSE היא במגמת ירידה. ∎

PCA

שאלה 2

> 2. Let $X:\Omega\to\mathbb{R}^d$ be a random variable with zero mean and covariance $\Sigma\in\mathbb{R}^{d\times d}$. Show that for any $\mathbf{v}\in\mathbb{R}^d$, where $\|\mathbf{v}\|_2=1$, the variance of $\langle\mathbf{v},X\rangle$ is not larger then variance obtained by the PCA embedding of X into a one-dimension subspace (assume that the PCA uses the actual Σ).

הוכחה: לפי ההגדרה, Σ (מטריצת ה־covariance של X) היא מטריצה סימטרית. לכן לפי המשפט הספקטרלי קיים בסיס $\mathcal{U}=\{u_1,\dots,u_d\}$ **אורתונורמלי** של $\mathbb{R}^d$ שמורכב מוקטורים עצמיים של Σ. נסמן את הערכים העצמיים המתאימים להם ב־ $\lambda_1,\dots,\lambda_d\in\mathbb{R}$ ונניח בה"כ ש־ $\lambda_1\geqslant\ldots\geqslant\lambda_d$. נסמן ב־$U$ את המטריצה שעמודותיה הן וקטורי הבסיס הנ"ל. אז לכל $v\in\mathbb{R}^d$ מתקיים ש־ $U^\top v=[v]_{\mathcal{U}}$ (כלומר זה הייצוג של v ביחס לבסיס הא"נ הזה).

נשים לב שמהנתון ש־ $\mathbb{E}\left(X\right)=0$ ומהגדרת Σ נובע ש־

$$\Sigma=\mathbb{E}\left(\left(X-\mathbb{E}\left(X\right)\right)\left(X-\mathbb{E}\left(X\right)\right)^\top\right)=\mathbb{E}\left(XX^\top\right)\qquad\text{(I)}$$

לפי ההגדרה של אלגוריתם ה־PCA, הוא מחזיר עבור הקלט $(X,1)$ (כאשר הוא משתמש ב־Σ המקורית) בדיוק את וקטור הבסיס u_1 שמתאים לערך העצמי λ_1, שהוא הערך העצמי המקסימלי של Σ.

לכל $v\in\mathbb{R}^d$ מתקיים ש־

$$\text{Var}\left(\langle v,X\rangle\right)\overset{\text{הגדרה}}{=}\mathbb{E}\left(\langle v,X\rangle^2\right)-\left(\mathbb{E}\left(\langle v\mid X\rangle\right)\right)^2=\mathbb{E}\left(\langle v,X\rangle^2\right)-\left(\mathbb{E}\left(\sum_{i=1}^d v_ix_i\right)\right)^2=\qquad\text{(II)}$$

$$=\mathbb{E}\left(\langle v,X\rangle^2\right)-\left(\sum_{i=1}^d v_i\mathbb{E}\left(x_i\right)\right)^2=\mathbb{E}\left(\langle v,X\rangle^2\right)-\left(\langle v\mid\mathbb{E}\left(X\right)\rangle\right)^2\underset{\mathbb{E}(X)=0}{=}\mathbb{E}\left(\langle v,X\rangle^2\right)-\left(\langle v\mid 0\rangle\right)^2=\mathbb{E}\left(\langle v,X\rangle^2\right)$$

בפרט, עבור $\text{PCA}\left(X,1\right)=u_1$ נקבל ש־

$$\text{Var}\left(\langle u_1\mid X\rangle\right)=\mathbb{E}\left(\langle u_1,X\rangle^2\right)=\mathbb{E}\left(\left(u_1^\top X\right)^2\right)=\mathbb{E}\left(u_1^\top Xu_1^\top X\right)=\mathbb{E}\left(u_1^\top XX^\top u_1\right)=u_1^\top\mathbb{E}\left(XX^\top\right)u_1\overset{\text{(I)}}{=}u_1^\top\Sigma u_1=$$

$$\underset{\Sigma\ \text{סימטרית}}{=}\langle\Sigma u_1\mid u_1\rangle=\langle\lambda_1u_1\mid u_1\rangle=\lambda_1\langle u_1\mid u_1\rangle\underset{\text{הבסיס א"נ}}{=}\lambda_1\qquad\text{(III)}$$

נשים לב שהטמעת ה־PCA (ה־embedding) של X על תת־מרחב ממימד 1 (כלומר על $\mathbb{R}$) היא, כפי שראינו בכיתה - $\langle u_1 \mid X \rangle$. כלומר לפי החישוב ב־(III) השונות של הטמעת ה־PCA על $\mathbb{R}$ היא בדיוק λ_1 שהוא הערך העצמי המקסימלי של Σ.

יהי כעת $v \in \mathbb{R}^d$ המקיים ש־ $\|v\|_2 = 1$. נסמן $v = \sum\limits_{i=1}^{d} \alpha_i u_i$. כלומר - $U^\top v = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_d \end{bmatrix}$. נחשב את השונות של $\langle v, X \rangle$:

$$\operatorname{Var}\left(\langle v, X\rangle\right) \overset{\text{(II)}}{=} \mathbb{E}\left(\langle v, X\rangle^2\right) = \mathbb{E}\left(v^\top X v^\top X\right) = \mathbb{E}\left(v^\top X X^\top v\right) = v^\top \mathbb{E}\left(X X^\top\right) v \overset{\text{(I)}}{=} v^\top \Sigma v = \langle \Sigma v \mid v \rangle =$$

$$= \left\langle \sum_{i=1}^{d} \alpha_i \Sigma u_i \mid \sum_{j=1}^{d} \alpha_j u_j \right\rangle = \sum_{i=1}^{d}\sum_{j=1}^{d} \alpha_i \alpha_j \langle \Sigma u_i \mid u_j \rangle = \sum_{i=1}^{d}\sum_{j=1}^{d} \alpha_i \alpha_j \lambda_i \langle u_i \mid u_j \rangle =$$

$$\underset{\substack{\uparrow \\ \text{הבסיס א"נ}}}{=} \sum_{i=1}^{d} \alpha_i^2 \lambda_i \leqslant \sum_{i=1}^{d} \alpha_i^2 \lambda_1 = \lambda_1 \cdot \sum_{i=1}^{d} \alpha_i^2 = \lambda_1 \cdot \left\|U^\top v\right\|_2^2 \overset{☼}{=} \lambda_1 \cdot \|v\|_2^2 = \lambda_1 \overset{\text{(III)}}{=} \operatorname{Var}\left(\langle u_1 \mid X \rangle\right)$$

☼: מכיוון ש־$\mathcal{U}$ הוא בסיס א"נ אז $U^\top$ היא מטריצה אורתוגונלית, ולכן (כפי שמוכיחים בליניארית 2) היא משמרת מרחקים.

קיבלנו שלכל $v \in \mathbb{R}^d$ עם $\|v\|_2 = 1$ השונות של $\langle v, X \rangle$ לא גדולה מהשונות של הטמעת ה־PCA של X על $\mathbb{R}$, כנדרש. ■

Kernels

שאלה 3

> 3. Let $k(\mathbf{x}, \mathbf{x}')$ be a valid PSD kernel. Provide a valid PSD kernel $\tilde{k}(\mathbf{x}, \mathbf{x}')$, constructed from k, which is guaranteed to be normalized. That is, for all $\mathbf{x}$ it holds that $\tilde{k}(\mathbf{x}, \mathbf{x}) = 1$. Prove your answer.

הוכחה: ראינו בכיתה ש־$k_1 : \mathcal{X} \to \mathbb{R}$ המוגדר ע"י $k_1(x, x') = 1$ לכל $x, x' \in \mathcal{X}$ הוא קרנל PSD ולידי.

ראינו גם שסכום של קרנלים ולידיים הוא ולידי, ולכן $k' : \mathcal{X} \to \mathbb{R}$ המוגדר ע"י

$$k'(x, x') = k(x, x') + 1$$

הוא ולידי. מאפיון מרסר נובע שקיימת $\psi : \mathcal{X} \to \mathcal{F}$ כך שלכל $x, x' \in \mathcal{X}$ מתקיים:

$$k'(x, x') = \left(\psi(x)\right)^\top \psi(x')$$

מכיוון ש־k הוא קרנל PSD ולידי אז לכל $x \in \mathcal{X}$ מתקיים ש־ $k(x, x) \geqslant 0$ ולכן:

$$1 \leqslant k(x, x) + 1 = k'(x, x) = \left(\psi(x)\right)^\top \psi(x) = \|\psi(x)\|^2$$

נגדיר $\tilde{k} : \mathcal{X} \to \mathbb{R}$ ע"י:

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

נקבל שלכל $x \in \mathcal{X}$ מתקיים ש־

$$\tilde{k}(x, x) = \frac{k'(x, x)}{\sqrt{\left(k'(x, x)\right)^2}} = 1$$

בנוסף, עם הפונקציה $f(x) = \frac{1}{\|\psi(x)\|}$ מתקיים לכל $x, x' \in \mathcal{X}$ ש־

$$\tilde{k}(x, x') = \frac{k'(x, x')}{\sqrt{k'(x, x)\, k'(x', x')}} = \frac{k'(x, x')}{\sqrt{\|\psi(x)\|^2 \|\psi(x')\|^2}} = \frac{k'(x, x')}{\|\psi(x)\| \|\psi(x')\|} = f(x)\, k'(x, x')\, f(x')$$

וראינו שמהולידיות של k' זה גורר שגם $\tilde{k}$ הוא קרנל PSD ולידי. ■

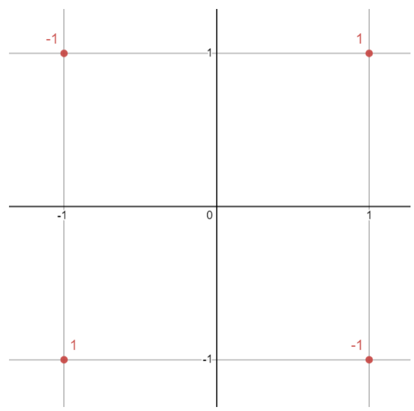
שאלה 4

4. Consider a data set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$, and a feature map $\psi : \mathbb{R}^d \to \mathcal{F}$ where $\mathcal{F}$ is some feature space. Give an example of a data set S and a feature map ψ such that S is not linearly separable in $\mathbb{R}^d$ (for $d \geq 2$) but that the transformed data set $S_\psi = \{(\psi(\mathbf{x}_i), y_i)\}_{i=1}^m$ is linearly separable in $\mathcal{F}$.

דוגמא:

נבחר $\mathbb{R}^2$, $m = 4$, את הפונקציה $\psi : \mathbb{R}^2 \to \mathbb{R}^3$ המוגדרת ע"י $\psi(x, y) = (x, y, xy)$ ואת הקבוצה

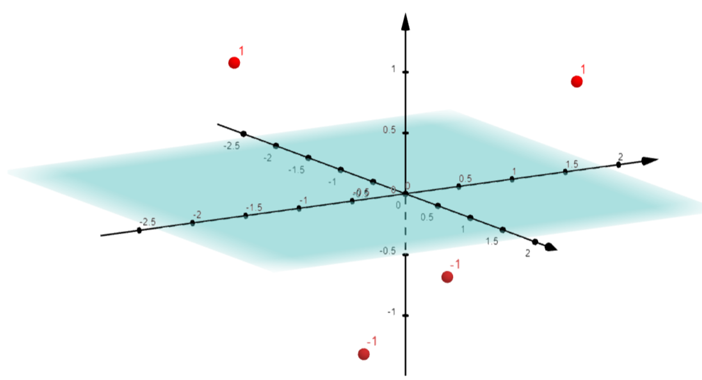
$$S = \{(x_i, y_i)\}_{i=1}^4 = \left\{ \left(\left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right), 1 \right), \left(\left(\begin{smallmatrix} -1 \\ -1 \end{smallmatrix}\right), 1 \right), \left(\left(\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right), -1 \right), \left(\left(\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right), -1 \right) \right\} \subseteq \mathbb{R}^2$$



S בבירור אינה ניתנת להפרדה ליניארית (ע"י קו ישר). אבל הקבוצה שמתקבלת מהפעלת ψ:

$$S_\psi = \{(\psi(x_i), y_i)\}_{i=1}^4 = \left\{ \left(\left(\begin{smallmatrix} 1 \\ 1 \\ 1 \end{smallmatrix}\right), 1 \right), \left(\left(\begin{smallmatrix} -1 \\ -1 \\ 1 \end{smallmatrix}\right), 1 \right), \left(\left(\begin{smallmatrix} -1 \\ 1 \\ -1 \end{smallmatrix}\right), -1 \right), \left(\left(\begin{smallmatrix} 1 \\ -1 \\ -1 \end{smallmatrix}\right), -1 \right) \right\}$$

ניתנת להפרדה ע"י המישור $z = 0$:



שאלה 5

5. For each of the following functions, prove it is a valid PSD kernel or show a counter example:
   (a) $k(\mathbf{x},\mathbf{y}) = \exp(||\mathbf{x}-\mathbf{y}||^2)$
   (b) $k(\mathbf{x},\mathbf{y}) = k_1(\mathbf{x},\mathbf{y}) - k_2(\mathbf{x},\mathbf{y})$ for any two valid kernels $k_1(\cdot,\cdot)$ and $k_2(\cdot,\cdot)$
   (c) $k(\mathbf{x},\mathbf{y}) = k_a(\mathbf{x}_a,\mathbf{y}_a) + k_b(\mathbf{x}_b,\mathbf{y}_b)$ for any two valid kernels $k_a(\cdot,\cdot)$ and $k_b(\cdot,\cdot)$, where

$$\mathbf{x} := \left[\frac{\mathbf{x}_a}{\mathbf{x}_b}\right], \quad \mathbf{y} := \left[\frac{\mathbf{y}_a}{\mathbf{y}_b}\right]$$

(a) הפונקציה $k(x,y) = \exp\left(\|x-y\|^2\right)$ היא **לא** קרנל PSD ולידי. דוגמא נגדית:
ניקח את הקבוצה $S = \{\left(\begin{smallmatrix}1\\1\end{smallmatrix}\right), \left(\begin{smallmatrix}0\\0\end{smallmatrix}\right)\} \subseteq \mathbb{R}^2$. מטריצת הגראם של k המתקבלת מ־S היא:

$$K = \begin{bmatrix} k\left(\left(\begin{smallmatrix}1\\1\end{smallmatrix}\right),\left(\begin{smallmatrix}1\\1\end{smallmatrix}\right)\right) & k\left(\left(\begin{smallmatrix}1\\1\end{smallmatrix}\right),\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right)\right) \\ k\left(\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right),\left(\begin{smallmatrix}1\\1\end{smallmatrix}\right)\right) & k\left(\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right),\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right)\right) \end{bmatrix} = \begin{bmatrix} \exp(0) & \exp(2) \\ \exp(2) & \exp(0) \end{bmatrix} = \begin{bmatrix} 1 & e^2 \\ e^2 & 1 \end{bmatrix}$$

נשים לב שעבור $\mathbb{R}^2 \ni v := \left(\begin{smallmatrix}-1\\1\end{smallmatrix}\right)$ מתקיים ש־

$$v^\top K v = \begin{pmatrix}-1 & 1\end{pmatrix} \begin{bmatrix} 1 & e^2 \\ e^2 & 1 \end{bmatrix} \begin{pmatrix}-1\\1\end{pmatrix} = \begin{pmatrix}-1 & 1\end{pmatrix} \begin{pmatrix}-1+e^2\\-e^2+1\end{pmatrix} = 1 - e^2 - e^2 + 1 = 2 - 2e^2 \overset{\overset{e^2>1}{\downarrow}}{<} 0$$

כלומר K היא לא מטריצת PSD, ולכן k היא לא קרנל PSD ולידי.

(b) הפונקציה $k(x,y) = k_1(x,y) - k_2(x,y)$, כאשר k_1, k_2 קרנלים ולידיים, היא **לא** קרנל ולידי. דוגמא נגדית:
נקח את $\mathcal{X} = \mathbb{R}$. ראינו בתרגול שתבניות קבועות חיוביות הן קרנלים ולידיים. לכן נוכל לקחת את $k_1(x,y) = 1$ ו־ $k_2(x,y) = 2$.
נקבל ש־k כפי שהוגדר לעיל הוא לא ולידי כי עבור $x = 7$ למשל נקבל ש־

$$k(x,x) = k_1(x,x) - k_2(x,x) = 1 - 2 = -1$$

ובאופן כללי לכל קבוצה ב־$\mathbb{R}$ נקבל שמטריצת הגראם של k שתתקבל ממנה תכיל רק ערכים שליליים על האלכסון הראשי, ולכן לא תהיה מטריצת PSD.

(c) הפונקציה $k(x,y) = k_a(x_a,y_a) + k_b(x_b,y_b)$ היא קרנל PSD ולידי.

הוכחה: יהי $m \in \mathbb{N}$ ותהי $\{x^i\}_{i=1}^m \subseteq \mathcal{X}$. לכל $i \in [m]$ נסמן:

$$x^i := \begin{pmatrix} x_a^i \\ x_b^i \end{pmatrix}$$

ונתבונן בקבוצות $\{x_a^i\}_{i=1}^m, \{x_b^i\}_{i=1}^m$. נתון ש־$k_a, k_b$ הם קרנלים ולידיים. נסמן ב־K^a, K^b את מטריצות הגראם שלהם ביחס לקבוצות הללו בהתאמה. לפי ההגדרה ־ אלו מטריצות PSD.

לכל $i,j \in [m]$ הכניסה ה־i,j של מטריצת הגראם K של k מקיימת:

$$K_{i,j} = k\left(x^i, x^j\right) = k_a\left(x_a^i, x_a^j\right) + k_b\left(x_b^i, x_b^j\right) = K_{i,j}^a + K_{i,j}^b$$

כלומר היא סכום הכניסות ה־i,j של K^a, K^b. לכן $K = K^a + K^b$. נקבל לכל $v \in \mathbb{R}^m$ ש־

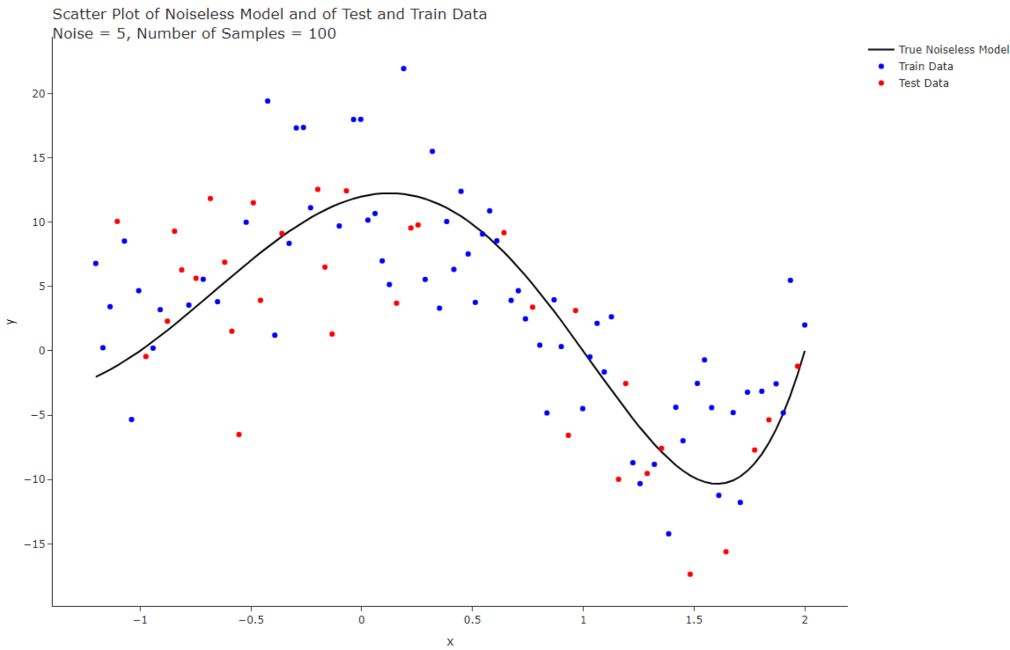
$$v^\top K v = v^\top \left(K^a + K^b\right) v = v^\top K^a v + v^\top K^b v \overset{\overset{K^a,K^b \succeq 0}{\downarrow}}{\geqslant} 0$$

ולכן K היא מטריצת PSD. מכיוון שזה מתקיים לכל $m \in \mathbb{N}$ ו־$\{x^i\}_{i=1}^m \subseteq \mathcal{X}$ אז לפי ההגדרה k היא קרנל ולידי. ■

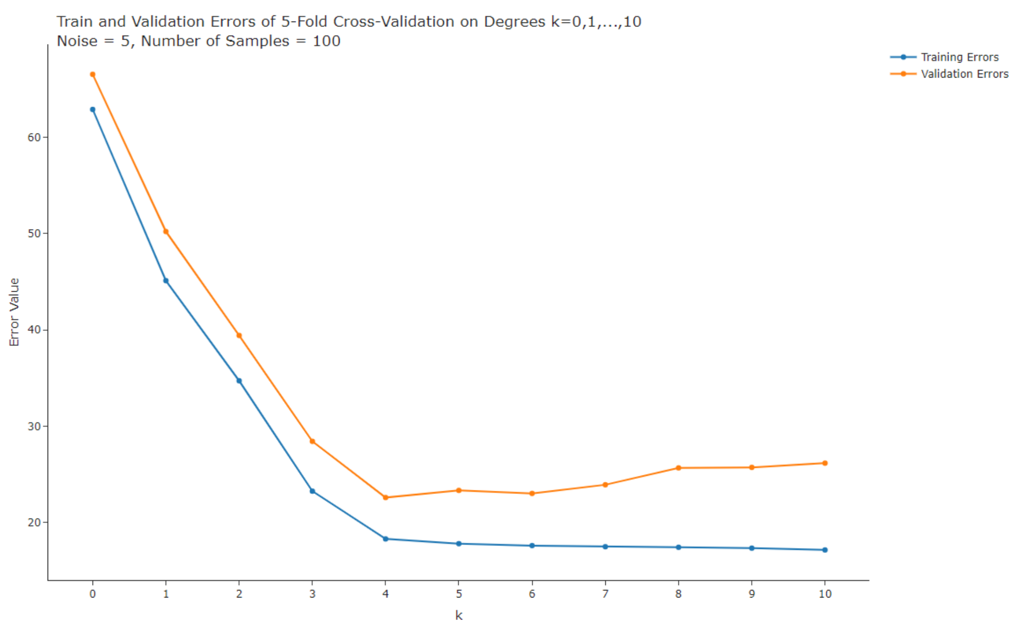
חלק פרקטי

Cross Validation For Selecting Polynomial Degree

שאלה 1



שאלה 2



ניתן לראות שככל שהערך של k (כלומר דרגת הפולינום) גדל, השגיאה של המודל על סט האימון קטנה. בנוגע לסט הולידציה - השגיאה קטנה כאשר k גדל בין 0 ל-4, אבל עבור ערכי k גדולים מדי (החל מ-$k = 5$) רואים בגרף שהשגיאה של המודל על סט הולידציה מתחילה לגדול. כלומר - עבור ערכי k בין 0 ל-4 המודל נעשה מורכב יותר וגם מצליח להכליל על דאטא שהוא לא ראה, אבל עבור ערכי k גדולים מדי הוא מתחיל לעשות Overfitting לסט האימון ולא מכליל טוב.

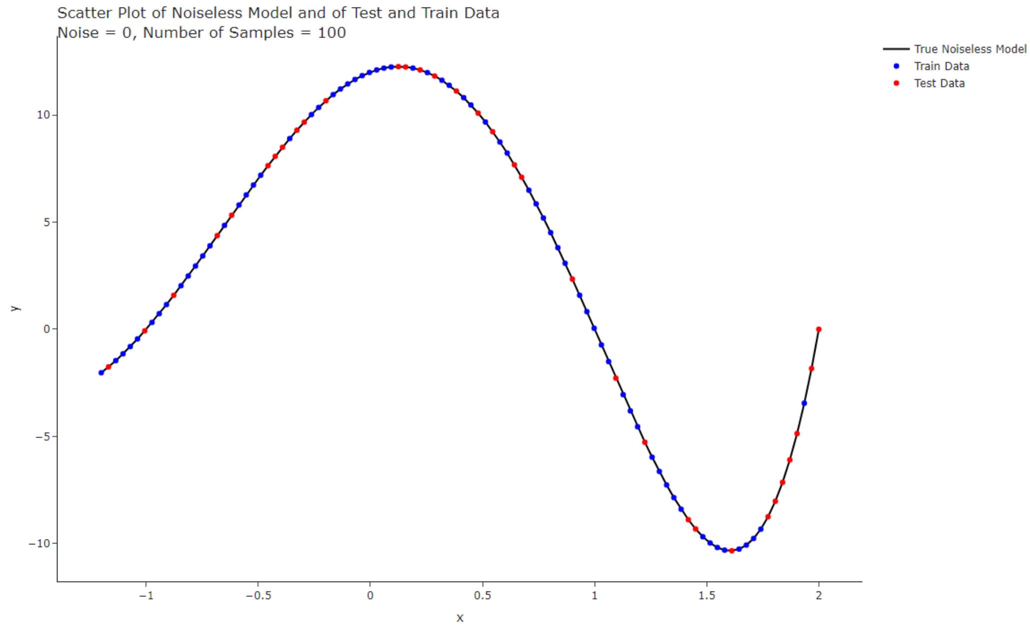
שאלה 3

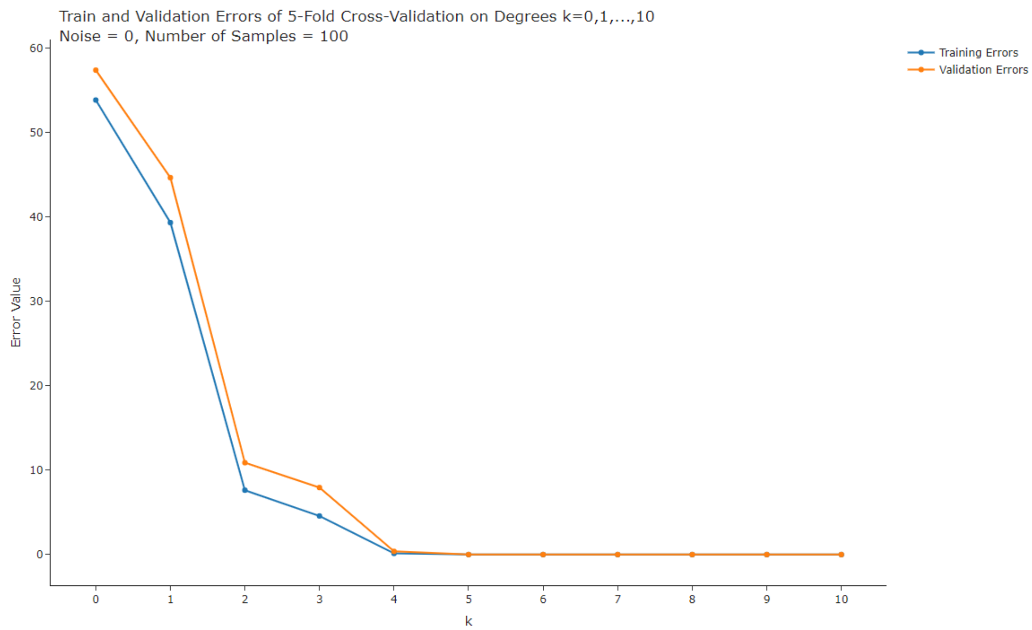
הערך של $k \in \{0,1,\ldots,10\}$ שעבורו התקבלה שגיאת הולידציה הנמוכה ביותר היה $k^*=4$, והשגיאה של המודל על סט האימון המלא היתה 29.09.

ערך השגיאה שהתקבל על סט הולידציה במהלך הריצה של ה־Cross-Validation על $k=4$ היה 22.60 - זהו ערך נמוך יותר מזה שהתקבל על סט האימון המלא.

שאלה 4

חזרה על התהליך של 3 השאלות הקודמות עבור רעש= 0:

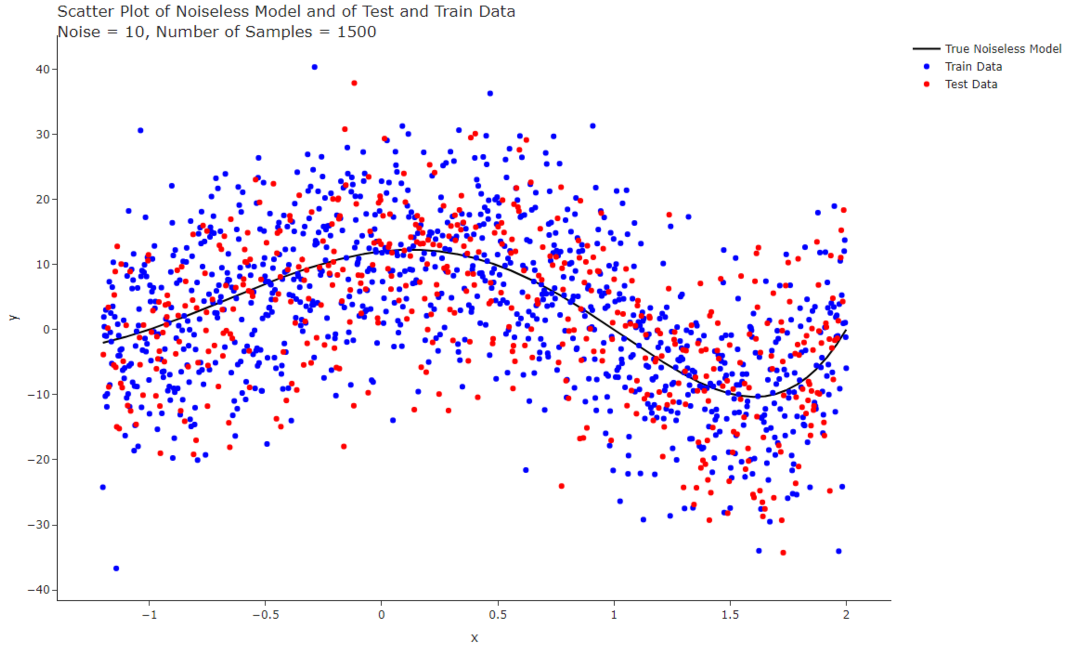


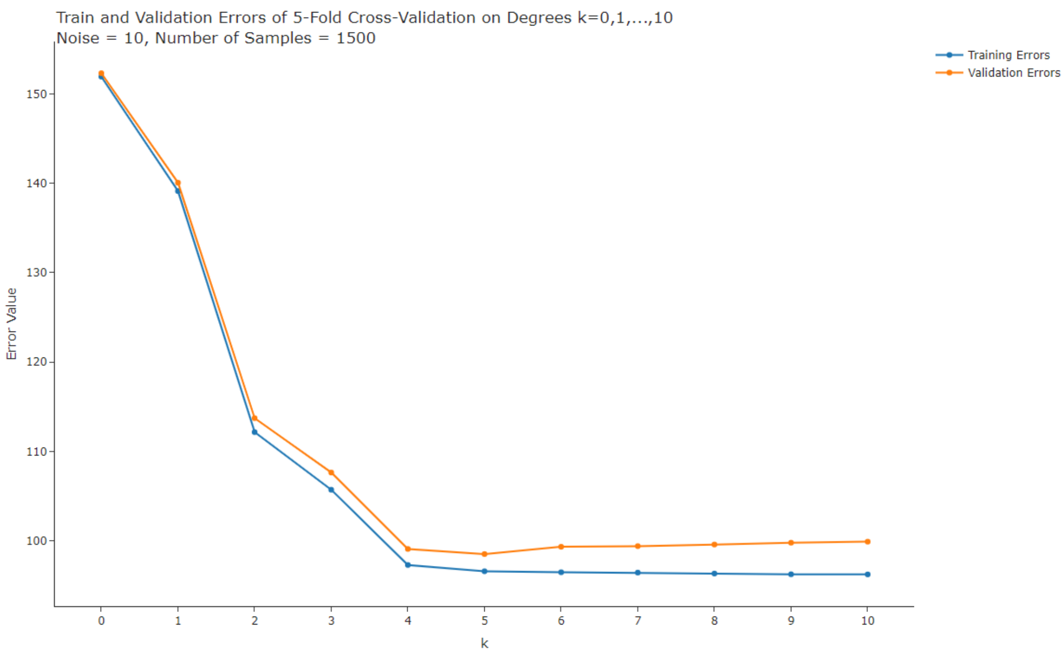


כאשר ערך הרעש הוא 0, רואים בגרף השני שכאשר עושים Cross-Validation עבור $k \geqslant 5$ השגיאה על סט האימון וגם על סט הולידציה היא 0. זה מסתדר עם כך שהפולינום שממנו נדגמו הנקודות הוא מדרגה 5, ושהן נדגמו ללא רעש. כלומר, אין סכנה שהמודל יעשה overfitting כאשר הוא משתמש בערכי k גדולים מ־5, כי הנקודות שהוא מתאמן עליהן ובודק את עצמו עליהן נמצאות בדיוק על הפולינום האמיתי.
הדרגה שעבורה התקבלה השגיאה המינימלית היתה כמובן $k^* = 5$, כמו הדרגה האמיתית של הפולינום, והשגיאה שהתקבלה היא 0.

שאלה 5

חזרה על התהליך של 3 השאלות הקודמות עבור רעש= 10 ועבור 1500 נקודות:





ניתן לראות בגרף השני שעבור 1500 דגימות השגיאה על סט האימון והשגיאה על סט הולידציה קרובות יותר לעומת השגיאות שקיבלנו עבור 100 דגימות בלבד.

בנוסף, הדרגה שהשיגה שגיאה מינימלית בתהליך ה־Cross-Validation היתה $k^* = 5$ (כמו של הפולינום האמיתי) והשגיאה שהתקבלה על סט **הולידציה** היתה 98.5. השגיאה שהתקבלה על סט **המבחן המלא** היתה 97.3. כלומר היחס בין שגיאת הולידציה לשגיאת המבחן הכללית קטן יותר עבור מספר גדול יותר של נקודות.

המגמה הזאת נתמכת ע"י החוק החלש של המספרים הגדולים, לפיו הערך של השגיאה האמפירית שואף בהסתברות לשגיאת ההכללה כאשר מספר הדגימות גדל, כי אומדים קונסיסטנטיים מקיימים אותו ואנחנו יודעים שהאומד שאיתו אנחנו עושים PolynomialFitting הוא קונסיסטנטי.

Choosing Regularizaion Parameters Using Cross Validation

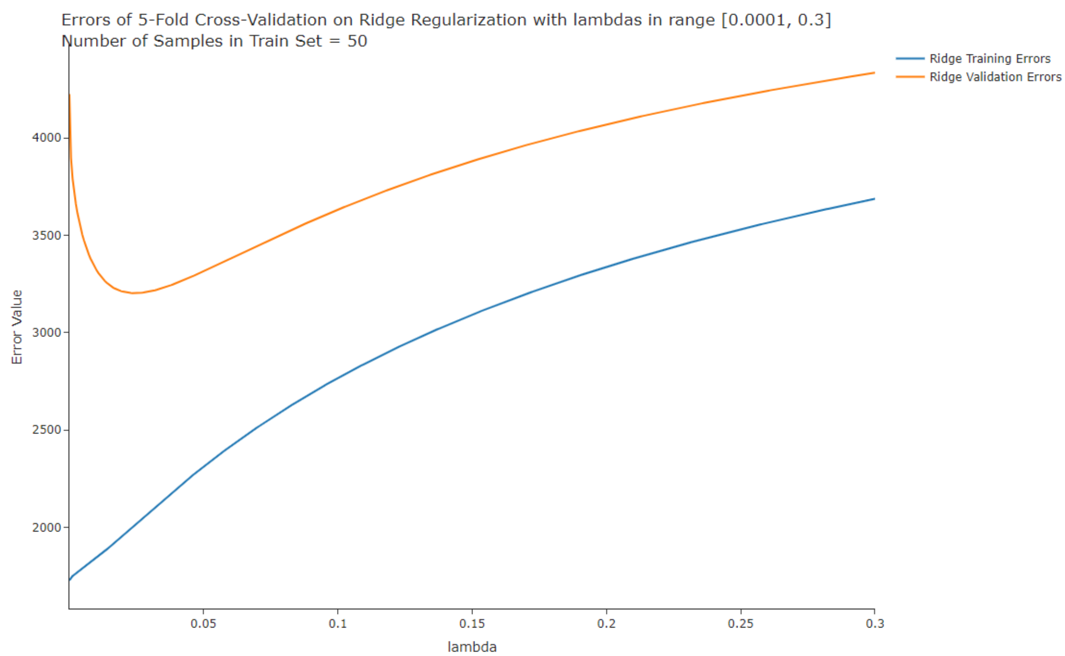
שאלה 6

בקוד

שאלה 7

הטווח של ערכי פרמטר הרגולריזציה λ שנראה לי הכי משמעותי עבור תהליך ה־Cross-Validation היה $[0,1]$ עבור ה־Lasso-Regression ו־$[0,0.3]$ עבור ה־Ridge-Regression. הסיבה לכך היא שבטווחים הללו ראיתי שיש מגמה "מעניינת" בערכי השגיאה של המודלים, כלומר שינוי במגמה. עבור ערכי λ גדולים מהטווח עבור כל אחד מהמודלים מגמת השגיאה היא פשוט עליה מונוטונית, וזה פחות רלוונטי עבור ה־Cross-Validation.

גרף המציג את שגיאת האימון והולידציה בתהליך ה־Cross-Validation על מודל Ridge Regression:

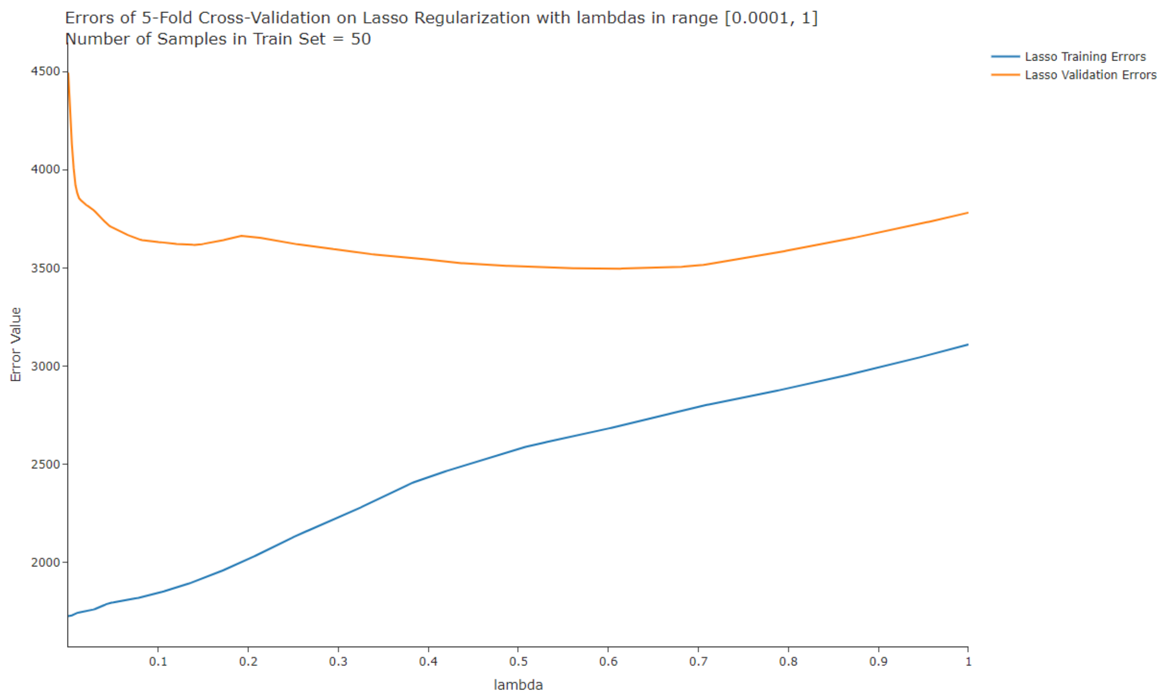


ניתן לראות בגרף שכאשר הערך של פרמטר הרגולריזציה λ גדל, השגיאה על סט האימון גדלה. המגמה הזו ברורה - עבור ערכי λ גדולים יותר המודל "מעניש" יותר כללי החלטה מורכבים, ולכן היכולת שלו להתאים את עצמו לסט האימון קטנה.

לעומת זאת, עבור ערכי λ בין 0 ל־0.03 בערך ניתן לראות שהשגיאה על סט הולידציה דווקא קטנה. ניתן להסיק מכך שהגבלת המורכבות של כלל ההחלטה שהמודל מייצר מאפשרת לו עבור הערכים שללו להכליל טוב יותר.

עבור ערכי λ גדולים יותר השגיאה על סט הולידציה עולה בדיוק כמו השגיאה על סט האימון. ניתן להסיק מכך שהמגבלה על מורכבות כלל ההחלטה היא קשוחה מדי ולא מאפשרת למודל ללמוד כמו שצריך את הדאטא.

גרף המציג את שגיאת האימון והולידציה התהליך ה־Cross-Validation על מודל Lasso-Regression:



ניתן לראות שהשגיאה על סט האימון גדלה ככל שהערך של λ גדל, וניתן להסביר זאת בדיוק כמו שהסברתי במקרה הקודם.
בנוגע לסט הולידציה - ניתן לראות שהשגיאה עליו היא במגמת ירידה עבור ערכי λ בין 0 ל־0.5 בערך, ושלאחר מכן השגיאה מתחילה לעלות עד שמגמת העליה שלה נהיית דומה לזו המתקבלת על סט האימון. ההסבר לכך זהה לזה שנתתי במקרה הקודם.

ההבדל בין שני האלגוריתמים:

שגיאת הולידציה המינימלית שהתקבלה בשני המודלים היא בטווח $3500-3200$. ערכי ה־λ שהביאו את המינימום **שונים** - עבור אלגוריתם ה־Ridge ערך ה־λ האופטימלי היה ~ 0.025 ולעומת זאת עבור אלגוריתם ה־Lasso הערך האופטימלי היה ~ 0.597. הסיבה לכך יכולה להיות העובדה שעבור וקטור משקולות $w \in \mathbb{R}^d$, Ridge עושה שימוש בנורמת ℓ_2 **בריבוע** (כלומר מחשב את $\|w\|_2^2$). לכן הוא נותן "משקל גדול יותר" לקואורדינטות של w (כי הוא מעלה אותן בריבוע). לעומת זאת Lasso עושה שימוש בנורמת ℓ_1, ולכן פחות מעניש קואורדינטות גדולות. ההבדל הזה מוביל לכך שערכי λ קטנים יחסית יחד עם נורמת ℓ_2 בריבוע יוצרים פרמטר רגולריזציה ש"מעניש" את ה"גודל" של w באופן דומה לערכי λ גדולים יותר יחד עם נורמת ℓ_1.

שאלה 8

```
Result of Cross-Validation on Ridge,Lasso and Least-Squares Regression Models:
Ridge Regression Model:
    Best Lambda = 0.024
    Error = 3247.307
Lasso Regression Model:
    Best Lambda = 0.597
    Error = 3641.194
Least-Squares Regression Model:
    Best Lambda = 0
    Error = 3612.25
```