תרגיל 5 IML - מאי ביבי

שאלה 1:

א. נוכיח שעבור $h^* \in \underset{\mathcal{H}_k}{\mathrm{ERM}}\left(S_{\mathrm{all}}\right)$ מתקיים בהסתברות של לפחות $1-\delta$:

$$L\left(h^*\right) \leq \min_{h \in \mathcal{H}_k} L\left(h\right) + \sqrt{\frac{2\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{m}}$$

בתרגול הוכחנו שלכל היפותזה h ולכל δ מתקיים

$$\mathbb{P}\left(|L_{S_{all}}(h) - L_{\mathcal{D}}(h)| \leqslant \sqrt{\frac{\ln(2/\delta)}{2m}}\right) \geqslant 1-\delta$$

בפרט עבור כל $h_i \in \mathcal{H}_k$, עבור $\frac{\delta}{|\mathcal{H}_k|}$ מתקיים

$$\mathbb{P}\left(|L_{S_{all}}\left(h_i\right) - L_{\mathcal{D}}\left(h_i\right)| \geq \sqrt{\frac{\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{2m}}\right) \leq \frac{\delta}{|\mathcal{H}_k|}$$

מחסם האיחוד, לכל $h_i \in \mathcal{H}_k$

$$\mathbb{P}\left(|L_{S_{all}}\left(h_i\right) - L_{\mathcal{D}}\left(h_i\right)| \geq \sqrt{\frac{2\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{2m}}\right) \leq \mathbb{P}\left(\bigcup_{h_i \in \mathcal{H}_k} |L_{S_{all}}\left(h_i\right) - L_{\mathcal{D}}\left(h_i\right)| \geq \sqrt{\frac{\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{2m}}\right)$$

$$\leq \sum_{h_i \in |\mathcal{H}_k|} \mathbb{P}\left(|L_{S_{all}}\left(h_i\right) - L_{\mathcal{D}}\left(h_i\right)| \geq \sqrt{\frac{\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{2m}}\right) \leq \sum_{h_i \in |\mathcal{H}_k|} \frac{\delta}{|\mathcal{H}_k|} = |\mathcal{H}_k| \frac{\delta}{|\mathcal{H}_k|} = \delta$$

סה"כ קיבלנו שלכל $h_i \in \mathcal{H}_k$ מתקיים עם הסתברות של לפחות $1-\delta$:

$$|L_{S_{all}}\left(h_i\right) - L_{\mathcal{D}}\left(h_i\right)| \geq \sqrt{\frac{2\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{2m}}$$

לכן עבור $h_i \in \mathcal{H}_k$, מתקיים עם הסתברות של לפחות $1-\delta$:

$$L_{\mathcal{D}}\left(h^*\right) \leqslant L_{S_{all}}\left(h^*\right) + \sqrt{\frac{\ln\left(2\left|\mathcal{H}_k\right|/\delta\right)}{2m}} \leqslant L_{S_{all}}\left(h_i\right) + \sqrt{\frac{\ln\left(2\left|\mathcal{H}_k\right|/\delta\right)}{2m}}$$

$$L_{\mathcal{D}}\left(h_i\right) + 2\sqrt{\frac{\ln\left(2\left|\mathcal{H}_k\right|/\delta\right)}{2m}} = L_{\mathcal{D}}\left(h_i\right) + \sqrt{\frac{2\ln\left(2\left|\mathcal{H}_k\right|/\delta\right)}{m}}$$

ובפרט עבור $h_i = \min_{h \in \mathcal{H}_k} L\left(h\right)$ נקבל

$$\mathbb{P}\left(L\left(h^*\right) \leq \min_{h \in \mathcal{H}_k} L\left(h\right) + \sqrt{\frac{2\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{m}}\right) \geqslant 1-\delta$$

כדרוש.

ב. מהסעיף הקודם

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

ולכל $i \in [k]$ מתקיים

$$\mathbb{P}\left(L_{\mathcal{D}}\left(h_i\right) \leqslant \min_{h\in\mathcal{H}_i} L_{\mathcal{D}}(h) + \sqrt{\frac{2\ln\left(4\left|\mathcal{H}_i\right|/\delta\right)}{(1-\alpha)m}}\right) \geqslant 1 - \frac{\delta}{2}$$

לכן, עם הסתברות של לפחות $1-\delta$:

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

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

שאלה 2:

כיוון ש־ $X^TX = I_d$, X הפיכה, ולכן הפתרון ל־LS הוא

$$\hat{w}^{LS} = \left(X^TX\right)^{-1} X^Ty = (=I_n)^{-1} X^Ty = X^Ty$$

א. בהרצאה ראינו שניתן להגיע למשוואה הבאה בדרך לפתרון ridge:

$$X^Ty = \left(X^TX + \lambda I\right)\hat{w}^{ridge} = (I+\lambda I)\,\hat{w}^{ridge}\hat{w}^{ridge}$$

$$\Longrightarrow \hat{w}^{ridge} = \frac{1}{1+\lambda}X^Ty = \frac{\hat{w}^{LS}}{1+\lambda}$$

ב. יהי $w \in \mathbb{R}^d$. מתקיים

$$\|y - Xw\|^2 = \left\|X^Ty - X^TXw\right\|^2 = \left\|\hat{w}^{LS} - w\right\|^2 = \sum_{i=1}^{d}\left(\hat{w}_i^{LS} - w_i\right)^2$$

בעיית ה־subset selection היא

$$\arg\min_{w\in\mathbb{R}^d} \|y-Xw\|^2 + \lambda\|w\|_0 = \arg\min_{w\in\mathbb{R}^d}\sum_{i=1}^{d}\left(\hat{w}_i^{LS}-w_i\right)^2 + \lambda\|w_i\|_0 = \arg\min_{w\in\mathbb{R}^d}\sum_{i=1}^{d}\left(\hat{w}_i^{LS}-w_i\right)^2 + \lambda\cdot\mathbf{1}_{w_i\neq 0}$$

עבור $i \in [d]$

$$\arg\min_{w_i\in\mathbb{R}}\left(\hat{w}_i^{LS}-w_i\right)^2 + \lambda\cdot\mathbf{1}_{w_i\neq 0}$$

נשים לב שעבור $\left(\hat{w}_i^{LS}\right)^2 > \lambda$ אנחנו נרצה $w_i = \hat{w}_i^{LS}$ ואז

$$\left(\hat{w}_i^{LS}-w_i\right)^2 + \lambda\cdot\mathbf{1}_{w_i\neq 0} = \lambda$$

וזה באמת המינימום במקרה זה. עבור $\left(\hat{w}_i^{LS}\right)^2 \leq \lambda$, נרצה $w_i = 0$, ואז נקבל

$$\left(\hat{w}_i^{LS}-w_i\right)^2 + \lambda\cdot\mathbf{1}_{w_i\neq 0} = \hat{w}_i^{LS}$$

וזה באמת המינימום במקרה זה. ונשים לב ששני המרים שתיארו הם בדיוק:

$$\eta_{\sqrt{\lambda}}\left(\hat{w}^{LS}\right)_i = \begin{cases}\hat{w}_i^{LS} & \left|\hat{w}_i^{LS}\right| \geq \sqrt{\lambda}\\ 0 & \text{else}\end{cases}$$

שאלה 3:

א. בהרצאה ראינו:

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

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

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

מתקיים:

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

כדרוש

ב. A_λ לא מקרית ולכן

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

עבור $\lambda > 0$ מתקיים $A_\lambda w = \left(X^T X + \lambda I_d\right)^{-1} \left(X^T X\right) w \neq w$.

ג. אנו יודעים שמתקיים:

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

מתקיים

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

ד. בהרצאה ראינו שמתקיים

$$\mathbb{E}\left(\|\hat{y} - y^*\|^2\right) = \mathbb{E}\left(\|\hat{y} - \overline{y}\|^2\right) + \|\overline{y} - y^*\|^2 = \mathrm{Var}\left(\hat{y}\right) + \mathrm{Bias}\left(\hat{y}\right)^2$$

כאשר y^* הם ערכי האמת, $\hat{y}$ הוא הפתרון שבוחר האלגוריתם ו־ $\overline{y} = \mathbb{E}\left(\hat{y}\right)$.
אצלנו מתקיים $y^* = w$, $\hat{y} = \hat{w}(\lambda)$, $\overline{y} = \mathbb{E}\left(\hat{w}(\lambda)\right)$. נקבל

$$\mathrm{Bias}(\lambda)^2 = \|\mathbb{E}\left(\hat{w}(\lambda)\right) - w\|^2 = \|A_\lambda w - w\|^2 = \|\left(A_\lambda - I\right) w\|^2$$

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

לכן

$$\mathrm{MSE}(\lambda) = \mathrm{Bias}(\lambda)^2 + \mathrm{Var}(\lambda) = \|\left(A_\lambda - I\right) w\|^2 + \sigma^2 \mathrm{Tr}\left(A_\lambda \left(X^T X\right)^{-1} A_\lambda^T\right)$$

נגזור ונציב $\lambda = 0$:

$$\left.\frac{d}{d\lambda} \mathrm{Bias}(\lambda)^2\right|_{\lambda=0} = \left.\frac{d}{d\lambda}\left(\|\left(A_\lambda - I\right) w\|^2\right)\right|_{\lambda=0} = \left.\frac{d}{d\lambda}\left(\sum_i \sum_j \left(\left(A_\lambda - I\right)_{i,j} w_j\right)^2\right)\right|_{\lambda=0}$$

$$= 2\sum_i \sum_j (A_\lambda - I)_{i,j} w_j \Big|_{\lambda=0} \cdot \frac{d}{d\lambda} \sum_j (A_\lambda - I)_{i,j} w_j \Big|_{\lambda=0} = 2\sum_i \sum_j 0_{i,j} w_j \Big|_{\lambda=0} \cdot \frac{d}{d\lambda} \sum_j (A_\lambda - I)_{i,j} w_j \Big|_{\lambda=0} = 0$$

$$\frac{d}{d\lambda}\mathrm{Var}(\lambda)\Big|_{\lambda=0} = \frac{d}{d\lambda}\sigma^2 \mathrm{Tr}\left(A_\lambda \left(X^T X\right)^{-1} A_\lambda^T\right)\Big|_{\lambda=0} = \frac{d}{d\lambda}\sigma^2 \sum_i \left(A_\lambda \left(X^T X\right)^{-1} A_\lambda^T\right)_{i,i}\Big|_{\lambda=0}$$

$$= \frac{d}{d\lambda}\sigma^2 \sum_i \left(\left(X^T X + \lambda I_d\right)^{-1} \left(X^T X\right) \left(X^T X\right)^{-1} \left(\left(X^T X + \lambda I_d\right)^{-1} \left(X^T X\right)\right)^T\right)_{i,i}\Big|_{\lambda=0}$$

$$= \frac{d}{d\lambda}\sigma^2 \sum_i \left(\left(X^T X + \lambda I_d\right)^{-1} \left(X^T X\right)^T \left(X^T X + \lambda I_d\right)^{-1T}\right)_{i,i}\Big|_{\lambda=0}$$

$$= \frac{d}{d\lambda}\sigma^2 \sum_i \left(\left(X^T X + \lambda I_d\right)^{-1} \left(X^T X\right) \left(X^T X + \lambda I_d\right)^{-1}\right)_{i,i}\Big|_{\lambda=0} \ldots < 0$$

לא כל כך ידעתי איך להמשיך..

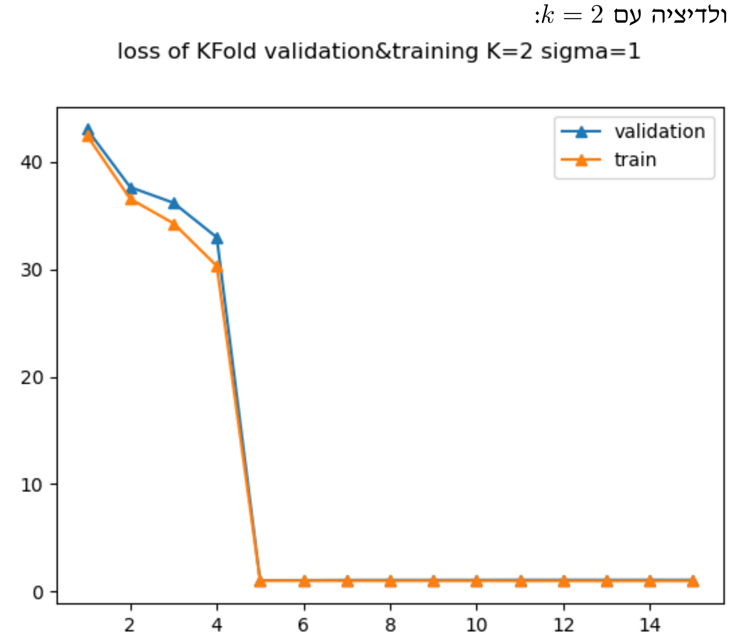
$$\frac{d}{d\lambda}\mathrm{MSE}(\lambda)\Big|_{\lambda=0} = \frac{d}{d\lambda}\Big|_{\lambda=0} \mathrm{Bias}(\lambda)^2 + \frac{d}{d\lambda}\mathrm{Var}(\lambda)\Big|_{\lambda=0} < 0$$

ה. הראינו שהנגזרת של $\mathrm{MSE}(\lambda)$ בנקודה 0 היא שלילית, כלומר הפונקציה בירידה ולכן קיים $\lambda > 0$ עבורו $\mathrm{MSE}(\lambda) < \mathrm{MSE}(0)$.

שאלה 4:

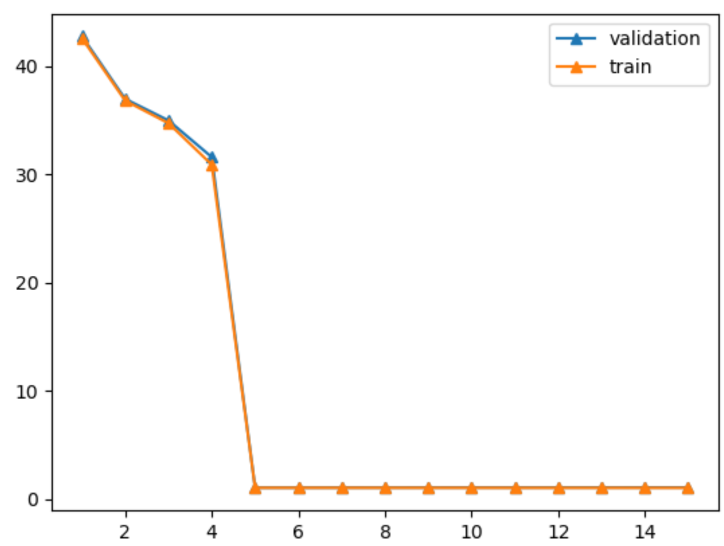
רעש עם $\sigma = 1$:

K-fold ולדיציה עם $k = 2$:



K-fold ולדיציה עם $k = 5$:

loss of KFold validation&training K=5 sigma=1

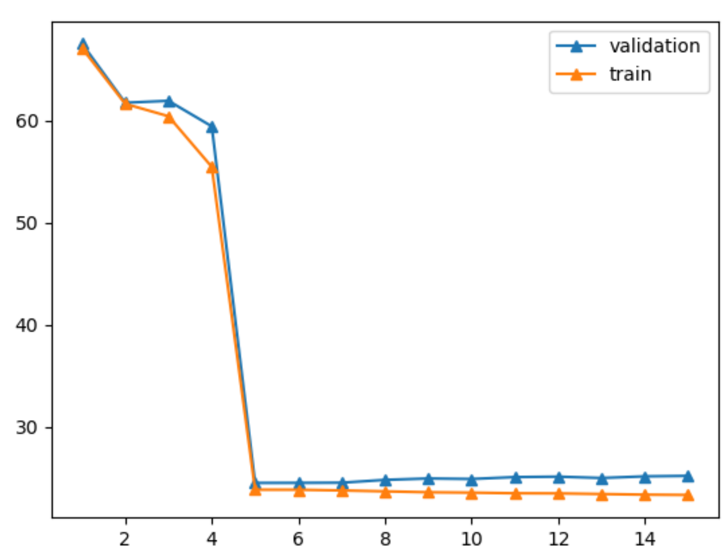


הדרגה עם השגיאה הכי נמוכה היא $d = 5$, וזה הגיוני כי הפולינום שלנו מדרגה 5.
שגיאת הולדיציה עבור $d = 5$ היא 1.0435
שגיאת ה־test עבור $d = 5$ היא 22.1962

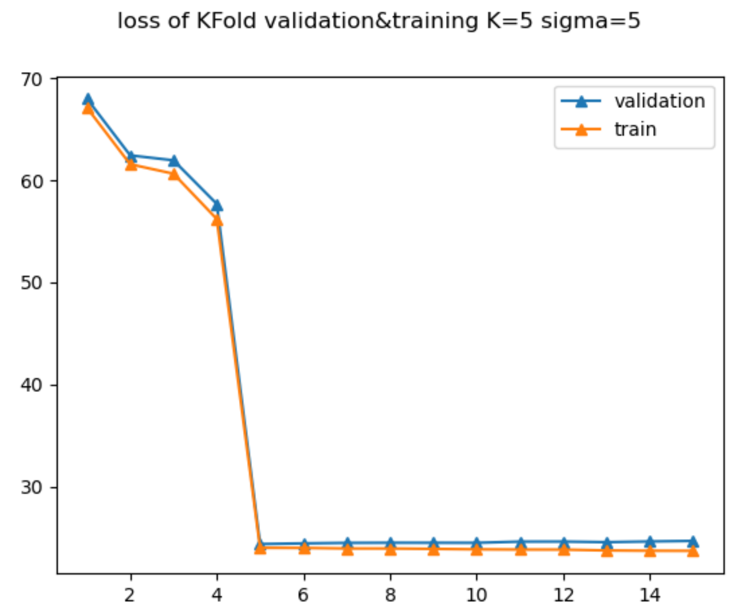
רעש עם $\sigma = 5$:

K-fold ולדיציה עם $k = 2$:

loss of KFold validation&training K=2 sigma=5



K-fold ולדיציה עם $k = 5$:

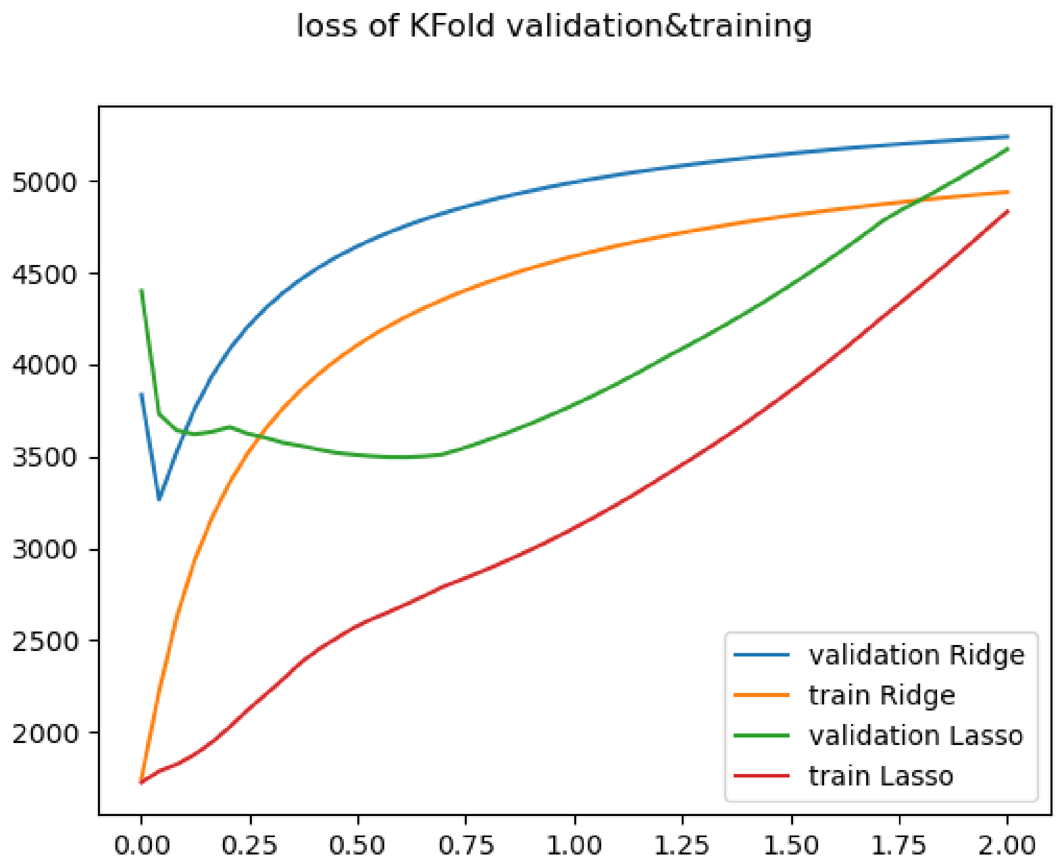


הדרגה עם השגיאה הכי נמוכה היא $d = 5$, וזה הגיוני כי הפולינום שלנו מדרגה 5.
שגיאת הולדיציה עבור $d = 5$ היא 24.5348
שגיאת ה־test עבור $d = 5$ היא 107.4876

כצפוי לדאטה עם יותר רעש, קיבלתי שגיאות גדולות יותר. אך עדיין קיבלנו $d = 5$

שאלה 5:

שיחקתי עם הערכים כדי לראות מתי השגיאה נמוכה יותר. בדקתי ערכי λ קטנים מאוד, כלומר מעט רגולריזציה, וגם ערכי $\lambda > 1$, הרבה רגולריזציה. לכן בדקתי ערכים בין 0 ל־2.




```
best lambda for ridge: 0.04179591836734694
best lambda for lasso: 0.612938775510204
```

```
Ridge test error:  3191.397109707721
Lasso test error:  3652.376475971041
LinearRegression test error:  3612.249688324898
```

ניתן לראות כי ה־λ הכי טובה עבור ridge היא קטנה מאוד, כלומר אנחנו מאוד מתקרבים לרגרסיה לינארית רגילה. אך השגיאה של ridge קטנה יותר משל הרגרסיה הרגילה, כלומר ridge כן משפר את השגיאה.