IML (67577) | תרגיל 5

שם: יואב שפירא| ת`ז: 312492838

8 ביוני 2022

תאורטי

.1

(א) בתרגיל ראינו שעבור מטריצה $X' \in \mathbb{R}^{d\times m}$ (כלומר טרנספוס של X שלנו) מתקיים ש

$$\hat{w}_\lambda = \left(XX^T + \lambda I_d\right)^{-1} Xy$$

לכן בשאלה זו נרצה להגיע להראות ש:

$$\hat{w}_\lambda = \left(X^TX + \lambda I_d\right)^{-1} X^Ty = A_\lambda \hat{w}$$

נזכיר כי $\hat{w} = X^{\dagger^T} y$. בתרגיל 2 הוכחנו ש אם X^TX הפיכה אז מתקיים $X^{\dagger^T} = \left(X^TX\right)^{-1} X^T$. נציב את זה ואת הביטוי של A_λ ונקבל:

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

כדרוש.

(ב) נחשב את התוחלת:

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

ידוע ש $\mathbb{E}\left[\hat{w}\right] = w$ ושאר הביטוי הוא קבוע ביחס ל w ולכן:

$$\mathbb{E}\left[\hat{w}_\lambda\right] = \left(X^TX + \lambda I_d\right)^{-1} \left(X^TX\right) \hat{w}$$

נבחין שהביטוי הזה שווה ל$\hat{w}$אם"ם $\lambda = 0$, שכן אז הביטוי $\left(X^TX\right)^{-1} \left(X^TX\right)$ מצטמצם.

(ג) בהתאם לסעיף הקודם נחשב:

$$Var\left(\hat{w}_{\lambda}\right)=Var\left(A_{\lambda}\hat{w}\right)$$

נוציא את A_{λ} לפי כללי הווריאנס:

$$Var\left(\hat{w}_{\lambda}\right)=A_{\lambda}Var\left(\hat{w}\right)A_{\lambda}^{T}$$

בהתאם לרמז נציב את הווריאנס של $\hat{w}$:

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

נסדר ונקבל את הדרוש:

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

(ד) בסעיף הזה לצורך נוחות אסמן תוחלת ע"י $\langle\cdot\rangle$ ולא ע"י $\mathbb{E}$.
נרצה למצוא פירוק ש $bias^{2}+var$ עבור הביטוי הבא:

$$MSE\left(\hat{w}_{\lambda}\right)=\left\langle ||\hat{w}_{\lambda}-w||^{2}\right\rangle$$

נוסיף ונחסר מהביטוי הפנימי את $\langle\hat{w}_{\lambda}\rangle$ ואז נפתח את הביטוי לאט לאט:

$$\begin{aligned}
MSE\left(\hat{w}_{\lambda}\right) &= \left\langle ||w-\langle\hat{w}_{\lambda}\rangle+\langle\hat{w}_{\lambda}\rangle-\hat{w}_{\lambda}||^{2}\right\rangle \\
&= \left\langle\left(w-\langle\hat{w}_{\lambda}\rangle+\langle\hat{w}_{\lambda}\rangle-\hat{w}_{\lambda}\right)^{T}\left(w-\langle\hat{w}_{\lambda}\rangle+\langle\hat{w}_{\lambda}\rangle-\hat{w}_{\lambda}\right)\right\rangle \\
&= \langle w^{T}w+2\langle\hat{w}_{\lambda}\rangle^{2}+\hat{w}_{\lambda}^{T}\hat{w}_{\lambda}-2w^{T}\langle\hat{w}_{\lambda}\rangle+2w^{T}\langle\hat{w}_{\lambda}\rangle \\
&\quad -2w^{T}\hat{w}_{\lambda}-2w\langle\hat{w}_{\lambda}\rangle^{2}+2\hat{w}_{\lambda}^{T}\langle\hat{w}_{\lambda}\rangle-2\hat{w}_{\lambda}^{T}\langle\hat{w}_{\lambda}\rangle\rangle \\
&= \left\langle ||w-\langle\hat{w}_{\lambda}\rangle||^{2}\right\rangle+\left\langle ||\langle\hat{w}_{\lambda}\rangle-\hat{w}_{\lambda}||^{2}\right\rangle+2\left\langle\left(w-\langle\hat{w}_{\lambda}\rangle\right)\left(\langle\langle\hat{w}_{\lambda}\rangle-\hat{w}_{\lambda}\rangle\right)\right\rangle
\end{aligned}$$

כשבשלב האחרון (בצבעים) קיבצתי את האיברים והפרדתי מליניאריות התוחלת. נזכיר כי $\hat{w}_{\lambda}=A_{\lambda}\hat{w}$ וכעת נסתכל על האיבר הצבוע בשחור:

$$\begin{aligned}
2\left\langle\left(w-\langle\hat{w}_{\lambda}\rangle\right)\left(\langle\langle\hat{w}_{\lambda}\rangle-\hat{w}_{\lambda}\rangle\right)\right\rangle &= 2\left\langle\left(w-\langle\hat{w}\rangle\right)\left(\langle\langle A_{\lambda}\hat{w}\rangle-A_{\lambda}\hat{w}\rangle\right)\right\rangle \\
&= 2\left\langle\left(w-A_{\lambda}\langle\hat{w}\rangle\right)\left(\langle A_{\lambda}\langle\hat{w}\rangle-A_{\lambda}\hat{w}\rangle\right)\right\rangle \\
&= 2\left\langle\left(w-A_{\lambda}\langle\hat{w}\rangle\right)\cdot 0\right\rangle=0
\end{aligned}$$

נשים לב שהאיבר האדום הוא למעשה $Var(\hat{w}_{\lambda})$ שמצאנו אותו בסעיף הקודם.
נבחן את האיבר האחרון, שהוא ה$bias^{2}(\hat{w}_{\lambda})$:

$$bias^{2}(\hat{w}_{\lambda})=\left\langle ||w-\langle\hat{w}_{\lambda}\rangle||^{2}\right\rangle=\left\langle ||w-A_{\lambda}\langle\hat{w}\rangle||^{2}\right\rangle$$

ניזכר ש $w = \langle\hat{w}\rangle$ ונציב. נשים לב שכעת התוחלת לא תלויה בשום איבר בפנים ונוכל להיפטר ממנה. נקבל:

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

נבחין שאנחנו מעוניינים בשונות של וקטור, וכרגע יש לנו את מטריצת הקו־ווריאנס שלו מסעיף ב. השונות, לא השונות המשותפת, היא בדיוק העקבה של מטריצת הקווריאנס ולכן:

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

קיבלנו את הפירוק שרצינו:

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

(ה) נתחיל באבחנה הבאה: עבור $\lambda = 0$ מתקיים ש $A_\lambda = I$:

$$A_\lambda = \left(X^T X - \lambda I\right)^{-1} X^T X = \left(X^T X\right)^{-1} X^T X = I$$

נמשיך בלגזור את $bias^2$:

$$\frac{\partial}{\partial\lambda} bias^2(\hat{w}_\lambda) = \frac{\partial}{\partial\lambda} \underbrace{w^T \left(A_\lambda - I\right)^T}_{X} \underbrace{\left(A_\lambda - I\right) w}_{Y}$$

נגזור לפי כלל השרשת:

$$\begin{aligned} \frac{\partial}{\partial\lambda} bias^2(\hat{w}_\lambda) &= \left(\frac{\partial}{\partial\lambda} X\right) Y + X \left(\frac{\partial}{\partial\lambda} Y\right) \\ &= \left(\frac{\partial}{\partial\lambda} w^T \left(A_\lambda - I\right)^T\right) \left(A_\lambda - I\right) w + w^T \left(A_\lambda - I\right)^T \left(\frac{\partial}{\partial\lambda} \left(A_\lambda - I\right) w\right) \end{aligned}$$

נסתפק בזה, כי נשים לב שאחרי כל הגזירות נצטרך להציב $\lambda = 0$ והביטוי הבא יתאפס: $\left(A_\lambda - I\right) = 0$ (לשים לב שהביטוי הזה נשאר, כמו שהוא ולא גזור ־ גם בביטוי הנגזרת הסופי). מכיוון שכך כל הביטוי שווה ל0.

$$\frac{\partial}{\partial\lambda} bias^2(\hat{w}_\lambda) = 0!!$$

נגזור את הווריאנס בנקודה $\lambda = 0$. לפי חוקי עקבה:

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

לפי חוקי גזירה של עקבה:

$$\frac{\partial}{\partial\lambda}_{|\lambda=0} Var(\hat{w}_\lambda) = \frac{\partial}{\partial\lambda}_{|\lambda=0} \sigma^2 Tr\left(A_\lambda \left(X^T X\right)^{-1} A_\lambda^T\right) = \sigma^2 Tr \frac{\partial}{\partial\lambda}_{|\lambda=0} \left(\underbrace{A_\lambda}_{X} \underbrace{\left(X^T X\right)^{-1} A_\lambda^T}_{Y}\right)$$

נגזור שוב לפי כלל השרשרת:

$$\frac{\partial}{\partial\lambda_{|\lambda=0}}Var(\hat{w}_\lambda)=\sigma^2Tr\left[\left(\frac{\partial}{\partial\lambda_{|\lambda=0}}X\right)Y+X\left(\frac{\partial}{\partial\lambda_{|\lambda=0}}Y\right)\right]$$
$$=\sigma^2Tr\left[\left(\frac{\partial}{\partial\lambda_{|\lambda=0}}A_\lambda\right)\left(X^TX\right)^{-1}A_\lambda^T+A_\lambda\left(\frac{\partial}{\partial\lambda_{|\lambda=0}}\left(X^TX\right)^{-1}A_\lambda^T\right)\right]$$

כעת נגזור את A_λ. ניעזר בחוקי גזירה שפורסמו בפורום:

$$\frac{\partial}{\partial\lambda_{|\lambda=0}}A_\lambda=\frac{\partial}{\partial\lambda_{|\lambda=0}}\left(X^TX-\lambda I\right)^{-1}X^TX=-\left(X^TX-\lambda I\right)^{-1}I\left(X^TX-\lambda I\right)^{-1}\left(X^TX\right)$$

נציב $\lambda=0$ כדי לקבל את הנגזרת בנקודה:

$$\frac{\partial}{\partial\lambda_{|\lambda=0}}A_\lambda=-\left(X^TX-0\cdot I\right)^{-1}I\left(X^TX-0\cdot I\right)^{-1}\left(X^TX\right)$$
$$=-\left(X^TX\right)^{-1}\left(X^TX\right)^{-1}\left(X^TX\right)=-\left(X^TX\right)^{-1}$$

הנגזרת של A^T זהה מטעמי סימטריה של A. נציב כעת בנגזרת של השונות:

$$\frac{\partial}{\partial\lambda_{|\lambda=0}}Var(\hat{w}_\lambda)=\sigma^2Tr\left[-\left(X^TX\right)^{-1}\left(X^TX\right)^{-1}A_\lambda^T+A_\lambda\left(\left(X^TX\right)^{-1}-\left(X^TX\right)^{-1}\right)\right]$$
$$=\sigma^2Tr\left[\left(X^TX\right)^{-1}\left(X^TX\right)^{-1}I+I\left(\left(X^TX\right)^{-1}-\left(X^TX\right)^{-1}\right)\right]$$
$$=\sigma^2Tr\left[-2\left(X^TX\right)^{-1}\left(X^TX\right)^{-1}\right]=-2\sigma^2Tr\left[\left(X^TX\right)^{-1}\left(X^TX\right)^{-1}\right]$$

מכיוון שX^TX היא סימטרית, אז $\left(X^TX\right)^{-1}$ היא סימטרית כלומר $\left(X^TX\right)^{-1}=\left(\left(X^TX\right)^{-1}\right)^T$, כלומר המטריצה שבתוך העקבה היא PSD, נסמן אותה בP:

$$P=\left(X^TX\right)^{-1}\left(X^TX\right)^{-1}=\left(X^TX\right)^{-1}\left(\left(X^TX\right)^{-1}\right)^T$$

נבחין שמכיוון ש X^TX היא הפיכה, אז כל הע"ע שלה שונים מ0, ולכן כל הע"ע העצמיים של P שונים מאפס. ניזכר בכך ש $Tr(P)$ היא בדיוק שווה לסכום של כל הע"ע של P, ובזה סיימנו: $Tr(P)>0$, ולכן

$$\frac{\partial}{\partial\lambda_{|\lambda=0}}Var(\hat{w}_\lambda)=-2\sigma^2Tr\left(P\right)<0$$

(ו) $MSE\left(\hat{w}_\lambda\right)$ מתאר את ההפרש בין המשקולות האופטימליים $\hat{w}_\lambda$ המוערך שלנו. הנגזרת של $MSE\left(\hat{w}_\lambda\right)$ מתארת את קצב השינוי לפי הפרמטר λ, ובפרט אם הנגזרת של $MSE\left(\hat{w}_\lambda\right)$ שלילית אז הפונקצייה עצמה יורדת. במקרה שלנו הפונקצייה עצמה זה ערך השגיאה. הרגע הראנו ש $\frac{\partial}{\partial\lambda_{|\lambda=0}}MSE\left(\hat{w}_\lambda\right)<0$ ולכן נובע מכך שעבור ערכים של λ שקרובים ל0, פונקציית השגיאה יורדת. מסקנה־ שימוש בפרמטר רגולריזציה λ קרוב ל0 ישיג מודל עם שגיאה יותר נמוכה.

2. ראינו בתרגול שלכל קרנל $k(x,y)$ ופונקצייה f מתקיים ש הקרנל $f(x)k(x,y)f(y)$ הוא קרנל חוקי.
נגדיר את הפונקצייה הבאה:

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

ונגדיר את הקרנל החוקי הבא $\tilde{k}(x,y)$:

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

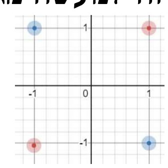
נבחין שהקרנל מנורמל, כלומר לכל x מתקיים:

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

3. נגדיר את הדאטא הבא:

$$X=\begin{bmatrix}1 & 1\\ 1 & -1\\ -1 & 1\\ -1 & -1\end{bmatrix},y=\begin{pmatrix}1\\ -1\\ -1\\ 1\end{pmatrix}$$

זה למעשה מגדיר את בעיית הקלסיפיקציה הבאה (XOR):



ברור שבמימד 2 הבעיה לא ניתנת להפרדה ליניארית. נשתמש בקרנל פולינומי, ונגזור ממנו את ψ ואת המרחב הגבוה יותר. נזכיר כי קרנל פולינומי מדרגה d הוא:

$$k(x,y)=\left(1+x^{T}y\right)^{d}$$

אנחנו נשתמש ב$d=2$. נפתח את המשוואה של הקרנל עבור $k(x,y)$:

$$\begin{aligned}k(x,y)&=\left(1+x^{T}y\right)^{2}=1+2\left(x^{T}y\right)+\left(x^{T}y\right)^{2}\\&=1+2\left(x_{1}y_{1}+x_{2}y_{2}\right)+2x_{1}y_{1}x_{2}y_{2}+\left(x_{1}y_{1}\right)^{2}+\left(x_{2}y_{2}\right)^{2}\\&=1+2x_{1}y_{1}+2x_{2}y_{2}+2x_{1}y_{1}x_{2}y_{2}+\left(x_{1}y_{1}\right)^{2}+\left(x_{2}y_{2}\right)^{2}\end{aligned}$$

נרצה שהביטוי הזה יגדיר מכפלה פנימית כלומר $k(x,y)=\langle\psi(x),\psi(y)\rangle$. נגדיר את ψ כך:

$$\psi(v)=\left(1,\sqrt{2}v_{1},\sqrt{2}v_{2},\sqrt{2}v_{1}v_{2},v_{1},v_{2}\right)^{T}$$

ואכן נקבל ש:

$$\langle\psi(x),\psi(y)\rangle = \left\langle \left(1,\sqrt{2}x_1,\sqrt{2}x_2,\sqrt{2}x_1x_2,x_1,x_2\right),\left(1,\sqrt{2}y_1,\sqrt{2}y_2,\sqrt{2}y_1y_2,y_1,y_2\right)\right\rangle$$
$$= 1+2x_1y_1+2x_2y_2+2x_1y_1x_2y_2+(x_1y_1)^2+(x_2y_2)^2 = k(x,y)$$

מצאנו את ψ מפורשות, וקיבלנו ש $\mathcal{F}=\mathbb{R}^6$ (מספר האלמנטי בכל וקטור ψ), וכעת נראה שאכן הדאטא ניתן להפרדה ליניארית במרחב הזה. כדי להראות את זה, נחשב את $S_\psi(X)$ המטריצה שמתארת את X במימד הגבוה, ונגדיר וקטור משקולות כך שיתקיים $S_\psi(X)w=y$. מטריצת ה$S_\psi(X)$:

$$S_\psi(X)_i = k(x_i,x_i) = \langle\psi(x_i),\psi(x_i)\rangle$$

נגדיר כעת את וקטור המשקולות הבא:

$$w=\left(0,0,0,\frac{1}{\sqrt{2}},0,0\right)$$

ונראה שהוא אכן מפריד את הדאטא:

$$[S_\psi(X)w]_1 = k(x_1,x_1)w = \frac{1}{\sqrt{2}}\sqrt{2}\cdot 1\cdot 1 = 1 = y_1$$

$$[S_\psi(X)w]_2 = k(x_2,x_2)w = \frac{1}{\sqrt{2}}\sqrt{2}\cdot (-1)\cdot 1 = -1 = y_2$$

$$[S_\psi(X)w]_3 = k(x_3,x_3)w = \frac{1}{\sqrt{2}}\sqrt{2}\cdot 1\cdot (-1) = -1 = y_3$$

$$[S_\psi(X)w]_4 = k(x_4,x_4)w = \frac{1}{\sqrt{2}}\sqrt{2}\cdot (-1)\cdot (-1) = 1 = y_4$$

אכן $S_\psi(X)w=y$ כלומר הדאטא ניתן להפרדה ליניארית במימד החדש.

4. זה מאוד דומה למה שנעשה בתרגול רק שהפעם נפרק את Σ לEVD שלה.
נסתכל על התוחלת של $\langle v,X\rangle$:

$$\mathbb{E}\left[\langle v,X\rangle\right] = \mathbb{E}\left[v^TX\right] = v^T\mathbb{E}[X] = 0$$

נחשב בעזרתה את השונות:

$$Var\left(\langle v,X\rangle\right) = \mathbb{E}\left[\left(\langle v,X\rangle - \mathbb{E}\left[\langle v,X\rangle\right]\right)^2\right] = \mathbb{E}\left[\left(v^TX-0\right)^2\right]$$
$$= \mathbb{E}\left[v^TXX^Tv\right] = v^T\mathbb{E}\left[X^TX\right]v = v^T\Sigma v$$

כשהמעבר האחרון נובע מכך ש$\mathbb{E}\left[X\right]=0$ ולכן $\mathbb{E}\left[X^TX\right]=\Sigma$. Σ סימטרית ולכן נוכל להסתכל על פירוק הEVD שלה:

$$\Sigma = UDU^T \Longrightarrow Var\left(\langle v,X\rangle\right) = v^TUDU^Tv$$

כשU אורתונורמלית וD אלכסונית עם הע"ע של Σ. לכן נוכל לפרק לעוד גורמים:

$$Var\left(\langle v, X\rangle\right) = v^T\sum_{i=1}^{d}u_i\lambda_i u_i^T v = \sum_{i=1}^{d}\lambda_i v^T u_i u_i^T v = \sum_{i=1}^{d}\lambda_i\left\langle u_i, v\right\rangle^2$$

כעת מכיוון ש$||v|| = 1$ ומכך ש U אורתונורמלית, נובע ש $\sum_{i=1}^{d}\langle u_i, v\rangle^2 = ||U^T v|| = 1$, ובנוסף לכל i מתקיים $\langle u_i, v\rangle^2 \geq 0$. בתרגול ראינו שזו בעיה קמורה והערך המקסימלי מתקבל עבור $v = u_1$, שזה בדיוק הפיתרון שPCA נותן לנו. מכיוון שזו בעיה קמורה יש מקסימום גלובלי אחד ולכן בוודאות PCA משיג שונות גבוהה או שווה לערך הנ"ל.

.5

(א) **לא חוקי. דוגמא נגדית:**
נגדיר:

$$x_1 = 0,\ x_2 = 1,\ x_3 = 2$$

ונגדיר את הוקטור

$$v = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$$

נסתכל על מטריצת הקרנל שיוצאת עבור $k\left(x_i x_j\right)$:

$$K = \begin{bmatrix} 1 & e & e^4 \\ e & 1 & e \\ e^4 & e & 1 \end{bmatrix}$$

נבדוק האם היא PSD:

$$\begin{aligned} v^T K v &= \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}^T \begin{bmatrix} 1 & e & e^4 \\ e & 1 & e \\ e^4 & e & 1 \end{bmatrix} \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \\ &= \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}^T \begin{pmatrix} 1 + e - e^4 \\ 1 \\ e^4 + e - 1 \end{pmatrix} \\ &= 1 + e - e^4 + 1 - \left(e^4 + e - 1\right) = 3 - 2e^4 < 0 \end{aligned}$$

המטריצה לא PSD ולכן זהו לא קרנל חוקי.

(ב) **לא חוקי. דוגמא נגדית:**

נגדיר את k_1וk_2 להיות הקרנלים החוקיים:

$$k_1(x,z) = \langle x,z \rangle \quad k_2(x,z) = (\langle x,z \rangle + 1)^2$$

נגדיר את המשתנה היחיד $x = 1$. נבנה את המטריצה K שבמקרה זה היא פשוט סקלר:

$$K = k(x,x) = k_1(1,1) - k_2(1,1) = 1 - 4 = -3$$

כמובן שלכל $a \in \mathbb{R} \backslash \{0\}$ מתקיים ש $aKa = -3 \cdot a^2 < 0$ וכן K לא PSD ולכן k לא קרנל חוקי.

(ג) **חוקי. הוכחה:**

יהיו n וקטורים $x_1 ... x_n \in \mathbb{R}^d$ מהצורה:

$$x_i = \begin{bmatrix} x_{i_a} \\ \hline x_{i_b} \end{bmatrix} = \begin{bmatrix} x_{i_1} \\ \vdots \\ x_{i_k} \\ \hline x_{i_{k+1}} \\ \vdots \\ x_{i_d} \end{bmatrix}$$

ויהיה וקטור $v \in \mathbb{R}^n$.

נראה סימטריה של הקרנל:

$$\begin{aligned} k(x_i, x_j) &= k_a(x_{i_a}, x_{j_a}) + k_b(x_{i_b}, x_{j_b}) \\ &= k_a(x_{j_a}, x_{i_a}) + k_b(x_{j_b}, x_{i_b}) \\ &= k(x_j, x_i) \end{aligned}$$

כשהמעבר האמצעי נובע מכך k_a, k_b קרנל חוקיים ולכן סימטריים.

כעת נראה ש $v^T K v \geq 0$ כלומר המטריצת היא PSD. נסתכל על איבר במטריצת הקרנל K:

$$K_{ij} = k(x_i, x_j) = k_a(x_{i_a}, x_{j_a}) + k_b(x_{i_b}, x_{j_b})$$

נתבונן במכפלה:

$$\begin{aligned} v^T K v &= \sum_i^n \sum_j^n v_i v_j K_{ij} \\ &= \sum_i^n \sum_j^n v_i v_j \left(k_a(x_{i_a}, x_{j_a}) + k_b(x_{i_b}, x_{j_b}) \right) \\ &= \sum_i^n \sum_j^n v_i v_j k_a(x_{i_a}, x_{j_a}) + \sum_i^n \sum_j^n v_i v_j k_b(x_{i_b}, x_{j_b}) \\ &= \underbrace{v^T K_a v}_{\geq 0} + \underbrace{v^T K_b v}_{\geq 0} \geq 0 \end{aligned}$$

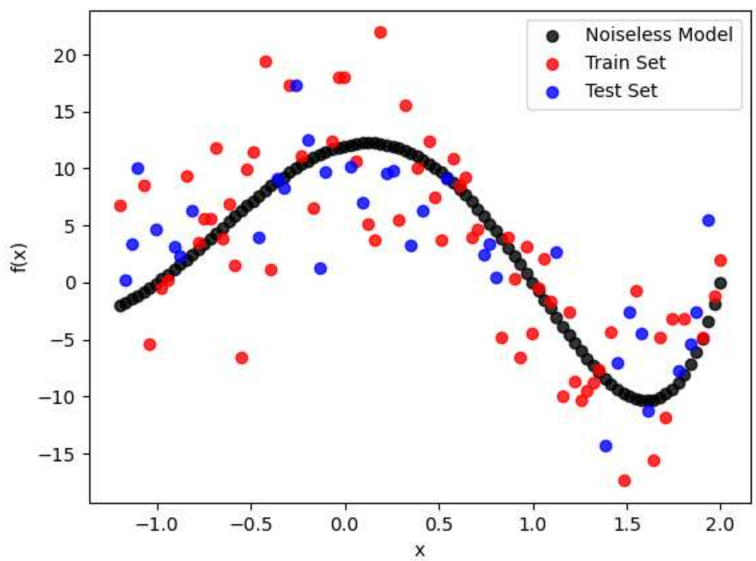
כשהמעבר האחרון נובע מהיותם של k_a, k_b קרנלים חוקיים.

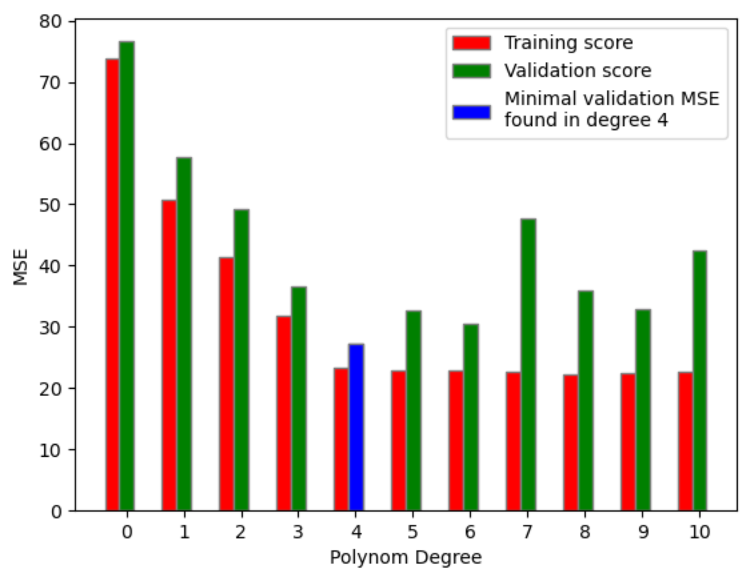
מעשי

שאלה 1

בשאלה זו נלקחו הדגימות $X=\{x_i\}_{i=1}^{n_{samples}}$ לפי $X=np.linspace(-1.2,2,n_{samples})$, ורעש של $\varepsilon\sim N\left(0,\sigma^2\right)$. לפיהם חושבו הדאטא הרועש $y=f(x)+\varepsilon$ והמודל הנקי $f(x)$.

1. כאן $\sigma^2=5$ ו$n_{sampes}=100$:





2.

ניתן לראות בגרף שרגרסיה פולינומית מדרגה 4 הגיעה לתוצאת ה$Validation$ הכי טובה. אפשר לראות כאן בבירור

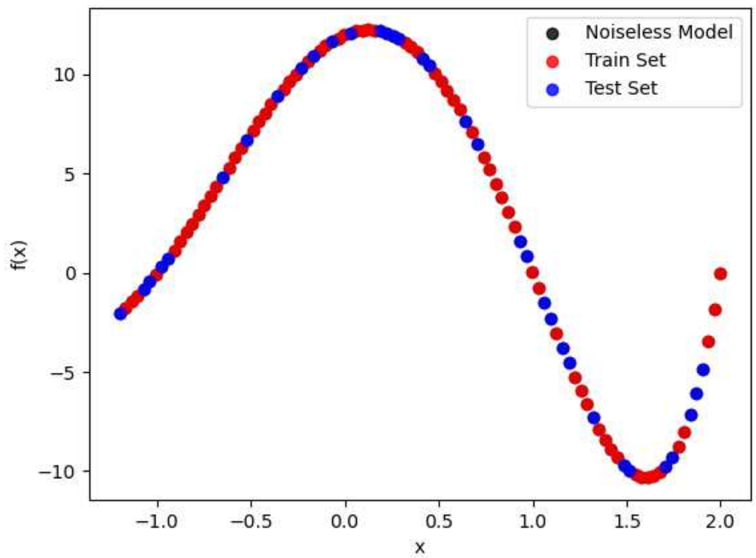
את הביאס-ווריאנס טרייד כגרף של ה$complexity$ שכאן היא דרגת הפולינום, ואת ה$sweet\ spot$ שמתקבל אכן עבור $k^* = 4$.

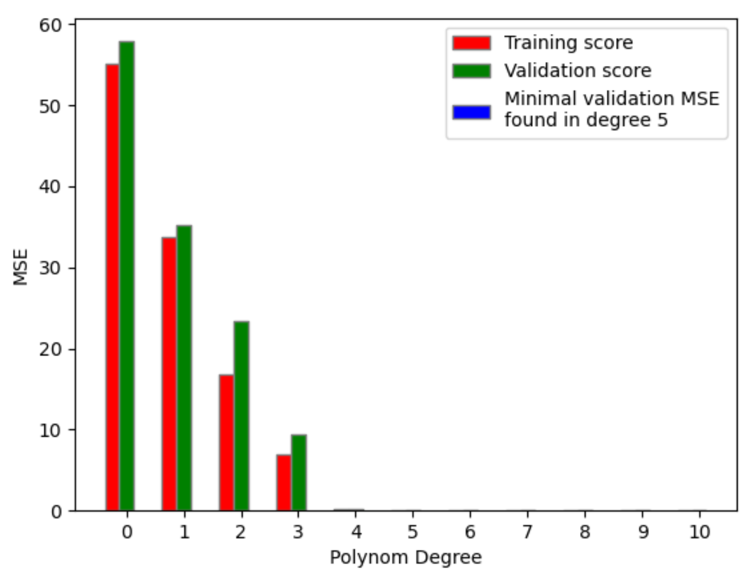
3. הערך של k^* כאמור הוא **4**. תוצאות השגיאה:

$test\ err : 17.29$.

$validation\ err : 27.17$.

4. חזרה על התהליך עבור $\sigma^2 = 0$ ו$n_{sampes} = 100$:

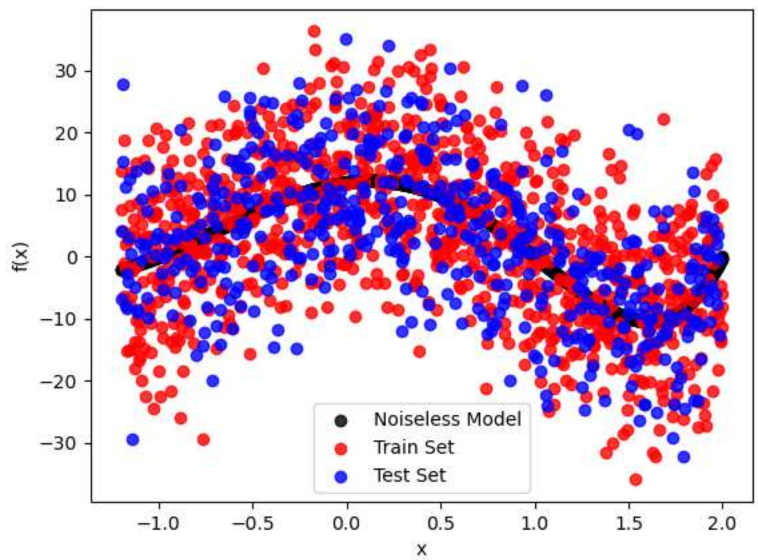


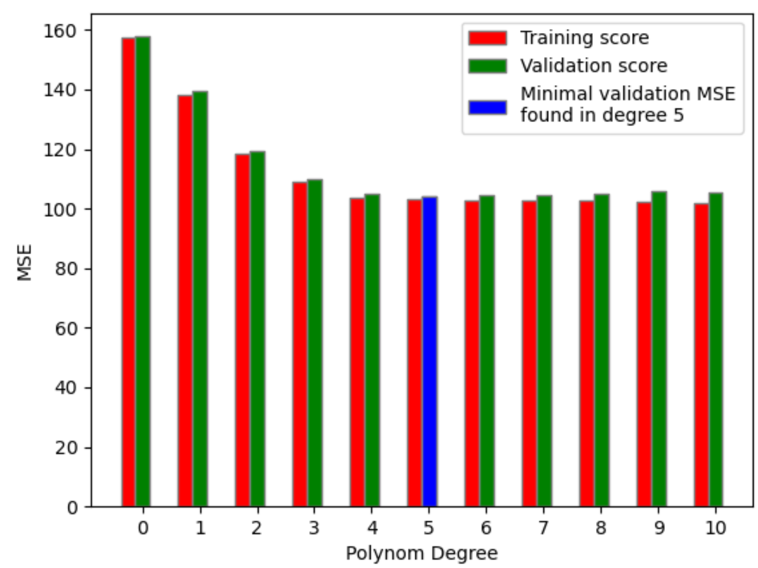


בגרף העליון ניתן לראות שמכיוון שאין רעש על האינפוט, כל הדאטא נמצא בדיוק על המודל ומכסה אותו. מכיוון שאין רעש, והמודל שלנו הוא פולינומי וגם המעריך שלנו הוא פולינומי, טבעי שהוא יימצא היפותזה שמקרבת את

הדאטא מאוד מאוד. ואכן ניתן לראות שהוא מוצא פולינום מדרגה 5, והשגיאות היו נמוכות מאוד ושאפו ממש ל0.

5. חזרה על התהליך עבור $\sigma^2 = 10$ ו$n_{sampes} = 1500$:





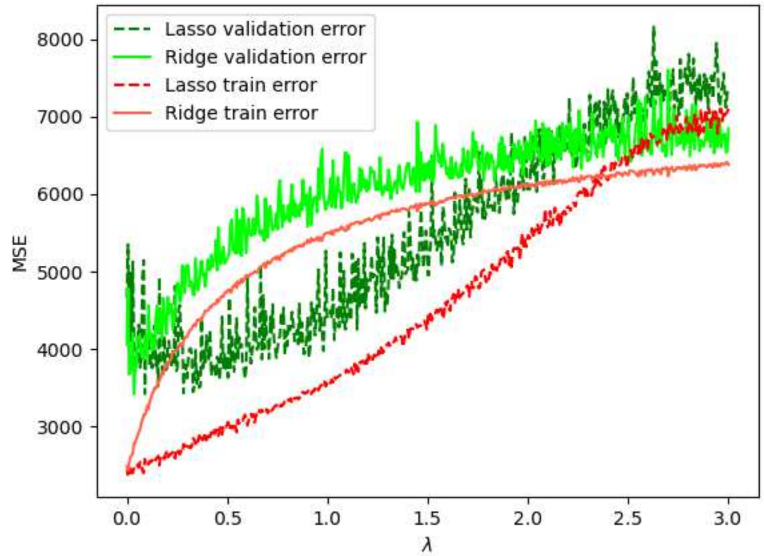
גם הפעם יצא $k^* = 5$. ניתן לראות שגרף הטרייד בין הביאס־ווריאנס קצת יותר מתון, או במילים אחרות קצת יותר קשה למצוא את ה$sweet - spot$ באופן חד משמעי. בנוסף ניתן לראות שהשגיאה האמפירית גדולה יותר מבמקרה הראשון, מכיוון שבמודל הזה נוסף יותר רעש. השגיאות:

$test\ err : 94.37$.

$validation\ err : 104.30$.

6. קוד

7. מניסיונות שביצעתי ראיתי שבטווח $\lambda \in [0,3]$ יש שינויים משמעותיים בין התוצאות, אולם עבור $\lambda > 3$ הגרפים כבר מתייצבים. לכן, הגרף הבא מציג את ה$validation\ error$ ואת ה $train\ error$ עבור שני האלגוריתמים, כשפרמטר הרגולריזציה נע בין $\lambda \in [0,3]$:



8.

(א) $Lasso$: הפרמטר λ הכי טוב יצא על $\lambda = 0.084$, עם שגיאה של $test\ error \cong 3988$.

(ב) $Ridge$: הפרמטר λ הכי טוב יצא על $\lambda = 0.078$, עם שגיאה של $test\ error \cong 3792$.