(67577) מבוא למערכות לומדות | תרגיל 2

שם: נמרוד בר גיורא | ת"ז: 207090622

חלק תאורטי

Let $\mathbf{X}$ be the design matrix of a linear regression problem with m rows (samples) and d columns (variables/features). Let $\mathbf{y} \in \mathbb{R}^m$ be the response vector corresponding the samples in $\mathbf{X}$. Recall that for some vector space $V \subseteq \mathbb{R}^d$ the orthogonal complement of V is: $V^\perp := \left\{\mathbf{x} \in \mathbb{R}^d \mid \langle \mathbf{x}, \mathbf{v}\rangle = 0 \quad \forall \mathbf{v} \in V\right\}$

פתרונות למשוואות הנורמליות

שאלה 1

1. Prove that: $Ker(\mathbf{X}) = Ker\left(\mathbf{X}^\top\mathbf{X}\right)$

הוכחה: נראה שיש הכלה דו־כיוונית בין הקבוצות וזה יוכיח את השיוויון:

($\supseteq$) יהי $v \in \ker\left(X^\top X\right)$. אזי־ $X^\top X v = 0$, ולכן:

זו הנורמה האוקלידית

$$0 = v^\top \cdot 0 = v^\top\left(X^\top X v\right) = \left(v^\top X^\top\right) X v = (Xv)^\top (Xv) = \|Xv\|^2$$

מכיוון שהנורמה היא חיובית בהחלט אז $Xv = 0$ בהכרח, כלומר ־ $v \in \ker(X)$, ובאופן כללי קיבלנו ש־ $\ker(X) \supseteq \ker\left(X^\top X\right)$.

($\subseteq$) יהי $v \in \ker(X)$. אזי־ $Xv = 0$, ולכן:

$$X^\top X v = X^\top \cdot 0 = 0$$

ולכן $v \in \ker\left(X^\top X\right)$, ובאופן כללי ־ $\ker(X) \subseteq \ker\left(X^\top X\right)$. ■

שאלה 2

2. Prove that for a square matrix A: $Im\left(A^\top\right) = Ker(A)^\perp$

נוכיח קודם 2 טענות עזר:

למה 1: לכל תמ"ו $V \subseteq \mathbb{R}^n$ מתקיים ש־ $\left(V^\perp\right)^\perp = V$.

הוכחה: יהי $v \in V$, אז לכל $v^\perp \in V^\perp$ מתקיים ש־ $\langle v \mid v^\perp \rangle = 0$, ולכן לפי הגדרת המרחב הניצב־ $v \in \left(V^\perp\right)^\perp$. כלומר ־ $V \subseteq \left(V^\perp\right)^\perp$.

בקורס ליניארית 2 מוכיחים שמרחב והמרחב הניצב לו הם סכום ישר של המרחב כולו. כלומר מתקיים ש־

$$\mathbb{R}^n = V \oplus V^\perp$$

ומצד שני מתקיים גם ש־

$$\mathbb{R}^n = V^\perp \oplus \left(V^\perp\right)^\perp$$

ולכן:

$$\dim V + \dim V^{\perp} = \dim V^{\perp} + \dim \left(V^{\perp}\right)^{\perp} \quad \Longrightarrow \quad \dim V = \dim \left(V^{\perp}\right)^{\perp}$$

כעת יחד עם ההכלה שהראינו מתקיים ש־ $V = \left(V^{\perp}\right)^{\perp}$. $\square$

<u>למה 2</u>: אם $U, V \subseteq \mathbb{R}^n$ תמ"וים כך ש־ $U^{\perp} \subseteq V$, אז $V^{\perp} \subseteq U$.
<u>הוכחה</u>: יהי $v^{\perp} \in V^{\perp}$. אז לכל $v \in V$ מתקיים ש־ $\langle v^{\perp} \mid v \rangle = 0$.
בפרט, מכיוון ש־ $U^{\perp} \subseteq V$, אז לכל $u^{\perp} \in U^{\perp}$ מתקיים ש־ $\langle v^{\perp} \mid u^{\perp} \rangle = 0$, ולכן

$$v^{\perp} \in \left(U^{\perp}\right)^{\perp} \overset{\text{לפי למה 1}}{\overset{\downarrow}{=}} U$$

כלומר מתקיים ש־ $V^{\perp} \subseteq U$. $\square$
עכשיו נוכיח את הטענה הנדרשת:

הוכחה: נוכיח הכלה דו־כיוונית:
$(\subseteq)$ יהי $v \in \operatorname{Im}\left(A^{\top}\right)$. אז קיים $w \in \mathbb{R}^n$ כך ש־ $A^{\top} w = v$. נשים לב שלכל $u \in \ker(A)$ מתקיים:

$$\langle v \mid u \rangle = v^{\top} u = \left(A^{\top} w\right)^{\top} u = w^{\top} A u = w^{\top} \cdot 0 = 0$$

ולכן $v \in (\ker(A))^{\perp}$. כלומר - $\operatorname{Im}\left(A^{\top}\right) \subseteq (\ker(A))^{\perp}$.
$(\supseteq)$ יהי $v \in \left(\operatorname{Im}\left(A^{\top}\right)\right)^{\perp}$. מכיוון שלכל $w \in \mathbb{R}^n$ מתקיים ש־ $A^{\top} w \in \operatorname{Im}\left(A^{\top}\right)$ אז לכל $w \in \mathbb{R}^n$ מתקיים:

$$0 = \langle v \mid A^{\top} w \rangle = v^{\top} A^{\top} w = (Av)^{\top} w$$

בפרט, זה נכון עבור $w = Av$, כלומר־

$$0 = (Av)^{\top} (Av) = \|Av\|^2$$

ולכן מחיוביות בהחלט של הנורמה האוקלידית נובע ש־ $Av = 0$, ולכן $v \in \ker(A)$.
כלומר הראינו ש־ $\left(\operatorname{Im}\left(A^{\top}\right)\right)^{\perp} \subseteq \ker(A)$ ולכן לפי למה 2 - $(\ker(A))^{\perp} \subseteq \operatorname{Im}\left(A^{\top}\right)$.
בסך הכל משתי ההכלות נובע ש־ $\operatorname{Im}\left(A^{\top}\right) = (\ker(A))^{\perp}$, כנדרש. $\blacksquare$

שאלה 3

3. Let $\mathbf{y} = \mathbf{X}\mathbf{w}$ be a non-homogeneous system of linear equations. Assume that $\mathbf{X}$ is square and not invertible. Show that the system has ∞ solutions $\Leftrightarrow y \perp Ker\left(\mathbf{X}^{\top}\right)$.

הוכחה: $(\Leftarrow)$ נניח ש־ $y \perp \ker\left(X^{\top}\right)$. כלומר - $y \in \left(\ker\left(X^{\top}\right)\right)^{\perp}$. לפי ההוכחה בשאלה הקודמת:

$$\left(\ker\left(X^{\top}\right)\right)^{\perp} = \operatorname{Im}\left(\left(X^{\top}\right)^{\top}\right) = \operatorname{Im}(X)$$

כלומר - $y \in \operatorname{Im}(X)$, ולכן **קיים** למערכת המשוואות $Xw = y$ פתרון. נסמן אותו ב־p, כלומר - $Xp = y$.
מכיוון ש־X לא הפיכה, אז למערכת המשוואות ההומוגנית $Xw = 0$ קיימים ∞ פתרונות.
נשים לב שלכל פתרון v כזה, המקיים ש־ $Xv = 0$, מתקיים ש־

$$X(p+v) = Xp + Xv = Xp + 0 = Xp = y$$

כלומר ניתן ליצור ∞ פתרונות למערכת $Xw = y$ בעזרת הפתרון p ואינסוף הפתרונות של המערכת ההומוגנית $Xw = 0$.

$(\Rightarrow)$ נניח שלמערכת המשוואות $Xw = y$ קיימים ∞ פתרונות. נובע מכך ש־ $y \in \operatorname{Im}(X)$, ולכן לפי ההוכחה בשאלה הקודמת:

$$y \in \left(\ker\left(X^\top\right)\right)^\perp$$

כלומר $y \perp \ker\left(X^\top\right)$. ■

שאלה 4

4. Consider the (normal) linear system $\mathbf{X}^\top\mathbf{X}\mathbf{w} = \mathbf{X}^\top\mathbf{y}$. Using what you have proved above prove that the normal equations can only have a unique solution (if $\mathbf{X}^\top\mathbf{X}$ is invertible) or infinitely many solutions (otherwise).

הוכחה: נשים לב ש־ $X^\top X$ היא מטריצה ריבועית (מסדר $d \times d$), ולכן אם היא **הפיכה** אז לכל $b \in \mathbb{R}^d$ קיים פתרון **יחיד** למערכת $X^\top Xw = b$.
בפרט, עבור $b = X^\top y$ קיים פתרון יחיד, ואפשר לקבל אותו ע"י כפל במטריצה ההפוכה:

$$\left[X^\top X\right]^{-1} \cdot X^\top Xw = \left[X^\top X\right]^{-1} X^\top y \quad \Longrightarrow \quad w = \left[X^\top X\right]^{-1} X^\top y$$

אם $X^\top X$ לא הפיכה אז לפי ההוכחה בשאלה הקודמת - קיימים ∞ פתרונות למערכת $X^\top Xw = X^\top y$ אם ורק אם מתקיים ש־

$$X^\top y \perp \ker\left(X^\top X\right)$$

אבל לפי ההוכחה בשאלה 1 - $\ker\left(X^\top X\right) = \ker(X)$, ומתקיים לכל $v \in \ker(X)$ ש־

$$\left\langle X^\top y \mid v\right\rangle = \left(X^\top y\right)^\top v = y^\top Xv \overset{\overset{v\in\ker(X)}{\downarrow}}{=} y^\top \cdot 0 = 0$$

כלומר $X^\top y \perp \ker(X)$ ולכן $X^\top y \perp \ker\left(X^\top X\right)$, ולכן אם $X^\top X$ לא הפיכה אז קיימים ∞ פתרונות למערכת. ■

מטריצות הטלה

שאלה 5

5. In this question you will prove some properties of orthogonal projection matrices seen in recitation 1. Let $V \subseteq \mathbb{R}^d$, $dim(V) = k$ and let $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be an orthonormal basis of V. Define the orthogonal projection matrix $P = \sum_{i=1}^k \mathbf{v}_i\mathbf{v}_i^\top$ (notice this is an outer product)
   (a) Show that P is symmetric.
   (b) Prove that the eigenvalues of P are 0 or 1 and that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are the eigenvectors corresponding the eigenvalue 1.
   (c) Show that $\forall \mathbf{v} \in V \;\; P\mathbf{v} = \mathbf{v}$.
   (d) Prove that $P^2 = P$.
   (e) Prove that $(I - P)P = 0$.

(a) **הוכחה:** נשים לב שמתכונות של פעולת השחלוף (שמוכיחים בקורס אלגברה ליניארית) נובע ש־

$$P^\top = \left(\sum_{i=1}^k v_i v_i^\top\right)^\top = \sum_{i=1}^k \left(v_i v_i^\top\right)^\top = \sum_{i=1}^k \left(v_i^\top\right)^\top v_i^\top = \sum_{i=1}^k v_i v_i^\top = P$$

ולכן P מטריצה סימטרית. ■

(b) **הוכחה:** כפי שמוכיחים בליניארית 2 - ניתן להשלים את הבסיס האורתונורמלי הנתון לבסיס של $\mathbb{R}^d$, ואז ע"י הליך גרם־שמידט לקבל בסיס $(u_1, \ldots, u_k, u_{k+1}, \ldots, u_d)$ **אורתונורמלי** של $\mathbb{R}^d$, כך שנשמרת התכונה ש- $V = \mathrm{Span}\,(u_1, \ldots, u_k)$, ומתקיים ש-

$$\mathbb{R}^d = \mathrm{Span}\,(u_1, \ldots, u_k) \oplus \mathrm{Span}\,(u_{k+1}, \ldots, u_d) = V \oplus V^\perp$$

יהי $0 \neq v \in \mathbb{R}^d$. אם $v \in V$ אז קיימים $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ כך ש- $v = \sum_{j=1}^k \alpha_j v_j$, ולכן:

$$Pv = \left(\sum_{i=1}^k v_i v_i^\top\right) \cdot \sum_{j=1}^k \alpha_j v_j = \sum_{i=1}^k \left(v_i v_i^\top \sum_{j=1}^k \alpha_j v_j\right) = \sum_{i=1}^k \sum_{j=1}^k \alpha_j v_i v_i^\top v_j = \sum_{i=1}^k \sum_{j=1}^k \alpha_j v_i \langle v_i \mid v_j \rangle \overset{☼}{=} \sum_{i=1}^k \alpha_i v_i = v = 1 \cdot v$$

☼: כי $(v_1, \ldots, v_k)$ בסיס א"נ ולכן לכל $1 \leqslant i, j \leqslant k$ מתקיים ש- $\langle v_i \mid v_j \rangle = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$.

לכן 1 הוא ע"ע של P, וכל $0 \neq v \in V$ הוא וקטור עצמי שמתאים לע"ע הזה, בפרט - $v_1, \ldots, v_k$ הם וקטורים עצמיים שמתאימים לו.

אם $v \in V^\perp$ אז $\langle v_i \mid v \rangle = 0$ לכל $i \in [k]$, ולכן:

$$Pv = \left(\sum_{i=1}^k v_i v_i^\top\right) \cdot v = \sum_{i=1}^k v_i v_i^\top v = \sum_{i=1}^k v_i \langle v_i \mid v \rangle = \sum_{i=1}^k v_i \cdot 0 = 0 = 0 \cdot v$$

לכן 0 הוא ע"ע של P, וכל $0 \neq v \in V^\perp$ הוא וקטור עצמי שמתאים לע"ע הזה.

נשים לב שאלו הערכים העצמיים היחידים כי אם $v = u + w$ עם $0 \neq u \in V$ ו- $0 \neq w \in V^\perp$ אז לפי מה שהראינו:

$$Pv = P(u + w) = Pu + Pw = u + 0 = u$$

ולא קיים $\lambda \in \mathbb{R}$ כך ש-$Pv = u = \lambda v$, כי אז נקבל ש- $v \in V$ בסתירה לכך ש-$w \neq 0$. כלומר v אינו וקטור עצמי במקרה הזה. ■

(c) **הוכחה:** הראיתי את זה בסעיף הקודם. נשים את זה כאן שוב למקרה ש..:

יהי $v \in V$. אז קיימים $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ כך ש- $v = \sum_{j=1}^k \alpha_j v_j$, ולכן:

$$Pv = \left(\sum_{i=1}^k v_i v_i^\top\right) \cdot \sum_{j=1}^k \alpha_j v_j = \sum_{i=1}^k \left(v_i v_i^\top \sum_{j=1}^k \alpha_j v_j\right) = \sum_{i=1}^k \sum_{j=1}^k \alpha_j v_i v_i^\top v_j = \sum_{i=1}^k \sum_{j=1}^k \alpha_j v_i \langle v_i \mid v_j \rangle \overset{☼}{=} \sum_{i=1}^k \alpha_i v_i = v$$

☼: כי $(v_1, \ldots, v_k)$ בסיס א"נ ולכן לכל $1 \leqslant i, j \leqslant k$ מתקיים ש- $\langle v_i \mid v_j \rangle = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$. ■

(d) **הוכחה:** לכל $v \in V$ מתקיים ש-

$$P^2 v = P \cdot Pv = P \cdot \sum_{j=1}^k v_j v_j^\top v = \left(\sum_{i=1}^k v_i v_i^\top\right) \cdot \sum_{j=1}^k v_j v_j^\top v = \sum_{i=1}^k \left(v_i v_i^\top \sum_{j=1}^k v_j v_j^\top v\right) =$$
$$= \sum_{i=1}^k \sum_{j=1}^k v_i v_i^\top v_j v_j^\top v = \sum_{i=1}^k \sum_{j=1}^k v_i \langle v_i \mid v_j \rangle v_j^\top v \underset{\uparrow \text{כי הבסיס א"נ}}{=} \sum_{i=1}^k v_i v_i^\top v = Pv$$

ולכן $P^2 = P$. ■

(e) **הוכחה:** לפי הסעיף הקודם $P^2 = P$ ולכן:

$$(I - P)\,P = P - P^2 = P - P = 0$$

■

ריבועים פחותים

שאלה 6

6. Show that if $\mathbf{X}^\top\mathbf{X}$ is invertible, the general solution we derived in recitation equals to the solution you have seen in class. For this part, assume that $\mathbf{X}^\top\mathbf{X}$ is invertible.

נוכיח קודם טענת עזר:

<u>למה:</u> אם $O\in\mathbb{R}^{d\times r}$ מטריצה אורתוגונלית ו־ $A\in\mathbb{R}^{r\times r}$ מטריצה הפיכה אז:

$$\left[OAO^\top\right]^{-1}=OA^{-1}O^\top$$

<u>הוכחה</u>: מכך ש־O אורתוגונלית נובע ש־ $O\cdot O^\top=I$ ולכן:

$$\left(OA^{-1}O^\top\right)\left(OAO^\top\right)=OA^{-1}O^\top OAO^\top=OA^{-1}AO^\top=OO^\top=I$$

מכיוון ש־ $OAO^\top$ היא מטריצה ריבועית (מסדר $d\times d$) אז השיוויון הזה מספיק כדי להסיק ש־ $OA^{-1}O^\top=\left[OAO^\top\right]^{-1}$.
עכשיו נוכיח את הטענה הנדרשת:

הוכחה: פירוק ה־SVD של X הוא:

$$X=U\Sigma V^\top=\overbrace{\begin{bmatrix}| & & |\\ u_1 & \dots & u_m\\ | & & |\end{bmatrix}}^{m\times m}\overbrace{\begin{bmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_d\end{bmatrix}}^{m\times d}\overbrace{\begin{bmatrix}- & v_1^\top & -\\ & \vdots & \\ - & v_d^\top & -\end{bmatrix}}^{d\times d}$$

וראינו בתרגול שניתן להפוך את הפירוק הזה לפירוק compact SVD באופן הבא:
נסמן ב$r\leqslant d$ את מספר הערכים הסינגולרים השונים מ־0 של X, ונסמן ב־ $\tilde{\Sigma}\in\mathbb{R}^{r\times r}$ את המטריצה האלכסונית שעל האלכסון שלה מופיעים r הערכים הסינגולרים הללו. נסמן ב־ $\tilde{U}\in\mathbb{R}^{m\times r}$ את המטריצה המורכבת מ־r העמודות הראשונות של U, וב־ $\tilde{V}\in\mathbb{R}^{d\times r}$ את המטריצה המורכבת מ־r העמודות הראשונות של V. נקבל ש־

$$X=U\Sigma V^\top=\begin{bmatrix}| & & | & | & & |\\ u_1 & \dots & u_r & u_{r+1} & \dots & u_m\\ | & & | & | & & |\end{bmatrix}\begin{bmatrix}\sigma_1 & & & 0\\ & \ddots & & \vdots\\ & & \sigma_r & 0\\ 0 & \dots & 0 & 0\end{bmatrix}\begin{bmatrix}- & v_1^\top & -\\ & \vdots & \\ - & v_r^\top & -\\ - & v_{r+1}^\top & -\\ & \vdots & \\ - & v_d^\top & -\end{bmatrix}=\begin{bmatrix}| & & |\\ u_1 & \dots & u_r\\ | & & |\end{bmatrix}\begin{bmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_r\end{bmatrix}\begin{bmatrix}- & v_1^\top & -\\ & \vdots & \\ - & v_r^\top & -\end{bmatrix}=\tilde{U}\tilde{\Sigma}\tilde{V}^\top$$

הפתרון הכללי למשוואות הנורמליות שראינו בתרגול הוא $\hat{w}=X^\dagger y$ כאשר $X^\dagger=V\Sigma^\dagger U^\top$, ו־ $\Sigma^\dagger\in\mathbb{R}^{d\times m}$ אלכסונית המוגדרת ע"י:

$$\forall i\in[d]:\quad \Sigma^\dagger_{i,i}=\begin{cases}1/\sigma_i & \sigma_i\neq 0\\ 0 & \sigma_i=\end{cases}$$

נשים לב שניתן לכתוב את $X^\dagger$ באופן דומה ל־compact SVD:

$$X^\dagger=V\Sigma^\dagger U^\top=\begin{bmatrix}| & & | & | & & |\\ v_1 & \dots & v_r & v_{r+1} & \dots & v_m\\ | & & | & | & & |\end{bmatrix}\begin{bmatrix}\frac{1}{\sigma_1} & & & 0\\ & \ddots & & \vdots\\ & & \frac{1}{\sigma_r} & 0\\ 0 & \dots & 0 & 0\end{bmatrix}\begin{bmatrix}- & u_1^\top & -\\ & \vdots & \\ - & u_r^\top & -\\ - & u_{r+1}^\top & -\\ & \vdots & \\ - & u_d^\top & -\end{bmatrix}=\begin{bmatrix}| & & |\\ v_1 & \dots & v_r\\ | & & |\end{bmatrix}\begin{bmatrix}\frac{1}{\sigma_1} & & \\ & \ddots & \\ & & \frac{1}{\sigma_r}\end{bmatrix}\begin{bmatrix}- & u_1^\top & -\\ & \vdots & \\ - & u_r^\top & -\end{bmatrix}=\tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^\top$$

עכשיו, מכך ש־$X^\top X$ הפיכה נובע כפי שראינו בכיתה שפתרון של המשוואות הנורמליות הוא $\overline{w}=\left[X^\top X\right]^{-1}X^\top y$.
נראה ש־ $\hat{w}=\overline{w}$. נציב את ה־compact SVD של X בנוסחא של $\overline{w}$ ונקבל:

$$\overline{w}=\left[X^\top X\right]^{-1}X^\top y=\left[\left(\tilde{U}\tilde{\Sigma}\tilde{V}^\top\right)^\top\tilde{U}\tilde{\Sigma}\tilde{V}^\top\right]^{-1}\left(\tilde{U}\tilde{\Sigma}\tilde{V}^\top\right)^\top y=\left[\tilde{V}\tilde{\Sigma}^\top\tilde{U}^\top\tilde{U}\tilde{\Sigma}\tilde{V}^\top\right]^{-1}\tilde{V}\tilde{\Sigma}^\top\tilde{U}^\top y=$$

$$\text{אורתוגונלית } \tilde{U},\ \text{אלכסונית ריבועית } \tilde{\Sigma} \rightarrow =\left[\tilde{V}\tilde{\Sigma}^2\tilde{V}^\top\right]^{-1}\tilde{V}\tilde{\Sigma}\tilde{U}^\top y=$$

$$\text{לפי הלמה כי } \tilde{V} \text{ אורתוגונלית} \rightarrow =\tilde{V}\left[\tilde{\Sigma}^2\right]^{-1}\tilde{V}^\top\tilde{V}\tilde{\Sigma}\tilde{U}^\top y=$$

$$\text{אורתוגונלית } \tilde{V} \rightarrow =\tilde{V}\left[\tilde{\Sigma}^2\right]^{-1}\tilde{\Sigma}\tilde{U}^\top y=\bigstar$$

נשים לב ש־

$$\left[\tilde{\Sigma}^2\right]^{-1}\tilde{\Sigma}=\begin{bmatrix}\sigma_1^2 & & \\ & \ddots & \\ & & \sigma_r^2\end{bmatrix}^{-1}\begin{bmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_r\end{bmatrix}=\begin{bmatrix}\frac{1}{\sigma_1^2} & & \\ & \ddots & \\ & & \frac{1}{\sigma_r^2}\end{bmatrix}\begin{bmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_r\end{bmatrix}=\begin{bmatrix}\frac{1}{\sigma_1} & & \\ & \ddots & \\ & & \frac{1}{\sigma_r}\end{bmatrix}=\tilde{\Sigma}^{-1}$$

ולכן קיבלנו ש־

$$\bigstar=\tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^\top y=X^\dagger y=\hat{w}$$

כנדרש. ∎

שאלה 7

7. Show that $\mathbf{X}^\top\mathbf{X}$ is invertible if and only if $\operatorname{span}\{\mathbf{x}_1,\ldots,\mathbf{x}_m\}=\mathbb{R}^d$.

הוכחה:

$$X^\top X \text{ הפיכה} \iff \ker\left(X^\top X\right)=\{0\} \overset{\text{לפי שאלה 1}}{\iff} \ker(X)=\{0\} \iff \dim\ker(X)=0 \iff$$

$$\underset{\text{משפט המימדים}}{\iff} \dim\operatorname{Span}\{x_1,\ldots,x_m\}=\operatorname{Rank}(X)=\dim\mathbb{R}^d \underset{\operatorname{Span}\{x_1,\ldots,x_m\}\subseteq\mathbb{R}^d}{\iff} \operatorname{Span}\{x_1,\ldots,x_m\}=\mathbb{R}^d$$

∎

שאלה 8

8. Recall that if $\mathbf{X}^\top\mathbf{X}$ is not invertible then there are many solutions. Show that $\hat{\mathbf{w}}=\mathbf{X}^\dagger\mathbf{y}$ is the solution whose L_2 norm is minimal. That is, show that for any other solution $\overline{\mathbf{w}}$, $||\hat{\mathbf{w}}||\le||\overline{\mathbf{w}}||$.

הוכחה: יהי $X=U\Sigma V^\top$ פירוק ה־SVD של X. נסמן ב־r את מספר הערכים הסינגולרים השונים מ־0 של X ונסמן:

$$U=\begin{bmatrix}U_1 & U_2\end{bmatrix}\quad,\quad V=\begin{bmatrix}V_1 & V_2\end{bmatrix}\quad,\quad \Sigma=\begin{bmatrix}\tilde{\Sigma} & 0\\ 0 & 0\end{bmatrix}$$

כאשר U_1,V_1 הן r העמודות הראשונות של U,V בהתאמה. אז כפי שראינו בסעיף 6 (ובתרגול) -

$$X=U_1\tilde{\Sigma}V_2^\top$$

הוא פירוק ה־compact SVD של X, כאשר המימדים הם $U_1\in\mathbb{R}^{m\times r}$, $\tilde{\Sigma}\in\mathbb{R}^{r\times r}$, $V_1^\top\in\mathbb{R}^{r\times d}$.
רוצים להראות ש־ $\hat{w}=X^\dagger y$ הוא פתרון של $X^\top Xw=X^\top y$ עם נורמה מינימלית.

לכל פתרון $w \in \mathbb{R}^d$ נסמן $b = V^\top w$, וגם $b_1 = V_1^\top w$, $b_2 = V_2^\top w$. אז מתקיים ש-

$$b = V^\top w = \begin{bmatrix} V_1^\top \\ V_2^\top \end{bmatrix} w = \begin{bmatrix} V_1^\top w \\ V_2^\top w \end{bmatrix} = \begin{bmatrix} b_1 \\ b_1 \end{bmatrix}$$

ניזכר שפתרון של המשוואות הנורמליות הוא $\underset{w \in \mathbb{R}^n}{\operatorname{argmin}} \|Xw - y\|^2$.

ניזכר גם שמכך ש-U ו-V אורתוגונליות נובע שהן משמרות נורמה. לכן $\hat{w}$ הוא עם נורמה מינימלית <u>אם ורק אם</u> $\hat{b} = V^\top \hat{w}$ הוא עם נורמה מינימלית.

בנוסף, ניעזר באורתוגנליות של U כדי לקבל לכל פתרון w ש-

$$\begin{aligned}
\|Xw - y\|^2 &= \left\|U\Sigma V^\top w - y\right\|^2 = \left\|U\left(\Sigma b - U^\top y\right)\right\|^2 \overset{U \text{ אורת'}}{=} \left\|\Sigma b - U^\top y\right\|^2 = \\
&= \left\|\begin{bmatrix} \tilde{\Sigma} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} - \begin{bmatrix} U_1^\top y \\ U_2^\top y \end{bmatrix}\right\|^2 = \left\|\begin{bmatrix} \tilde{\Sigma} b_1 \\ 0 \end{bmatrix} - \begin{bmatrix} U_1^\top y \\ U_2^\top y \end{bmatrix}\right\|^2 = \left\|\begin{bmatrix} \tilde{\Sigma} b_1 - U_1^\top y \\ -U_2^\top y \end{bmatrix}\right\|^2 \overset{☼}{=} \\
&\overset{☼}{=} \left\|\tilde{\Sigma} b_1 - U_1^\top y\right\|^2 + \left\|U_2^\top y\right\|^2
\end{aligned}$$

☼: כי הנורמה בריבוע זה פשוט סכום הריבועים של הקואורדינטות של הוקטור $\begin{bmatrix} \tilde{\Sigma} b_1 - U_1^\top y \\ -U_2^\top y \end{bmatrix}$, ואפשר לחלק את הסכום אז לשני סכומים שנותנים את הנורמות של הוקטורים $\tilde{\Sigma} b_1 - U_1^\top y$ ו-$-U_2^\top y$ בנפרד.

מכיוון ש-$\left\|U_2^\top y\right\|^2$ קבוע אי-שלילי, אז $\overline{w} \in \mathbb{R}^d$ מקיים ש-$\overline{w} = \underset{w \in \mathbb{R}^n}{\operatorname{argmin}} \|Xw - y\|^2$ <u>אם ורק אם</u> $\overline{b_1} = V_1^\top \overline{w}$ מקיים ש-

$$\overline{b_1} = \underset{b_1 \in \mathbb{R}^n}{\operatorname{argmin}} \left\|\tilde{\Sigma} b_1 - U_1^\top y\right\|^2$$

כלומר אם ורק אם $\overline{b_1}$ הוא פתרון של המשוואות הנורמליות הבאות:

$$\begin{gathered}
\tilde{\Sigma}^\top \tilde{\Sigma} \overline{b_1} = \tilde{\Sigma}^\top U_1^\top y \\
\Updownarrow \text{ כי } \tilde{\Sigma} \text{ אלכסונית} \\
\tilde{\Sigma}^2 \overline{b_1} = \tilde{\Sigma} U_1^\top y \\
\Updownarrow \text{ כפל ב-} \left[\tilde{\Sigma}^2\right]^{-1} \\
V_1^\top \overline{w} = \overline{b_1} = \tilde{\Sigma}^{-1} U_1^\top y
\end{gathered}$$

כלומר $\overline{w} \in \mathbb{R}^d$ הוא פתרון של המשוואות הנורמליות $X^\top X w = X^\top y$ אם ורק אם הוא מקיים ש-$\overline{b_1} = V_1^\top \overline{w} = \tilde{\Sigma}^{-1} U_1^\top y$, וזה בכלל לא תלוי ברכיב השני - $\overline{b_2} = V_2^\top \overline{w}$!

ניזכר עכשיו שפתרון $\overline{w}$ הוא **עם נורמה מינימלית** אם ורק אם $V^\top \overline{w}$ הוא עם נורמה מינימלית, ולכן $\overline{w}$ מקיים בהכרח ש-$V_2^\top \overline{w} = 0$, כי אחרת אם $V_2^\top \overline{w} \neq 0$ נוכל פשוט לקחת את הוקטור $\begin{bmatrix} V_1^\top \overline{w} \\ 0 \end{bmatrix}$ ולקבל לפי מה שראינו לעיל שהוא עדיין פתרון והוא גם מקיים-

$$\left\|\begin{bmatrix} V_1^\top \overline{w} \\ 0 \end{bmatrix}\right\| < \left\|\begin{bmatrix} V_1^\top \overline{w} \\ V_2^\top \overline{w} \end{bmatrix}\right\|$$

לסיום נראה שהפתרון $\hat{w} = X^\dagger y$ מקיים בדיוק את התכונה הזו:

לצורך כך נשתמש בצורת ה-compact SVD של $X^\dagger$ שהראיתי בסעיף 6 (שם סימנתי $V_1 = \tilde{V}$, $U_1 = \tilde{U}$). נקבל ש-

$$\begin{aligned}
V_1^\top \hat{w} &= V_1^\top X^\dagger y = V_1^\top V_1 \tilde{\Sigma}^{-1} U_1^\top y \overset{(1)}{=} \tilde{\Sigma}^{-1} U_1^\top y \\
V_2^\top \hat{w} &= V_2^\top X^\dagger y = V_2^\top V_1 \tilde{\Sigma}^{-1} U_1^\top y \overset{(2)}{=} 0 \cdot \tilde{\Sigma}^{-1} U_1^\top y = 0
\end{aligned}$$

$(1), (2)$: שני השיוויונות נובעים מכך ש-V אורתוגונלית - כי זה אומר שעמודותיה הן בסיס אורתונורמלי של $\mathbb{R}^d$. לכן במכפלה $V_1^\top V_1$ נקבל בדיוק את I_r, ובמכפלה $V_2^\top V_1$ נקבל רק מכפלות מהצורה $v_i^\top v_j$, עם $i \in [r]$ ו-$j \in [d] \setminus [r]$, השוות ל-0 מהאורתונורמליות (כש-v_i היא העמודה ה-i של V).

לכן $\hat{w}$ הוא הפתרון עם הנורמה המינימלית למשוואות הנורמליות, כנדרש. ■

חלק מעשי

רגרסיה ליניארית

שאלה 1

- איזה פיצ'רים שמרתי ואיזה לא?

<u>תשובה</u>: מחקתי את הפיצ'רים id,date,lat,long (אחרי שעשיתי שימוש בנתונים שיש בעמודת date), שמרתי את שאר הפיצ'רים.

- איזה פיצ'רים מורכבים מערכים קטגוריים ואיך התמודדתי איתם?

<u>תשובה</u>: הפיצ'ר היחיד שהתייחסתי אליו כבעל ערכים קטגוריים הוא ה־zipcode. השתמשתי בשיטת one-hot encoding כדי להפוך אותו לערך של 0 או 1.

- איזה פיצ'רים נוספים הוספתי ומה היה ההיגיון מאחורי זה?

<u>תשובה</u>: הוספתי את הפיצ'רים הבאים:

  - age: הוספתי לכל רשומה בדאטא ערך שמציין את ה"גיל" של הבית, ומוגדר ע"י:

$$\text{sale\_year} - \max(\text{yr\_built}, \text{yr\_renovated})$$

עשיתי זאת כי לדעתי זו דרך טובה לתת ערך מספרי (עם יחס סדר) לתאריכים שיש לכל רשומה בדאטא - פחות ברור מה הקשר בין בית שנמכר ב2018 ובית שנמכר ב2015 מאשר הקשר בין בית שנמכר 20 שנים לאחר ששופץ לאחרונה ובין בית שנמכר 2 שנים לאחר ששופץ לאחרונה.

  - sqft_living_ratio: פיצ'ר שמודד את הגודל היחסי של הבית ביחס לגודל הממוצע של 15 הבתים הקרובים אליו ביותר. מוגדר ע"י:

$$\frac{\text{sqft\_living15}}{\text{sqft\_living}}$$

  - sqft_lot_ratio: פיצר שמודד את הגודל היחסי של הנכס ביחס לגודל הממוצע של 15 הנכסים הקרובים ביותר. מוגדר ע"י:

$$\frac{\text{sqft\_lot15}}{\text{sqft\_lot}}$$

- איך התמודדתי עם ערכים חסרים או לא הגיוניים?

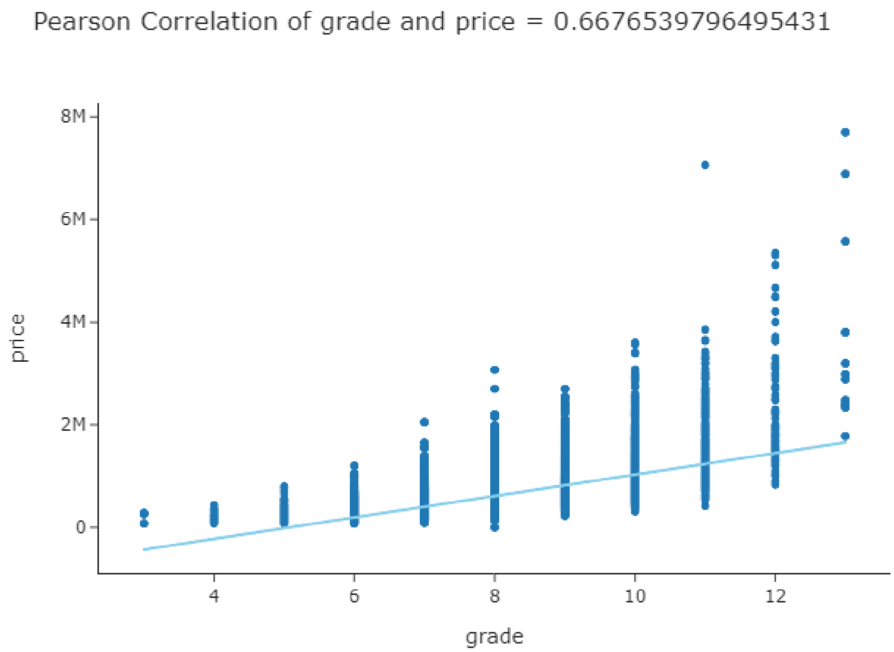
<u>תשובה</u>: מכיוון שלא היו הרבה ערכים חסרים או לא הגיוניים - מחקתי את כל הרשומות (השורות) שבהן היו ערכים כאלו. בסך הכל לא מחקתי יותר מ20 רשומות.

- דברים נוספים שעשיתי:

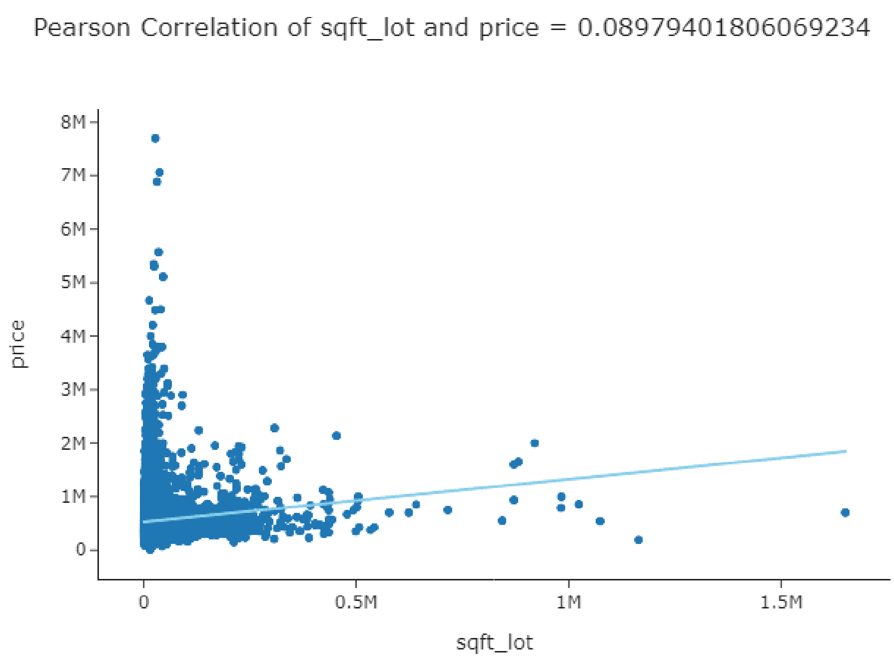
בנוסף, מחקתי דגימות עם מחירי בתים גבוהים במיוחד (מעל 5 מיליון), כי ראיתי שבאופן יחסי אלו מקרי קצה והם מנפחים מאוד את ערך ה־MSE כפי שראינו בתרגול.

שאלה 2

ערך שנראה שמועיל למודל הוא ערך ה־grade. הנה גרף שמציג אותו ביחס למחיר של כל בית בדאטא, ואת הקורלציה ביניהם:



ערך שנראה שפחות מועיל למודל הוא sqft_lot:



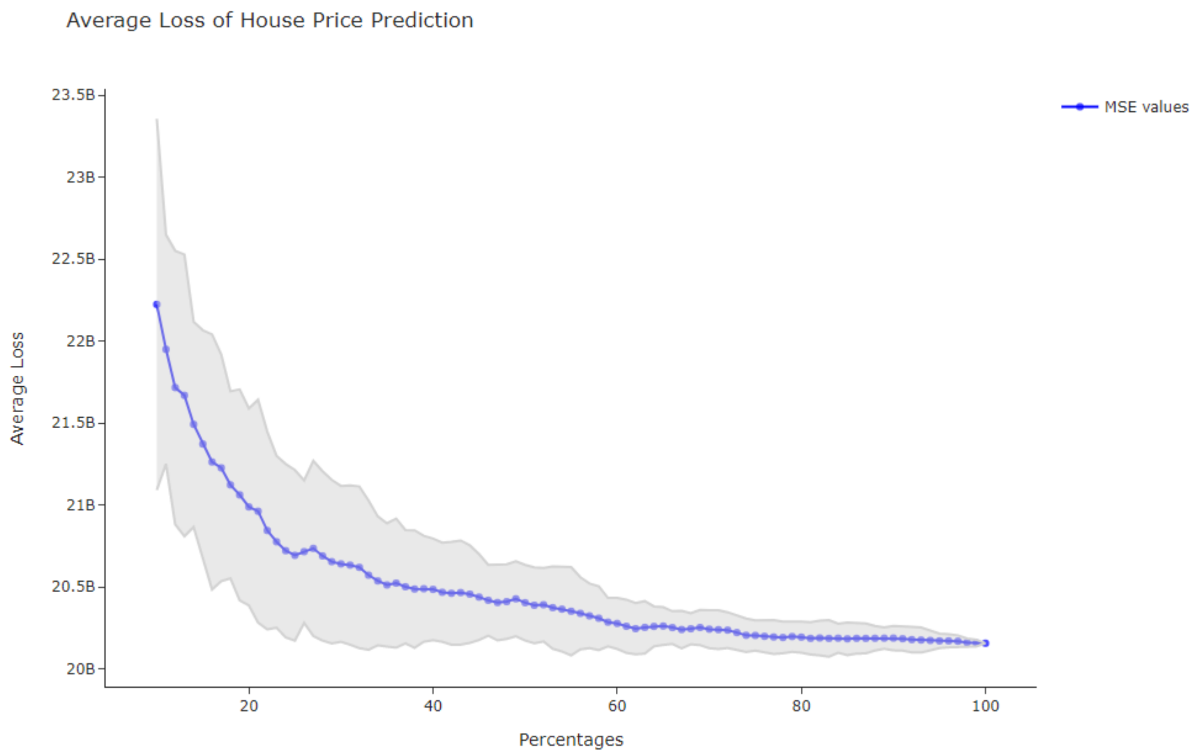
הסיבה שהסקתי שהראשון מועיל והשני פחות מועיל היא שהקורלציה בין ה־grade למחיר של הבית גבוהה בהרבה מהקורלציה בין ה־sqft_lot למחיר (שקרובה מאוד ל־0).

שאלה 3

בקוד.

שאלה 4

הגרף הנדרש:



<u>הסבר:</u>

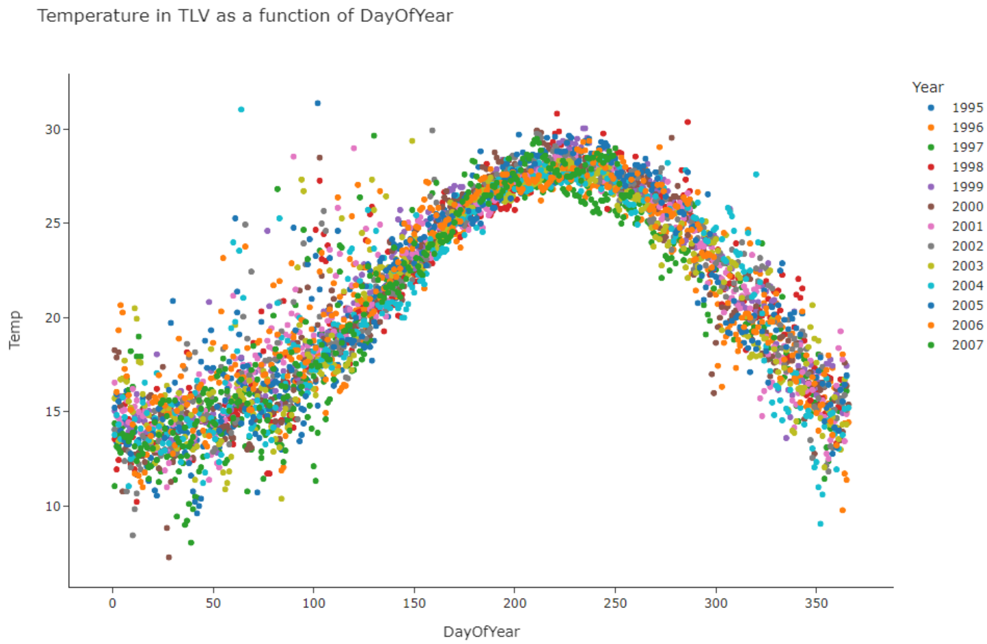
ככל שאחוז הדגימות מתוך סט האימון **גדול יותר**, ערך ה-Loss של המודל **קטן** (כאשר הוא מחושב כממוצע של האומדן עבור 10 דגימות אקראיות מסט ה-test), וה-Confidence interval הולך ו"מתהדק" - כלומר השונות של האומדן הולכת **וקטנה** ככל שהמודל מתאמן על אחוז **גדול** יותר מתוך סט האימון. הירידה הזו בשגיאה מעידה על כך שהאומד שלנו משתפר ככל שמספר הדגימות שהוא מתאמן עליהן עולה, והוא הולך ונעשה פחות מוטה (biased) ועם שונות נמוכה יותר (variance).

הערה נוספת שראוי להוסיף על ערך ה-confidence interval, כלומר על השונות היא שכשמתקרבים ל-100% מתוך סט האימון ומריצים את המודל 10 פעמים, ההבדל בתוצאות נהיה קטן מאוד כי החפיפה בין השורות שנבחרות בכל הרצה הולכת וגדלה. זה יכול להסביר גם את הירידה בשונות.

Polynomial Fitting

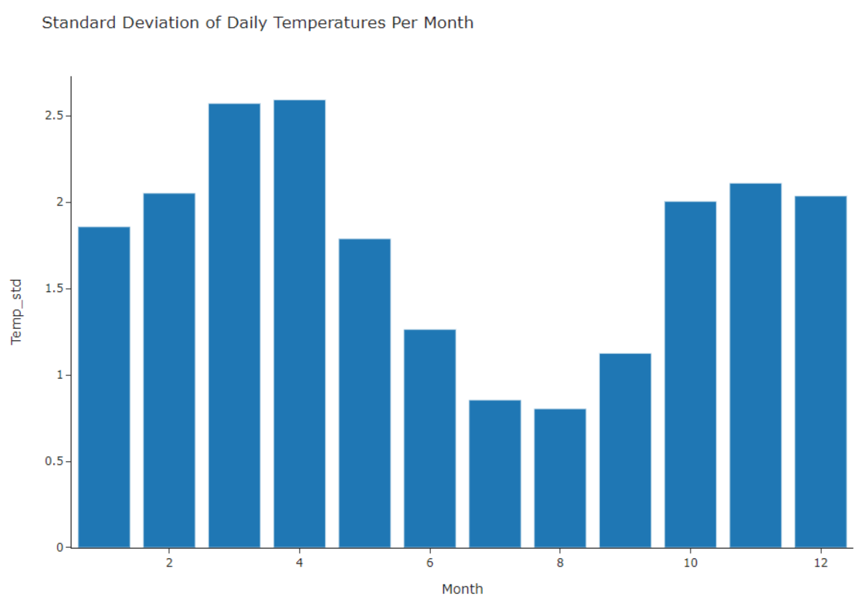
שאלה 2

- גרף שמציג את הטמפרטורה בישראל (בפועל בתל־אביב) בפונקציה של היום בשנה:



לדעתי פולינום מדרגה 3 או 4 יספיק כדי להתאים לנתונים האלו.

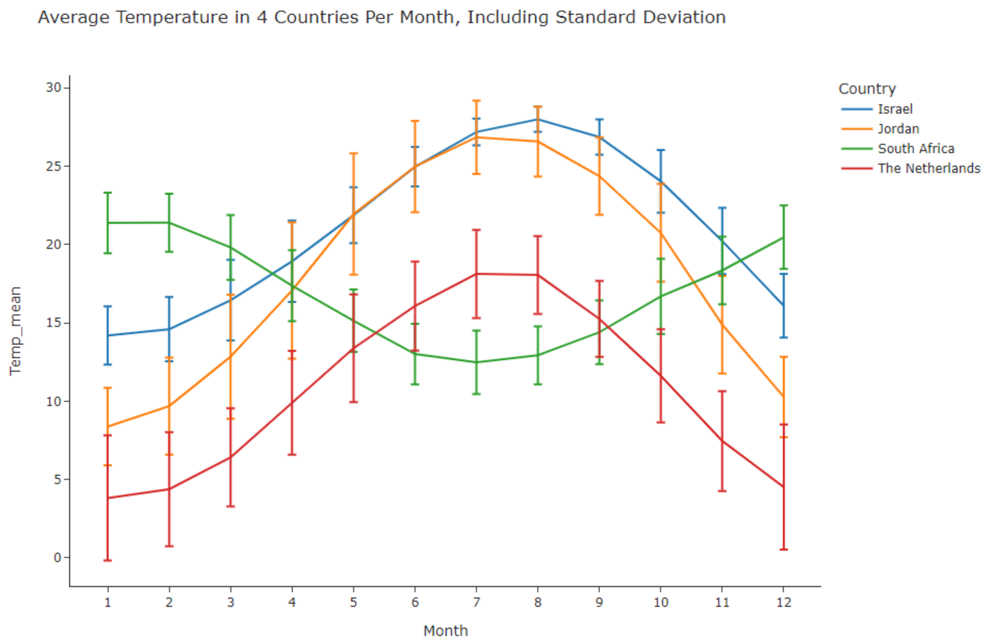
- גרף שמציג את סטיית התקן של הטמפרטורות היומיות בכל חודש:



ע"פ הגרף הזה אני מצפה שמודל פולינומיאלי יהיה מדויק יותר בחודשים $7, 8$ שבהם סטיית התקן של הטמפרטורות נמוכה יותר.

שאלה 3

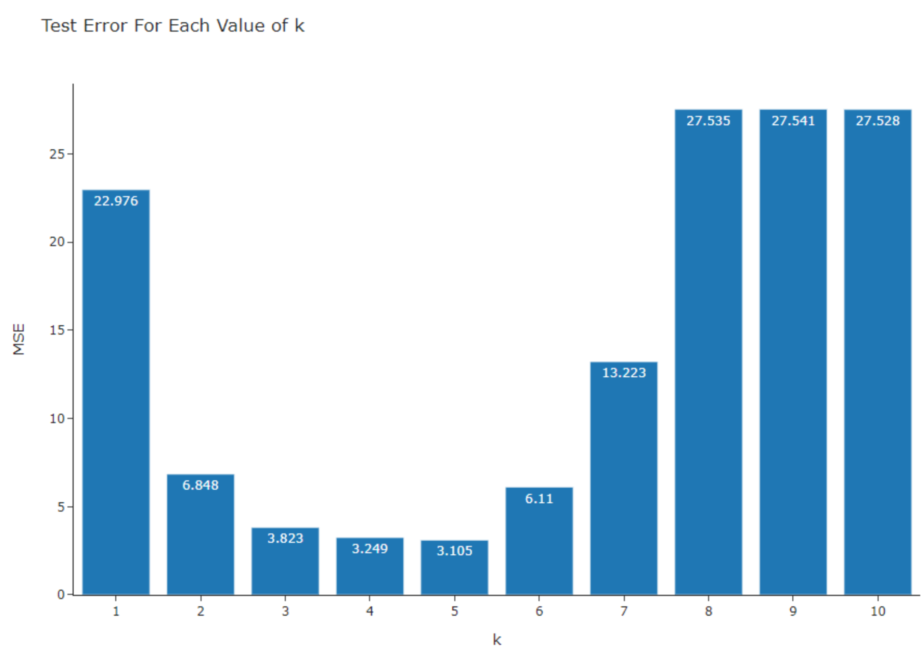
גרף המציג את הטמפרטורה הממוצעת בכל חודש בישראל, ירדן, דרום אמריקה והולנד, כולל סטיית התקן:



רואים בגרף לעיל שתבנית הטמפרטורה לפי חודשים של ישראל וירדן די דומה, והערכים די קרובים.
אני חושב שמודל שלמד את הדאטא של ישראל בלבד יעבוד הכי טוב (באופן יחסי) על הדאטא של **ירדן**.

שאלה 4

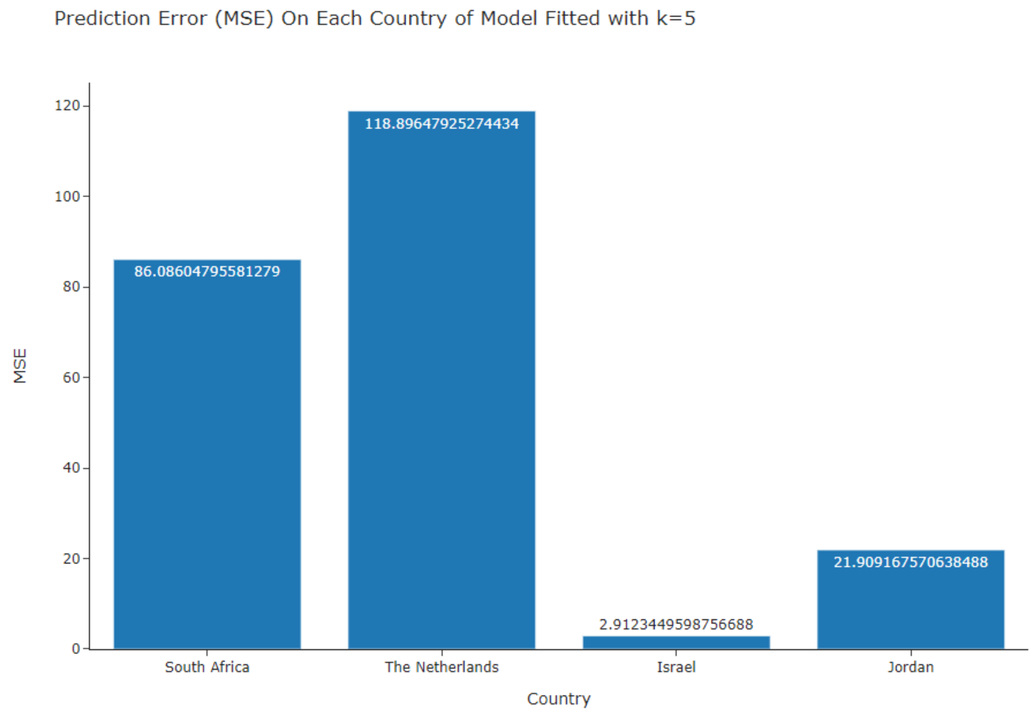
גרף המציג את הטעות (MSE) של המודל עבור כל ערך של k בין 1 ל-10:



לפי התוצאות האלו רואים שהערך של k עבורו המודל התאים ביותר לדאטא הוא 5.
התוצאות שהתקבלו עבור $k = 4$ ו־ $k = 3$ דומות מאוד, ולכן יתכן שעדיף לבחור באחד מהערכים הללו עבור k כי זה יתן לנו מודל פשוט יותר.

שאלה 5

גרף המציג את השגיאה (MSE) של מודל שלמד על הדאטא של ישראל בלבד, עם $k = 5$, על דאטא של כל אחת מ־4 המדינות שבדאטא־סט כולו:



כצפוי, השגיאה הקטנה ביותר היא על הדאטא של ישראל, כי עליו המודל התאמן.
כמו שצפיתי בתשובה לשאלה 3 ־ השגיאה שהתקבלה עבור הדאטא של ירדן היא הנמוכה מבין 3 המדינות הנוספות.