מבוא למערכות לומדות – 67577 – תרגיל 1

אופק אבידן – 318879574

2. חלק תיאורטי:

2.1. רקע מתמטי:

2.1.1 אלגברה לינארית:

1. Calculate the SVD of the following matrix A. That is, find the matrices $U, \Sigma, V^\top$ where U, V are orthogonal matrices and Σ diagonal.

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix}$$

Recall, that to find the SVD of A we can calculate $A^\top A$ to deduce V,Σ and then calculate $AA^\top$ to deduce U. Equivalently, once we deduced V, Σ we can fine U using the equality $AV = U\Sigma$.

למדנו כי לכל מטריצה ממשית A, קיימים V,U א"ג ו-Σ אלכסונית כך ש-

$$A = U\Sigma V^T$$

ונבחין כי מתקיים

$$1. A^TA = (V\Sigma U^T)(U\Sigma V^T) = V\Sigma^2V^T$$

$$2. AA^T = (U\Sigma V^T)(V\Sigma U^T) = U\Sigma^2U^T$$

מתקיים כי

$$AA^T = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 6 \end{bmatrix} = U\Sigma^2U^T$$

לכן נסיק כי

$$U = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \Sigma = \begin{matrix} \sqrt{6} & 0 & 0 \\ 0 & \sqrt{2} & 0 \end{matrix}$$

כעת נצמד להוראות ונבצע חישוב A^TA:

$$B =_{\text{סימון}} A^TA = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 2 & -2 & 4 \end{bmatrix}$$

כעת נוכל להסיק את V. נמצא ערכים עצמיים של המטריצה B

$$\det(B - \lambda I) = 0$$

$$\det\left(\begin{bmatrix}2 & 0 & 2\\ 0 & 2 & -2\\ 2 & -2 & 4\end{bmatrix} - \begin{bmatrix}\lambda & 0 & 0\\ 0 & \lambda & 0\\ 0 & 0 & \lambda\end{bmatrix}\right) = 0$$

$$\det\left(\begin{bmatrix}2-\lambda & 0 & 2\\ 0 & 2-\lambda & -2\\ 2 & -2 & 4-\lambda\end{bmatrix}\right) = 0$$

נפתח לפי שורה ראשונה:

$$\begin{aligned}\det\left(\begin{bmatrix}2-\lambda & 0 & 2\\ 0 & 2-\lambda & -2\\ 2 & -2 & 4-\lambda\end{bmatrix}\right) &= (2-\lambda) * \begin{vmatrix}2-\lambda & -2\\ -2 & 4-\lambda\end{vmatrix} - 0 + 2 * \begin{vmatrix}0 & 2-\lambda\\ 2 & -2\end{vmatrix}\\ &= (2-\lambda) * \big((2-\lambda)*(4-\lambda) - (-2)(-2)\big) + 2 * \big(-2*(2-\lambda)\big)\\ &= (2-\lambda)*(8-6\lambda+\lambda^2-4) + 2*(-4+2\lambda)\\ &= (2-\lambda)*(4-6\lambda+\lambda^2) - 8 + 4\lambda\\ &= 8 - 12\lambda + 2\lambda^2 - 4\lambda + 6\lambda^2 - \lambda^3 - 8 + 4\lambda = \boldsymbol{-\lambda^3 + 8\lambda^2 - 12\lambda}\end{aligned}$$

$$-\lambda^3 + 8\lambda^2 - 12\lambda = -\lambda(\lambda^2 - 8\lambda + 12) = -\lambda(\lambda-6)(\lambda-2)$$

כלומר הע״ע הם $\boldsymbol{\lambda_1 = 0, \lambda_2 = 2, \lambda_3 = 6}$.

נמצא ו״ע:

עבור $\lambda_1 = 0$:

$$\begin{bmatrix}2 & 0 & 2\\ 0 & 2 & -2\\ 2 & -2 & 4\end{bmatrix} = 0 \to \begin{bmatrix}1 & 0 & 1\\ 0 & 1 & -1\\ 0 & 0 & 0\end{bmatrix} = 0 \to \begin{cases}v_1 + v_3 = 0\\ v_2 - v_3 = 0\\ -\end{cases}$$

אפשר לבחור כל תוצאה שבה $v_2 = v_3$, נבחר $v_2 = v_3 = 1$. מכאן שמהמשוואה הראשונה נסיק כי $v_1 = -1$. כלומר $v_{\lambda_1} = \begin{pmatrix}-1\\1\\1\end{pmatrix}$.

עבור $\lambda_2 = 2$:

$$\begin{bmatrix}2 & 0 & 2\\ 0 & 2 & -2\\ 2 & -2 & 4\end{bmatrix} - \begin{bmatrix}2 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 2\end{bmatrix} = 0 \to \begin{bmatrix}0 & 0 & 2\\ 0 & 0 & -2\\ 2 & -2 & 2\end{bmatrix} = 0 \to \begin{bmatrix}1 & -1 & 0\\ 0 & 0 & 0\\ 0 & 0 & 1\end{bmatrix} = 0$$

$$\to \begin{cases}v_3 = 0\\ -\\ v_1 - v_2 = 0\end{cases}$$

נבחר עבור $v_1 = v_2$, את הערך $v_1 = 1$, ומכאן ש-$v_2 = 2$, כלומר $v_{\lambda_2} = \begin{pmatrix}1\\1\\0\end{pmatrix}$.

עבור $\lambda_3 = 6$:

$$\begin{bmatrix}2 & 0 & 2\\0 & 2 & -2\\2 & -2 & 4\end{bmatrix} - \begin{bmatrix}6 & 0 & 0\\0 & 6 & 0\\0 & 0 & 6\end{bmatrix} = 0 \rightarrow \begin{bmatrix}-4 & 0 & 2\\0 & -4 & -2\\2 & -2 & -2\end{bmatrix} = 0 \rightarrow \begin{bmatrix}1 & 0 & -\frac{1}{2}\\0 & 1 & \frac{1}{2}\\0 & 0 & 0\end{bmatrix} = 0$$

$$\rightarrow \begin{cases}v_1 - 0.5v_3 = 0\\v_2 + 0.5v_3 = 0\\-\end{cases}$$

נבחר עבור $v_1 = 0.5v_3$, את הערך $v_1 = 1$, ומכאן ש-$v_3 = 2$, ומכאן ש-$v_2 = -1$. כלומר $v_{\lambda_3} = \begin{pmatrix}1\\-1\\2\end{pmatrix}$.

נחלק כל וקטור בנורמה שלו, נרכיב אותם ביחד ובסה"כ נקבל:

$$\boldsymbol{V} = \begin{bmatrix}\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}}\\-\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}}\\\frac{2}{\sqrt{6}} & \mathbf{0} & \frac{1}{\sqrt{3}}\end{bmatrix}$$

סה"כ קיבלנו:

$$A = U\Sigma V^T = \begin{bmatrix}0 & 1\\1 & 0\end{bmatrix} \cdot \begin{bmatrix}\sqrt{6} & 0 & 0\\0 & \sqrt{2} & 0\end{bmatrix} \cdot \begin{bmatrix}\frac{\mathbf{1}}{\sqrt{\mathbf{6}}} & -\frac{\mathbf{1}}{\sqrt{\mathbf{6}}} & \frac{\mathbf{2}}{\sqrt{\mathbf{6}}}\\\frac{\mathbf{1}}{\sqrt{\mathbf{2}}} & \frac{\mathbf{1}}{\sqrt{\mathbf{2}}} & \mathbf{0}\\-\frac{\mathbf{1}}{\sqrt{\mathbf{3}}} & \frac{\mathbf{1}}{\sqrt{\mathbf{3}}} & \frac{\mathbf{1}}{\sqrt{\mathbf{3}}}\end{bmatrix}$$

כנדרש.

2. Show that the outer product of two vectors $\mathbf{v} \in \mathbb{R}^n, \mathbf{u} \in \mathbb{R}^m$, which is denoted by $\mathbf{v} \otimes \mathbf{u}$ or $\mathbf{v} \cdot \mathbf{u}^\top$ is a matrix $A \in \mathbb{R}^{n \times m}$ with $rank(A) = 1$. That is, show that all rows (or columns) in A are linearly dependent.

יהיו ווקטורים $v \in \mathbb{R}^n, u \in \mathbb{R}^m$. אז המכפלה החיצונית של שני הוקטורים, שהיא $v * u^T$, תוצאתה מטריצה בגודל $n\ X\ m$, מקורס אלגברה לינארית 1. נסמן את המטריצה ב-A. אזי, מתכונות של לינארית 1, מתקיים $A_{i,j} = v_i * u_j$. הדרגה של מטריצה היא המימד של המרחב הווקטורי המתפרש על ידי העמודות (או השורות), השווה גם למספר המירבי של עמודות (שוב, או שורות) בלתי תלויות לינארית. כדי להוכיח ש-A מדרגה 1, כלומר שמתקיים $rank(A) = 1$, עלינו להראות שכל העמודות או השורות תלויות לינארית.

נסתכל על העמודה ה-i של המטריצה A, נסמנה A_i. מלינארית 1, העמודה הזו מורכבת מהכפל $v_i * u$ (האיבר ה-i של הוקטור v כפול כל הוקטור u). לכן, כל עמודה במטריצה A מורכבת מכפל בסקלר של אותו הוקטור u. לכן כל העמודות במטריצה A הן אותה עמודה רק בכפל בסקלר שונה. כלומר, כל העמודות תלויות לינארית (יכלנו גם להוכיח זאת באותה דרך על השורות, אך אין צורך).

מכך שכל העמודות תלויות לינארית, נסיק כאמור כי הדרגה של המטריצה A היא 1, ונסיים.

3. Show that for any otrthonormal basis $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ and any aribtrary vector $\mathbf{x} \in \mathbb{R}_n$ such that $\mathbf{x} = \sum_{i=1}^n a_i \cdot \mathbf{u}_i$, it holds that $a_i = \langle \mathbf{x}, \mathbf{u}_i \rangle$ for any $i \in [1, n]$. That is, show that the i'th coefficient of representing $\mathbf{x}$ in the basis $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$, is the inner product between $\mathbf{x}$ and $\mathbf{u}_i$.

יהי בסיס אורנורמלי $b = (u_1, \ldots, u_n)$, ויהי וקטור שרירותי $x \in \mathbb{R}$, כך שמתקיים $x = \Sigma_{i=1}^n a_i * u_i$, כאשר $a_i = \langle x, u_i \rangle$ לכל $i \in [1, n]$. נראה כי המקדם ה-i של x (כאשר הוא מיוצג על ידי הבסיס b), הוא המכפלה הפנימית בין x ל-u_i. כלומר נראה כי $a_i = \langle x, u_i \rangle$ לכל $i \in [1, n]$.

מכך ש-b אורתונורמלי, מתקיים כי (*)

$$\langle u_i, u_j \rangle = \begin{cases} 1\ if\ i = j \\ 0\ if\ i\ \neq j \end{cases}$$

כעת, נשים לב כי

$$\langle x, u_i \rangle = \langle \Sigma_{j=1}^n a_j * u_j, u_i \rangle =_{\text{לינאריות מכפלה פנימית}} \Sigma_{i=1}^n a_j * \langle u_j, u_i \rangle$$

כעת נשתמש ב (*), ונקבל $\langle x, u_i \rangle = \Sigma_{i=j} a_j * 1 + \Sigma_{i \neq j} a_j * 0$. מכך ש-$i=j$ רק עבור ערך אחד (j אינדקס רץ, i קבוע), נקבל בסה״כ $\langle x, u_i \rangle = a_i$.

4. In (a-e) 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. So the orthogonal projection matrix onto V is defined as $P = \sum_{i=1}^{k} \mathbf{v}_i \mathbf{v}_i^\top$ (notice this is an outer product).
   (a) Show that P is symmetric.

א. נוכיח כי P סימטרית, כלומר כי $P = P^T$.

$$P^T = \left(\Sigma_{i=1}^{k} v_i \cdot v_i^T\right)^T = \Sigma_{i=1}^{k} v_i^{T^T} \cdot v_i^T = \Sigma_{i=1}^{k} v_i \cdot v_i^T = P$$

(b) Prove that 0 and 1 are the eigenvalues of P and show that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are the eigenvectors corresponding the eigenvalue 1.

נראה כי הוקטורים $v_1, \ldots, v_k$ הם ו"ע עם ע"ע 1:

יהי וקטור $v_j \in \{v1, \ldots, v_k\}$ ונראה כי $Pv_j = v_j$.

$$Pv_j = \Sigma_{i=1}^{k} (v_i \cdot v_i^T) v_j =$$

(נשתמש בלינאריות של סכום ומכפלה חיצונית):

$$= \Sigma_{i=1}^{k} v_i \cdot (v_i^T \cdot v_j) = v_j \cdot (v_j^T \cdot v_j) = v_j$$

כאשר המעבר הראשון נובע מכך שאם i שונה מ-j, המכפלה תהיה 0 (מכך שוקטורי הבסיס אורתונורמליים ובפרט ניצבים אחד לשני), והמעבר השני נובע מכך שנורמה של כל וקטור בסיס הוא 1 כשמדובר בבסיס אורתונורמלי, ואכן v_j הוא וקטור כלשהו מהבסיס האורתונורמלי הנתון.

נראה כי כל וקטור שרירותי $v \in V^T$ הוא ו"ע עם ע"ע 0:

עבור וקטור $v, v \in V^T$ ניצב לכל אחד מוקטורי הבסיס של V. ולכן:

$$Pv = \Sigma_{i=1}^{k} (v_i \cdot v_i^T) v = \Sigma_{i=1}^{k} v_i \cdot (v_i^T \cdot v) = \Sigma_{i=1}^{k} v_i \cdot \vec{0} = \vec{0}$$

(c) Show that $\forall \mathbf{v} \in V \ P\mathbf{v} = \mathbf{v}$.

יהי $v \in V$. מכך ש-$b = (v_1, \dots, v_k)$, הוא בסיס אורתונומלי של V, כל וקטור v ב-V, ניתן לבטא כצירוף לינארי של וקטורי הבסיס הללו, ונסמן את המקדמים ב-α_i של כל וקטור בבסיס בהתאמה, כלומר $v = \Sigma_{i=1}^{k} \alpha_i \cdot v_i$.

כעת נחשב את Pv:

$$Pv = P\left(\Sigma_{i=1}^{k} \alpha_i \cdot v_i\right) = \Sigma_{i=1}^{k} \alpha_i \cdot Pv_i = \Sigma_{i=1}^{k} \alpha_i \cdot (v_i \cdot v_i^T) \cdot v_i = \Sigma_{i=1}^{k} \alpha_i \cdot 1 \cdot v_i$$
$$= \Sigma_{i=1}^{k} \alpha_i \cdot v_i$$

כלומר אנו מקבלים כי $Pv = v$, כנדרש.

(d) Prove that $P^2 = P$.

ראשית, נחשב את P^2:

$$P^2 = (\Sigma_{i=1}^{k} v_i \cdot v_i^T) \cdot (\Sigma_{j=1}^{k} v_j \cdot v_j^T) = \Sigma_{i=1}^{k} \Sigma_{j=1}^{k} (v_i \cdot v_i^T) \cdot (v_j \cdot v_j^T)$$
$$= \Sigma_{i=1}^{k} \Sigma_{j=1}^{k} v_i \cdot (v_i^T \cdot v_j) \cdot v_j^T$$

כעת נשים לב שמכך שהבסיס אורתונורמלי, הביטוי $(v_i^T \cdot v_j)$, הוא 1 אם ורק אם $i=j$, ו-0 אחרת. לכן הביטוי $\Sigma_{i=1}^{k} \Sigma_{j=1}^{k} v_i \cdot (v_i^T \cdot v_j) \cdot v_j^T$ יתאפס למעט המקרה שבו $i=j$. נביט במקרה שבו $i=j$: $v_i \cdot (v_i^T \cdot v_j) \cdot v_j^T = v_i \cdot (v_i^T \cdot v_i) \cdot v_i^T = v_i \cdot 1 \cdot v_i^T = v_i \cdot v_i^T$.

לכן בסה"כ נקבל:

$$P^2 = \Sigma_{i=1}^{k} \Sigma_{j=1}^{k} v_i \cdot (v_i^T \cdot v_j) \cdot v_j^T = \Sigma_{i=1}^{k} v_i \cdot v_i^T = P$$

כנדרש.

(e) Prove that $(I-P)P=0$.

$$(I-P)P = (\mathrm{I} - \Sigma_{i=1}^{k} v_i \cdot v_i^T) \cdot (\Sigma_{j=1}^{k} v_j \cdot v_j^T) = (\Sigma_{i=1}^{k}(I - v_i \cdot v_i^T)) \cdot (\Sigma_{j=1}^{k} v_j \cdot v_j^T)$$

כעת נבחן את הביטוי $(I - v_i \cdot v_i^T)$:

$$(I - v_i \cdot v_i^T) = I - (v_i \cdot v_i^T)$$

מכך ש-vi הוא וקטור כחלק מבסיס אורתונורמלי, $v_i \cdot v_i^T$ היא מטריצת הטלה על תת המרחב הנפרש על ידי vi. כמו כן, $I - (v_i \cdot v_i^T)$ הוא המשלים האורתוגונלי של הטלה זו, כלומר, הוא מטיל על המשלים האורתוגונלי של תת המרחב המתפרש על ידי vi.

כשיש לנו $I - (v_i \cdot v_i^T)$, הוא מסיר כל חלק בוקטור שנמצא בכיוון של vi, ומשאיר רק את החלק שהוא אורתוגונלי אל vi. אם הווקטור כבר אורתוגונלי ל-vi. הוא נשאר ללא שינוי. אם הווקטור נמצא כולו בכיוון של vi, הוא הופך לאפס.

לכן, $(I - v_i \cdot v_i^T)$ מוחק כל וקטור בתת המרחב הנפרש על ידי vi, משאיר אותו ללא שינוי אם הוא אורתוגונלי ל-vi, ואפס אם הוא נמצא בתת-המרחב שמתפרש על ידי vi.

לכן, כאשר מכפילים $(I-P)P$, כל גורם מלבד אלה שבהם $i=j$ יביא לאפס. לאותם גורמים אשר עבורם מתקיים $i=j$, התוצאה תהיה $(v_i \cdot v_i^T)$ בגלל ש- $(I - v_i \cdot v_i^T)$ מוטל על המרחב האורתוגונלי המשלים הנפרש על ידי vi , וכאשר אנחנו מכפילים $(v_i \cdot v_i^T)$ זה יביא לאפס גם כן.

לכן בסה״כ נקבל $(I-P)P = 0$.

5. Use the chain rule to calculate the gradient of $h(\sigma) = \frac{1}{2}||f(\sigma) - y||^2$, where $\sigma \in \mathbb{R}^d$ and f is some arbitrary function from $\mathbb{R}^d$ to $\mathbb{R}^d$. 2.1.2

כדי למצוא את הגרדיאנט של הפונקציה $h(\sigma) = \frac{1}{2}\left|\left|f(\sigma) - y\right|\right|^2$ ביחס ל-$\sigma \in \mathbb{R}^d$, נשתמש בכלל השרשרת. תהי f פונקציה שרירותית מ-$\mathbb{R}^d$ ל-$\mathbb{R}^d$, ויהי y וקטור קבוע ב-$\mathbb{R}^d$.

נבחין כי

$$h(\sigma) = \frac{1}{2}\left|\left|f(\sigma) - y\right|\right|^2 = \frac{1}{2}(f(\sigma) - y)^T(f(\sigma) - y)$$

כדי למצוא את הגרדיאנט $\nabla h(\sigma)$, נשתמש כאמור בכלל השרשרת. נגדיר $g(\sigma) = f(\sigma) - y$.
כלומר $h(\sigma) = \frac{1}{2}g(\sigma)^T g(\sigma)$.

כעת, הגרדיאנט של h ביחס ל-σ הוא:

$$\nabla h(\sigma) = \frac{\partial h}{\partial g}\frac{\partial g}{\partial \sigma}$$

נתחיל ב-$\frac{\partial h}{\partial g}$:

$$h(g) = \frac{1}{2}g^T g$$

הגרדיאנט של $\frac{1}{2}g^T g$ ביחס ל-g הוא:

$$\frac{\partial h}{\partial g} = g =_{\text{הגדרת } g} f(\sigma) - y$$

אז, הנגזרת של g ביחס ל-σ היא פשוט מטריצת היעקוביאן של f:

$$\frac{\partial g}{\partial \sigma} = \frac{\partial f(\sigma)}{\partial \sigma} = J_f(\sigma)$$

נשלב את כל אלו באמצעות כלל השרשרת ונקבל:

$$\nabla h(\sigma) = (f(\sigma) - y)^T J_f(\sigma)$$

כדי לבטא את הגרדיאנט כוקטור עמודה (שזו הצורה הסטנדרטית להציג גרדיאנטים), אנחנו צריכים לבצע טרנספוז לתוצאה, כלומר בסה״כ נקבל:

$$\nabla h(\sigma) = J_f(\sigma)^T(f(\sigma) - y)$$

כאשר $J_f(\sigma)$ היא מטריצת היעקוביאן של המטריצה f לפי σ.

6. In recitation 2 we saw the softmax function $S:\mathbb{R}^k \to [0,1]^k$, which is defined as follows:

$$S(\mathbf{x})_j = \frac{e^{x_j}}{\sum_{l=1}^{k} e^{x_l}}$$

This function takes an input vector $x \in \mathbb{R}^d$ and outputs a probability vector (non-negative entries that sum up to 1), corresponding to the weight of original entries of x.
Question: Calculate the Jacobian of the softmax function S.

כדי לחשב את מטריצת היעקוביאן של S, אנחנו צריכים למצוא את הנגזרות החלקיות של כל קומפוננט של פלט פונקציית הסופטמקס ביחס לכל קומפוננט של וקטור הקלט.

נסמן את וקטור הקלט כ- $x = [x_1, \dots, x_k]$ ואת וקטור הפלט כ- $S(x) = [S_1(x), \dots, S_k(x)]$.

כאמור פונקציית הסופטמקס S מוגדרת כ- $S(x)_j = \frac{e^{x_j}}{\Sigma_{l=1}^{k} e^{x_l}}$.

עבור $j=1,2,..,k$.

1. **נחשב את הנגזרת החלקית של $S(x)_j$ ביחס ל- x_l:**

אנחנו צריכים לחשב את $\frac{\partial S(x)_j}{\partial x_l}$ לכל j ו- l.

נבחין כי $S(x)_j = \frac{e^{x_j}}{Z}$ כך ש- $Z = \Sigma_{i=1}^{k} e^{x_i}$.

2. **המקרה שבו $l=j$:**

$$\frac{\partial S(x)_j}{\partial x_l} = \frac{\partial S(x)_j}{\partial x_j} = \frac{\partial}{\partial x_j}\left(\frac{e^{x_j}}{Z}\right)$$

נשתמש בכלל המנה $\frac{u}{v}$ כאשר $u = e^{x_j}$, ו- $v = Z$.

$$\frac{\partial S(x)_j}{\partial x_j} = \frac{(e^{x_j} \cdot Z) - (e^{x_j} \cdot e^{x_j})}{Z^2} = \frac{e^{x_j}(Z - e^{x_j})}{Z^2} = S(x)_j(1 - S(x)_j))$$

3. **המקרה שבו $l \neq j$:**

$$\frac{\partial S(x)_j}{\partial x_l} = \frac{\partial}{\partial x_l}\left(\frac{e^{x_j}}{Z}\right) =_{\text{כלל המנה}} \frac{(e^{x_j} \cdot 0) - (e^{x_j} \cdot e^{x_l})}{Z^2} = \frac{-(e^{x_j} \cdot e^{x_l})}{Z^2}$$

$$= -\frac{e^{x_j}}{Z} \cdot \frac{e^{x_l}}{Z} = -S_j(x) \cdot S_l(x)$$

4. **נשלב את שני המקרים 2,3:**

$$[J_S(x)]_{jl} = \begin{cases} S_j(x)\left(1 - S_j(x)\right) & if\ j = l \\ -S_j(x)S_l(x) & if\ j \neq l \end{cases}$$

5. **ובצורת מטריצה:**

בצורת מטריצה, היעקוביאן יכול להיות מבוטא כך:

$$J_S(x) = diag\left(S(x)\right) - S(x)S(x)^T$$

כאשר $diag\big(S(x)\big)$ היא מטריצה אלכסונית עם ההסתברויות $S_1(x), \dots, S_k(x)$ על האלכסון, ו-$S(x)S(x)^T$ הוא המכפלה החיצונית של הוקטור $S(x)$ עם עצמו.

לכן, מטריצת היעקוביאן של פונקציית ה-*softmax* היא:

$$J_S(x) = diag\big(S(x)\big) - S(x)S(x)^T$$

2.2 רגרסיה לינארית

Let $\mathbf{X}$ be the input 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 $\mathbf{X}$.

2.2.1

1. In (a-d) you will incrementally prove several important properties regarding the solutions of the normal equations.
   (a) Prove that: $Ker(\mathbf{X}) = Ker\left(\mathbf{X}^\top \mathbf{X}\right)$

נוכיח באמצעות הכלה דו כיוונית:

ראשית נוכיח כי $Ker(X) \subseteq Ker(X^TX)$:

יהי $v \in Ker(X)$, ונוכיח כי $v \in Ker(X^TX)$. מההגדרה של Ker, מתקיים $Xv = 0$.
כעת נשים לב כי מתקיים $(X^TX)v = X^T(Xv) = X^T(0) = 0$ ולכן מההגדרה של Ker, מתקיים $v \in Ker(X^TX)$. מכך ש-v וקטור שרירותי, מתקיים $Ker(X) \subseteq Ker(X^TX)$.

כעת נוכיח כי $Ker(X^TX) \subseteq Ker(X)$:

יהי $v \in Ker(X^TX)$, ונוכיח כי $v \in Ker(X)$. מההגדרה של Ker, מתקיים $(X^TX)v = 0$. נשים לב כי $(X^TX)v = X^TXv$ ולכן מתקיים

$$||Xv||^2 = (Xv)^T(Xv) = v^TX^TXv = v^T(X^TXv) = v^T(0) = 0$$

אנו יודעים כי הנורמה של וקטור היא 0 אמ״מ מדובר בוקטור האפס, ולכן $Xv = 0$. ולכן מההגדרה של Ker, מתקיים $v \in Ker(X)$. מכך ש-v וקטור שרירותי, מתקיים $Ker(X^TX) \subseteq Ker(X)$.

מהכלה דו כיוונית נקבל שיויון ונסיים.

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

נוכיח באמצעות הכלה דו כיוונית:

1. כל וקטור ב-$Im(A^T)$ הוא אורתוגונלי לכל וקטור ב-$Ker(A)$
2. כל וקטור אורתוגונלי ל-$Ker(A)$ הוא ב-$Im(A^T)$

נוכיח ראשית כי $Im(A^T) \subseteq Ker(A)^{\perp}$:

יהי $y \in Im(A^T)$. כלומר מההגדרה קיים (לפחות אחד) וקטור v כך שמתקיים $y = A^T v$. אנחנו צריכים להראות כי y ניצב לכל וקטור ב-$Ker(A)$. יהי $u \in Ker(A)$. כלומר מההגדרה מתקיים $Au = 0$. נביט במכפלה הפנימית של $y \cdot u$:

$$y \cdot u = (A^T v) \cdot u = v \cdot (Au) = v \cdot 0 = 0$$

לכן y ניצב ל-u. מכך ש-u הוא וקטור שרירותי בקרנל של A, אז y ניצב ל-$Ker(A)$, ולכן $y \in Ker(A)^{\perp}$ ומכאן ש-$Im(A^T) \subseteq Ker(A)^{\perp}$.

כעת נוכיח כי $Ker(A)^{\perp} \subseteq Im(A^T)$. יהי $y \in Ker(A)^{\perp}$. כלומר y ניצב לכל וקטור בקרנל של A. יהי $u \in Ker(A)$, אז $y \cdot u = 0$.

אנחנו צריכים למצוא v כך שיתקיים $A^T v = y$ (כדי ש-y יהיה בתמונה של A^T).

מכך ש-$y \in Ker(A)$, זה אומר ש-y במרחב השורות של A. זה אומר ש-y במרחב העמודות של A^T, כלומר בתמונה של A^T. לכן $y \in Im(A^T)$ ומכך $Ker(A)^{\perp} \subseteq Im(A^T)$ כנדרש. מהכלה דו"כ נקבל שיוויון ונסיים.

(c) 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 $\Leftrightarrow y \perp Ker\left(\mathbf{X}^{\top}\right)$.

נוכיח כיוון ראשון – אם למערכת המשוואות יש אינסוף פתרונות, אז $y \perp Ker(X^T)$.
מכיוון ש-X היא ריבועית ולא הפיכה, אז היא סינגולרית. זה אומר של-X יש גרעין לא טריוויאלי, כלומר קיים וקטור (לפחות אחד) שאינו וקטור האפס, נאמר v, כך שמתקיים $Xv = 0$. מכך שלמערכת המשוואות $y=Xw$ יש אינסוף פתרונות, חייב להיות לפחות פתרון אחד למשוואה. נסמן ש-w_0 הוא פתרון מסוים ל-$y=Xw$. כל פתרון אחר יכול להיות רשום כ- $w = w_0 + v$ כאשר $v \in Ker(X)$.

נחליף את $w = w_0 + v$ ב-$y = Xw$:

$$y = X(w_0 + v) = Xw_0 + Xv$$

מכיוון ש-$w0$ הוא פתרון מסויים למשוואה, נקבל כי $y = Xw_0$. בנוסף, מכיוון ש-$v \in Ker(X)$, כלומר $Xv = 0$, מתקיים:

$$y = Xw_0 + 0 = Xw_0$$

כדי שיהיו אינסוף פתרונות, y חייב להיות במרחב העמודות של X. בנוסף, y חייב להיות אורתוגונולי לכל וקטור במרחב הניצב למרחב העמודות של X. המרחב הניצב

למרחב העמודות של X הוא הקרנל של X^T. לכן y חייב להיות אורתוגונלי לכל וקטור ב-$Ker(X^T)$. מכך ש-$y \perp Ker(X^T)$.

כעת נוכיח את הכיוון השני, כלומר אם $y \perp Ker(X^T)$, אז למערכת יש אינסוף פתרונות.

מכך ש-$y \perp Ker(X^T)$, זה אומר ש-y נמצא במרחב העמודות של X . מכך ש-X היא ריבועית ולא הפיכה, מרחב העמודות שלה לא פורש את כל המרחב $\mathbb{R}^m$, אלא תת-מרחב שלו.

אם y הוא במחב העמודות של X, אז קיים לפחות פתרון אחד w_0 כך ש-$y = Xw_0$. אולם, מכך ש-X סינגולרית (ולא הפיכה), קיימים אינסוף פתרונות בגלל שכל וקטור במרחב האפס של X יכול להיות נוסף ל-$w0$ ולייצר פתרון נוסף. באופן פורמלי, אם w_0 הוא פתרון מסוים, אז הפתרון הכללי יכול להיות כתוב כך:

$$w = w_0 + v \;\; for\ any \;\; v \in Ker(X)$$

בגלל שהגרעין של X לא טריוויאלי, בטוח קיים v כזה שאינו וקטור האפס, ולכן קיימים אינסוף w כאלה.

לכן, למערכת $y=Xw$ יש אינסוף פתרונות. ומשני הכיוונים, נקבל אמ"מ, וזה מסיים את ההוכחה.

(d) 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).

נחלק למקרים - $(X^T)X$ הפיכה, ולא הפיכה.

מקרה 1 - אם X^TX הפיכה:

נסמן $rank(X^TX) = d$. מכך שX^TX הפיכה, הגרעין שלה טריוויאלי, כלומר $Ker(X^TX) = \{0\}$. מהסעיף הראשון זה גם אומר ש-$Ker(X) = \{0\}$. בגלל ש-X^TX הפיכה, אז ניתן למצוא את w באופן ישיר:

$$w = (X^TX)^{-1}(X^T)y$$

וזהו הפתרון הייחודי.

מקרה 2 – אם X^TX לא הפיכה:

כלומר למשוואה $X^TXw = (X^T)y$ יש אינסוף פתרונות, שכן:

- אם X^TX לא הפיכה אזי $rank(X^TX) < d$
- הגרעין אינו טריוויאלי, כלומר $Ker(X^TX) \neq \{\mathbf{0}\}$.
- מכאן שמסעיף א' $Ker(X) \neq \{\mathbf{0}\}$.

כעת נביט במשוואה $X^TXw = (X^T)y$. בזכות העובדה ש-$Im(X^T) = Ker(X)^{\perp}$, אנו יודעים ש-$y \perp Ker(X^T)$. לכן y נמצא במרחב העמודות של X.

קבוצת הפתרונות של $X^TXw = (X^T)y$ יכולה להיות מתוארת באופן הבא:

- יהי w_0 פתרון מסוים ל-$X^TXw = (X^T)y$.
- כל פתרון w יכול להיות כתוב כך $w = w_0 + v$, כאשר $v \in Ker(X^TX)$.

מכיוון ש-$Ker(X) = Ker(X^TX)$, והעובדה ש-$Ker(X)$ הוא לא טריוויאלי, יש אינסוף פתרונות כאלה v, שיובילו לאינסוף פתרונות עבור w. כלומר הוכחנו את הנדרש.

2.2.2 Least Squares

Given a sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, the ERM rule for linear regression w.r.t. the squared loss is

$$\hat{\mathbf{w}} \in \underset{\mathbf{w} \in \mathbb{R}^d}{\operatorname{argmin}} ||\mathbf{X}\mathbf{w} - \mathbf{y}||^2$$

where $\mathbf{X}$ is the input matrix of the linear regression with rows as samples and y the vector of responses. Let $\mathbf{X} = U\Sigma V^\top$ be the SVD of $\mathbf{X}$, where U is a $m \times m$ orthonormal matrix, Σ is a $m \times d$ diagonal matrix, and V is an $d \times d$ orthonormal matrix. Let $\sigma_i = \Sigma_{i,i}$ and note that only the non-zero σ_i-s are

singular values of $\mathbf{X}$. Recall that the pseudoinverse of $\mathbf{X}$ is defined by $\mathbf{X}^\dagger = V\Sigma^\dagger U^\top$ where $\Sigma^\dagger$ is an $d \times m$ diagonal matrix, such that

$$\Sigma^\dagger_{i,i} = \begin{cases} \sigma_i^{-1} & \sigma_i \neq 0 \\ 0 & \sigma_i = 0 \end{cases}$$

2. Assuming that $\mathbf{X}^\top\mathbf{X}$ is invertible, show that the general solution we derived in recitation 3 $\left(\mathbf{X}^{\dagger\top}\mathbf{y}\right)$ equals to the solution you have seen in lecture 1 $\left([\mathbf{X}^\top\mathbf{X}]^{-1}\mathbf{X}^\top\mathbf{y}\right)$.

נבטא את $X^\dagger$ דרך שימוש ב-SVD:

בהנתן $X = U\Sigma V^T$, הפסודו-הופכית של X מוגדרת כך:

$$X^\dagger = V\Sigma^\dagger U^T$$

כעת נבטא את X^TX דרך שימוש ב-SVD:

$$X^TX = (U\Sigma V^T)^T(U\Sigma V^T) = V\Sigma^T U^T U\Sigma V^T =_{(U\ is\ orthonormal)} V\Sigma^T\Sigma V^T$$

כמו כן נשים לב כי

$$(X^TX)^{-1} = (V\Sigma^T\Sigma V^T)^{-1} =_{V\ is\ orthonormal\ and\ thus\ V^T=V^{-1}} V(\Sigma^T\Sigma)V^T$$

וכן מתקיים

$$(X^TX)^{-1}X^Ty = V(\Sigma^T\Sigma)^{-1}V^T(V\Sigma^TU^Ty) =_{V^TV=I} V(\Sigma^T\Sigma)^{-1}(\Sigma^TU^Ty)$$

נפשט את $(\Sigma^T\Sigma)^{-1}\Sigma^T$:

המטריצה סיגמא היא אלכסונית עם ערכים סינגולריים σ_i על האלכסון. יהי Σ_d להיות המטריצה האלכסונית עם ערכים סינגולריים (עם d ערכים שאינם 0):

$$\Sigma^T\Sigma = \Sigma_d^2$$

לכן

$$(\Sigma^T\Sigma)^{-1} = \Sigma_d^{-2}$$

ומכאן:

$$(\Sigma^T\Sigma)^{-1}\Sigma^T = \Sigma_d^{-2}\Sigma_d = \Sigma_d^{-1}$$

בסה"כ נקבל:

$$(X^TX)^{-1}X^Ty = V\Sigma_d^{-1}U^Ty$$

נזכר כי

$$\boldsymbol{X}^\dagger\boldsymbol{y} = V\Sigma^\dagger U^Ty =_{\Sigma^\dagger=\Sigma_d^{-1}} V\Sigma_d^{-1}U^Ty = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

לפיכך, הפתרון הכללי המתקבל באמצעות הפסודו-אינברס שווה לפתרון המופק באמצעות $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ כאשר X^TX היא הפיכה.

3.1.3

In the `Answers.pdf` file, describe in details the analysis process that lead you to the decisions of:

- Which features to keep and which not?
- Which features are categorical how how did you treat them?

- What other features did you design and what is the logic behind creating them?
- How did you treat invalid/missing values?
- Explain any additional processing performed on the data.

R Why do we want the split before preprocessing? We aim to construct a model capable of predicting on new samples it has not encountered previously, a concept known as generalization. To evaluate the success of our model, we require the test set to remain concealed and independent from our design choices. If we were to include the test set in preprocessing tasks, such as exploring feature correlations with the response, we would risk contaminating our training process. Consequently, our assessments of our model's generalization performance would be compromised, leading to inaccurate conclusions.

לתיאור תהליך הניתוח שהוביל להחלטות שהוזכרו:

בחירת פיצ'רים:

ראשית, בחרתי למחוק שורות שבהן תאריך המכירה של הבית (*date*) הוא ריק, שכן שורות אלה עלולות להיות בעייתיות.

אחרי זה קראתי לפונקציה *preprocess_test* כדי לחסוך כפל קוד. שם בחרתי למחוק את הפיצ'רים שניתן לראות די בקלות (פתיחת הטבלה באקסל, עין אנושית והיגיון בריא) שאין קשר בינהם לבין המחיר, ואלו: $id, date, long, lat$. בהמשך פונקציה זו בחרתי למחוק את *condition*, שכן בהמשך התרגיל ראיתי שמתאם הפירסון שלו עם המחיר נמוך מאוד. בסוף הפונקציה בחרתי ליצור עמודות למשתנים קטגוריים שעליהם אדבר בהמשך.

לאחר מכן, בחרתי לסנן שורות כפולות. ככתוב ב-*PDF*, גם סיננתי שורות לא הגיוניות שבהן שנת השיפוץ ישנה יותר מהשנה שבו הבית נוסד. בחרתי גם לסנן שורות בהן *sqft_living* קטן מ-400, שכן 37 מטר רבוע נשמע לי קצת לא הגיוני למאפיין זה. כמו כן מעל 15 שירותים לא נשמע הגיוני למדי, לכן בחרתי בסינון זה גם כן.

טיפול בפיצ'רי קטגוריות:

בחרתי ליצור קטגוריה של "ישן/חדש/אמצע" – כלומר בתים שנבנו בין 1900 ל1950, בתים שנבנו אחרי 2000, בתים שנבנו בין 1950 ל2000. כמובן שבחרתי ליצור 3 עמודות נפרדות לכל אחד מהם. כפי שלמדנו בתרגול, נפצל את זה ל3 משקולות. כך

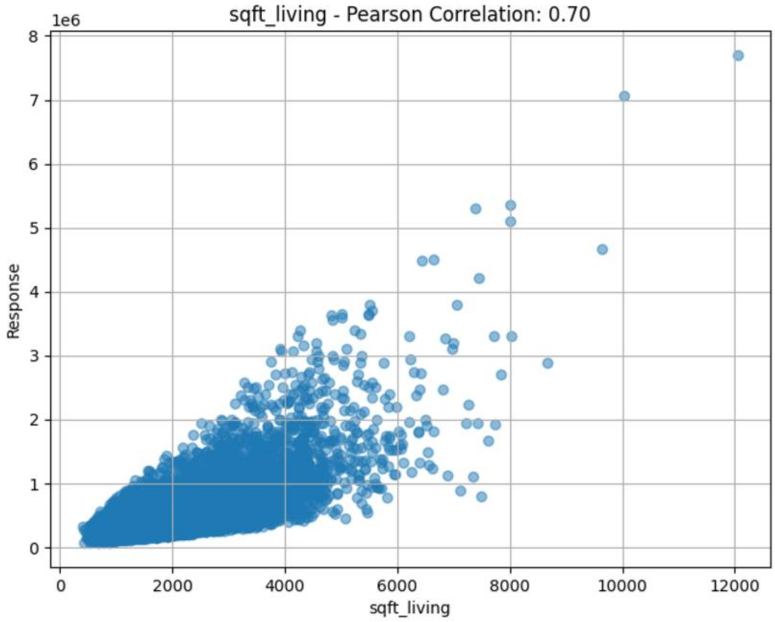
הפכנו את זה לאינדיקטור ו-3 עמודות בת"ל אחת בשניה. (בנוסף, כך אם הבית חדש נוכל להקפיץ את המחיר בהתאם למשקולת).

עיצוב של פיצ'רים חדשים:

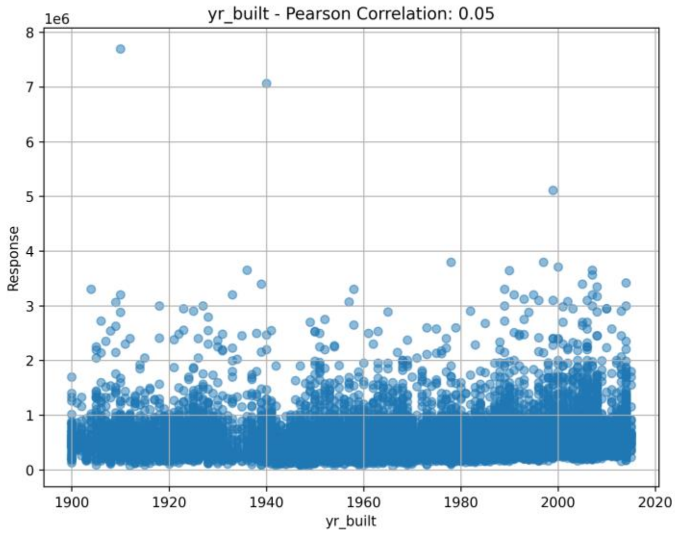
כאמור, 3 עמודות חדשות שמסמלות את "חדשנות" הבית.

3.1.4

For the Answers.pdf, choose two features: one that seems to be beneficial for the model and one that does not. Include in Answers.pdf the plots of the chosen two features, and explain how do you conclude if they are beneficial or not.



בחרתי במאפיין (פיצ'ר) זה לטובה, כי זה מראה קשר בין גודל הבית למחיר הבית. קורלציה של 0.7 היא קורלציה גבוהה, ולמעשה זה הפיצ'ר עם הקורלציה הגבוהה ביותר למחיר הבית. מה שמרמז שככל ששטח המגורים גדל, מחיר הבית נוטה לעלות גם כן. הקשר הליניארי החזק הזה הופך את *sqft_living* לתכונה מועילה לחיזוי מחירי בתים.



זהו מאפיין שלא נותן הרבה אינדיקציה על הבית. לתכונת *yr_built* יש מתאם נמוך עם מחירי הדירות (0.05), מה שמרמז על כך שהשנה שבה נוסד הבית אינה

Add the plot to the `Answers.pdf` file and explain what is seen. Address both trends in loss and in confidence interval as function of training size.

משפיעה באופן משמעותי על המחיר. הקשר הליניארי החלש הופך את *yr_built* לתכונה לא מועילה לניבוי מחירי בתים.

3.1.6

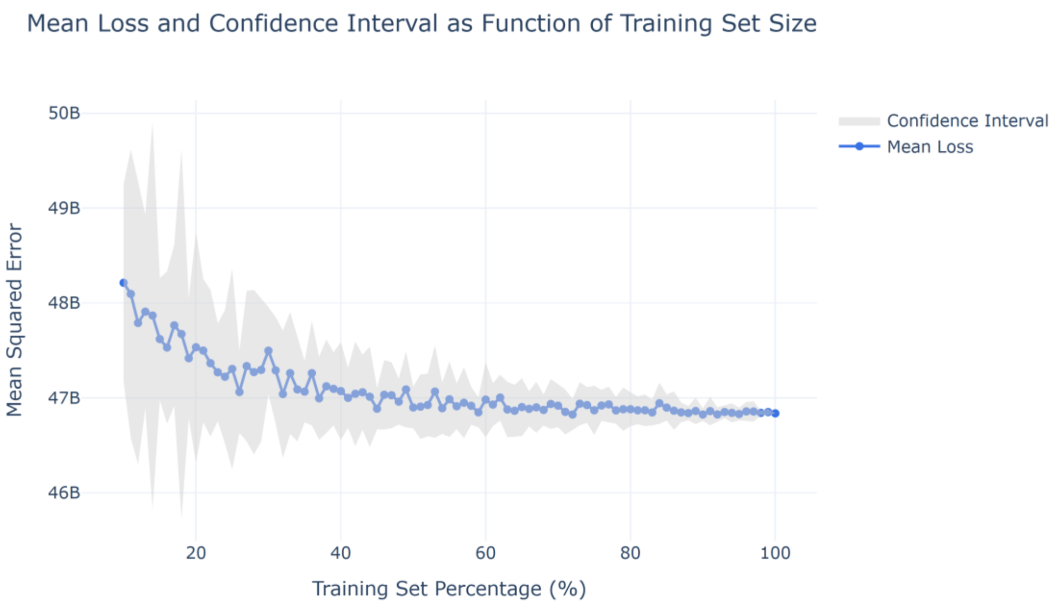
6. Fit a linear regression model over increasing percentages of the *training set*, and measure the loss over the *test set*:
   - Iterate for every percentage $p = 10\%, 11\%, ..., 100\%$ of the training set.
   - Sample $p\%$ of the train set. You can use the `pandas.DataFrame.sample` function.
   - Repeat sampling, fitting and evaluating 10 times for each value of p.

   Plot the mean loss as a function of $p\%$, as well as a confidence interval of $mean(loss) \pm 2 * std(loss)$. If implementing using the `Plotly` library, see how to create the confidence interval in Chapter 2 - Linear Regression code examples.

   Add the plot to the `Answers.pdf` file and explain what is seen. Address both trends in loss and in confidence interval as function of training size.

   R In this question, our objective is to assess the impact of enlarging the training set on the accuracy of the test set. By iteratively sampling the data 10 times, we aim to enhance the reliability of our conclusions. A single sampling instance may inadequately capture the data's variability, potentially leading to skewed results. However, by conducting the sampling procedure repeatedly, we mitigate the influence of outlier samples and ensure a more reliable conclusions regarding our model.

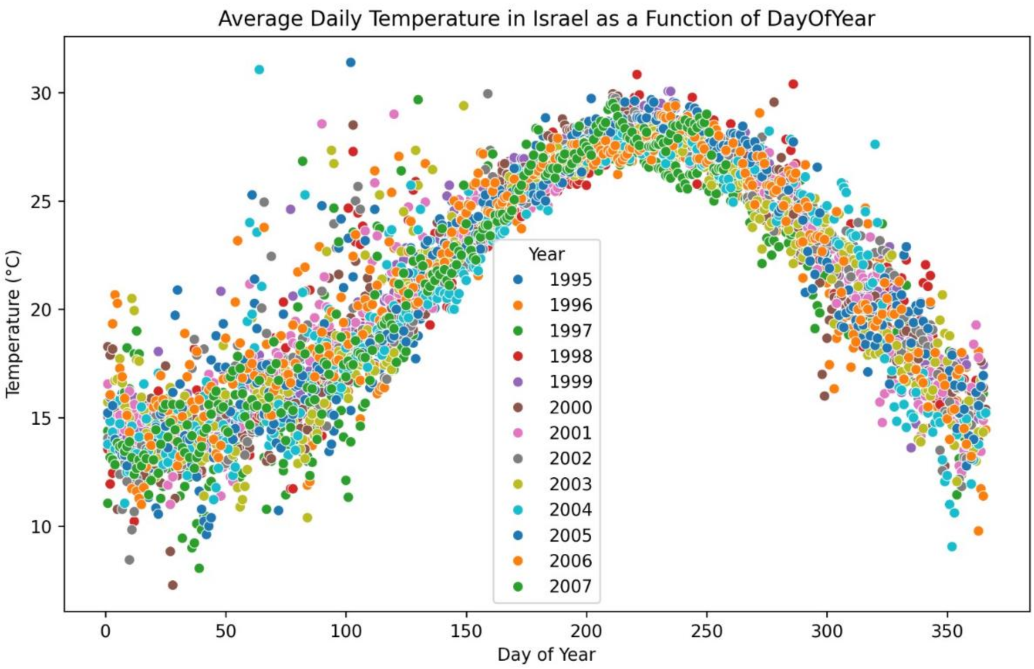
   Implementation clarification:
   (a) Notice that when predicting on the training set, you need to be consistent with some of the preprocessing



אנו יכולים לראות מגמה בגרף זה - כשאנו מאמנים את המודל על אחוזים גבוהים יותר של הטריינינג סט, אנו מקבלים ערכים קטנים יותר של הלוס, וזה בהחלט מה שציפינו לו בעקבות התרגולים בשבועות האחרונים (אלא אם נקבל אוברפיטינג, אבל כאן זה לא קרה, כנראה בשל סט אימון מספיק מגוון ורחב או לא מספיק חזרות בשביל שזה יקרה). כמו כן ניתן לראות שגם הטווח האפור מצטמצם (ה- *confidence interval*), ולאחר התייעצות קצרה עם צ'אט ג'יפיטי והאינטרנט מה נוכל להסיק מכך, ניתן להבין שאם הוא מצטמצם, זה אומר שהמודל יציב יותר.

:3.2.3

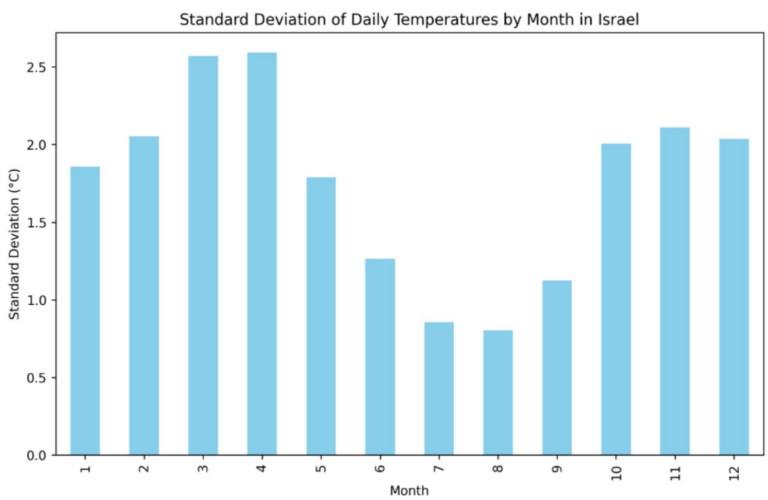
3. Filter the dataset to contain samples only from the country of Israel. Investigate how the average daily temperature (‘`Temp`‘ column) change as a function of the ‘`DayOfYear`‘.
   - Plot a scatter plot showing this relation, and color code the dots by the different years (make sure color scale is discrete and not continuous). What polynomial degree might be suitable for this data?



ניתן לראות שמדובר בגרף בצורה גלית, שמייצג עלייה עד לנקודה של 225 ימים, ולאחר מכן ירידה. מצד שני, ישנה ירידה קלה בין היום ה-0 ל-50 (לפחות בחלק מהשנים), כלומר מייצג פולינום מדרגה 3 ומעלה ולא פרבולה פשוטה. הנקודות מפוזרות על הגל ואין קשר ישיר או אינדקציה למידע נוסף בעזרת העין האנושית.

- Group the samples by ‘Month‘ (have a look at the pandas ‘groupby‘ and ‘agg‘ functions) and plot a bar plot showing for each month the standard deviation of the daily temperatures. Suppose you fit a polynomial model (with the correct degree) over data sampled uniformly at random from this dataset, and then use it to predict temperatures from random days across the year. Based on this graph, do you expect a model to succeed equally over all months or are there times of the year where it will perform better than on others? Explain your answer.

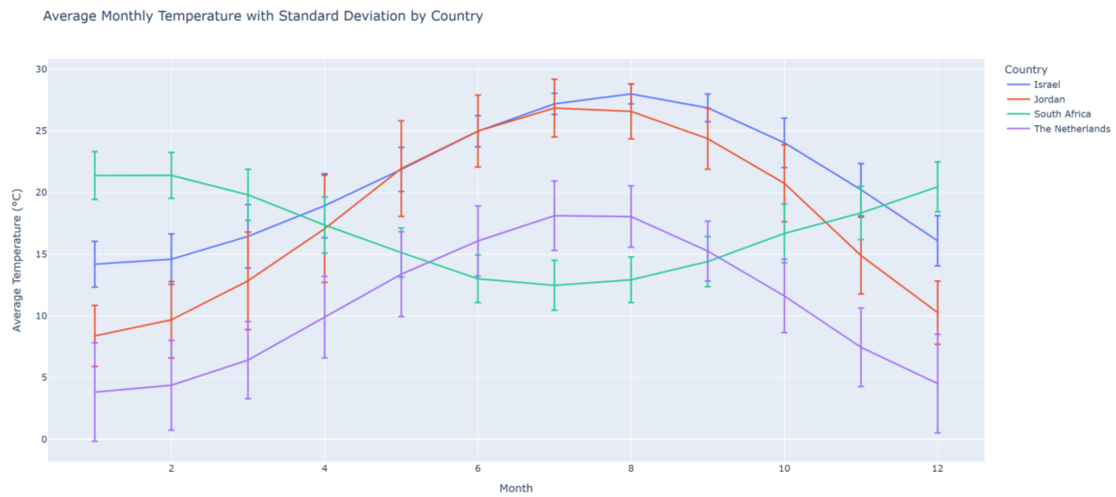
Add plots and answers to the `Answers.pdf` file.



לפי גרף זה, אני מצפה שהמודל לא יצליח במידה שווה על כל החודשים. ניתן לראות לדוגמא, שלכמה חודשים יש ערך גבוה יותר של סטיית תקן וזה אומר שהנתונים של הטמפרטורות בחודשים אלה יכולים להשתנות יותר במערך הנתונים שלנו. לדוגמא, חודשים מספר 3 ו-4 בגרף (מרץ ואפריל), עלולים לגרום למודל לקבל דגימות הרחוקות יותר מהממוצע ולכן לגרום למודל להיות פחות מדויק. אם הגרף היה נראה יותר יציב ונמוך כמו ביולי-אוגוסט, הייתי מצפה שהמודל היה מצליח במידה שווה על כל החודשים (בשל יציבותו), ואפילו במידה טובה באופן יחסי (בשל סטיית תקן נמוכה יחסית).

:3.2.4

4. Returning to the full dataset (including all 4 original countries), group the samples according to ‘Country‘ and ‘Month‘ and calculate the average and standard deviation of the temperature. Plot a line plot of the average monthly temperature, with error bars (using the standard deviation) color coded by the country. If using `plotly.express.line` have a look at the `error_y` argument.
Based on this graph, do all countries share a similar pattern? For which other countries is the model fitted for Israel likely to work well and for which not? Explain your answers.



ראשית, על השאלה הראשונה, לא כל המדינות חולקות דפוס דומה. לגרף של הולנד יש מגמה דומה לזו הישראלית אך הדגימות רחוקות יותר מזו ולכן המודל הישראלי יתאים פחות להולנד, והטמפרטורה הממוצעת שם נמוכה ב-10 מעלות צלזיוס מן הטמפרטורה הממוצעת בישראל באותו החודש (לכל חודש). מבחינת ערכים אין ספק שירדן היא הקרובה ביותר לישראל, אם כי הגרף מתנהג באופן קצת שונה (ולראייה, טמפ' כמעט זהה בחודשים 5 ו-6 לישראל, אבל לאחר מכן ירידה תלולה יותר מאשר בישראל). זה גם הגיוני מבחינה גיאוגרפית כמובן. דרום אפריקה הפוכה לישראל שכן לגרף יש צורה הפוכה ולא כל כך קרוב לדגימות הישראליות, לכן בהכרח המודל לא יעבוד טוב עליה – הטמפ' שם עולה מתי שהטמפ' בישראל יורדת

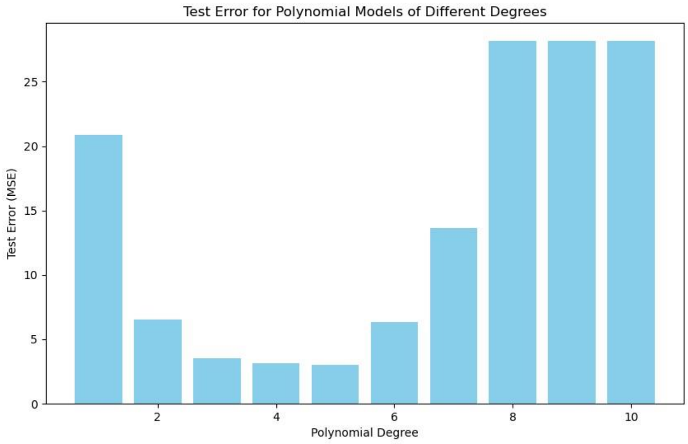
ולהפך. לא מפתיע ביחס למיקום הגיאוגרפי ועונות השנה שנגרמות מהסיבוב סביב השמש וכמות הקרינה בכל עונה לכל מדינה (ישראל – החצי הצפוני, דרום אפריקה – החצי הדרומי).

3.2.5:

5. Over the subset containing observations (i.e., samples) only from Israel perform the following:
   - Randomly split the dataset into a training set (75%) and test set (25%).
   - For every value $k \in [1, 10]$, fit a polynomial model of degree k using the training set.
   - Record the loss of the model over the test set, rounded to 2 decimal places.

   Print the test error recorded for each value of k. In addition plot a bar plot showing the test error recorded for each value of k. Based on these which value of k best fits the data? In the

   case of multiple values of k achieving the same loss select the simplest model of them. Are there any other values that could be considered?

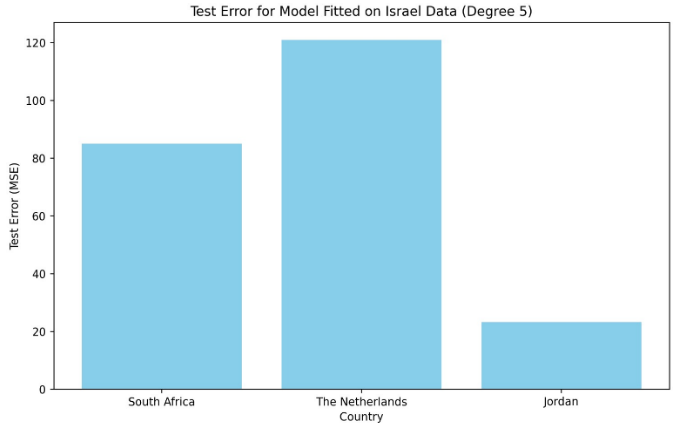



```
Degree 1: Test Error = 20.89
Degree 2: Test Error = 6.56
Degree 3: Test Error = 3.51
Degree 4: Test Error = 3.12
Degree 5: Test Error = 2.98
Degree 6: Test Error = 6.32
Degree 7: Test Error = 13.66
Degree 8: Test Error = 28.14
Degree 9: Test Error = 28.15
Degree 10: Test Error = 28.16
```

לפי התוצאות על מחשבי האוניברסיטה, $k = 5$ הוא המודל המתאים ביותר עם שגיאה של 2.98. גם 4 ו-3 קרובים לשגיאה של 5 ומביאים שגיאה קטנה יחסית עם 3.12 ו-3.51 בהתאמה.

:3.2.6

6. Fit a model over the entire subset of records from Israel using the *k* chosen above. Plot a bar plot showing the model's error over each of the other countries. Explain your results based on this plot and the results seen in question 3.



הגרף הזה לא אמור להפתיע אותנו. בין חודש 5 לחודש 9 הערכים של ישראל ודרא"פ היו יחסית קרובים, ואמנם התרחקו לאחר מכן – אבל בהולנד המרחק תמיד היה באיזור ה-10 מעלות צלזיוס. כלומר בניגוד לדרא"פ לא היו חודשים קרובים לישראל מבחינת טמפ'. כמובן שלמרות שדרא"פ קיבלה טעות נמוכה יותר מהולנד, היא עדיין לא קרובה בכלל לתוצאה המצוינת של ירדן (שוב, לא מפתיע בשל המיקום הגיאוגרפי, ותכונות מערכת השמש כמו עונות השנה וקרינה רבה יותר בקיץ).
לסיכום המודל עבד בצורה הטובה ביותר עבור חיזוי הטמפ' של ירדן ואין זו הפתעה בשל קירבתה לישראל והנתונים הדומים לה בסט האימון.

איפה השתמשתי בצ'אט ג'יפיטי בתרגיל:

לצערי, השתמשתי בו המון בקוד. אין לי ניסיון ב*pandas* או ב*numpy*... ניסיתי תחילה ללמוד דרך יוטיוב, אבל לא הצלחתי לסנן את המידע כפי שרציתי. עד עכשיו לא השתמשתי בו בתרגילים בקורסים כי בסוף זה פוגע בי במבחן, אבל כאן הרגשתי שאם אני חושב על סינון הדאטה בעצמי, הוספת עמודות בעצמי, משתנים קטגוריאלים וכו', אין סיבה שלא אעזר בו לכתיבת הקוד, בטח כשמדובר בסינטקס של ספריה שעדיין לא יצא לי להשתמש בה או ביצירת גרפים כאלה ואחרים.