תרגיל 2 | IML **(67577)**

שם: יואב שפירא| ת`ז: 312492838

6 באפריל 2022

חלק תאורטי

2.1

1. **נראה הכלה דו כיוונית:**

   (א) $\Longrightarrow$**: יהי** $v \in Ker(X)$ **כלומר** $Xv = 0$**, אז ברור כי גם** $X^TXv = 0$ **כלומר** $v \in Ker(X^TX)$**.**

   (ב) $\Longleftarrow$**:יהי** $v \in Ker(X^TX)$ **אז:**

$$\begin{aligned} X^TXv = 0 &\iff v^TX^TXv = 0 \\ &\iff (Xv)^TXv = 0 \\ &\iff Xv = 0 \end{aligned}$$

   כלומר $v \in Ker(X$**.**

2. **טענת עזר:** לכל מטריצה A **וקטורים** v, w **מתקיים:**

$$\langle Av, w\rangle = \left\langle v, A^Tw\right\rangle$$

הוכחה:

$$\begin{aligned} \langle Av, w\rangle &= w^TAv \\ {}_1 &= (A^Tw)^Tv \\ {}_2 &= v^T(A^Tw) \\ {}_3 &= \left\langle v, A^Tw\right\rangle \end{aligned}$$

כשהמעבר הראשון הוא מהגדרת מכפלה פנימית (והפכתי את סדר הוקטורים), 1 קיבוץ והגדרת טרנספוס, 2 היפוך סדר המכפלה וטרנספוס בהתאם, 3 הגדרת מכפלה פנימית.

כעת להוכחה הדרושה:

$$\begin{aligned} v \in Ker(A) &\iff Av = 0 \\ {}_1 &\iff \langle Av, w\rangle = 0 \ \ \forall w \\ {}_2 &\iff \left\langle v, A^T w\right\rangle = 0 \ \ \forall w \\ {}_3 &\iff v \in Im(A^T)^\perp \end{aligned}$$

כש1 מהגדרת מכפלה פנימית, 2 מהטענה למעלה, 3 מהגדרת מרחב ניצב.
כלומר, קיבלנו כי שני הת"מ הם ניצבים זה לזה. להשלמת ההוכחה הדרושה, ניתן גם לרשום בצורה שקולה: $v \in Ker(A)^\perp \iff v \in Im(A^T)$

3. X **לא הפיכה ולכן יש למערכת או 0 פתרונות או אינסוף פתרונות. נניח כי קיים פתרון,** y **, כלומר** $y \in Im(X)$**. כחלק מהוכחה בסעיף 2 הראינו שעבור** A **ריבועית מתקיים** $Ker(A) = Im(A^T)^\perp$**. מכיוון ש** A **ריבועית אז מתקיים גם מתכונת הטרנספוס ש** $Im(A) = Ker(A^T)^\perp$**. כלומר במקרה שלנו** $y \in Im(X) \iff y \perp Ker(X^T)$ **כדרוש.**

4. **ברור מליניארית שאם** X^TX **הפיכה אז יש פתרון יחיד למערכת. נראה שיש למערכת המשוואות אינסוף פתרונות אם"ם** X^TX **היא לא הפיכה:**
נניח שX^TX **לא הפיכה. אז מסעיף קודם, למערכת יש אינסוף פתרונות אם"ם** $X^Ty \perp Ker(X^TX)$ **. מסעיף א נובע שזה קורה אם"ם** $X^Ty \perp Ker(X)$**.ניקח וקטור** $v \in Ker(X)$**, כלומר כך** $Xv = 0$**. אזי בוודאי ש** $\langle Xv, y\rangle = 0$**. מהטענה בסעיף 2 מתקיים כי** $\langle Xv, y\rangle = \left\langle v, X^Ty\right\rangle$ **וקיבלנו כי אכן** $X^Ty \perp Ker(X) = Ker(X^TX)$**. כלומר** $X^TX\ invertible \iff \infty\ solutions$**.**

2.2

אני אוכיח את המשפטים בסדר קצת שונה כדי להקל את העניינים:

- P **היא סימטרית:**
מהגדרת המטריצה מתקיים:

$$P_{ij} = \sum_{l=1}^{k} v_{l_i} v_{l_j}$$

v_{l_i}, v_{l_j}**הם סקלרים אז מקומוטטיביות מתקיים:**

$$= \sum_{l=1}^{k} v_{l_j} v_{l_i} = P_{ji}$$

כלומר P **היא סימטרית.**

- $P=P^2$:

נבצע מכפלה לפי ההגדרה:

$$\begin{aligned} P^2=\left(\sum_{i=1}^{k} v_i v_i^T\right)^2 &= \\ &= v_1v_1^Tv_1v_1^T+v_1v_1^Tv_2v_2^T+...+v_1v_1^Tv_kv_k^T+ \\ &+.....+ \\ &+v_kv_k^Tv_1v_1^T+........................+v_kv_k^Tv_kv_k^T= \\ &=\sum_{i=1}^{k}\sum_{j=1}^{k} v_iv_i^Tv_jv_j^T \\ &=\sum_{i=1}^{k}\sum_{j=1}^{k} v_iv_j^Tv_jv_i^T \end{aligned}$$

המעבר האחרון הוא מקומוטטיביות. מכיוון ש הבסיס אורתונורמלי אז $v_j^Tv_j=1$ וקיבלנו את הדרוש:

$$P^2=\sum_{i=1}^{k}\sum_{j=1}^{k} v_i\cancel{v_j^Tv_j}v_i^T=\sum_{i=1}^{k} v_iv_i^T=P$$

- $\forall v\in V$ מתקיים $Pv=v$:

כל השורות בP הם בת"ל (וגם העמודות כי P סימטרית אבל לא חשוב) ולכן $Rank(P)=k$. לכן נובע כי $dim\left(Im(P)\right)=k$, ולכן נובע כי $Im(P)=V$ היא בעצם מכסה את כל התת מרחב. מכאן נובע שלכל $v\in V$ קיים $x\in V$ כך ש

$$Px=v$$

מסעיף קודם נובע כי

$$P^2x=Px$$

נשלב הכל ונקבל את הדרוש:

$$Pv=PPx=P^2x=Px=v$$

- $(1-P)\,P=0$:

מסעיף 2 זה נובע מיידית:

$$(1-P)P=P-P^2=P^2-P^2=0$$

- בעזרת הסעיפים הקודמים נוכיח יותר ממה שצריך: לP יש בדיוק k ע"ע שווים ל1. תחילה מהגדרת ע"ע:

$$Pv = \lambda v$$

מסעיף 2:

$$P \cdot Pv = P^2 = Pv$$

אבל גם מהגדרת הע"ע:

$$P^2 v = P\lambda v = \lambda Pv = \lambda^2 v$$

כלומר קיבלנו כי $\lambda^2 = \lambda$ ולכן $\lambda \in \{0,1\}$. כמובן ש 0 מתאים לו"ע w כך ש $w \in Ker(P)$. הראינו קודם ש $V = Im(P)$ ומשיקולי מימד נובע כי $dim\left(Ker(P)\right) = 0$, כלומר אין ו"ע שמתאים להם הערך 0. כלומר לP יש רק ע"ע ששווים ל1, ולכן יש בדיוק k כאלה. כמקרה פרטי של סעיף 3, לכל $i \le k$ מתקיים כי $Pv_i = v_i$ כלומר אכן הבסיס מהווה בסיס של ו"ע המתאימים לע"ע 1.

2.3

1. אנחנו נדרשים להוכיח את הטענה: עבור X כך ש X^TX הפיכה, צ"ל:

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

טענת עזר: תהי $D = diag(\sigma_1^2, ..., \sigma_d^2)$ המטריצה האלכסונית וV אורתונורמלית כך ש $X^TX = VDV^T$ (מתרגיל 1). אז מתקיים:

$$D^{-1}\Sigma^T = \Sigma^{\dagger}$$

הוכחה: D אלכסונית עם ערכים שונים מ0 כי X^TX הפיכה ולכן אין לה ע"ע 0. לכן D הפיכה ומתקיים כי $D^{-1} = diag(\frac{1}{\sigma_1^2}.....\frac{1}{\sigma_d^2})$. נתבונן על $D^{-1}\Sigma^T$:

$$\left[D^{-1}\Sigma^T\right]_{ij} = \sum_j D^{-1}_{ji}\Sigma^T_{ij}$$

נבחין כי $D^{-1}\Sigma^T \in \mathbb{R}^{d\times m}$, וכי עבור $i > d$ מתקיים כי $\Sigma^T_{ij} = 0$, ולכן הערך של הסכום שווה ל0. עבור כל $i \le d$ מתקיים:

$$\sum_j D^{-1}_{ji}\Sigma^T_{ij} = \sum_j D^{-1}_{ii}\Sigma^T_{ii} = \frac{1}{\sigma_i^2}\cdot\sigma_i = \frac{1}{\sigma_i}$$

ומכאן קיבלנו בדיוק את ההגדרה של $\Sigma^{\dagger}$.

כעת להוכחה הדרושה:

נציג את X^TX בעזרת פירוק הSVD שלה:

$$\begin{aligned}X^TX &= \left(U\Sigma V^T\right)^T\left(U\Sigma V^T\right)\\ {}_1&= V\Sigma^TU^TU\Sigma V^T\\ {}_2&= V\Sigma^T\Sigma V^T\\ {}_3&= VDV^T\end{aligned}$$

כש1 מתכונה הטרנספוס, 2 מכך שU אורתונורמלית, ו3 מתרגיל 1.
עכשיו נשתמש בזה ונסתכל על $\left(X^TX\right)^{-1}X^T$:

$$\begin{aligned}\left(X^TX\right)^{-1}X^T &= \left(VDV^T\right)^{-1}X^T\\ {}_1&= \left(VDV^T\right)^{-1}\left(U\Sigma V^T\right)^T\\ {}_2&= VD^{-1}V^TV\Sigma^TU^T\\ {}_3&= VD^{-1}\Sigma^TU^T\\ {}_4&= V\Sigma^{\dagger}U^T = X^{\dagger}\end{aligned}$$

כשב1 הצבתי את הSVD של X, 2 נובע מתכונת האורתונורמליות של V להיפוך, 3 נובע מאורתונרלמיות של V לכפל עם עצמו, 4 מהטענת עזר.

2. X היא למעשה טרנספורמציה ליניארית: $X:\mathbb{R}^m \longrightarrow \mathbb{R}^d$. כאן אנחנו נדרשים למעשה להוכיח, ש$dim\left(Ker(X^TX)\right)=0$ (כלומר X^TX הפיכה) אמ"ם $Rank(X)=d$ - כלומר, הטרנספורמציה 'מכסה' את כל התת מרחב $\mathbb{R}^d$ (ובפרט הגרעין ריק ולכן היא חח"ע). הוכחה פשוטה:
אם $Rank(X)=d$ אז $dim(Im(X))=d$ ואז $dim(Ker(X))=0$ כלומר $dim(Ker(X^TX))=0$ (נובע מהשאלה הראשונה בתרגיל זה) כלומר X^TX הפיכה.

3. אם X^TX לא הפיכה, יש אינסוף פתרונות. בהתאם לרמז: האינסוף פתרונות נובעים מכך ש לכל $i>d$ ניתן לבחור הרבה ערכים עבור $\hat{w}_i$. לכל $i\leq d$, יש לנו פתרון יחיד מינימלי עבור $\hat{w}_i$. כלומר, לכל $\overline{w}$ שמקיים מינימליות, מתקיים כי $\overline{w}_i=\hat{w}_i$ עבור $i\leq d$. עבור שאר ה $i>d$, הגדרנו כי $\hat{w}_i=0$ (כלומר אלה פיצ'רים בדאטא ש'לא מעניינים' אותנו). כלומר, $||\hat{w}||^2=\sum_{i=1}^{d}\overline{w}_i^2+0$. מכאן נובע ש:

$$||\overline{w}||^2=\sum_{i=1}^{m}\overline{w}_i^2=\sum_{i=1}^{d}\overline{w}_i^2+\sum_{i=d+1}^{m}\overline{w}_i^2=||\hat{w}||^2+C\geq||\hat{w}||^2$$

כשC קבוע כלשהו. מחיוביות בהחלט של נורמה נובע כי $||\hat{w}||\leq||\overline{w}||$ לכל $\overline{w}$ שמקיים את ה$argMin$.

חלק פרקטי

3.1

1. מפורט כאן התהליך שביצעתי עבור הדאטא ב$Preprocessing$. עבור פיצ׳רים קטגוריים, ביצעתי את שיטת ה$Dummies\ variables$, שמוסיפה פיצ׳ר בוליאני לכל דגימה, עבור כל ערך של הקטגוריה.

   (א) Na: החלפתי את כל ערכי הNa בערך הממוצע של הפיצ׳ר, כך שהם לא ישפיעו על השונות של הפיצ׳ר.

   (ב) השמטתי את כל השורות עבורן $price \leq 0$. בסופו של דבר נשארתי עם 21609 דגימות.

   (ג) $Categorials$: את הפיצ׳רים הקטגוריים פיצלתי ל$Dummy\ variables$ - נניח ש F הוא פיצ׳ר קטגורי שיכול לקבל ערך אחד מתוך $\{\alpha_1....\alpha_n\}$, וS היא סאמפל כך ש$S_F = \alpha_i$. מחקתי את F והוספתי n עמודות בוליאניות (אינדיקטורים) בשם F_i (לכל $i \leq n$) כך ש

$$S_{F_i} = \begin{cases} 1 & S_F = \alpha_i \\ 0 & else \end{cases}$$

   ככה המידע לא נאבד מצד אחד, ומצד שני הוא לא ׳מבלבל׳ את המודל בערכים שאין ביניהם שום יחס סדר.

   i. $zipcode$: ערך קטגורי קלאסי, פיצלתי אותו ל$dummy\ vars$.

   ii. הפיצ׳רים של $lat, long$ הם מיותרים, שכן הפיצ׳ר של מיקוד מכיל את אותו המידע - מיקום הנכס. החלטתי להשמיט את ה$lat, long$ ולהתייחס רק אל המיקוד.

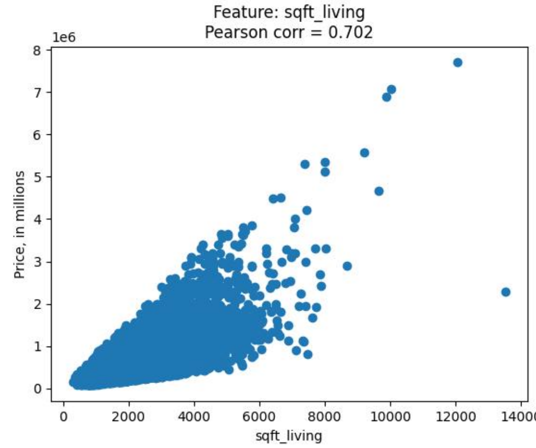
   iii. הפיצ׳ר $date$: פירשתי את הפיצ׳ר כקטגורי, כי אין משמעות לחודש שהוא "גדול" מחודש אחר. כך שחילקתי אותו לקטגוריות של שנה חודש, ויום. לא רציתי להתעלם כי לתאריך בשנה של הרכישה עשוי להיות השפעה על המחיר (אפילו ברמת היום בשבוע).

   (ד) **הוספתי:**

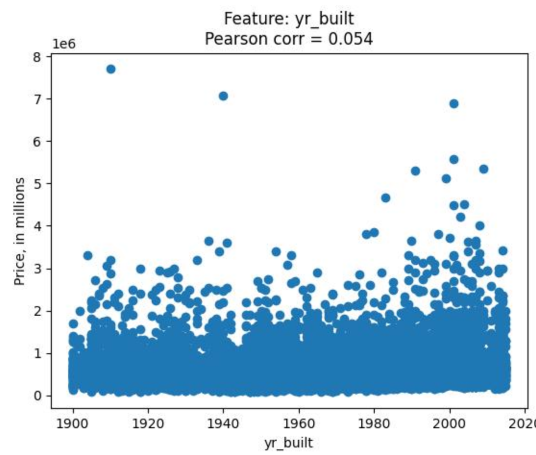
   i. היחס בין חדרי השינה לחדרי המקלחת נשמע לי חשוב אז הוספתי פיצ׳ר של היחס ביניהם: $bathroom\ \ bedroom\ ratio$.

   ii. קראתי באינטרנט ש$sqft\ living15$ הוא "הגודל הממוצע של הסלון ב 15 הבתים הקרובים ביותר". בדומה כך גם $sqrt\ lot15$. הוספתי פיצ׳ר שהוא היחס בין $sqrt\ living$ ל$sqrt\ living15$. הפיצ׳רים האלה הם סוג של ׳מעמד חברתי׳ בשכונה ולכן קראתי להם $social\ status\ living$ ו$social\ status\ lot$. (בחישוב הקורלציה של פירסון יצא להם ציון של 0.6 בערך, לא רע)

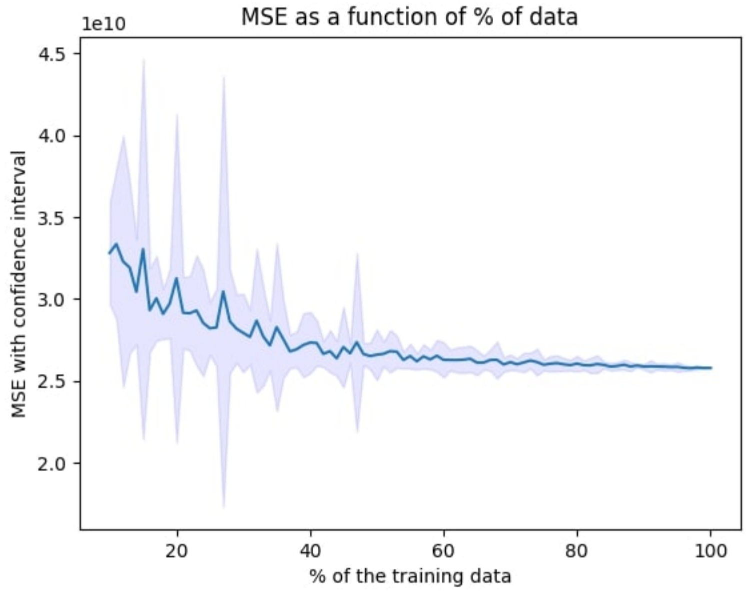
2. פיצ׳ר שנראה חשוב למודל: $sqrt\ living$. ניתן לראות בגרף שאפשר לשרטט קו ליניארי שיסביר את המגמה, ובנוסף $Pearson\ correlation$ קיבל ציון גבוה של קורלציה חיובית (1 משמעותו קורלציה חיובית מלאה). זה הפיצ׳ר שמתאר את גודל הסלון, וברור למה הוא משפיע באופן ישיר על המחיר:



פיצ׳ר שנראה לא כל כך חשוב למודל, הוא $yr\ built$. ניתן לראות בגרף שהנקודות די מפוזרות וגם הציון של הקורלציה נמוך מאוד (0 משמעותו אין קורלציה בכלל). זה הפיצ׳ר שמתאר את שנת הבנייה של הבית. וברור למה אין קורלציה - באותה שנה יכולים להבנות בתים פשוטים ובתי פאר. אולי **מתוך בתים שמסכימים על שאר הפיצ׳רים**, שנת הבנייה היא פקטור - שכן בית ישן יותר יהיה שווה פחות (אלא אם כן חודש). (לכן השארתי אותו כפיצ׳ר רגיל ולא קטגורי).



3. הגרף שיצא לי עבור הMSE (צריך להדגיש שלערכי הy אין כל משמעות ממשית כאן. שכחתי להוריד אותם):



ניתן ללמוד מספר תובנות:

(א) ככל שיש יותר דאטא לאימון, אז השגיאה יותר נמוכה. בנוסף, עבור כמות דאטא נמוכה ההתנהגות לא צפויה (בא לידי ביטוי בקפיצות).

(ב) ככל שיש יותר דאטא לאימון, **הביטחון** בחיזוי הוא יותר גבוה. ניתן לראות כי השונות של הMSE (עבור 10 אינטרוולים בכל פעם) הייתה מאוד גדולה עבור כמות נמוכה של דאטא והצטמצמה מאוד ככל שהוא גדל.

(ג) הגרף, בגדול, לא קעור (כלומר אין לו 'בטן') דבר שמעיד על כך שאין כאן $overfit$.

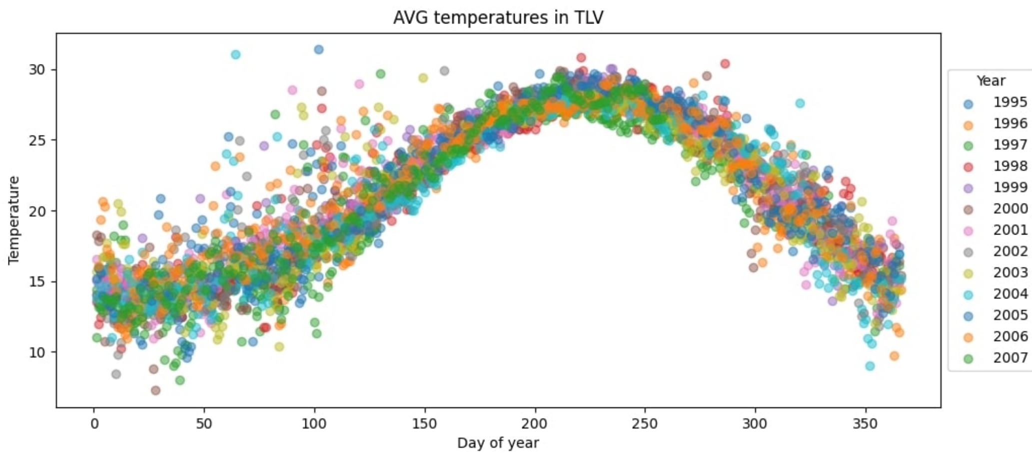
(ד) אם היינו קצרים בדאטא: אפשר לראות שהמודל מתייצב כבר ב75%, ולכן היינו יכולים לתת לו פחות דאטא לאימון ועדיין היינו מקבלים תוצאות מספקות. בכל זאת, זה היה עולה לנו בביטחון שכן השונות שם גדולה יותר.

3.2

1. מימוש קוד

2.

(א) הגרף הבא מציג את הטמפרטורה בת"א בכל יום מימות השנה, כשלכל שנה צבע שונה:

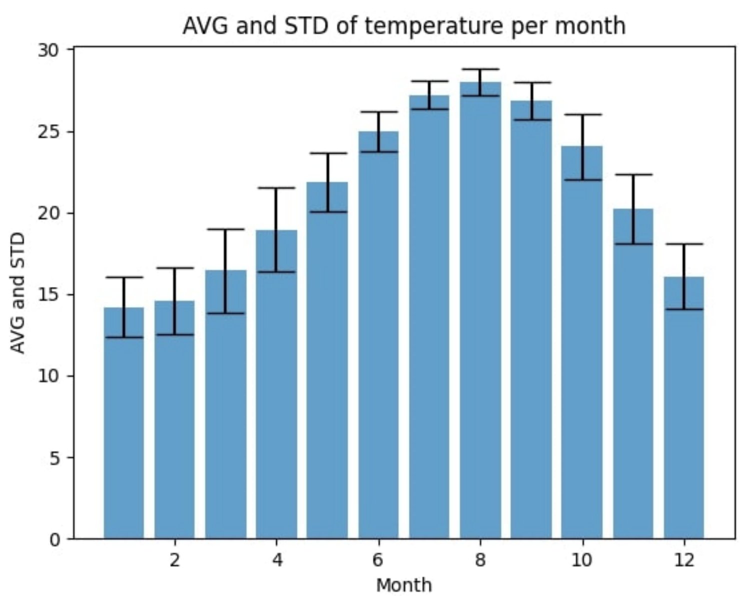


ניתן לראות כי יחסית, אין הרבה 'ברחנים', וגם אם יש - הם שונים בצבעם, כלומר בכל שנה ההתנהגות של הגרף די דומה. ניתן גם לראות שונות גדולה יותר בחודשי החורף מאשר בחודשי הקיץ כיאה למזג האוויר הנהדר בלבנט.

כמו כן אם מסתכלים טוב אפשר לשים לב שיש בגרף שני קיצונים וגם שתי נקודות פיתול (ביום ה125, וביום ה300, בערך).

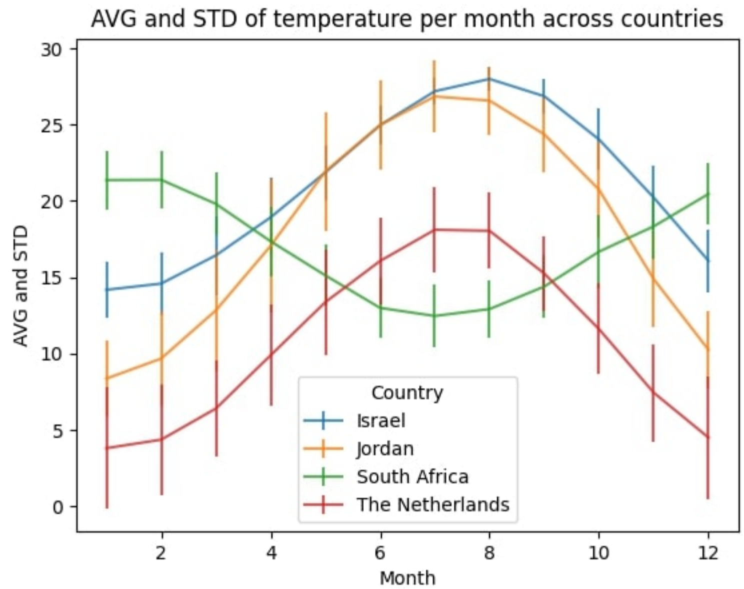
בגלל הנקודות האלה אני חושב שכדאי לבחור דרגה אי זוגית לפולינום (אחרת הצורה שלו תהיה פרבולית יחסית), ולדעתי 5 תהיה בחירה טובה.

(ב) הגרף הבא מדגיש את נושא השונות: לכל חודש הוא מציג את הטמפרטורה הממוצעת (גובה הבר) ואת הסטיית תקן (גודל הגביע). ניתן לראות בבירור שבחודשים הקרים יש יותר שונות בטמפרטורה:



בגלל השונות המשתנה, האומדן יהיה קצת מוטה: אומד 'מוצלח' יחזה את הדאטא לא רע - אבל בכל זאת יטעה יותר עבור ימים שנלקחו מהחורף לעומת אלה שנלקחו מהקיץ.

(ג) הגרף הבא מציג את אותו מידע מסעיף קודם רק עם פילוח לפי מדינות:



מה שאפשר ללמוד מהגרף:

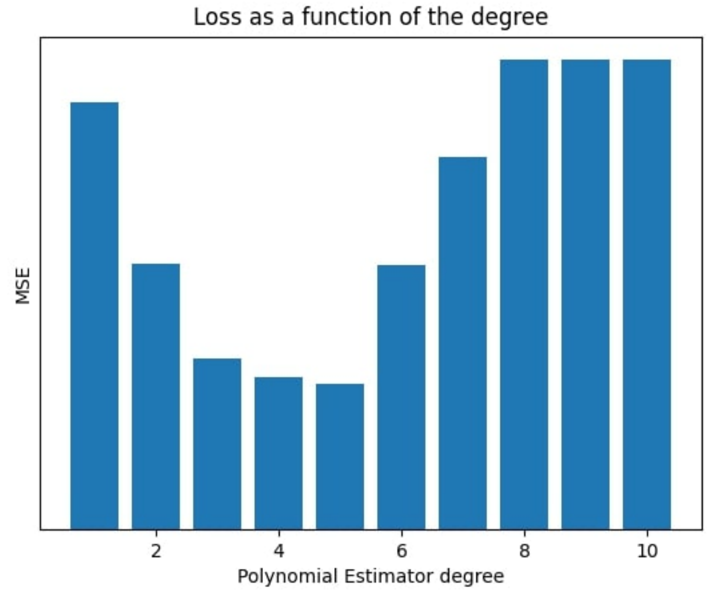
i. דרום אפריקה באופן מפתיע נמצאת בחצי כדור שונה מכל השאר

ii. ירדן וישראל חולקות מזג אוויר דומה ולכן מודל שאומן על ישראל יצליח לחזות די טוב את הדאטא מירדן

iii. מזג האוויר הכי שונה **בממוצע** מישראל נמצא בהולנד. למרות המגמה המאוד דומה, הערכים הם תמיד שונים מאוד, ולכן מודל שאומן על ישראל לא יצליח לחזות טוב את הדאטא מהולנד.

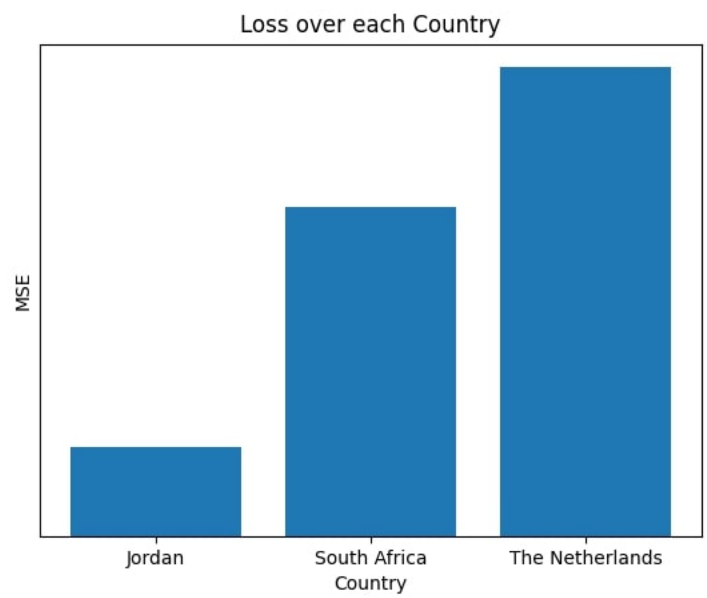
iv. לגבי דרום אפריקה - בגלל עונות המעבר שבהן הטמפרטורות יחסית דומות בין ישראל לדרום אפריקה, יכול להיות שהמודל שאומן על ישראל, **בממוצע** (MSE) ישיג תוצאה סבירה. יותר טובה משל הולנד בטוח.

(ד) הגרף הבא מציג את השגיאה הממוצעת עבור **דאטא מישראל** לכל דרגת פולינום מ1 ל10.
הערה: הדאטא מוצג פה ב$log\ scale$ ההדפסות בקוד נשארו בערכיהם המקוריים.



התחזית הייתה נכונה ואכן מתקבלת נקודת מינימום ב$k = 5$. החל משם כבר מתחיל $Overfit$.

(ה) הגרף הבא מציג את הMSE עבור 3 המדינות, כשהמודל אומן על דאטא מישראל:



ניתן לראות שבאמת המודל חזה את ירדן הכי טוב, ואת דרום אפריקה יותר טוב מהולנד - תודות לעונות המעבר שבהן הטמפרטורה של ישראל ודרום אפריקה יחסית דומות.