Introduction to Machine Learning (67577)

Exercise 2
Linear Regression

Second Semester, 2022

Contents

IML 2 תרגיל 317564112 עמרי מתאמת

1. $\ker(X) = \ker(X^TX)$

נוכיח: הכלה לשני הכיוונים - $\ker(X)\subseteq\ker(X^TX)$ ו $\ker(X^TX)\subseteq\ker(X)$

- נניח $v\in\ker(X)$. כלומר $Xv=0$

נכפיל משמאל ב X^T ונקבל $X^TXv=0$

מוגדרת היטב

כלומר $v\in\ker(X^TX)$

- נניח $v\in\ker(X^TX)$ - כלומר $X^TXv=0$

נסתכל על $\langle Xv,Xv\rangle = (Xv)^TXv = v^TX^TXv = v^T0_v = 0$

מההנחה.

מהגדרה מ"פ $\langle Xv,Xv\rangle=\|Xv\|_2^2$ ומאפיוני נורמה $\|v\|=0 \iff v=v_0$

לכן $Xv=0$ כלומר $v\in\ker(X)$

2. x A: $Im(A^\top)=Ker(A)^\perp$

נוכיח הכלה לשני הכיוונים — $Im(A^T)\subseteq Ker(A)^\perp$ ו $(Ker(A))^\perp\subseteq Im(A^T)$

- נניח $v\in Im(A^T)$ כלומר קיים u_1 כך ש $A^Tu_1=v$

יהי $u_2\in Ker(A)$ - $\langle u_2,v\rangle = \langle u_2, A^Tu_1\rangle = (A^Tu_1)^Tu_2 = u_1^TAu_2 = u_1^T\cdot 0_v = 0$

הגדרת v, פירוק v לפי הגדרת Im, טרנספוז, הגדרת u_2

לכל $u_2\in Ker(A)$ $\langle u_2,v\rangle=0$

$\Downarrow$

$v\in Ker(A)^\perp$

- נניח $v \in \ker(A)^\perp$ כלומר לכל $u \in \ker(A)$ $\langle u, v\rangle = 0$

יהי $u \in \ker(A)$ $\langle u, v\rangle = 0$

כפל במטריצה פיתגורס מ"פ

$$\Downarrow$$

$$\langle A^T A v, u\rangle = 0$$

פירוק והרכבה מ"פ

$$\Downarrow$$

$$(A^T A v)^T u = 0$$

$$v^T A A^T u = 0$$

$$\langle A^T v, A^T u\rangle = 0$$

$$v \in \ker(A A^T)$$

מטענה קודמת

$$\Downarrow$$

$$v \in \ker(A^T)$$

נזכיר כי $\overline{Im(A)^\perp} \subseteq \overline{\ker(A^T)}$ ולכן נסיק $\ker(A^T) \subseteq Im(A)^\perp$

ולכן נסיק $v \in Im(A)^\perp \Rightarrow v \in \ker(A^T)$. (כי העתקות נכונות מהמטריצה לינארית $A \to B$)

נראה $\overline{Im(A)^\perp} \subseteq \overline{\ker(A^T)}$:

יהי $v \in \overline{Im(A)^\perp} \Rightarrow$ נניח להפך $v \in (Im(A)^\perp)^\perp \Rightarrow v \in Im(A)$

נמצא $u \in Im(A^T)^\perp$ כך ש $\langle u, v\rangle \neq 0$, ומההנחה קיים u

כזה $\ker(A^T)$ אינו טריוויאלי. מתשלים את מכפלת נאורך ונקבל $A^T A v \in Im(A^T)$

$$\langle v, A^T A v\rangle = 0$$

$$\langle v, A^T A v\rangle = v^T A^T A v = (v^T A^T)^T (Av) = \langle Av, Av\rangle = \|Av\|^2 = 0$$

מחלקה נורמה $\|\cdot\| = 0$ אמ"מ $\cdot = 0$ כלומר $Av = 0$ גורר

$v \in \ker(A)$ - מכאן אפשר להמשיך $v \notin \overline{\ker(A)} \Leftarrow v \notin \ker(A^T)$

$$\Downarrow$$

$$v \in \overline{\ker(A^T)}$$

כעת נוכיח את השלמה.

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

נתון: $X \in \mathbb{R}^{m \times m}$

X סינגולרית

$y = Xw$

צ"ל: $y \perp Ker(X^T) \iff$ למערכת אינסוף פתרונות

כיוון ראשון: $y \perp Ker(X^T)$

מהסעיף הקודם $Ker(X^T)^\perp = Im(X^T)$, כלומר $y \in Im(X^T)$

כלומר קיים v_1 כך ש $X^T v_1 = y$.

מכיוון שהמטריצה אינה הפיכה $Ker(X) \neq 0_v$ כלומר קיים

$0_v \neq u \in Ker(X)$, ונוכיח שמתקיים $X(v+u) = y$ כי

$$Xv + Xu = y + 0_v = y$$

באופן אינדוקטיבי נוכל להכפיל את u בכל סקלר $\lambda \in \mathbb{R}$

ולכן קיבלנו אינסוף פתרונות.

כיוון שני: למערכת אינסוף פתרונות.

קיים $w_0 \in \mathbb{R}^m$ כך ש $Xw = y$, כלומר $y \in Im(X)$

מהסעיף הקודם $y \in Ker(X^T)^\perp$ ולכן $y \perp Ker(X^T)$

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).

לזכרון

$$X^TXw = X^Ty$$

1. פתרון יחיד - w $\Leftarrow$ X^TX סינגולרית

2. אינסוף פתרונות $\Leftarrow$ X^TX לא סינגולרית (הפיכה)

הוכחה

1. נובע מהתכונה של שלילת ההפיכה לפי חומר של לינארית 1.

2. מסעיפים 1,2 $Ker(X^TX)^\perp = Ker(X)^\perp = Im(X^T)$

לכן $Im(X^T) \perp Ker(X^TX) \iff Im(X^T) = Ker(X^TX)^\perp$

מהנתון $X^TXw = X^Ty$, לכן $X^Ty \in Im(X^T)$ לכן $X^Ty \perp Ker(X^TX)$

ומסעיף 3 נסיק שקיימים אינסוף פתרונות.

J. recitation 1. Let $V \subseteq \mathbb{R}^a$, $dim(V) = k$ 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. $$[P]_{i,j} = \left[\sum_{l=1}^{k} v_l v_l^T\right]_{i,j} =$$

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad (1\,1\,1\,1)$$

1) מתכונות מכפלה חיצונית

$$\sum_{l=1}^{k} v_{l,i} \cdot v_{l,j} =$$

$$\begin{bmatrix} 2 \\ 2 \\ 2 \\ 2 \end{bmatrix} \quad (2\,2\,2\,2)$$

2) כפל מספרי סימטרי

$$\sum_{l=1}^{k} v_{l,j} \cdot v_{l,i} =$$

$$\left[\sum v_l v_l^T\right]_{j,i} = [P]_{j,i}$$

$$\Downarrow$$

המטריצה סימטרית.

b. $$Pv_j = \sum^{k} v_i v_i^T v_j = \sum v_i \underbrace{(v_i^T \cdot v_j)} = v_j$$

וקטורים א"נ לכן $v_i^T v_j = \begin{cases} 0 & i=j \\ 1 & i \neq j \end{cases}$

לכן $(v_1 \ldots v_k)$ הם ו"ע של ע"ע 1.

לכל $u \notin Span(v_1 \ldots v_k)$ מתקיים $u \perp Span(v_1 \ldots v_k)$ לכן

$$Pv = \sum^{k} v_i v_i^T \cdot u = \sum u_i (v_i^T \cdot u) = \sum u_i \cdot 0 = 0_v$$

כי $v_i \subseteq Span(v_1 \ldots v_k)$ ו $u \perp Span(..)$.

לכן כל ה u האחרים הם ו"ע של ע"ע 0.

c. $u \in V$

$\Downarrow$

פירוק וקטור מבסיס $(v_1 \dots v_k)$

$$u = \sum a_i v_i \quad (A)$$

$$Pu = \sum^{k} v_i v_i^T u = \sum^{k} v_i v_i^T \cdot \sum^{k} a_i v_i = \sum^{k} v_i \sum^{k} a_i v_i^T v_i =$$

פירוק הכפל   שימוש ב (A)   מכפלה וקטורים אנך   לינאריות

$$= \sum^{k} v_i a_i = \underline{\underline{u}}$$

הגדרת u

d. נרחיב את המרחב הוקטורי $V_1=(v_1 \dots v_k)$ ל $\mathbb{R}^d$ ע"י הוקטורים $V_2=(v_{k+1} \dots v_d)$ משלים ל $V_1 \oplus V_2$ משלים השלמה, ונוכל לכתוב כל $u \in \mathbb{R}^d$ כצירוף של וקטורי הבסיס V_1, V_2.

$$Pv = P\Big(\underbrace{\sum_1^{d} a_i v_i}_{\text{צ"ל של } V_1} + \underbrace{\sum_{k}^{d} a_i v_i}_{\text{צ"ל של } V_2}\Big) = P\Big(\sum_{k=1}^{d} a_i v_i\Big) + \underbrace{\sum^{k}\Big(v_i v_i^T \cdot \sum_{k+1}^{d} a_i v_i\Big)}_{} = P\Big(\sum_1^{k} a_i v_i\Big) = u$$

מכפלה של וקטורים מאונכים יצביעו שורה אפס.

לפי סעיף $Pv = P \cdot c$.

$$P^2 v = P(Pv) = Pv = P$$

לפי השורה הקודמת   לפי סעיף c

$\Downarrow$

$$\forall v \in \mathbb{R}^d \quad P^2 v = Pv = v$$

$\Downarrow$

$$P^2 = P$$

e. $$(I-P)P = IP - P^2 = P - P^2 = P - P = 0_V$$

לפי סעיף d

6. $w_1 = (X^TX)^{-1}X^Ty$ הפתרון מהתרגול!

$w_2 = X^+y$ הפתרון מההרצאה!

$$w_1 = (X^TX)^{-1}X^Ty = X^{-1}\underbrace{(X^T)^{-1}X^T}_{I}y = X^{-1}y$$

נתון כי X^TX הפיכה לכן X הפיכה ומכאן ו $A^{-1} = A^T$, $\Sigma^{-1} = \Sigma^T$ ו $X^+ = U\Sigma^+U^T$.

$$(X^TX)^{-1} = \left((U\Sigma U^T)^T \cdot (U\Sigma U^T)\right)^{-1} =$$

[illegible] -

U, U^T - מטריצות [illegible].

$$\left(U\Sigma^TU^T \cdot U \cdot \Sigma \cdot U^T\right)^{-1} =$$

$$(U \cdot \Sigma^T \cdot \Sigma \cdot U^T)^{-1} = U\Sigma^{-1} \cdot (\Sigma^T)^{-1} \cdot U^T$$

$$(X^TX)^{-1}X^Ty = \left(U\Sigma^{-1} \cdot (\Sigma^T)^{-1} \cdot U^T\right)U\Sigma^TU^Ty = U\Sigma^{-1}U^Ty = U\Sigma^+U^Ty = X^+y$$

פירוקים למטריצות

מכפלה של משנית אלכ

$\underbrace{(\Sigma^T)^{-1}, \Sigma^T}$, $\underbrace{U^T, U^T}$

7. הוכחה: X^TX הפיכה $\iff$ $Span(\underbrace{x_1 \dots x_d}_{\text{עמודות}}) = \mathbb{R}^d$

X^TX הפיכה אמ"מ X הפיכה

$X_{[d\times d]}$ הפיכה רק אם מימד השורות שווה ל d

לכן כל הוקטורים ממימד הלינארי שווה למימד העמודות

ופורסו אותו מרחב כמו מרחב העמודות ולכן $Span(x_1 \dots x_m) = \mathbb{R}^d$.

8.

נזכיר ש $\|A^T x\| = \|x\|$ (תרגול ראשון)

ונכתוב את פירוק ה SVD של x בצורה מלאה

$$U = [U_1, U_2] \quad V^T = [V_1\ V_2] \quad \Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}$$

Σ_1 - מטריצה ריבועית אלכסונית, U_1U_2, V_1V_2 - ועמודות אורתונורמליות, המימד יהיה r הדרגה של x.

נסמן למען הנוחות $U^T w = z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$

$(z_1 = U_1^T w\ ,\ z_2 = U_2^T w)$

נחשב את המינימום לפי z

$$\|y - x^T w\|_2^2 = \|y - (U\Sigma V^T)^T w\|_2^2 = \|y - V\Sigma^T U^T w\|_2^2 = \|V(V^T y - \Sigma U^T w)\|_2^2 =$$

פירוק SVD טרנספוזיציה $\Sigma^T = \Sigma$, $VV^T = I$

$$\|V_1 V_2 y - \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}\|_2^2 =$$

$$\underbrace{\|V_1^T y - \Sigma_1 z_1\|_2^2}_{\text{נשאר למצוא המינימום}} + \underbrace{\|V_2 y\|_2^2}_{\text{לא תלוי ב } w}$$

$$\min\{\|V_1^T y - \Sigma z_1\|_2^2\} = \|V_1^T y - \Sigma z_1\|_2^2 = 0$$

$$\Downarrow$$

$$V_1^T y = \Sigma_1 z_1$$

$$\Downarrow$$

$$z_1 = \Sigma_1^{-1} V_1^T y$$

נראה כי $\hat{w} = x^{\dagger} y$ מקיים את התנאי

כפל שמחילים את

$$U_1^T \hat{w} = U_1^T x^{\dagger} y = U_1^T U_1 \Sigma^{\dagger} V_1^T y = \Sigma^{-1} V_1^T y$$

$$U_2^T \hat{w} = U_2^T x^{\dagger} y = U_2^T U_2 \Sigma_1^{\dagger} V_1^T y = 0$$

כפל דומיננטים אחרים

קיבלנו ש $\hat{w}$ מקיים את התנאי למינימום וגם כל פתרון אחר ה"מקיים" $\|w\|_2^2 \geq \|\hat{w}\|_2^2$

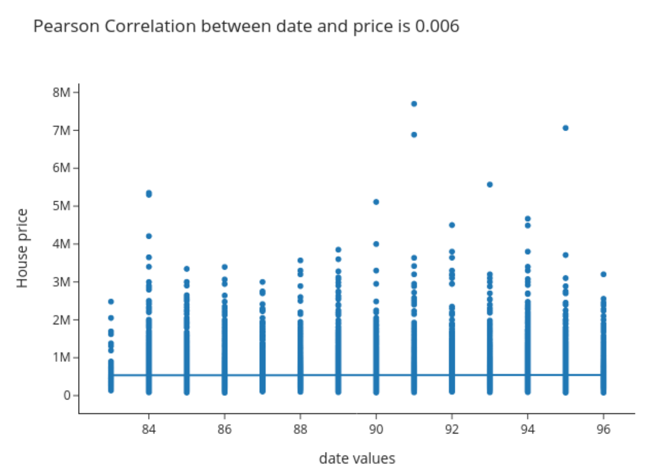
חלק מעשי

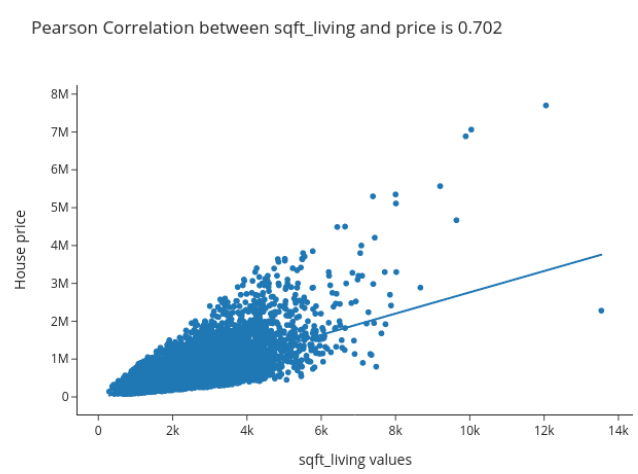
1. כדי לספק רק נתונים פרטיים שמשמעותם למשתנה מסבירים, הורדתי נתונים
שהם מזהי קלה מהם מספר חדרי השינה
גודל החצר
גודל החניה
גדול מידי

הורדתי דגימות עם ערכים לא תקניים ואת המיקום הגאוגרפי שאינו לינארי (ראו מקרה של פארק צמוד מול שכונה לשכונות מצוקה)

- העמודה של המיקוד צומצמה למשתנה בינארי {0,1} לפי zipcode.
- התאריך שונה לערך מספרי כאשר תחילת המילה היא חודשים כדי לקחת קצת מידה שמתייחס את המציאות. (שינוי ערך נקלט נמדד בחודשים במקום מתכונים).
- תאריך הבנייה חושב מחדש לפי תאריך בנייה מקורי ותאריך שיפוץ, כאשר
- דירה שעברה שיפוץ קיבלה ערך בנייה של תאריך הבנייה ועוד $\frac{3}{4}$ מהזמן שהדירה חיכתה לשיפוץ.
- העמודה הראשונה של כל ערכה שינוי מחדש לעמודת intercept עם ערך 1.
- עמודת המחירים שונתה לעמודת ה y.

2.





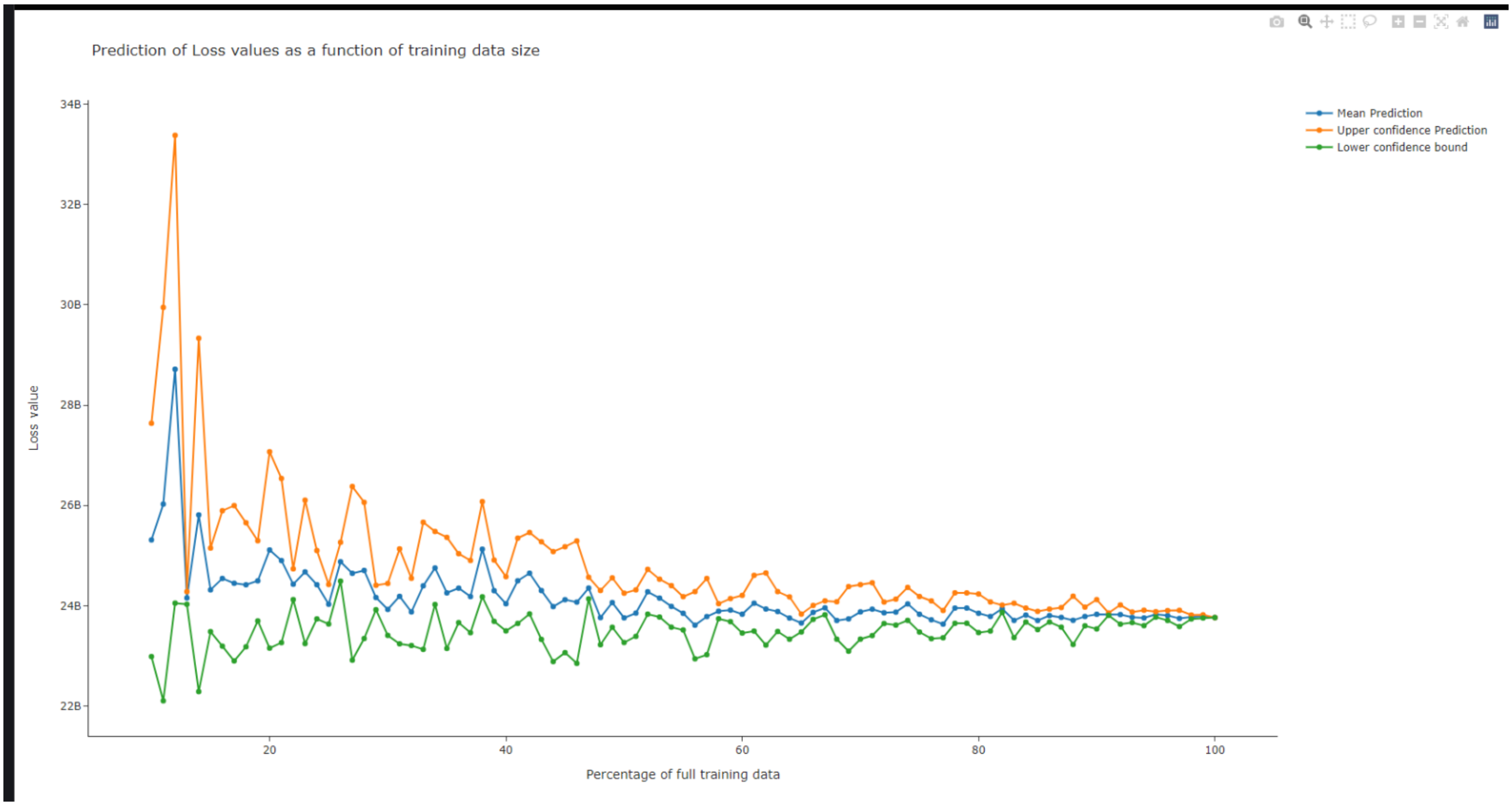
קורלציה מינימלית　　　　　קורלציה מקסימלית

קל להבחין שככל שהדירה גדולה יותר תוחלת המחיר שלה מסוים שעבר שנים מנייה, התוחלת נשארת דומה לאורך השנים. מכאן אפשר להסיק איזה פיצ'ר משמעותי יותר בהיצע המחיר לפי נתונים.

היישר מחבר גם התאמה הלינארית בין הפיצ'ר למחיר הקפוני כאשר שיפוע הקו יותר מאוזן, המודל אפשרי יותר לשינוי בפיצ'ר.

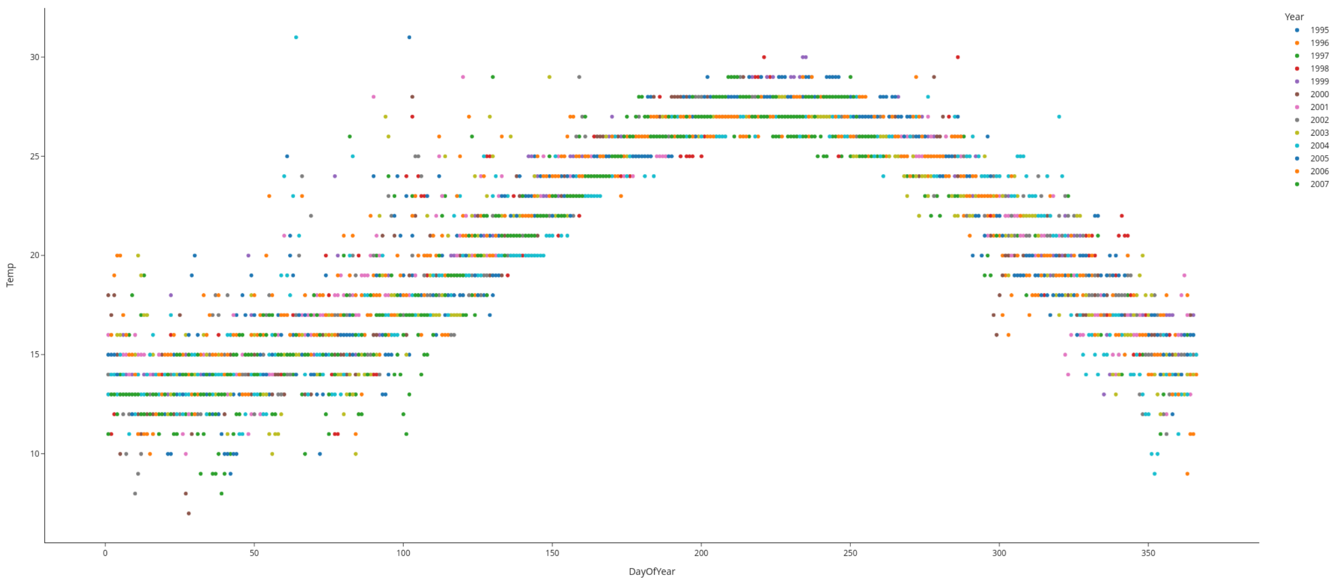
```
Worst beneficial value is date with value 0.005611648368041766
Best beneficial feature is sqft_living with value 0.7020680698104694
```

4. ניתן לראות שהמגמה הכללית היא שככל שה DS גדול יותר אז
תוחלת הטעות קטנה יותר (מתכנסת ל 24B) והשונות קטנה גם כן.

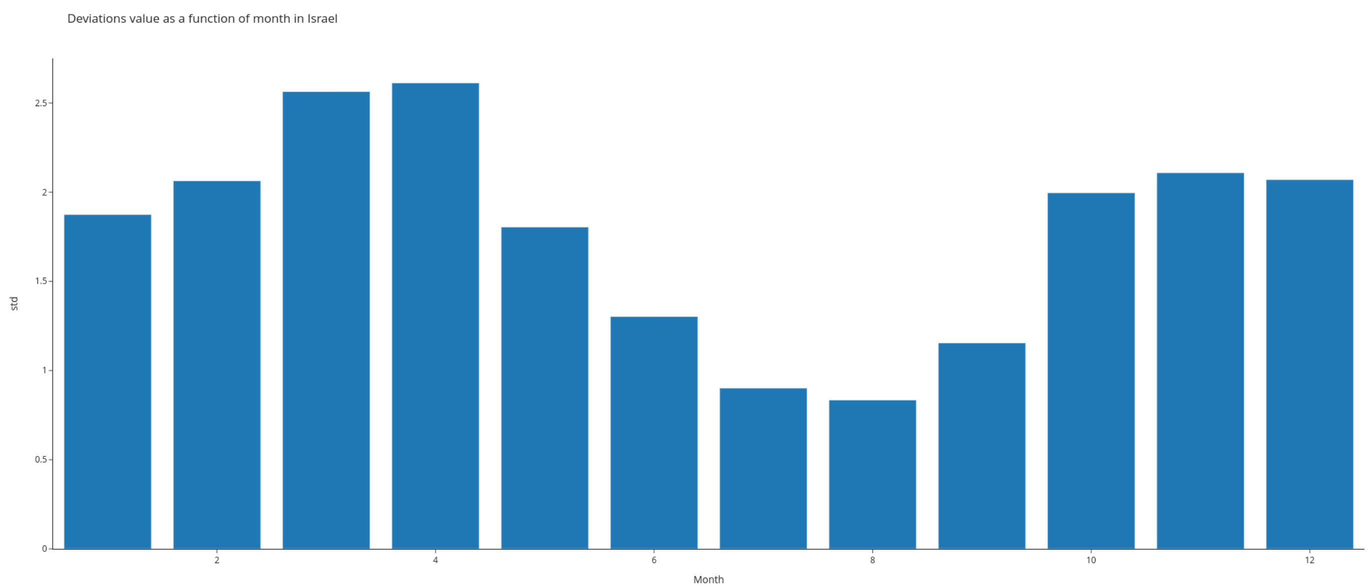


Polynomial Fitting

2. ניתן להסיק שחודשים שבהם השונות קטנה יותר המודל יחזה תוצאות בצורה מדויקת יותר. השונות הקטנה נובעת מכך שבקיץ הטמפרטורות בכל הימים היו מרוכזות יותר בהשוואה לחודשים אחרים. לכן התאמתנו למודל שלנו הרבה מה תחזית תהיה קטנה יותר.



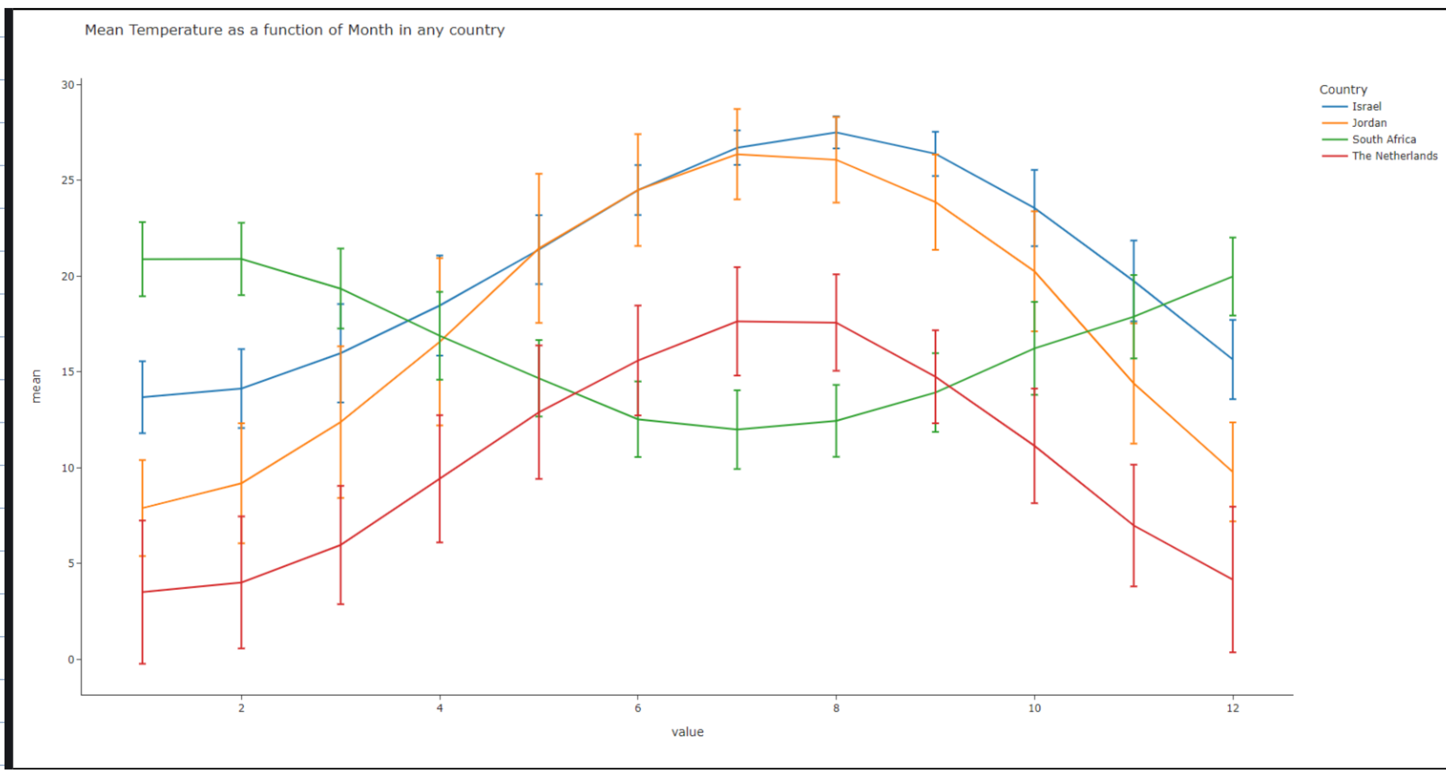
נבחין משינוי קמירות 3 פעמים לכן נשער שפולינום מדרגה 4 לפחות יצליח לבטא את התנודות הנקודות.



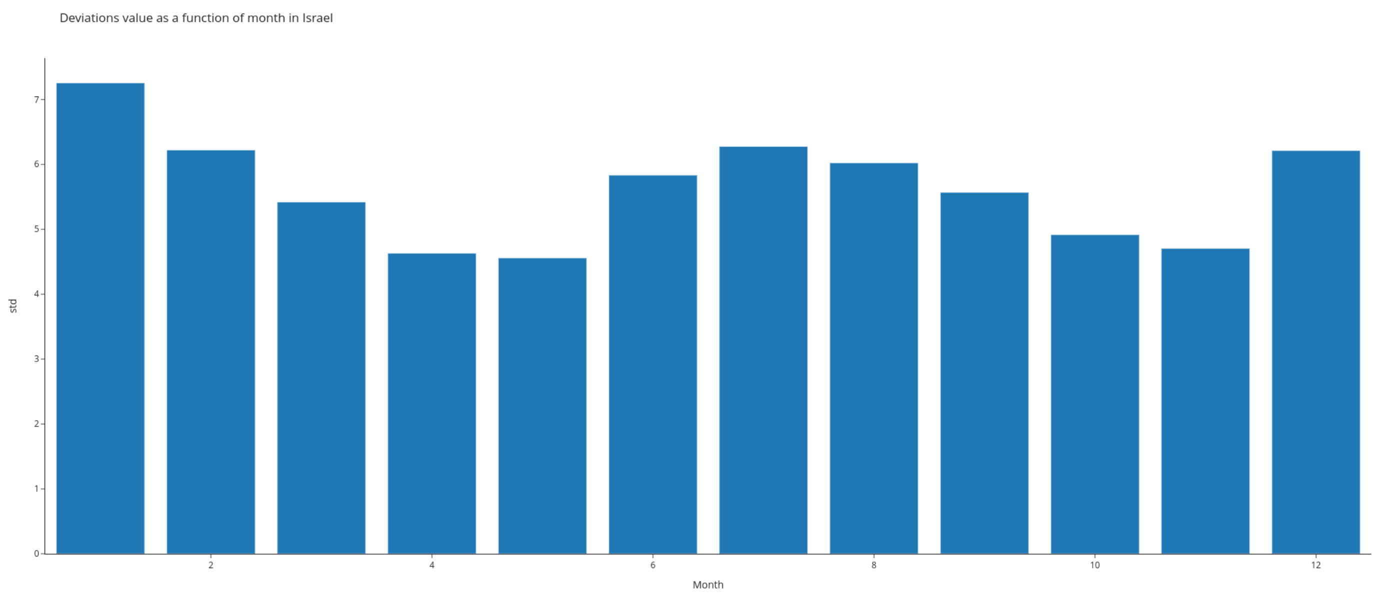
נבחין כי מחודשים 7,8,9 השונות היא קטנה באופן יחסי ולכן החיזוי בחודשים אלו יהיה טוב יותר.

3. ניתן להסיק מההתפלגות המתוארות בגרף שירדן הינה המדינה עם הטמפ' הדומות ביותר לישראל בתוך השנה.

הגרף של דרום אפריקה מראה התנהגות חורף-קיץ בזמנים שונים של השנה, לכן חיזוי לפי ישראל שבה עונות השנה מתרחשות בזמן שונה מדר"אפ יוביל לתוצאה מרוחקת. עם זאת, עונות המעבר אפריל ונובמבר עם ערכים דומים בדר"אפ וישראל לכן מהגרפים אלו הדיוק יהיה גבוה.
הולנד מתנהגת כמו ישראל מבחינת חורף-קיץ אך הטמפ' הממוצעת שלה תמיד נמוכה בכמה מעלות מישראל בעונות החודשים.



4. ניתן לראות שעבור $k=5$, הפולינום עם הערכה המקסימלית, המודל
הצליח לחזות בצורה הטובה יותר.



5. ניתן לראות שהמודל עם הכי מעט טעות ישראל - שכן המודל התאמן לפי
הנתונים מישראל. אחריו ירדן שהייתה דומה בתוצאות לישראל,
אח"כ דרא"פ נמצאת הקטין בשמ"פ מעונות המעבר ולמרות ההפרשים
הגדולים בשנה הקיץ חורף. המקום האחרון נשמר להולנד שהתחזית
של השגיאה היא $10°C$ עבור כל חודש בערך ויהיה המודל הלקוחה
גדולה יותר.
$MSE = 12\cdot\frac{10^2}{12}=100$