מבוא למערכות לומדות - תרגיל 2

גיא קורנבליט, ת.ז 308224948

חלק תיאורטי

יהיו $X \in M_{p\times n}(\mathbb{R})$ מטריצת המדגם, $y \in \mathbb{R}^n$ וקטור התיוג.

פתרונות המשוואות הנורמליות

(1) **טענה:** $Ker\left(X^T\right) = Ker\left(XX^T\right)$.

הוכחה: יהי $\vec{v} \in Ker\left(X^T\right)$, כך ש- $X^T\vec{v} = \vec{0}$. מתקיים $XX^T\vec{v} = X\left(X^T\vec{v}\right) = X\vec{0} = \vec{0}$ ולכן $\vec{v} \in Ker\left(XX^T\right)$. בכיוון השני, יהי $\vec{w} \in Ker\left(XX^T\right)$ כך ש- $XX^T\vec{w} = 0$. מתקיים -

$$v^TXX^Tv = \left(X^Tv\right)^T\left(X^Tv\right) = \left\|X^Tv\right\|^2 = 0$$

מחיוביות בהחלט של הנורמה, נסיק כי $X^Tv = \vec{0}$ ולכן $v \in Ker\left(X^T\right)$. הראינו הכלה דו-כיוונית ומכאן השוויון. ■

(2) תהי $A \in \mathbb{R}^{p\times p}$ מטריצה ריבועית, אזי

$$Im\left(A^\top\right) = Ker(A)^\perp = \{x \in \mathbb{R}^p : \langle x, v\rangle = 0, \forall v \in Ker\left(A\right)\}$$

הוכחה: נסמן ב-V את המ"ו עליו פועל האופרטור המושרה על ידי A.

$\boldsymbol{Im\left(A^\top\right) \subseteq Ker(A)^\perp}$: יהיו $\vec{v} \in Im\left(A^\top\right)$ ו-$\vec{u} \in Ker(A)$, כלומר $A^\top\vec{w} = \vec{v}$ עבור $w \in V$, וגם $Au = \vec{0}$. מתקיים

$$\langle v, u\rangle = v^\top u = \left(A^\top w\right)^\top u = w^\top\left(Au\right) = w^\top\vec{0} = 0$$

ולכן $v \in Ker(A)^\perp$.

$\boldsymbol{Im\left(A^{\top}\right)\supseteq Ker(A)^{\perp}}$: יהי $\vec{v}\in\left(Im\left(A^{\top}\right)\right)^{\perp}$, לכן מתקיים $\vec{v}^{T}A^{T}\vec{u}=0$ לכל $\vec{u}\in V$, ובפרט עבור $\vec{u}=A\vec{v}$ נקבל

$$\vec{v}^{T}A^{T}A\vec{v}=0\iff\left(A\vec{v}\right)^{T}A\vec{v}=0$$

מחיוביות בהחלט של המכפלה הפנימית נקבל כי $A\vec{v}=\vec{0}$ ובפרט $\vec{v}\in Ker\left(A\right)$. כלומר, הראינו כי $\left(Im\left(A^{\top}\right)\right)^{\perp}\subseteq Ker\left(A\right)$. מתכונות המרחב הניצב, מתקיים

$$\left(Ker\left(A\right)\right)^{\perp}\subseteq\left(Im\left(A^{\top}\right)\right)^{\perp\perp}$$

ובנוסף לכל מרחב סופי מתקיים השוויון $\left(Im\left(A^{\top}\right)\right)^{\perp\perp}=Im\left(A^{\top}\right)$ ומכאן הטענה. ■

(3) תהי $y=X^{\top}w$ מערכת משוואות ליניארית לא הומוגנית (כלומר $y\neq\vec{0}$), כאשר $X^{\top}$ ריבועית וסינגולרית. נראה כי למערכת יש אינסוף פתרונות $\iff$ $y\perp Ker\left(X\right)$.

הוכחה: ראשית, למערכת משוואות כזו יש אינסוף פתרונות כאשר $dim\left(Ker\left(X^{T}\right)\right)\neq0$ וגם $y\in Im\left(X^{T}\right)$. נשים לב כי התנאי הראשון מתקיים מההנחה כי X^{T} סינגולרית, לכן מספיק שיתקיים $y\in Im\left(X^{T}\right)$ כדי שלמערכת יהיו ∞ פתרונות. כעת, מהסעיף הקודם מתקיים כי

$$y\perp Ker\left(X\right)\iff y\in Ker\left(X\right)^{\perp}\iff y\in Im\left(X^{T}\right)$$

ומכאן הטענה. ■

(4) נתבונן במערכת המשוואות הנורמלית $XX^{T}w=Xy$. נוכיח כי למערכת קיים פתרון יחיד, או אינסוף פתרונות.

הוכחה:. נחלק למקרים - כאשר XX^{T} הפיכה אז מתקיים $w=\left(XX^{T}\right)^{-1}Xy$, ולכן w הנו פתרון יחיד. פתרון כלומר w פתרון יחיד לבעיה. אחרת, XX^{T} מטריצה ריבועית וסינגולרית. מהטענה הקודמת, למערכת יש אינסוף פתרונות אם ורק אם

$$Xy\perp Ker\left(\left(XX^{T}\right)^{T}\right)=Ker\left(XX^{T}\right)$$

ומסעיף 1 נסיק כי מספיק להראות שמתקיים $Xy\perp Ker\left(X^{T}\right)$, ואכן לכל $u\in Ker\left(X^{T}\right)$

$$\left(Xy\right)^{T}u=y^{T}X^{T}u=y^{T}\cdot\vec{0}=\vec{0}$$

ולכן יש אינסוף פתרונות במקרה בו XX^T סינגולרית. כנדרש. ■

מטריצת ההטלה

(5) יהי $V \subseteq \mathbb{R}^d$ מ"ו, כך ש- $dim\left(V\right) = k$, ויהא $B = \left(v_1, v_2, ..., v_k\right)$ בסיס או"נ של V.
מטריצת ההטלה האורתוגונלית מוגדרת ע"י $P = \sum_{i=1}^{k} v_i v_i^T$.

(א) P סימטרית. **הוכחה:**

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

(ב) הערכים העצמיים של P הם 0 או 1, ו- $v_1, ..., v_k \in V_1$ (מתאימים לע"ע 1). **הוכחה:**
ראשית, נשים לב כי עבור וקטור $v_r \in B$, מתקיים

$$Pv_r = \sum_{i=1}^{k} v_i v_i^T v_r = \underbrace{\sum_{i=1}^{r-1} v_i v_i^T v_r}_{\delta_{ij}} + v_r v_r^T v_r + \underbrace{\sum_{i=r+1}^{k} v_i v_i^T v_r}_{\delta_{ij}}$$

$$= 0 + v_r \left\|v_r\right\| + 0 = v_r \cdot 1$$

הבסיס או"נ ולכן וקטורי הבסיס אנכים זה לזה, ובעלי נורמה 1. מכאן, לכל $v_i \in B$ מתקיים $Pv_i = v_i$, כלומר וקטורי הבסיס הם וקטורים עצמיים של P המתאימים לערך עצמי 1.

כעת, נסמן $U = \mathbb{R}^d$, אז מתקיים $U = V \bigoplus V^{\perp}$. יהי $w \in U$. **נחלק למקרים**. אם $w \in V^{\perp}$ אזי לכל $v \in V$ מתקיים $v \perp w$ ולכן $Pw = \sum_{i=1}^{k} v_i \left(v_i^T w\right) = 0$, כלומר 0 הינו ע"ע של P. במקרה ו-$w \in V$, נוכל לייצג את w כצירוף ליניארי של איברי הבסיס, כלומר $w = \sum_{i=1}^{k} a_i v_i$, ולכן

$$Pw = \sum_{i=1}^{k} a_i P v_i = \sum_{i=1}^{k} a_i v_i = w$$

כלומר, הערכים העצמיים של P הם 0 ו-1, כאשר הוקטורים העצמיים המתאימים לע"ע 1 הם וקטורי הבסיס האו"נ.

(ג) $Pv=v\ \forall v\in V$. **הוכחה:** יהי $v\in V$, נוכל להציג את v כצ"ל של איברי הבסיס B. כלומר $v=\sum_{i=1}^{k}a_iv_i$ עבור $\{a_i\}_{i=1}^{k}$ סדרת ערכים כלשהי לא כולם אפס. אזי $Pv=\sum_{i=1}^{k}a_iPv_i$, אבל $Pv_i=v_i$ לכל $i\in[k]$ מהסעיף הקודם, ולכן $Pv=v$ כנדרש.

(ד) $P^2=P$. הוכחה: ראינו כי P סימטרית מעל $\mathbb{R}$ ולכן בפרט אורתוגונלית, כלומר קיים פירוק EVD כך שמתקיים $P=U\Sigma U^T$ עבור U מטריצה או"ג ו-Σ מטריצה אלכסונית, כאשר על האלכסון מופיעים הערכים העצמיים של P. מתקיים

$$P^2=U\Sigma U^TU\Sigma U^T=U\Sigma^2U^T=P$$

השוויון האחרון נובע מהטענה לפיה הערכים העצמיים של P הם 0 או 1, אז בהכרח $\Sigma^2=\Sigma$.

(ה) $(I-P)P=0$. **הוכחה**: נובע מיידית מהטענה האחרונה - $P^2=P \iff P-P^2=0 \iff P(I-P)=0$. ■

הפרש הריבועים

בשאלה זו נניח כי $X\in\mathbb{R}^{d\times m}$.

(6) נניח כי XX^T הפיכה.

- **טענה:** $\left(XX^T\right)^{-1}=UD^{-1}U^T$ כאשר $D=\Sigma\Sigma^T$ (לפי פירוק SVD של X).

הוכחה: מתקיים $\left(XX^T\right)^{-1}=\left(U\Sigma V^TV\Sigma^TU^T\right)^{-1}=\left(UDU^T\right)^{-1}$ אבל מפני ש-U מטריצה או"ג, היא צמודה לעצמה, כלומר $U^T=U^{-1}$ ובפרט הפיכה. כמו כן, D ריבועית ואלכסונית ולכן הפיכה. מכאן, מתקיים

$$\left(UDU^T\right)^{-1}=\left(U^T\right)^{-1}D^{-1}U^{-1}=UD^{-1}U^T$$

כנדרש.

- **מסקנה:** מתקיים $\left(XX^T\right)^{-1}X = X^{T\dagger}$. **הוכחה:**

$$\begin{aligned}\left(XX^T\right)^{-1}X &= UD^{-1}U^TX\\ &= \left(UD^{-1}U^T\right)\left(U\Sigma V^T\right)\\ &= UD^{-1}\Sigma V^T\end{aligned}$$

מאחר ו-D^{-1} הינה מטריצה ריבועית ואלכסונית, לכל $i \in [d]$ מתקיים $D_{ii} = \sigma_i^{-2}$, ולכן $\left(D^{-1}\Sigma\right)_{ii} = \sigma_i^{-1}$. מכאן, נובע כי

$$\left(XX^T\right)^{-1}X = UD^{-1}\Sigma V^T = U\Sigma^\dagger V^T = X^{T\dagger}$$

■

(7) **טענה:** XX^T הפיכה $\iff$ $Span\left\{x_1,..,x_m\right\} = \mathbb{R}^d$ (עבור $\boldsymbol{x_i}$ וקטורי הדגימות ב-X).

הוכחה:

$$\begin{aligned} XX^T \in \mathbb{R}^{d\times d} \text{ is invertible} &\iff rank\left(XX^T\right) = d \iff \left(\Sigma\Sigma^T\right)_{dd} = \sigma_d^2 > 0\\ (1) &\iff \Sigma_{dd} = \sigma_d > 0 \iff X \text{ is invertible} \iff rank\left(X\right) = d\\ (2) &\iff dim\left(Col\left(X\right)\right) = dim\left(Span\left\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m\right\}\right) = d\\ &\iff Span\left\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m\right\} = \mathbb{R}^d \end{aligned}$$

מעבר (1) נובע מפירוק ה-SVD של X, והקשר בין פירוק זה לפירוק EVD של XX^T.

מעבר (2) מתקיים כי דרגת השורות שווה לדרגת העמודות במטריצה. ■

(8) **טענה:** בהנחה שקיימים אינסוף פתרונות למערכת המשוואות הנורמלית, כלומר XX^T לא הפיכה, אזי $\hat{w} = X^{T\dagger}y$ הוא הפתרון בעל הנורמה המינימלית.

הוכחה: יהי $\bar{w} \in \mathbb{R}^d$ פתרון כלשהו למערכת המשוואות $X^T\bar{w} = y$. יהא $X = U\Sigma V^T$ פירוק SVD של X. $V \in \mathbb{R}^{m\times m}$ מטריצה או"ג, ועמודותיה $\left(v_1,...,v_m\right)$ מהוות בסיס אורתונורמלי ל-$\mathbb{R}^m$, לכן נוכל לכתוב $y = \sum_{i=1}^m a_iv_i$ עבור $\vec{0} \neq a \in \mathbb{R}^m$ וקטור מקדמים

כלשהו. באותו אופן, $U \in \mathbb{R}^{d\times d}$ מטריצה או"ג, ועמודותיה $(u_1, .., u_d)$ מהוות בסיס אורתונורמלי ל- $\mathbb{R}^d$. לכן, נוכל לכתוב $\bar{w} = \sum_{i=1}^d b_i u_i$ עבור $\vec{0} \neq b \in \mathbb{R}^d$ וקטור מקדמים.

מההנחה כי XX^T אינה הפיכה, נובע כי X אינה הפיכה, ומפני שתכונה זו נקבעת ע"י כמות הערכים הסינגולריים של X, נסיק כי קיים $1 \leq r < d$ עבורו $\sigma_r > 0$ ו- $\sigma_j = 0$ לכל $r+1 \leq j \leq d$, כאשר $\sigma_i = \Sigma_{ii}$ מפירוק SVD המתואר לעיל. מכאן, מתקיים

$$\begin{aligned} U^T\hat{w} &= U^T X^{T\dagger} y = U^T U \Sigma^\dagger V^T y \\ &= \Sigma^\dagger V^T y = \Sigma^\dagger \left(\sum_{i=1}^m a_i V^T v_i \right) \\ &= \Sigma^\dagger \left(\sum_{i=1}^m a_i e_i \right) = \sum_{i=1}^r \frac{a_i}{\sigma_i} \cdot e_i \end{aligned}$$

מפני שלכל פתרון מתקיים $X^T\bar{w} = y$, אז מוכרח להתקיים $\bar{w}_i = \hat{w}_i$ לכל $1 \leq i \leq r$. באשר ל- $d-r$ הקוארדינטות בוקטור הפתרון, הראינו כי לכל $r+1 \leq j \leq d$ מתקיים מהגדרה $\hat{w}_j = 0$. וכל פתרון $\bar{w}$ מוגדר באופן אחר על ידי $d-r$ הקוארדינטות הללו. כלומר -

$$\|\hat{w}\|_2 = \sqrt{\sum_{i=1}^r \hat{w}_i^2} \leq \sqrt{\sum_{i=1}^r \hat{w}_i^2 + \sum_{i=r+1}^d \bar{w}_i^2} = \|\bar{w}\|_2$$

חלק מעשי

שאלה 12 - Preprocessing

בתחילת התהליך, ביצעתי אקספלורציה על כל אחד מהפיצ'רים במטרה לאפיין את ההתנהגות של כל פיצ'ר ביחס למדגם. למשל, מה השונות בין הבתים שנדגמו מבחינת מספר החדרים, או מדוע החדרים וחדרי האמבטיה מסומנים במספר דצימלי ולא טבעי (תקן אמריקאי לסוג המתקנים בכל חלל). בדיקות אלו אפשרו לזהות מידע חריג או זבל.

באופן שיטתי, מפני שמרחב הדגימות שולט ממש על מרחב הפיצ'רים, העדפתי למחוק דגימות בעלות מידע חסר או פגום, באופן שעשוי להעיד על בעיה בדגימה כולה (למשל, מזהה בית לא תקין). בפרט, תהליך הניקוי כלל סינון רשומות המכילות מידע לא הגיוני ביחס לייצוג של הפיצ'ר (מחיר שלילי, שטח אפס וכו'). השלב האחרון, כלל עיבוד של פיצ'רים לפיצ'רים חדשים, שהנחתי

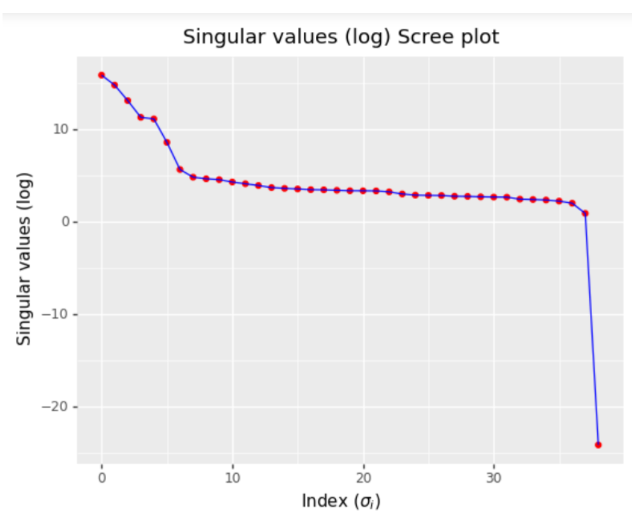
על בסיס היוריסטיקות שיסבירו באופן טוב יותר את מחיר הבתים. בתוך כך, פיצ'ר שמתאר את גיל הבית (מבניה ועד מכירה), והאם הבית שופץ או לא.

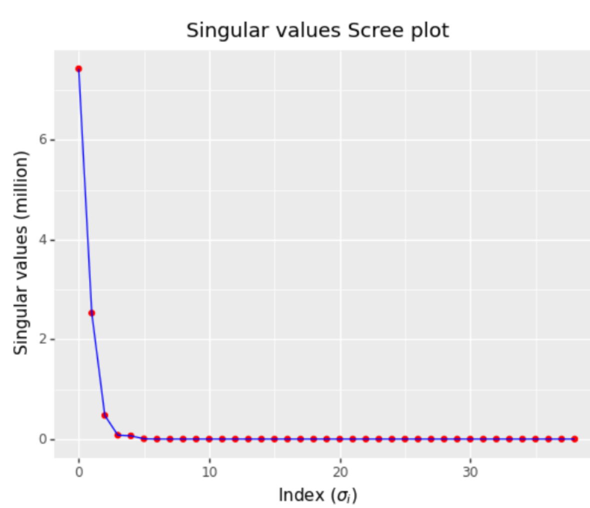
שאלה 13 - פיצ'ר קטגורי

מבחינת פיצ'רים קטגוריאליים, ראיתי לנכון להתייחס רק למיקוד, מפני ששאר הפיצ'רים היו רציפים או אורדינליים (ציון, מצב הנכס וכו') באופן שתאם את ההגיון למחיר הבית (ציון גבוה יותר יביא למחיר גבוה יותר וכדומה). בנוגע לפיצ'ר המיקוד, השתמשתי במידע כדי לשייך כל בית לעיר אליה משויך (כולם במדינת וושינגטון), כדי להשתמש במידע הגיאוגרפי באופן פשוט (מאשר קואורדינטות), וכדי להפחית את כמות הקטגוריות מ-71 ערכי מיקוד שונים, ל-25 ערים (כנראה בשל מיקוד שונה באותה העיר). בסוף התהליך, קודדתי את הערים בשיטת one-hot-encoding, מפני שהפיצ'ר קטגורי ללא יחס סדר נראה לעין. **מנגד, החיסרון הבולט בשיטה זו, הוא הטיית המודל לחיזוי בתים בערים אלו, לעומת חיזוי של בתים באיזורים אחרים.** כמו כן, לאחר בדיקה קצרה, החלטתי שלא להשתמש במידע כדי ליצור פיצ'ר של המחיר הממוצע לאיזור, כי התוצאה לא השתפרה משמעותית ביחס לקידוד הערים, ומפני שפיצ'ר כזה היה מעצים את ההשפעה של ההטייה ב-train על המודל.

שאלה 15

נתבונן בהתפלגות הערכים הסינגולריים של המטריצה $X \in \mathbb{R}^{40 \times m}$

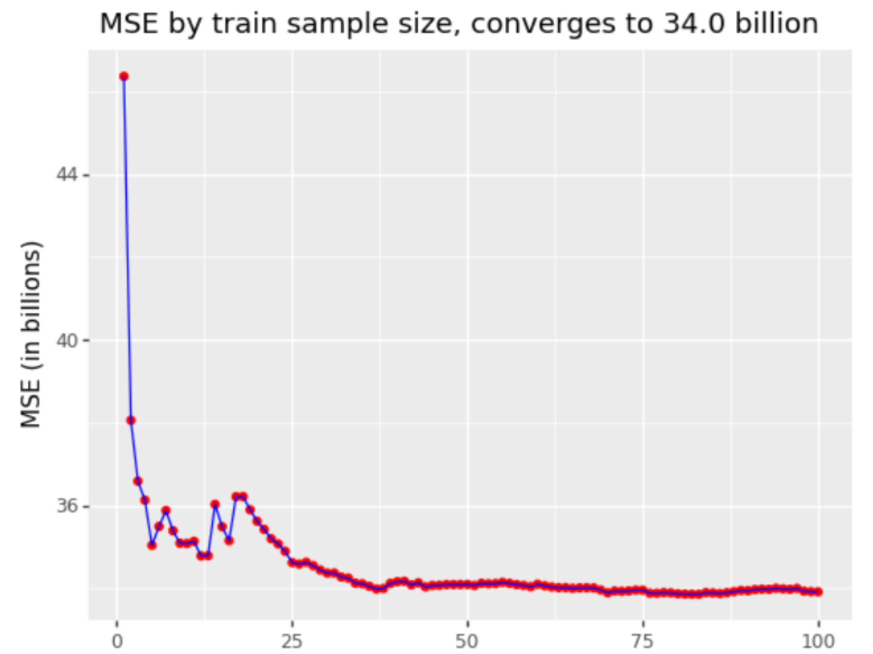




נבחין כי הערכים הסינגולריים מתייצבים בטווח ערכים יחסית נמוך (כפי שניתן לראות בגרף הלוגריתם). בפרט, σ_{d+1} קרוב מאוד לאפס, ולכן X כמעט סינגולרית. מהבנייה של פירוק SVD וקצב ההתכנסות המתואר בגרף, נסיק כי ניתן לתאר את הדגימות באמצעות מעט וקטורים עצמיים ב-U, המייצגים את ״הבתים הטיפוסיים״ (eigen-houses) שנוכל לשחזר בדיוק גבוה מתוכם את הבתים במדגם. בנוסף, מפני שהערכים הסינגולריים זהים גם עבור X^TX, נוכל להסיק באותו אופן, כי קיימת תלות לינארית (או כמעט תלות ליניארית) בין הפיצ׳רים, כך שניתן להתאים מודל טוב באותה מידה אם נאמן אותו על וקטורים עצמיים (eigen-features) אלו (כלומר וקטורים עצמיים של המטריצה V), ובאופן זה נקטין את סיבוכיות המודל ונפחית את מידת הרעש. במובן נוסף, אם קיימת תלות ליניארית בין הפיצ׳רים אז נוכל לקבל את אותה תוצאה בקירוב אם נוריד מ- X^T את הפיצ׳רים אשר כמעט תלויים ליניארית באחרים, והמרחב הנפרש על ידם יוותר ללא שינוי.

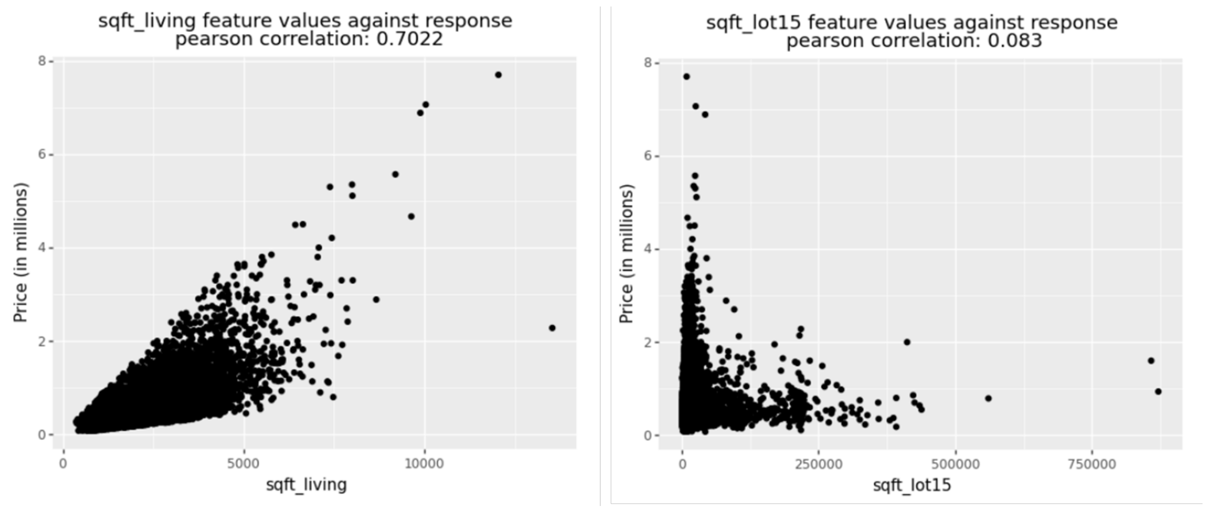
שאלה 16

נתבונן בהתפלגות הטעות הריבועית הממוצעת של מודל שאומן על חלק יחסי הולך וגדל מתוך train-set שהוגדר להיות $\frac{3}{4}$ מהמדגם. כל מודל אומן על אחוז הולך וגדל מה-train-set ונבחן ביחס ל-$\frac{1}{4}$ מהמידע המקורי. נבחין כי הגרף מונוטוני יורד, כלומר ככל שמאמנים את המודל על יותר מהמידע, החיזוי שמפיק המודל קרוב יותר למציאות. עם זאת, נראה שהשיפור באיכות החיזוי מוגבלת. כלומר מנקודה מסוימת התרומה של המידע לשיפור תוצאות המודל לא הייתה משמעותית. ככל שגדלה כמות הדאטה מתרחשים מספר תהליכים במקביל. ראשית, הגדלת מספר השורות במטריצה X, מגבירה את הסיכוי שהפיצ׳רים יהיו בלתי תלויים ליניארית, ולכן $rank(X) = min(m, d) = d$ עולה עד שהיא מגיעה לדרגה מלאה. במילים אחרות, מימד התמונה של X^T הולך וגדל ב-$\mathbb{R}^m$, סיבוכיות המודל עולה וההטיה של המודל יורדת. במובן נוסף, הגדלת כמות הדגימות חושפת את המודל פחות להטיות של outliers, מפני שהחשיבות שלהם הולכת ופוחתת מפני שהמשקל היחסי שלהם ביחס לשאר הדגימות יפחת. כמו כן, בהיבט הרעש של המדגם, מחוק המספרים הגדולים נובע כי ככל שהדגימה גדלה הרעש מתכנס לקבוע (לתוחלת) ולכן השפעתו גם הולכת וקטנה. כמו כן, מפני שהתרומה של הדגימות הולכת ופוחתת, ניתן להסיק כי השונות של הדגימות ביחס לכמות שלהן הייתה מוגבלת, ולכן הצלחנו יחסית בשלב מוקדם למצוא פונקציה שתתאר את משפחת הדגימות המתוארת בצורה טובה. באופן שתואם את התפלגות הערכים הסינגולריים של המטריצה X.



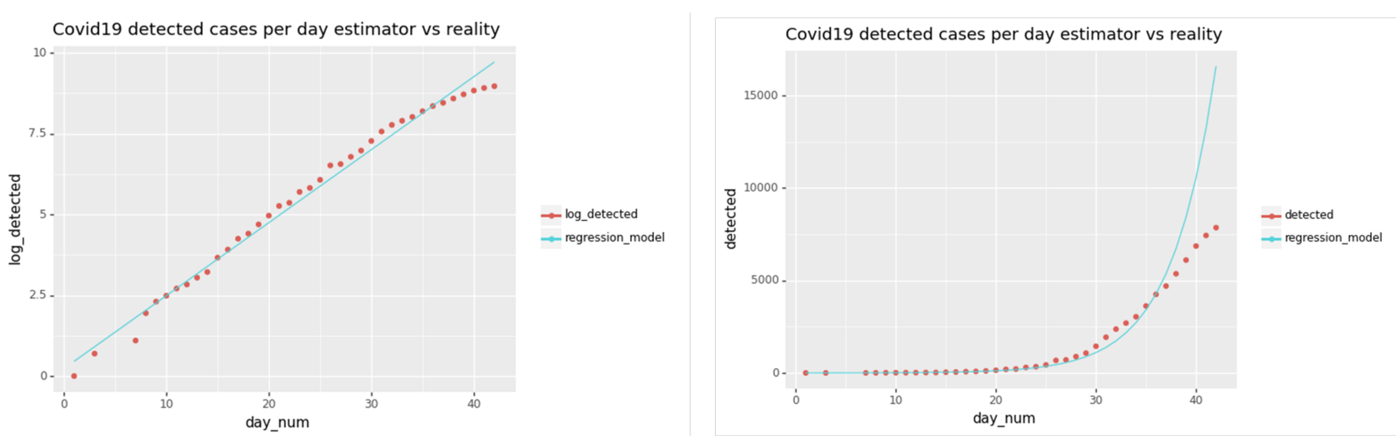
שאלה 17

נתבונן בהתפלגות הפיצ׳רים הבאים למול וקטור המחירים הנצפים, נטען כי sqft_living מועיל למודל, בעוד sqft_lot15 אינו מועיל למודל. ראשית, נבחן את מדד הקורלציה כאינדיקטור להשפעת המידע המקודד בפיצ׳ר על התפלגות וקטור המחיר. במילים אחרות, מדד קורלציה רחוק יותר מאפס - מצביע על שקיימת קורלציה, בעוד מדד קרוב לאפס מעיד על משתנים בלתי מתואמים. על בסיס הבנה זו, ניתן לראות כי ההשפעה של sqft_lot15 יחסית זניחה ביחס ל-sqft_living, ובפרט התפלגות הערכים של הפיצ׳רים ביחס לערך המחיר מעידה על קשר ליניארי בין sqft_living לוקטור התגובה, בעוד שהתפלגות sqft_lot15 לא מעידה על קשר מובהק בין הערכים.



איור 0.1: sqft_living is beneficial to the model, and sqft_lot15 isn't.

שאלה 21



איור 0.2: התפלגות מספר הנדבקים היומי בישראל, ביחס לשערוך ליניארי של מספר הנדבקים

שאלה 22

בהינתן דגימה (x, y), השתמשנו במודל האקספוננציאלי לעיל, ב-ERM בהתבסס על פונקציית ההפסד הבאה

$$L_{exp}\left(f_w, (x, y)\right) = \left(\langle w, \boldsymbol{x}\rangle - log(y)\right)^2$$

על מנת להתאים את המודל לחזות מידע בסדר גודל אקספוננציאלי, נידרש להגדיר פונקציית הפסד חדשה:

$$L\left(f_w,(x,y)\right)=\left(\exp\left(\langle w,\boldsymbol{x}\rangle\right)-y\right)^2$$

באופן זה, נשמור על היחס בין הגדלים. במקרה זה, כדי למצוא פתרון לבעיית ERM נגדיר Empirical risk בתור

$$RSS=\frac{1}{m}\sum_{i=1}^{m}\left(\exp\left(\langle w,\boldsymbol{x}_i\rangle\right)-y_i\right)^2$$

נגזור את הפונקציה לפי w ונמצא את נקודת המינימום. נשים לב כי בדומה למקרה הליניארי, קיבלנו תבנית ריבועית, כלומר פונקציה קמורה ולכן בעלת מינימום גלובאלי.