(20942) מבוא ללמידה חישובית | סיכום הרצאה 3

מנחה: שי מימון
סמסטר: 2022א'
נכתב על ידי: מתן כהן

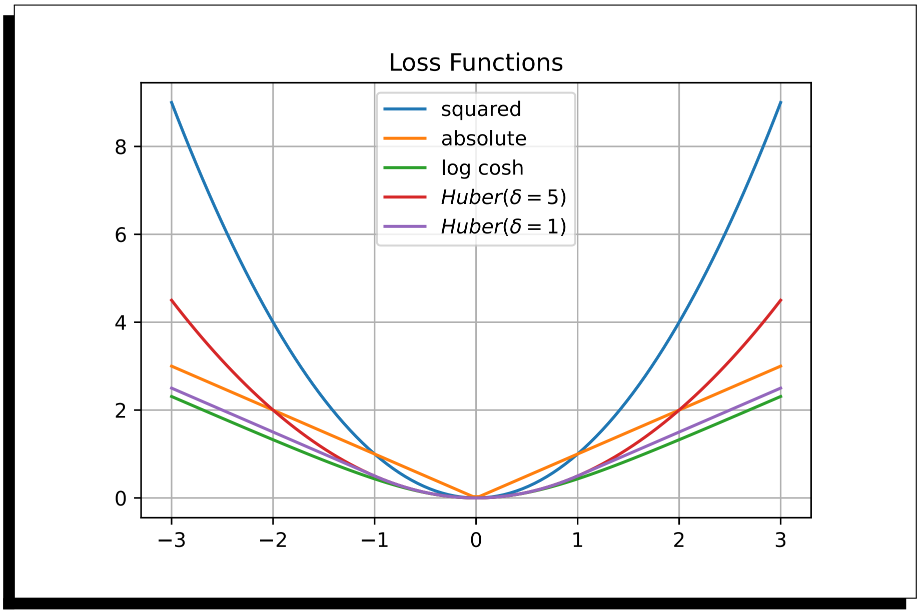
1 רגרסיה לינארית - Linear Regression

ההנחה הכללית שחוזרת - הלייבל y (פלט) מתנהג בצורה לינארית כפונקציה של המאפיינים $\mathbf{x}$
הנחה נוספת - y הוא משתנה רציף (בעית רגרסיה)

1.1 פונקצית המחיר - The Loss Function

כאשר נדבר על בעית רגרסיה, בדומה לכל בעית למידה אחרת - נרצה להגדיר את ה-Loss Function שלה (פונקציית המחיר).
בבעיית סיווג - הדבר פשוט, כיוון ששם חוזים ערך מסוים ונותנים לו תווית (כמו עם הפרספטרון שעבר לערכים ב $\{1, -1\}$).
אם סיווגנו נכון - השגיאה היא אפס ואין צורך **לשלם שום מחיר** על הסיווג, אם הסיווג שגוי - המחיר גבוה יותר.
עם זאת, בבעיית רגרסיה המתודולוגיה שונה - יש צורך לכמת את המרחק בין החיזוי של האלגוריתם לבין הערך האמיתי בעזרת פונקציית המחיר

האינטואיציה: כיצד אנחנו "מענישים" את עצמינו על חיזוי שגוי בהתאם **למרחק** בין החיזוי לבין הערך האמיתי



פונקציות מחיר מוכרות

- פונקציה ריבועית:

$$e(h(\mathbf{x}), y) = (h(\mathbf{x}) - y)^2$$

שמגדילה את השגיאה באופן ריבועי

- פונקציה של ערך מוחלט:

$$e(h(\mathbf{x}), y) = |y - h(\mathbf{x})|$$

יש צורך להבחין בעובדה שהפונקציה הריבועית "מענישה מאוד" על שגיאות גדולות לעומת הערך המוחלט, ולכן במצבים בהם יש outliers הפונקציה הריבועית יכולה לפגום בטיב האלגוריתם.

1.2 פונקצית המחיר הריבועית - Squared Loss Function

כאשר נדבר על הפונקציה הנ"ל נרצה תמיד למצוא פתרון לבעיית האופטימיזציה שמביאה למינימום את התוחלת של פונקצית השגיאה הריבועית:

$$E_{out}(h) = \mathbb{E}\left[(h(\mathbf{x}) - y)^2\right]$$

אך לנו לא נתון הפילוג של סט הדוגמאות! לכן נרצה להתמודד עם השגיאה הריבועית הממוצעת מתוך סט האימון שלנו כאשר נרצה להביא למינימום את E_{in}:

$$E_{in}(h) = \frac{1}{N}\sum_{n=1}^{N}\underbrace{(h(\mathbf{x}_n) - y_n)^2}_{squared-function}$$

הציפייה שלנו היא שהמינימום שנמצא יניב לנו חזאי מספיק טוב.

1.2.1 פיתוח הפונקציה - Least-Squares Derivation

עבור דוגמה כלשהי $\begin{pmatrix} 1 \\ x_{n_1} \\ \vdots \\ x_{n_d} \end{pmatrix} = \begin{pmatrix} x_{n_0} \\ x_{n_1} \\ \vdots \\ x_{n_d} \end{pmatrix} = \mathbf{x}_n$ ומשקלים $\begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix} = \mathbf{w}$ ועבור היפותזה $h(\mathbf{x}) = \mathbf{w}^T\mathbf{x}_n$ נוכל להתבונן ב in sample error שהוא **עיקרון הריבועים הפחותים**:

$$E_{in}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2$$

ונרצה להביא למינימום את $E_{in}(\hat{\mathbf{w}})$ על מנת למצוא את סט המשקלים $\hat{\mathbf{w}}$ האופטימלי:

$$\underset{\hat{\mathbf{w}}\in\mathbb{R}^{d+1}}{min} E_{in}(\hat{\mathbf{x}})$$

לשם כך נגדיר מטריצה חדשה ששורותיה יהיו הוקטורים שמייצגים את הפיצ'רים:

$$X_{N\times(d+1)} = \begin{pmatrix} - & \mathbf{x}_1^T & - \\ - & \mathbf{x}_2^T & - \\ \vdots & \vdots & \vdots \\ - & \mathbf{x}_N^T & - \end{pmatrix}$$

ונבחין כי העמודה הראשונה במטריצה היא עמודה $(1, 1, .., 1) \in \mathbb{R}^{d+1}$

בצורה דומה נגדיר וקטור $\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \mathbf{y}_{N\times 1}$ וכעת נרצה להתבונן בעיקרון הריבועים הפחותים על ידי כתיב מטריציאלי, לשם כך נרשום את E_{in} בצורה נוחה יותר:

$$E_{in}(\hat{\mathbf{w}}) = \frac{1}{N}\left\|\mathbf{y} - X\hat{\mathbf{w}}\right\|^2 = \frac{1}{N}\cdot\left[\|\mathbf{y}\|^2 + \hat{\mathbf{w}}^T X^T X\hat{\mathbf{w}} - 2\cdot\left(\mathbf{y}^T\cdot X\hat{\mathbf{w}}\right)\right]$$

נבחין כי $X\cdot\hat{\mathbf{w}}$ הוא וקטור שכל איבר k-י בו הוא סקלר שהתקבל מהכפל $\mathbf{x}_k^T\cdot\hat{\mathbf{w}}$.

כעת נרצה למצוא נקודה סטציונרית (חשודה לקיצון) ולשם כך נשתמש בגרדיאנט ונשווה לאפס:

$$\nabla_{\hat{\mathbf{w}}} E_{in}(\hat{\mathbf{w}}) = \frac{1}{N}\left[\underbrace{2X^TX\hat{\mathbf{w}}}_{(1)}\underbrace{-2\cdot X^T\mathbf{y}}_{(2)}\right] = \frac{2}{N}\cdot X^T(X\hat{\mathbf{w}}-\mathbf{y}) = \mathbf{o}$$

$$\Longleftrightarrow \boxed{X^T(X\hat{\mathbf{w}}-\mathbf{y}) = \mathbf{o}} \quad \backslash normal-equations$$

$$\Longleftrightarrow X^TX\hat{\mathbf{w}} = X^T\mathbf{y}$$

(1) - נסמן $X^TX = A$ ואז $\nabla_{\hat{\mathbf{w}}}(\hat{\mathbf{w}}^TA\hat{\mathbf{w}}) = (\underbrace{A+A^T}_{X^TX+XX^T=2X^TX})\hat{\mathbf{w}} = 2X^TX\hat{\mathbf{w}}$

(2) - נסמן $\mathbf{b} = -2\cdot\mathbf{y}^T\cdot X$ ואז $\nabla_{\hat{\mathbf{w}}}(\mathbf{b}^T\hat{\mathbf{w}}) = \mathbf{b} = -2\cdot X^T\mathbf{y}$

לפני שנמשיך למציאת הנקודה, נתבונן בנגזרת השנייה (הסיין):

$$\nabla^2_{\hat{\mathbf{w}}} E_{in}(\hat{\mathbf{w}}) = \frac{2}{N}X^TX$$

כעת נראה שהביטוי X^TX מתאר מטריצה אי-שלילית, לשם כך נשתמש בהגדרה ונגדיר וקטור כלשהוא $\mathbf{a}\in\mathbb{R}^N$ ונבדוק:

$$\mathbf{a}^TX^TX\mathbf{a} = (X\mathbf{a})^T(X\mathbf{a}) = \|X\mathbf{a}\|^2 \geq 0$$

לכן הנגזרת השנייה היא **אי-שלילית** ולכן הנקודות שייקימו את השוויון מעלה הן נקודות מינימום **גלובליות** וכמו-כן מדובר בפונקציה קונבקסית.

נמשיך בפתרון המשוואה: $X^TX\hat{\mathbf{w}} = X^T\mathbf{y}$. **נניח כי** X^TX הפיכה ונכפיל את הביטוי משמאל בהפכית:

$$X^TX\hat{\mathbf{w}} = X^T\mathbf{y} \overset{\backslash(X^TX)^{-1}}{\Longleftrightarrow} \boxed{\hat{\mathbf{w}} = (X^TX)^{-1}(X^T\mathbf{y})}$$

קיבלנו פתרון יחיד!

הערה: ההנחה ש X^TX הפיכה נובעת מכך שהמטריצה X תהיה מדרגה מלאה - דבר שלא יכול לקרות במידה וקיימת **קורלציה מלאה** בין הפי'צרים או לחלופין שמספר הדוגמאות קטן ממספר הפי'צרים - אלו יניבו לנו דרגה נמוכה.

- דוגמה למספר דוגמאות קטן ממספר פי'צרים - למידה מתמונות שכל פיקסל הוא פי'צר
- ניתן להתמודד עם מצב שכזה עם צמצום של הפיצ'רים על ידי טרנספורמציות וכד'

כעת, נחזור קצת אחורה ונתבונן בביטוי $X^T X\hat{\mathbf{w}} = X^T\mathbf{y}$ ונוכיח טענה:

טענה: X בעלת דרגה מלאה $\Leftarrow$ $X^T X$ הפיכה

הוכחה. נניח כי X בעלת דרגה מלאה

- נניח בשלילה כי $X^T X$ סינגולרית (לא הפיכה)
  - אז קיים וקטור $\mathbf{c} \neq \mathbf{0}$ כך ש:

$$(X^T X)\mathbf{c} = \mathbf{0}$$

  - לכן:

$$\begin{aligned}
&(X^T X)\mathbf{c} = \mathbf{0} \\
\overset{\backslash \mathbf{c}^T}{\Longleftrightarrow}\; &\mathbf{c}^T X^T (X\mathbf{c}) = 0 \\
\Rightarrow\; &\|X\mathbf{c}\|^2 = 0 \\
\Rightarrow &X\mathbf{c} = 0
\end{aligned}$$

  - לפיכך X לא בעלת דרגה מלאה
  - זו סתירה להנחה ולכן בהכרח $X^T X$ הפיכה

■

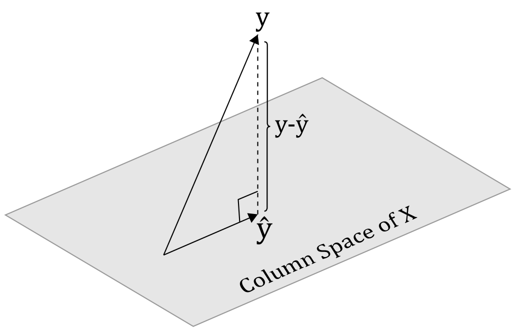
מסקנה: לאחר שהוכחנו ש $X^T X$ הפיכה - בהכרח אין לה ע"ע 0 ולכן בהכרח $X^T X \succ 0$!

1.3 קצת אינטואיציה

נתבונן לרגע במשוואה $E_{in}(\hat{\mathbf{w}}) = \frac{1}{N}\|\mathbf{y} - X\hat{\mathbf{w}}\|^2$ ובמשוואה הנורמלית $X^T \cdot (y - X\hat{\mathbf{w}}) = 0$.
כידוע, $X\hat{\mathbf{w}}$ זה בעצם התוצאה שהניב לנו המודל על סט הדוגמאות, לשם נוחות נסמן: $\hat{y} = X\hat{\mathbf{w}}$.
שאלה: מה נוכל להסיק על הקשר בין $\hat{y}$ לבין הפיצ'רים ב- X?
תשובה: נוכל להסיק ש- $\hat{y}$ הוא בעצם קומבינציה לינארית של עמודות X - או במילים אחרות $\hat{y}$ בעצם שייך למרחב שנפרש על ידי עמודות המטריצה X!
כמו-כן יש לשים לב שהפתרון שקיבלנו בסופו של דבר **קיים את המשוואות הנורמליות** - ואכן על מנת לקיים:

$$X^T \cdot (y - \hat{y}) = 0$$

יש צורך בכך ש X^T יהיה **אורתוגונלי** ל- $(y - \hat{y})$.
במילים אחרות: אנחנו רוצים למצוא $\hat{y}$ שתלוי במרחב העמודות של X שיהיה **הכי קרוב** ל- y מבחינת שגיאה ריבועית על ידי פתרון המשוואות הנורמליות. נעשה זאת על ידי בחירת $\hat{y}$ כך שהשגיאה בינו לבין y תהיה אורתוגונלית לעמודות X



נוכל כעת לחבר את הכל ולהשתמש בכך ש- $\hat{\mathbf{w}} = (X^TX)^{-1}(X^T\mathbf{y})$ ולהגדיר את מטריצת ההיטל האורתוגונלי H:

$$\hat{\mathbf{y}} = X\hat{\mathbf{w}} = \underbrace{X(X^TX)^{-1}X^T}_{H} \cdot \mathbf{y}$$

כאשר ל-H יש 2 מאפיינים חשובים:

(1) H סימטרית - $H^T = H$

(2) $H^2 = H$

כעת, נוכל להתבונן בנורמה בריבוע של המרחק בין $\hat{\mathbf{y}}$ ל- $\mathbf{y}$ ולקבל ביטוי דומה **למשפט פיתגורס**:

$$\begin{aligned}
\|\hat{\mathbf{y}} - \mathbf{y}\|^2 &= \|\mathbf{y} - X\hat{\mathbf{w}}\|^2 \\
&= (\mathbf{y} - X\hat{\mathbf{w}})^T(\mathbf{y} - X\hat{\mathbf{w}}) \\
X^T \cdot (y - \hat{y}) = 0 \quad &= \mathbf{y}^T \cdot (\mathbf{y} - X\hat{\mathbf{w}}) \\
&= \|\mathbf{y}\|^2 - \mathbf{y}^TX\hat{\mathbf{w}} \\
\mathbf{y}^TX \underbrace{=}_{normal-equation} (X^T\mathbf{y})^T = (X^TX\hat{\mathbf{w}})^T = \hat{\mathbf{w}}^TX^TX \quad &= \|\mathbf{y}\|^2 - \underbrace{\hat{\mathbf{w}}^TX^T}_{\hat{\mathbf{y}}^T}\underbrace{X\hat{\mathbf{w}}}_{\hat{\mathbf{y}}} \\
&= \|\mathbf{y}\|^2 - \|\hat{\mathbf{y}}\|^2
\end{aligned}$$

1.4 ומה קורה אם X לא בעלת דרגה מלאה?

כפי שהזכרנו יכול לקרות מצב כזה, וכאשר ננסה לפתור את המשוואות הנורמליות לא נוכל להשתמש בהנחה ש X^TX הפיכה ולכן נסיק שישנן אינסוף פתרונות ל- $\hat{\mathbf{w}}$.

עם זאת, הוקטור $\hat{\mathbf{y}} = X\hat{\mathbf{w}}$ הוא יחיד כיוון שהוא מייצג את ההיטל של $\mathbf{y}$ על מרחב העמודות של X.

הפער הוא ש**יש יותר מוקטור $\hat{\mathbf{w}}$ יחיד שמספק את $\hat{\mathbf{y}}$.**

הערה: אמנם יש הרבה פתרונות, אך הם לא שונים - ולכן אין שוני בין פתרון אחד לשני מבחינת קריטריון השגיאה שבחרנו לפי in sample error