

תרגול 2 - רגרסיה לינארית

PDF

Code

תקציר התיאוריה

למידה מונחית

הגדרה

- נתונים שני משתנים אקראיים x ו y בעלי פילוג לא ידוע.
- נתון לנו **מדגם** של זוגות של x ו y אשר יוצרו מ N דגימות בלתי תלויות:

$$\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^N$$

- מרחב החזאים \ השערות \mathcal{H} . במרחב זה נמצאים כל החזאים האפשריים.
- נסמן ב $\hat{y} = h(x)$ חזאיים אפשריים של y בהינתן x .
- נתונה לנו **פונקציית מחיר** $C(h)$ אשר מחשבת מחיר לכל חזאי. (C יכול להיות תלוי בפילוג).
- נרצה למצוא את החזאי h^* עם המחיר הנמוך ביותר.

הערות

- את המשתנים y מקובל לכנות **labels** (תגיות).
- את המשתנים x מקובל לכנות **observations \ measurements** (תצפיות / מדידות).
- גם x וגם y יכולים להיות וקטורים או סקלרים. המקרה הנפוץ הינו ש x הוא וקטור ו y סקלר.

רישום כבעיית אופטימיזציה

את הבעיה של מציאת החזאי האופטימאלי ניתן לרשום כ:

$$h^* = \arg \min_{h \in \mathcal{H}} C(h)$$

בעיה: לרוב, פונקציית המחיר C תהיה תלויה בפילוג הלא ידוע. לשם כך נאלץ להשתמש במדגם כתחליף לפילוג הלא ידוע. במהלך הקורס נכיר כמה שיטות לעשות זאת.

אבחנה בין שני מקרים

מקובל לחלק את הבעיות בלמידה מונחית לשתי קטגוריות:

בעיות רגרסיה - המקרה בו y הוא משתנה רציף.

בעיות סיווג - המקרה בו y הוא משתנה בדיד המקבל סט סופי של ערכים.

(בעיקרון יכולות להיות גם בעיות בהן y בדיד ולא סופי. בבעיות מסוג זה לרוב פשוט מניחים שע רציף והופכים את הבעיה לבעיית רגרסיה)

פונקציות הפסד וסיכון

דרך נפוצה להגיד את פונקציית המחיר היא כתוחלת על פונקציית הפסד באופן הבא:

- נגדיר פונקציה l אשר מחשבת לחיזוי בודד גודל המכונה **loss** (הפסד). זאת אומרת שבעבור דגימה בודדת, עם ערכי y ו \mathbf{x} כל שהם, ועם תוצאת חיזוי $\hat{y} = h(\mathbf{x})$. ההפסד מוגדרת להיות:

$$l(\hat{y}, y) = l(h(\mathbf{x}), y)$$

- בעזרת פונקציית loss ניתן להגדיר את פונקציית המחיר כתוחלת של ההפסד על פני הפילוג של \mathbf{x} ו y :

$$C(h) = \mathbb{E} [l(h(\mathbf{x}), y)]$$

במקרים כאלה, מוקבל לכנות את פונקציית המחיר, פונקציית **risk** (סיכון), ולסמנה באות R :

$$R(h) = \mathbb{E} [l(h(\mathbf{x}), y)]$$

(Empirical risk minimization (ERM

אחת הדרכים הנפוצות לנסות ולהתמודד עם חוסר הידיעה של הפילוג, היא להחליף את התוחלת על הפילוג הלא ידוע, בתוחלת אמפירית על המדגם. התוחלת האימפירית מוגדרת כמוצע על פני אוסף של דגימות (במקרה שלנו על המדגם). נסמן את התוחלת האימפירית על פני מדגם \mathcal{D} ב $\hat{\mathbb{E}}_{\mathcal{D}}$, ואת risk האמפירי ב $\hat{R}_{\mathcal{D}}$:

$$\hat{R}_{\mathcal{D}}(h) = \hat{\mathbb{E}}_{\mathcal{D}} [l(h(\mathbf{x}), y)] = \frac{1}{N} \sum_{i=0}^N l(h(\mathbf{x}^{(i)}), y^{(i)})$$

בעיית האופטימיזציה תהיה במקרה זה:

$$h_{\mathcal{D}}^* = \arg \min_{h \in \mathcal{H}} \frac{1}{N} \sum_{i=0}^N l(h(\mathbf{x}^{(i)}), y^{(i)})$$

(empirical risk minimization (ERM מוכנה

שימו לב: מכיוון שהתוחלת האמפירית היא רק קירוב של התוחלת האמיתית, הפתרון של בעיית ה ERM גם יהיה רק קירוב של הפתרון של הבעיה המקורית. זאת אומרת שבמקרה הכללי $h_{\mathcal{D}}^* \neq h^*$.

מודלים פרמטריים

לרוב אנו נגביל את החיפוש של החזאי למשפחה מצומצמת של חזאים בעלי צורה קבועה עד כדי מספר סופי של פרמטרים. את הפרמטרים של המודל נסמן בעזרת הוקטור θ . אנו נשתמש ב $h(\mathbf{x}; \theta)$ לתיאור של חזאי מהמשפחה עם פרמטרים θ .

דוגמא למשפחה פרמטרית:

$$h(\mathbf{x}; \theta) = \theta_1 \cos(\theta_2 x) e^{-\theta_3 x}$$

כאשר עובדים עם מודל פרמטרי, האופטימיזציה היא למעשה על על הפרמטרים, והבעיה הופכת להיות:

$$\theta_{\mathcal{D}}^* = \arg \min_{\theta} \frac{1}{N} \sum_{i=0}^N l(h(\mathbf{x}^{(i)}; \theta), y^{(i)})$$

מאפיינים

בהינתן מודל פרמטרי כל שהוא, ניתן בקלות לייצר מודלים פרמטריים חדשים על ידי ביצוע עיבוד מקדים כל שהוא ל \mathbf{x} לפני שהוא מוזן למודל. את העיבוד המקדים ניתן לתאר כאוסף של פונקציות φ_k אשר פועלות על \mathbf{x} . את המידע המעובד (המוצא של ה φ -ים) מקובל לכנות מאפיינים. כמו כן, לרוב נוה לאגד את כל הפונקציות φ לפונקציה אחת Φ אשר פועלת על \mathbf{x} ומחזירה את וקטור המאפיינים:

$$\Phi(\mathbf{x}) = [\varphi_1(\mathbf{x}), \varphi_2(\mathbf{x}), \dots, \varphi_M(\mathbf{x})]^T$$

המודל הפרמטרי החדש יהיה הרכבה של Φ ו h :

$$h_{\text{new}}(\mathbf{x}; \theta) = h(\Phi(\mathbf{x}); \theta)$$

דרך אחרת להסתכל על המאפיינים הינה שאנו כביכול מחליפים את המדגם שקיבלנו במדגם חדש באופן הבא:

$$\mathbf{x}_{\text{new}} = \Phi(\mathbf{x}_{\text{old}})$$

רגרסיה לינארית

רגרסיה לינארית עוסקת בבעיות רגרסיה שבהם המודל הינו **לינארי בפרמטרים שלו**. זאת אומרת, בעיות בהם המודל הינו מהצורה של:

$$h(\mathbf{x}; \boldsymbol{\theta}) = \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_d x_d = \mathbf{x}^\top \boldsymbol{\theta}$$

כפי שצינו קודם, וכפי שנראה בתרגיל, תמיד ניתן להשתמש במאפיינים על מנת לקבל פונקציות מורכבות יותר:

$$h(\mathbf{x}; \boldsymbol{\theta}) = \theta_1 \varphi_1(\mathbf{x}) + \theta_2 \varphi_2(\mathbf{x}) + \dots + \theta_M \varphi_M(\mathbf{x}) = \Phi(\mathbf{x})^\top \boldsymbol{\theta}$$

(Linear least squares (LLS

מקרה מיוחד הוא המקרה שבו משתמשים במודל לינארי יחד עם risk עם loss ריבועי:

$$l(\hat{y}, y) = (\hat{y} - y)^2$$

במקרה זה, מתקבל בעיית אופטימיזציה אשר ניתן לפתור אותה באופן אנליטי:

$$\boldsymbol{\theta}_D^* = \arg \min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=0}^N (h(\mathbf{x}^{(i)}; \boldsymbol{\theta}) - y^{(i)})^2 = \arg \min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=0}^N (\mathbf{x}^{(i)\top} \boldsymbol{\theta} - y^{(i)})^2$$

כתיב מטריצי

כדי לפתור את הבעיה, נוח יותר לרשום אותה בכתיב מטריצי.

- נגדיר את הוקטור \mathbf{y} כוקטור של כל התגיות במדגם:

$$\mathbf{y} = [y^{(1)}, y^{(2)}, \dots, y^{(N)}]^\top$$

- נגדיר את המטריצה X כמטריצה של כל ה- \mathbf{x} ים במדגם:

$$X = \begin{bmatrix} - & \mathbf{x}^{(1)} & - \\ - & \mathbf{x}^{(2)} & - \\ & \vdots & \\ - & \mathbf{x}^{(N)} & - \end{bmatrix}$$

בעזרת הגדרות אלו ניתן לרשום את בעיית האופטימיזציה של LLS באופן הבא:

$$\boldsymbol{\theta}_D^* = \arg \min_{\boldsymbol{\theta}} \frac{1}{N} \|X\boldsymbol{\theta} - \mathbf{y}\|_2^2$$

הפתרון של LLS

את בעיית האופטימיזציה הזו ניתן לפתור על ידי גזירה והשוואה ל-0, כפי שנעשה בתרגיל הראשון. הפתרון המתקבל הינו:

$$\boldsymbol{\theta}_D^* = (X^\top X)^{-1} X^\top \mathbf{y}$$

הפתרון הזה נכון כאשר $X^\top X$ הפיכה. תנאי הכרחי בכדי שזה יקרה הינו שמספר הדגימות N יהיה גדול מהמימד של \mathbf{x} (אשר נסמן כ D). כאשר המטריצה לא הפיכה יש לבעיה יותר מפתרון יחיד, כפי שנראה בהמשך.

(המטריצה $(X^\top X)^{-1} X^\top$ נקראת Moore-Penrose pseudo inverse)

הערה

כשאר משתמשים במאפיינים המטריצה X תהיה:

$$X = \begin{bmatrix} - & \Phi(\mathbf{x}^{(1)}) & - \\ - & \Phi(\mathbf{x}^{(2)}) & - \\ & \vdots & \\ - & \Phi(\mathbf{x}^{(N)}) & - \end{bmatrix}$$

תרגיל 2.1

הראו כי כאשר $X^T X$ הפיך, הפתרון של בעיית האופטימיזציה של LLS:

$$\boldsymbol{\theta}_D^* = \arg \min_{\boldsymbol{\theta}} \frac{1}{N} \|X\boldsymbol{\theta} - \mathbf{y}\|_2^2$$

נתון על ידי:

$$\boldsymbol{\theta}_D^* = (X^T X)^{-1} X^T \mathbf{y}$$

פתרון 2.1

נפתור על ידי גזירה והשוואה ל:0:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \left(\frac{1}{N} \|X\boldsymbol{\theta} - \mathbf{y}\|_2^2 \right) &= 0 \\ \Leftrightarrow \nabla_{\boldsymbol{\theta}} \left((X\boldsymbol{\theta} - \mathbf{y})^T (X\boldsymbol{\theta} - \mathbf{y}) \right) &= 0 \\ \Leftrightarrow \nabla_{\boldsymbol{\theta}} (\boldsymbol{\theta}^T X^T X \boldsymbol{\theta} - 2\mathbf{y}^T X \boldsymbol{\theta} + \|\mathbf{y}\|_2^2) &= 0 \end{aligned}$$

בכדי לחשב את הנגזרות נשתמש בשני הנגזרות המוכרות הבאות:

$$\nabla_{\mathbf{x}} (\mathbf{a}^T \mathbf{x}) = \mathbf{a}, \quad \nabla_{\mathbf{x}} (\mathbf{x}^T A \mathbf{x}) = 2A\mathbf{x}$$

על ידי שימוש בנגזרות אלו נקבל

$$\begin{aligned} \Leftrightarrow \nabla_{\boldsymbol{\theta}} (\boldsymbol{\theta}^T X^T X \boldsymbol{\theta} - 2\mathbf{y}^T X \boldsymbol{\theta} + \|\mathbf{y}\|_2^2) &= 0 \\ \Leftrightarrow 2X^T X \boldsymbol{\theta} - 2X^T \mathbf{y} &= 0 \\ \Leftrightarrow X^T X \boldsymbol{\theta} &= X^T \mathbf{y} \end{aligned}$$

זוהי בעיה של פתרון מערכת משוואות ליניארית מהצורה של $A\mathbf{x} = \mathbf{b}$ כאשר:

$$A = X^T X, \quad \mathbf{b} = X^T \mathbf{y}$$

כאשר המטריצה $X^T X$ הפיכה, הפתרון של בעיה זו נתון על ידי:

$$\boldsymbol{\theta} = (X^T X)^{-1} X^T \mathbf{y}$$

כאשר היא אינה הפיכה ישנם מספר פתרונות (למעשה קיים מרחב ליניארי של פתרונות אשר פותרים את הבעיה).

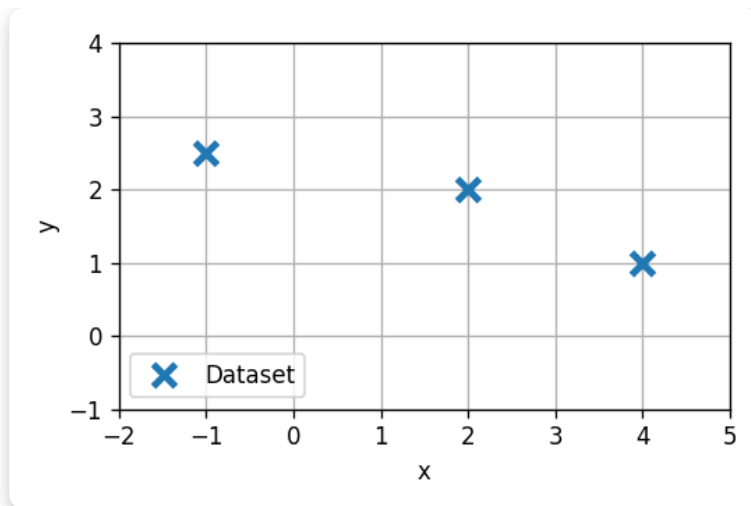
תרגיל 2.2

נתונה לנו בעיית LLS עם המדגם הבא:

$$\mathcal{D} = \{ \{x^{(1)} = -1, y^{(1)} = 2.5\}, \{x^{(2)} = 2, y^{(2)} = 2\}, \{x^{(3)} = 4, y^{(3)} = 1\} \}$$

להבא בקורס, נשמט את הרישום של x ו y בהגדרת המדגם ונרשום אותו בקצרה באופן הבא:

$$\mathcal{D} = \{ \{-1, 2.5\}, \{2, 2\}, \{4, 1\} \}$$



- (1)** נרצה כעת להשתמש במאפיינים בכדי לקבל מודל שהוא פונקציה לינארית (עם איבר היסט) ב.ש. רשמו את המאפיינים המתאימים ואת המודל המתקבל. מצאו את הפרמטרים של המודל האופטימאלי?
- (2)** נרצה כעת להשתמש במאפיינים בכדי לקבל מודל שהוא פולינום מסדר 2 ב.ש. רשמו את המאפיינים ואת המודל המתקבל ומצאו את הפרמטרים של המודל האופטימאלי?
- (3)** נרצה כעת להשתמש במאפיינים בכדי לקבל מודל שהוא פולינום מסדר 3 ב.ש. האם במקרה זה קיים פתרון יחיד? מצאו את הפתרונות למקרה שבו $\theta_1 = 0$ (איבר ההיסט מתאפס) ולמקרה שבו $\theta_3 = 0$ (המקדם של x^2 מתאפס)
- (4)** נרצה כעת להשתמש בפונקציות המאפיינים הבאות:

$$\varphi_m(x) = \exp\left(-\frac{(x - \mu_m)^2}{2\sigma_m^2}\right) \quad m \in 1, 2, 3$$

כאשר

$$\sigma_1 = 1.5, \sigma_2 = \sigma_3 = 1 \quad \mu_1 = -1, \mu_2 = 2, \mu_3 = 4$$

חשבו את הפרמטרים של המודל האופטימאלי בעבור מאפיינים אלו.

- (5)** בעבור כל אחד מהסעיפים חשבו את הסיכון האמפירי המתקבל. האם לדעתכם סיכון אמפירי קטן יותר בהכרח מעיד על סיכון (לא אמפירי) קטן יותר?

פתרון 2.2

(1)

אנו מעוניינים במודל מהצורה:

$$h(x; \theta) = \theta_1 + \theta_2 x$$

כפי שראינו בהרצאה, ניתן להוסיף את איבר ההיסט על ידי שימוש במאפיינים. אנו נעשה זאת על ידי שימוש במאפיינים הבאים:

$$\varphi_1(x) = 1, \quad \varphi_2(x) = x$$

או בכתיב וקטורי

$$\Phi(x) = [1, x]^T$$

פעולה זו למעשה פשוט מוסיפה את האיבר 1 ל- x וכביכול יוצר את המדגם הבא:

$$\mathcal{D} = \{ \{ [1, -1]^T, 2.5 \}, \{ [1, 2]^T, 2 \}, \{ [1, 4]^T, 1 \} \}$$

נרשום את X ו- y :

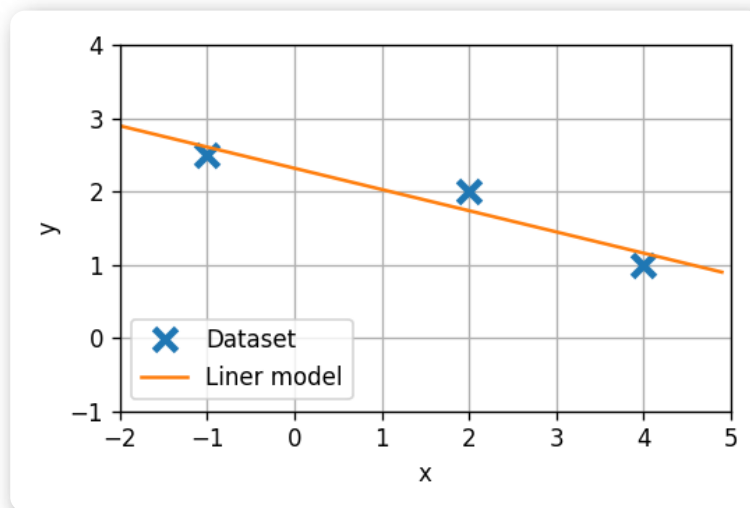
$$X = \begin{bmatrix} 1 & -1 \\ 1 & 2 \\ 1 & 4 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 2.5 \\ 2 \\ 1 \end{bmatrix}$$

הפרמטרים האופטימאליים של המודל אשר ממזערים את הסיכון האמפירי יהיו אם כן:

$$\begin{aligned} \theta_{\mathcal{D}}^* &= (X^T X)^{-1} X^T \mathbf{y} \\ &= \left(\begin{bmatrix} 1 & 1 & 1 \\ -1 & 2 & 4 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 2 \\ 1 & 4 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 \\ -1 & 2 & 4 \end{bmatrix} \begin{bmatrix} 2.5 \\ 2 \\ 1 \end{bmatrix} \\ &= \left(\begin{bmatrix} 3 & 5 \\ 5 & 21 \end{bmatrix} \right)^{-1} \begin{bmatrix} 5.5 \\ 4.5 \end{bmatrix} = \frac{1}{38} \begin{bmatrix} 21 & -5 \\ -5 & 3 \end{bmatrix} \begin{bmatrix} 5.5 \\ 4.5 \end{bmatrix} \\ &= \frac{1}{38} \begin{bmatrix} 88 \\ -11 \end{bmatrix} = \begin{bmatrix} 2.3158 \\ -0.2895 \end{bmatrix} \end{aligned}$$

מכאן שהמודל שלנו הינו:

$$h(x) = 2.3158 - 0.2895x$$



(2)

בדומה לסעיף הקודם נבחר את פונקציות המאפיינים הבאות:

$$\Phi(x) = [1, x, x^2]^T$$

על מנת לקבל את המודל הבא:

$$h(x; \theta) = \theta_1 + \theta_2 x + \theta_3 x^2$$

לאחר הפעלת פונקציות המאפיינים נקבל את המדגם הבא:

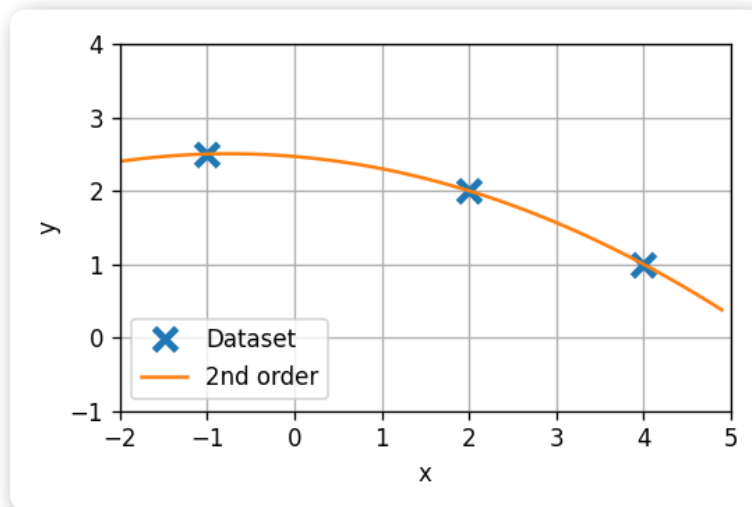
$$\mathcal{D} = \{ \{ [1, -1, 1]^T, 5 \}, \{ [1, 2, 4]^T, 2 \}, \{ [1, 4, 16]^T, 1 \} \}$$

הוקטור \mathbf{y} אינו מושפע מבחירת המאפיינים, אך המטריצה X תהיה כעת:

$$X = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \end{bmatrix}$$

הוקטור הפרמטרים המתקבל הינו: (את זה כבר עדיף לחשב במחשב)

$$\theta_D^* = (X^T X)^{-1} X^T \mathbf{y} = \frac{1}{30} [74, -3, -2]^T = [2.467, -0.1, -0.067]^T$$



(3)

נבחר את פונקציות המאפיינים הבאות:

$$\Phi(x) = [1, x, x^2, x^3]^T$$

המטריצה X תהיה כעת:

$$X = \begin{bmatrix} 1 & -1 & 1 & -1 \\ 1 & 2 & 4 & 8 \\ 1 & 4 & 16 & 64 \end{bmatrix}$$

כיוון שלמטריצה X יש יותר עמודות משורות (יש יותר פרמטרים מדגימות) המטריצה $X^T X$ בהכרח לא תהיה הפיכה. ולכן כפי שצינינו קודם יהיו לבעיה הרבה פתרונות. (ניתן להראות כי המטריצה לא הפיכה לפי העובדה שלא יכולים להיות ב $X^T X$ יותר מ-3 שורות בלתי תלויות ולכן המימד שלה הוא לכל היותר 3)

נסתכל על שני המקרים הפרטיים $\theta_1 = 0$ ו $\theta_3 = 0$.

$$\theta_1 = 0$$

במקרה זה המודל הפרמטרי מתנוון ל:

$$h(x; \theta) = \theta_2 x + \theta_3 x^2 + \theta_4 x^3$$

אנו למעשה יכולים לפתור את זה כבעיית LLS עם וקטור פרמטריים $\theta = [\theta_2, \theta_3, \theta_4]^T$ ומאפיינים:

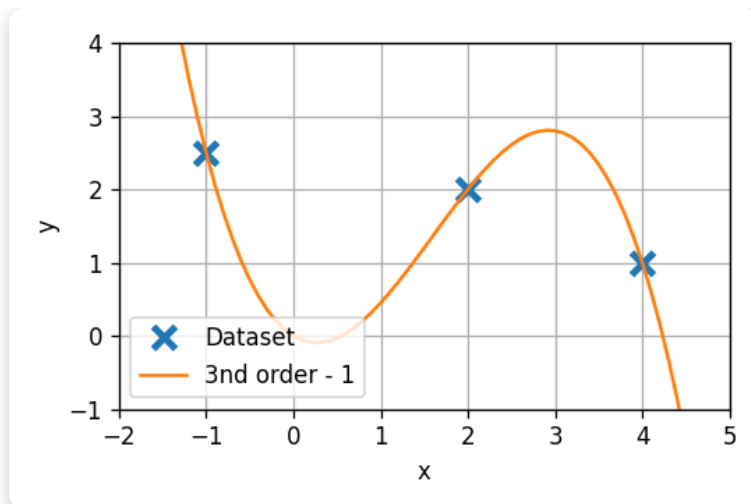
$$\Phi(x) = [x, x^2, x^3]^T$$

המטריצה X תהיה:

$$X = \begin{bmatrix} -1 & 1 & -1 \\ 2 & 4 & 8 \\ 4 & 16 & 64 \end{bmatrix}$$

והפתרון שיתקבל יהיה:

$$\theta_D^* = (X^T X)^{-1} X^T \mathbf{y} = \frac{1}{120} [-86, 177, 37]^T = [-0.7167, 1.475, -0.3083]^T$$



$$\theta_3 = 0$$

במקרה השני, המודל הפרמטרי מתנוון ל:

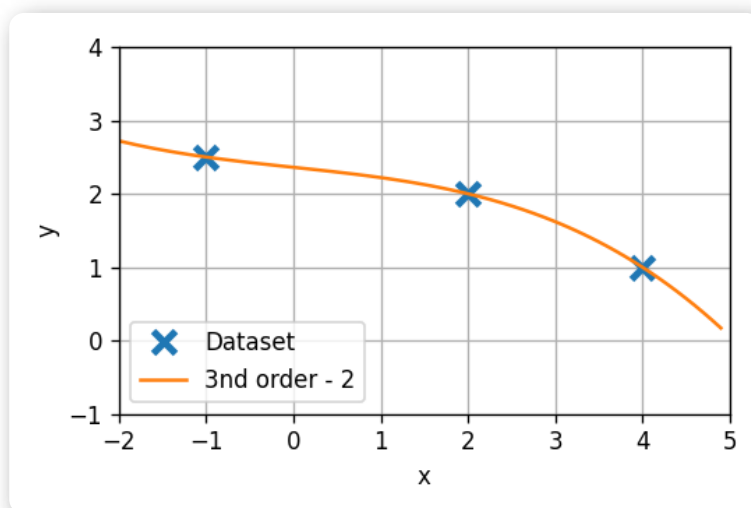
$$h(x; \theta) = \theta_1 + \theta_2 x + \theta_4 x^3$$

המטריצה X תהיה:

$$X = \begin{bmatrix} 1 & -1 & -1 \\ 1 & 2 & 8 \\ 1 & 4 & 64 \end{bmatrix}$$

והפתרון שיתקבל יהיה:

$$\theta_{\mathcal{D}}^* = (X^T X)^{-1} X^T \mathbf{y} = \frac{1}{150} [354, -19, -2]^T = [2.36, -0.1267, -0.0133]^T$$



(4)

העובדה שהמאפיינים הם לא חזקות של x לא משנה דבר. נפתור את הבעיה באופן דומה לסעיפים הקודמים:

בעבור המאפיינים:

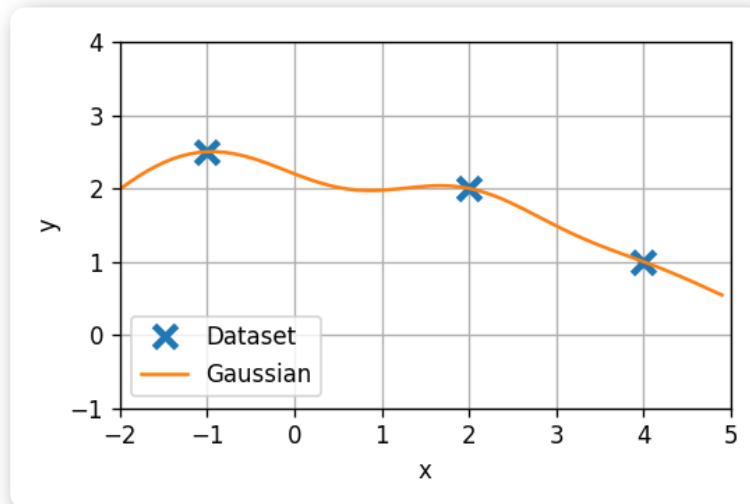
$$\varphi_1(x) = \exp\left(-\frac{(x+1)^2}{2 \cdot 1.5^2}\right), \varphi_2(x) = \exp\left(-\frac{(x-2)^2}{2}\right), \varphi_3(x) = \exp\left(-\frac{(x-4)^2}{2}\right)$$

המטריצה X תהיה:

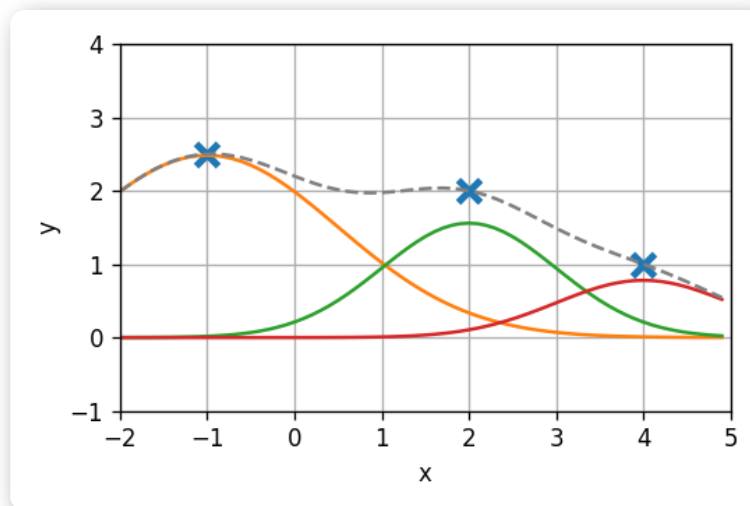
$$X = \begin{bmatrix} 1 & 0.0111 & 3.7 \times 10^{-6} \\ 0.1353 & 1 & 0.1353 \\ 0.0039 & 0.1353 & 1 \end{bmatrix}$$

והפתרון שיתקבל יהיה:

$$\theta_D^* = (X^T X)^{-1} X^T \mathbf{y} = [2.4827, 1.5585, 0.7794]^T$$

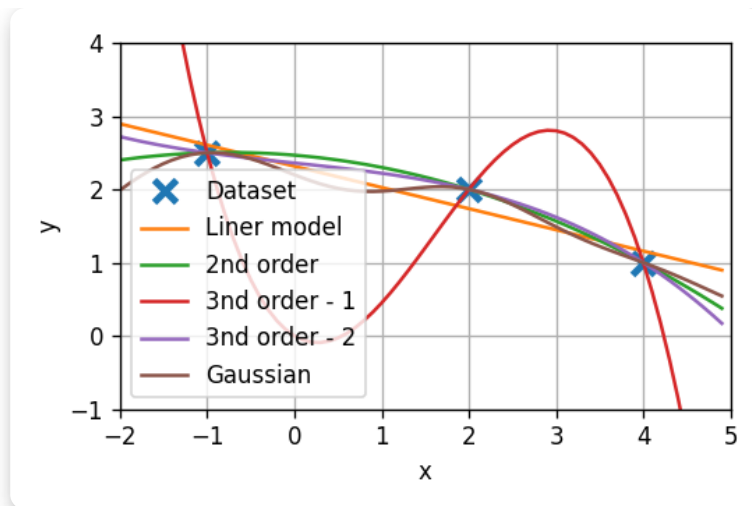


המודל הליניארי הוא פשוט קומבינציה ליניארית של פונקציות של x . לשם המחשה נראה כיצד התוצאה שהתקבלה היא למעשה קומבינציה ליניארית של שלושת הגאוסיאנים של המאפיינים:



(5)

נציג על גרף אחד את כל התוצאות שהתקבלו עד כה:



מלבד במקרה הליניארי, קיבלנו בכל המודלים שגיאה אמפירית 0.

הסבר: אנו למעשה מנסים למצוא וקטור פרמטרים כך ש $X\theta$ יהיה כמה שיותר קרוב ל y . כאשר יש יותר פרמטרים מדגימות ניתן תמיד (עד כדי מקרים שבהם יש תלויות לינאריות בין המאפיינים או הדגימות) למצוא וקטור θ כך שלבעיה הלינארית $X\theta = y$ יהיה פתרון (מצב של יותר נעלמים ממשוואות).

לגבי סיכון האמיתי, ללא מידע נוסף על הבעיה לא ניתן לדעת איזה מודל יכליל בצורה טובה יותר (יעבוד טוב יותר על דוגמאות לא מדגם), וכמובן שאין כל הכרח שמודלים עם שגיאה האמפירית הקטנה יותר יכלילו טוב יותר.

ניחושים

אנחנו כן נוכל לנסות להשתמש בניסיון שלנו מתכונות כלליות של מודלים בעולם האמיתי על מנת לשער איזה מודל יכליל טוב יותר. לדוגמא, הינו מצפים שהמודל לא "ישתולל", זאת אומרת שערכים קרובים של x יתנו ערכים יחסית דומים של y . מה שלא קורה במודל הראשון מסעיף 3 (הגרף האדום). ניתן לדוגמא לשער שהמודל הליניארי, שלא מניב שגיאה אמפירית של 0, יכליל בצורה טובה יותר מהמודל האדום.

בתרגול וההרצאה הקרובים, ננסה להשתמש בכלי בשם רגולריזציה בכדי לתרגם את הדרישה שהמודל לא "ישתולל" לדרישה מתמטית על מודל.

תרגיל 2.3

בעבור וקטור אקראי x באורך 2, $x = (x_1, x_2)^T$ (או בכתיב מתמטי $x \in \mathbb{R}^2$), מהם המאפיינים בהם יש להשתמש על מנת לקבל פולינום מסדר 2 בערכים של x

פתרון 2.3

אנו מעוניינים במודל מהצורה:

$$h(x; \theta) = \theta_1 + \theta_2 x_1 + \theta_3 x_2 + \theta_4 x_1 x_2 + \theta_5 x_1^2 + \theta_6 x_2^2$$

נוכל לייצר זאת על ידי בחירה של המאפיינים הבאים:

$$\Phi(x) = [1, x_1, x_2, x_1 x_2, x_1^2, x_2^2]^T$$

תרגיל 2.4

נסתכל כעת על בעיית רגרסיה של מודל ליניארי כללי $h(x; \theta) = x^T \theta$ ופנקציית סיכון עם הפסד של l_3 :

$$l(\hat{y}, y) = |\hat{y} - y|^3$$

1 רשמו את בעיית האופטימיזציה של ERM.

(2) האם ניתן לפתור בעיה זו בעזרת גזירה והשוואה ל-0?

(3) בעבור פתרון בעזרת gradient descent, רשמו את כלל העדכון של אלגוריתם (עם צעד גרדיאנט η).

פתרון 2.4

(1)

בעיית האופטימיזציה שיש לפתור הינה:

$$\begin{aligned}\theta_{\mathcal{D}}^* &= \arg \min_{\theta} \hat{R}_{\mathcal{D}}(h(\cdot; \theta)) \\ &= \arg \min_{\theta} \frac{1}{N} \sum_{i=0}^N l(h(\mathbf{x}^{(i)}; \theta), y^{(i)}) \\ &= \arg \min_{\theta} \frac{1}{N} \sum_{i=0}^N |\mathbf{x}^{(i)\top} \theta - y^{(i)}|^3\end{aligned}$$

(2)

נחשב את הנגזרת של פונקציית הסיכון האמפירי:

$$\begin{aligned}\nabla_{\theta} \hat{R}_{\mathcal{D}}(h(\cdot; \theta)) &= \nabla_{\theta} \left(\frac{1}{N} \sum_{i=0}^N |\mathbf{x}^{(i)\top} \theta - y^{(i)}|^3 \right) \\ &= \frac{1}{N} \sum_{i=0}^N \nabla_{\theta} |\mathbf{x}^{(i)\top} \theta - y^{(i)}|^3 \\ &= \frac{1}{N} \sum_{i=0}^N \nabla_{\theta} \left((\mathbf{x}^{(i)\top} \theta - y^{(i)})^3 \text{sign}(\mathbf{x}^{(i)\top} \theta - y^{(i)}) \right) \\ &= \frac{1}{N} \sum_{i=0}^N 3\mathbf{x}^{(i)} (\mathbf{x}^{(i)\top} \theta - y^{(i)})^2 \text{sign}(\mathbf{x}^{(i)\top} \theta - y^{(i)})\end{aligned}$$

כנראה שלא ניתן להשוות ביטוי זה ל-0 ולפתור אותו באופן אנליטי.

(3)

תזכורת, אלגוריתם הגרדיאנט מנסה למצוא מינימום לוקאלי על ידי התקדמות בכיוון ההפוך מהגרדיאנט של פונקציית המטרה (שאותה רוצים למזער). הוא עושה זאת בעזרת סדרה של צעדים באופן הבא:

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_{\theta} f(\theta^{(t)})$$

(הפרמטר η משפיע על גודל הצעד)

במקרה שלנו צעד העדכון של האלגוריתם יהיה:

$$\begin{aligned}\theta^{(t+1)} &= \theta^{(t)} - \eta \nabla_{\theta} \hat{R}(h(\cdot; \theta^{(t)})) \\ &= \theta^{(t)} - \eta \frac{1}{N} \sum_{i=0}^N 3\mathbf{x}^{(i)} (\mathbf{x}^{(i)\top} \theta^{(t)} - y^{(i)})^2 \text{sign}(\mathbf{x}^{(i)\top} \theta^{(t)} - y^{(i)})\end{aligned}$$

דוגמא מעשית - חיזוי זמן נסיעה של מוניות בניו יורק

Code

מדגם נסיעות המונית בעיר New York

כחלק מהמאמץ של העיר New York להנגיש לציבור את המידע אותו היא אוספת, עריית NYC מפרסמת בכל חודש את הפרטים של כל נסיעות המונית אשר בוצעו בעיר באותו חודש. בקורס זה, אנו נעשה שימוש ברשימת הנסיעות מחודש ינואר 2016. ניתן למצוא את הרשימה, [פה](#).

הרשימה המלאה כוללת מעל 10 מיליון נסיעות, בכדי לקצר את זמן החישוב, אנו נעשה שימוש רק ברשימה חלקית הכוללת רק 100 אלף נסיעות (אשר נבחרו באקראי אחרי ניקוי מסויים של הרשימה). את הרשימה החלקית ניתן למצוא [פה](#)

המדגם ושדותיו

בטבלה מלטה מוצגים עשרת השורות הראשונות ברשימה

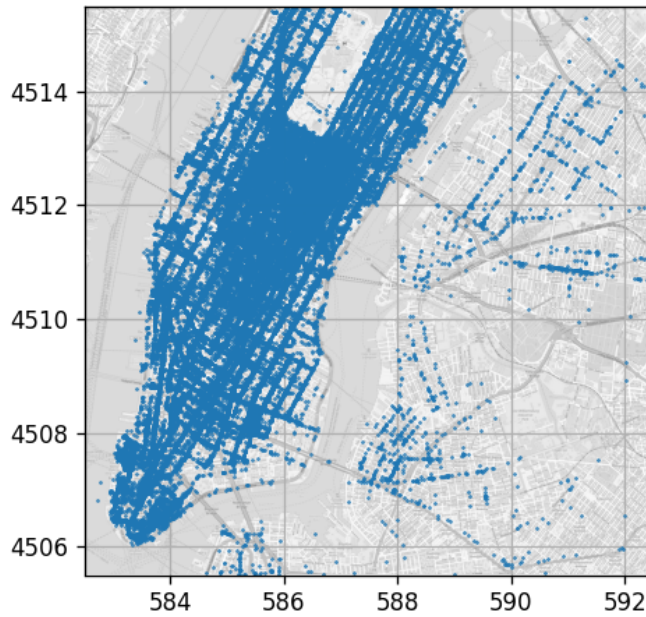
ay of ek	duration	dropoff northing	dropoff easting	pickup northing	pickup easting	tip amount	fare amount	payment type	trip distance	passenger count
3	11.5167	4515.18	588.155	4512.98	586.997	0	9.5	2	2.76806	2 0
6	12.6667	4512.63	584.85	4512.92	587.152	0	10	2	3.21868	1 1
0	5.51667	4513.17	585.434	4513.36	587.005	2.49	7	1	2.57494	1 2
1	9.88333	4512.55	586.672	4511.73	586.649	1.65	7.5	1	0.965604	1 3
2	8.68333	4511.76	585.262	4511.89	586.967	1.66	7.5	1	2.46229	1 4
3	9.43333	4511.54	585.169	4512.88	585.926	2.2	7.5	1	1.56106	5 5
5	7.95	4514.21	588.71	4515.08	586.731	1	8	1	2.57494	1 6
5	4.95	4509.55	585.844	4509.71	585.345	0	5	2	0.80467	1 7
5	11.0667	4507.74	583.671	4509.48	585.422	1.1	10	1	3.6532	1 8
3	4.21667	4513.71	587.701	4514.93	587.875	1.36	5.5	1	1.62543	6 9

הרשימה כוללת מספר שדות (עמודות), אך בתרגול זה אנו נתמקד רק בשדות הבאים:

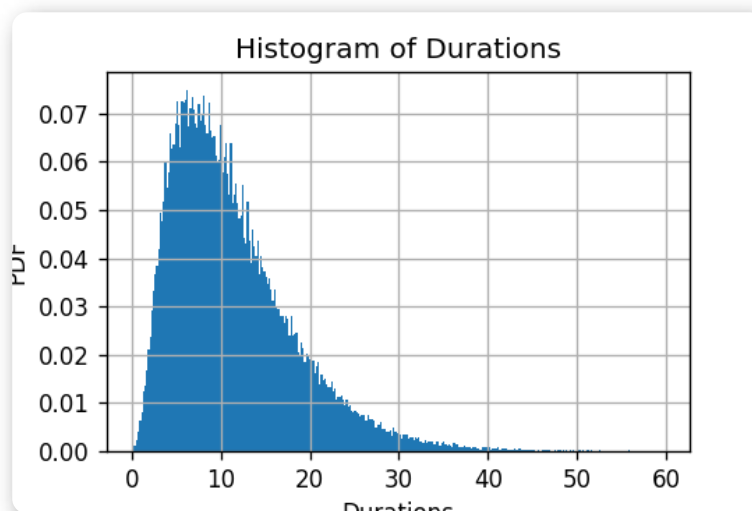
- **pickup_easting** - הקואורדינאטה האורכית (מזרח-מערב) של נקודת האיסוף.
- **pickup_northing** - הקואורדינאטה הרוחבית (צפון דרום) של נקודת האיסוף.
- **dropoff_easting** - הקואורדינאטה האורכית של נקודת ההורדה.
- **dropoff_northing** - הקואורדינאטה הרוחבית של נקודת ההורדה.
- **duration** - משך הנסיעה בדקות.

(למתעניינים, הקואורדינאטות נתונות בפרומט בשם WGS84-UTM שבהם הקואורדינאטות הם ביחידות של בערך קילומטרים)

יזואליזציה של נקודות ההורדה



הפילוג של זמן הנסיעה



הגדרה של הבעיה

בדוגמא זו ננסה לחזות את משך הנסיעה על פי נקודת האיסוף ונקודת ההורדה (הנקודות של תחילת הנסיעה וסיום הנסיעה). נסמן את המשתנה האקראי של משך הנסיעה ב y ואת ארבעת המשתנים של האיסוף וההורדה ב \mathbf{x} :

$$\mathbf{x} = [x_{\text{pick east}}, x_{\text{pick north}}, x_{\text{drop east}}, x_{\text{drop north}}]^T$$

בתור פונקציית מחיר ניקח פונקציית סיכון עם הפסד ריבועי:

$$R(h) = \mathbb{E} [(h(\mathbf{x}) - y)^2]$$

בתרגול זה נסתפק בלנסות למצוא את מודל אשר ימזער את הסיכון האמפירי

$$h_{\mathcal{D}}^* = \arg \min_{h \in \mathcal{H}} \hat{R}_{\mathcal{D}}(h) = \arg \min_{h \in \mathcal{H}} \frac{1}{N} \sum_{i=0}^N (h(\mathbf{x}^{(i)}) - y^{(i)})^2$$

בתרגול הבא ננסה לשפר את הבחירה של המודל כך שהוא גם יוכל להכליל טוב יותר. כמו כן, בתרגול זה נשתמש במודל ליניארי לפונקציית החיזוי:

$$h(\mathbf{x}; \boldsymbol{\theta}) = \Phi(\mathbf{x})^\top \boldsymbol{\theta}$$

$$\boldsymbol{\theta}_{\mathcal{D}}^* = \arg \min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=0}^N (h(\mathbf{x}^{(i)}; \boldsymbol{\theta}) - y^{(i)})^2$$

בחירת מאפיינים - נסיון ראשון

נתחיל בלנסות להשתמש במאפיין בודד. מכיוון שאנו מצפים שהדבר שהכי ישפיע על זמן הנסיעה הינו המרחק שאותו יש ליסוע, ננסה להשתמש במרחק האווירי בין נקודת האיסוף לנקודת ההורדה כמאפיין שלנו:

$$\varphi_{\text{dist}}(\mathbf{x}) = \sqrt{(x_{\text{pick east}} - x_{\text{drop east}})^2 + (x_{\text{pick north}} - x_{\text{drop north}})^2}$$

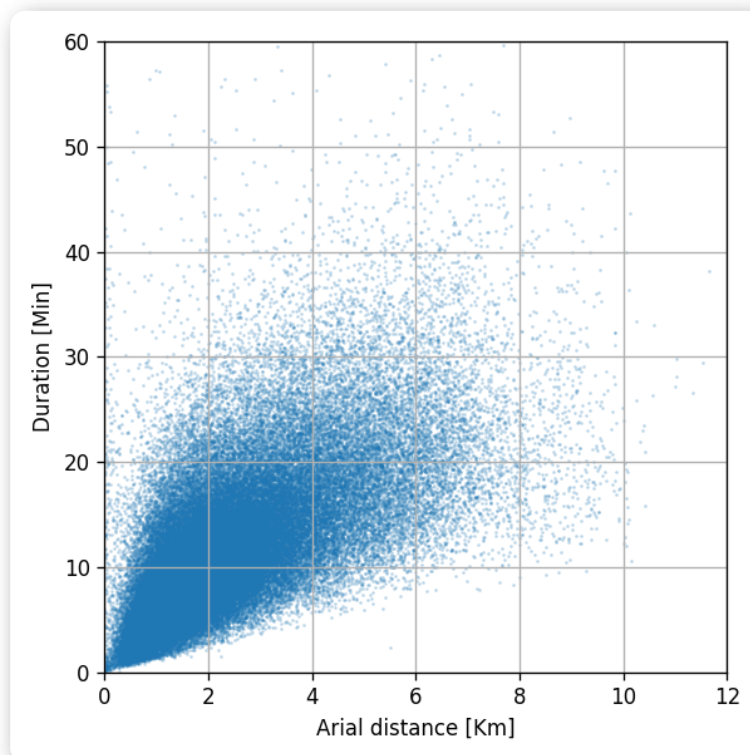
והמודל הליניארי יהיה:

$$h(\mathbf{x}; \theta) = \theta \varphi_{\text{dist}}(\mathbf{x}) = \theta \sqrt{(x_{\text{pick east}} - x_{\text{drop east}})^2 + (x_{\text{pick north}} - x_{\text{drop north}})^2}$$

זאת אומרת שהחיזוי שלנו יהיה פונקציה ליניארית (ללא היסט) של המרחק האווירי בין נקודת האיסוף להורדה.

מציאת הפרמטר של המודל

לפני שנחשב את הפרמטר נציג את התלות בין המרחק האווירי ומשך הנסיעה



ניתן לראות שישנו פיזור גדול ושהנקודות לא יושבות קרוב לקו ליניארי, אך עם זאת, ניתן לראות כי ישנה מגמה כללית של הנקודות. אנו נקווה שהמודל הליניארי ינסה ללמוד מגמה זו.

נחשב את הפרמטר של המודל מתוך הנוסחה:

$$\theta_D^* = (X^T X)^{-1} X y$$

כאשר X הוא הוקטור של המאפיין היחיד φ_{dist}

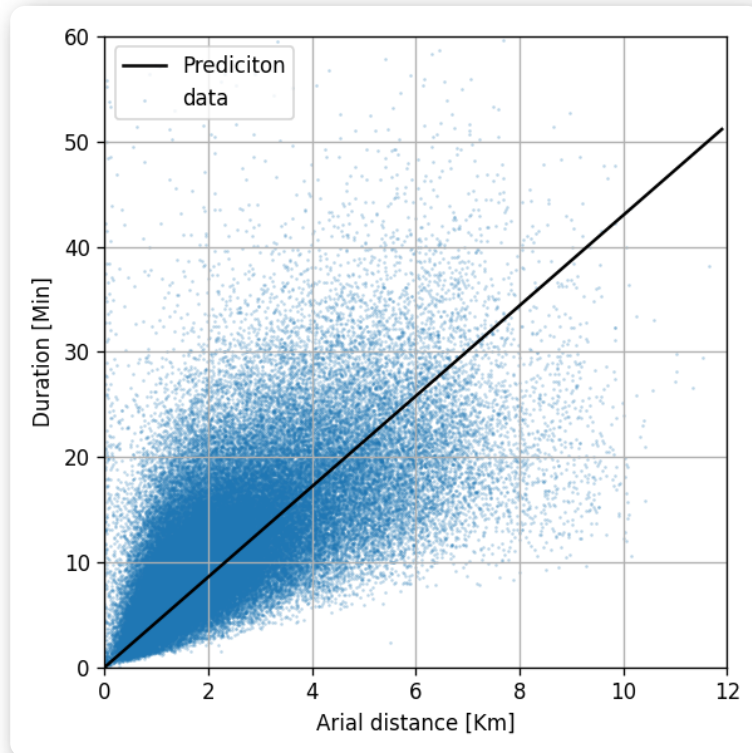
$$X = [\varphi_{\text{dist}}(\mathbf{x}^{(1)}), \varphi_{\text{dist}}(\mathbf{x}^{(2)}), \dots, \varphi_{\text{dist}}(\mathbf{x}^{(N)})]^T$$

הפרמטר המתקבל מחישוב זה הינו:

$$\theta = 4.23$$

זאת אומרת שתחת המודל הלינארי החיזוי של זמן הנסיעה שווה למרחק האווירי אותו יש לעבור כפול 4.23.

נוסיף את המודל הלינארי על גבי הגרף הקודם:



אכן המודל הלינארי למד את המגמה הכללית של הנקודות.

נחשב את הסיכון האמפירי המתקבל בעבור מודל זה:

$$\hat{R}_D(h) = \frac{1}{N} \sum_{i=0}^N (h(\mathbf{x}^{(i)}) - y^{(i)})^2 = 32.67 \text{min}^2$$

בחירת מאפיינים - נסיון שני

ננסה להוסיף למודל עוד מאפיינים. ננסה להוסיף למודל פולינום מסדר שני של נקודת האיסוף תחת ההנחה שיש איזורים עמוסים יותר ואיזורים עמוסים פחות בעיר. המודל שלנו כעת יהיה:

$$\begin{aligned} h(\mathbf{x}; \theta) = & \theta_1 \sqrt{(x_{\text{pick east}} - x_{\text{drop east}})^2 + (x_{\text{pick north}} - x_{\text{drop north}})^2} \\ & + \theta_2 + \theta_3 x_{\text{pick east}} + \theta_4 x_{\text{pick north}} \\ & + \theta_5 x_{\text{pick east}} x_{\text{pick north}} + \theta_6 x_{\text{pick east}}^2 + \theta_7 x_{\text{pick north}}^2 \end{aligned}$$

מכיוון שמדובר בכמות גדולה של מאפיינים כבר לא ניתן להציג את הנתונים ופונקציית החיזוי בגרף.

חישוב של הפרמטרים והסיכון האמפירי בשיטה דומה לקודם נותנת סיכון אמפירי של:

$$\hat{R}_D(h) = \frac{1}{N} \sum_{i=0}^N (h(\mathbf{x}^{(i)}) - y^{(i)})^2 = 26.42 \text{min}^2$$

עוד מאפיינים

ניתן להמשיך באופן דומה ולהוסיף עוד ועוד מאפיינים. לרוב, כל הוספה של מאפיין תקטין את השגיאה האפירית. בתרגול ובהרצאה הקרובים נעסוק בבעיות של הוספת כמות גדולה מידי של מאפיינים ונדבר על דרכים כיצד לעשות זאת באופן מבוקר.