

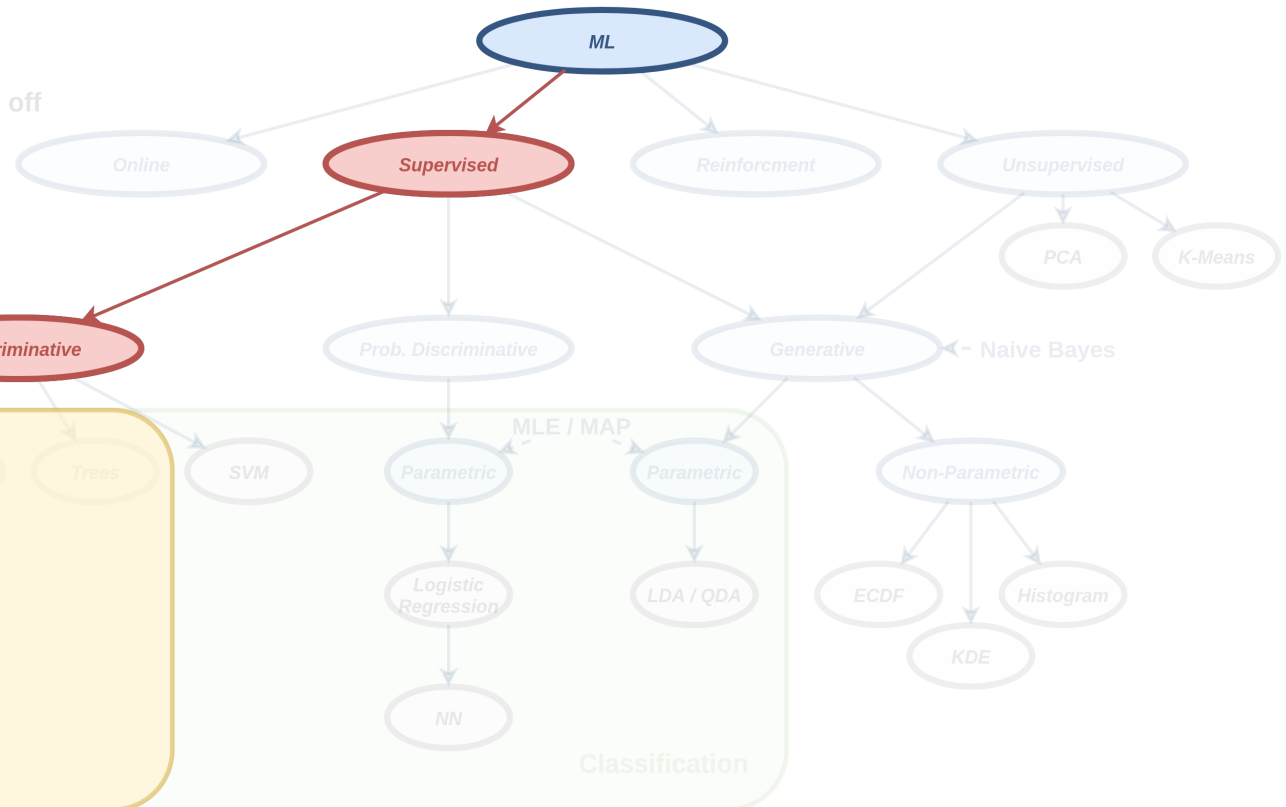
# הרצאה 2 - רגרסיה לינארית

PDF

## Subjects Covered in this Course

### General concepts:

- Cost / Risk / Loss functions
- Features
- Overfitting
- Approximation vs. Estimation trade off
- Bias vs. Variance trade off
- Cross validation
- Regularization
- Bagging
- Boosting
- Gradient Descent



# Supervised learning (למידה מונחית)

---

- בעיות supervised learning הם הבסיסיות ביותר בתחום והבנה טובה של בעיות אלו היא הבסיס להבנה של כל שאר הבעיות במערכות לומדות.
- בקורס זה אנו נעסוק בעיקר בבעיות מסוג זה.
- על מנת להבין מה הם בעיות supervised learning עלינו ראשית לחזור על הנושא של בעיות חיזוי.

- בבעיות חיזוי אנו מנסים לחזות את ערכו של משתנה אקראי לא ידוע, לרוב על סמך משתנים אקראיים ידועים.
- בעיות חיזוי הם **מאד** נפוצות ומופיעות במגוון רחב של תחומים בהנדסה ומדע.
- בהנדסת חשמל בעיות אלו מופיעות בתחומים כגון עיבוד אותות, תקשורת ספרתית ובקרה.
- בעיות חיזוי מלוות אותנו כמעט בכל פעולה יום יומית. לדוגמא האם לקחת מטריה כשיוצאים מהבית.
- ביום יום אנחנו לא מנסים לפתור את באופן מתמטי. אנו מחזיקים מודל של הקשרים הסטטיסטיים ומשתמשים בו בצורה איכותית.

# הקשר ל supervised learning

---

- בבעיות חיזוי קלאסיות, אנו מניחים שהפילוג ידוע.

- ב supervised learning (ובמערכות לומדות) אנו מניחים כי הפילוג אינו ידוע.

- במקום הפילוג יש לנו מדגם.

- את החזאי נאלץ כעת לבנות על סמך המדגם (במקום על סמך הפילוג).

- **Labels** (תויות / תגיות):  
 $y$  - המשתנה האקראי שאותו אנו מנסים לחזות. (לרוב סקלר)
- **Observations \ measurements** (תצפיות או מדידות):  
 $x$  - הוקטור האקראי אשר שעלפיו נרצה לבצע את החיזוי. (לרוב וקטור)
- $\hat{y}$  - תוצאת חיזוי.
- $\mathcal{H}$  - מרחב החזאים \ השערות
- $\hat{y} = h(x)$  - פונקציית החיזוי.  $h \in \mathcal{H}$ .
- $D$  אורך של הוקטור  $x$

# (המדגם) The dataset

---

המדגם יהיה מורכב מדגימות של הזוג  $x$  ו  $y$ :

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

כאשר  $N$  הוא מספר הדגימות שבמדגם.

## הנחת ה i.i.d.

אנו תמיד נניח כי הדגימות נוצרו כולם מאותו הפילוג באופן בלתי תלוי אחת בשניה.

זאת אומרת שבעבור זוג אינדקסים  $i \neq j$  הדגימה  $\{x^{(i)}, y^{(i)}\}$  הינה בלתי תלויה סטטיסטית בדגימה  $\{x^{(j)}, y^{(j)}\}$ .

# Regression vs. Classification

---

מוקבל לחלק את הבעיות ב supervised learning לשני תתי תחומים:

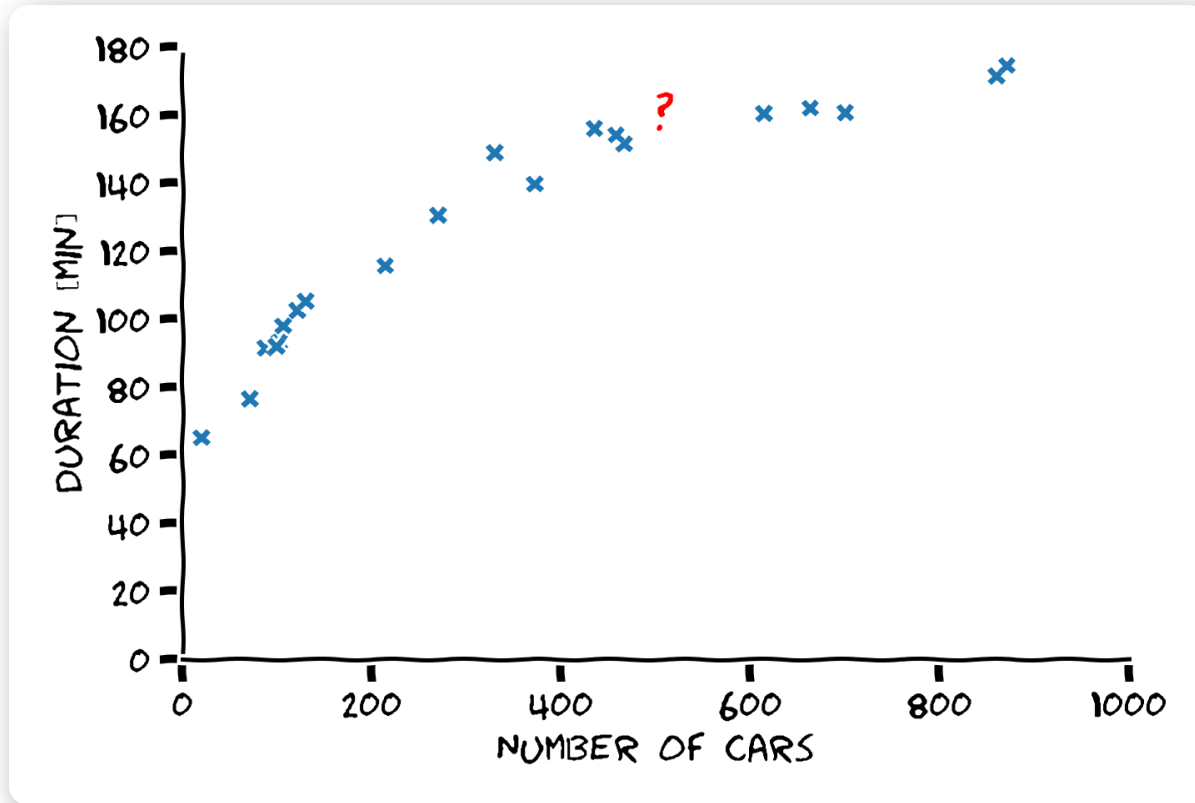
- בעיות regression (רגרסיה) -  $y$  רציף.

- בעיות classification (סיווג) -  $y$  בדיד עם סט ערכים סופי (לרוב קטן).



# דוגמא לבעיית רגרסיה

הבעיה של חיזוי משך הנסיעה



# ניסוח פורמלי - חיזוי משך הנסיעה

•  $y$  - Labels - המשתנה האקראי של זמן הנסיעה.

•  $x$  - Measurements - המשתנה האקראי של מספר המכוניות על הכביש.

•  $h$  - פונקציית החיזוי אשר מקבלת את מספר המכוניות על הכביש ומוציאה חיזוי של זמן הנסיעה.

•  $\mathcal{D}$  - מדגם הנתון של הזוגות של (מספר מכוניות, זמן נסיעה).

המטרה שלנו הינה להשתמש ב  $\mathcal{D}$  על מנת למצוא חזאי  $\hat{y} = h(x)$  אשר יהיה כמה שיותר מוצלח תחת קריטריון שאותו נצטרך להגדיר.

- כל פונקציה אשר ממפה מ  $x$  ל  $y$  היא פונקציית חיזוי חוקית.
- היינו מעוניינים למצוא חזאי אשר לעולם לא טועה.
- מכיוון ש  $y$  משתנה אקראי לא נוכל לחזותו במדוייק.
- אנו צריכים להגדיר דרך להשוות בין הטעויות שאותם מבצעים החזאים שונים. (לדוגמא, הרבה טעויות קטנות או מעט גדולות)

# השלבים הכלליים לפתרון הבעיה

---

1. נגדיר קריטריון מתמטי אשר מודד עד כמה מודל מסויים מצליח לבצע את המשימה.
2. נבחר משפחה רחבה של מודלים בתקווה שלפחות אחד מהם יהיה מוצלח מספיק.
3. נחפש מבין כל המודלים במשפחה את המודל המוצלח ביותר.

# The cost function (פונקציית המחיר)

---

- פונקציית המחיר  $C(h)$  מעניקה לכל חזאי ציון.

- ציון נמוך יותר = חזאי טוב יותר.

- החזאי האופטימאלי  $h^*$  הוא החזאי בעל הציון הנמוך ביותר:

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

- פונקציית המחיר אמורה לשקף את המחיר אותו "נשלם" על שימוש בחזאי כל שהוא.

- בפועל, משתמשים באחת מכמה פונקציות מחיר נפוצות.

# Risk and loss functions

## (פונקציות סיכון והפסד)

---

- פונקציית הcost נותנת ציון לחזאי.

- פונקציית הloss (הפסד)  $l$  נותנת ציון לחיזוי בודד.

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

- ניתן להגדיר את פונקציית הcost כתוחלת על פונקציית loss:

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

- במקרים כאלה, מוקבל להשתמש בשם risk ובסימון:

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

# פונקציות loss (risk) נפוצות

**Zero-one loss (misclassification rate):**

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

**נפוצה בבעיות classification.**

**$l_2$  loss (mean squared error (MSE)):**

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

**נפוצה בבעיות regression.  
בנוסף קיים גם (root mean squared error (RMSE).**

**$l_1$  loss (mean absolute error (MAE)):**

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

**גם כן נפוצה בבעיות regression.**

# בעיה: הפילוג של המשתנים לא ידוע

---

**-פונקציית הסיכון מוגדרת על ידי תוחלת על פני הפילוג של המשתנים האקראיים בבעיה שהוא כאמור לא ידוע.**

**• בעיה זו קיימת לא רק בפונקציות מחיר מסוג סיכון.**



# supervised learning

## בעיות לפתרון גישות

---

הגישה הגנרטיבית

מדגם



פילוג על סמך המדגם



חזאי אופטימאלי בהינתן הפילוג

הגישה הדיסקרימינטיבית

מדגם



חזאי בעל ביצועים טובים על המדגם

# שיערוך אמפירי של פונקציית המחיר / סיכון

---

- במקום לנסות ולחשב את המחיר באופן אנליטי, נוכל לנסות לשערך אותו על סמך אוסף של דוגמאות (מדגם).
- שיערוך על סמך דוגמאות מכונה שיערוך אמפירי.

# Empirical risk (סיכון אמפירי)

---

הסיכון האמפירי מוגדר על ידי החלפת התוחלת בפונקציית הסיכון בגרסא האמפירית שלה.

אנו נשתמש בסימון  $\hat{\mathbb{E}}_{\mathcal{D}}$  על מנת לסמן את תחולת האמפירית המבוססת על המדגם נתון  $\mathcal{D}$ .

$$\mathbb{E}[f(\mathbf{x})] \approx \hat{\mathbb{E}}_{\mathcal{D}}[f(\mathbf{x})] = \frac{1}{N} \sum_{i=1}^N f(\mathbf{x}^{(i)})$$

כאשר מספר הדגימות  $N$  הולך לאין סוף התוחלת האמפירית מתכנסת לתוחלת האמיתית במובן הסתברותי.

# Empirical risk (סיכון אמפירי) - המשך

---

הסיכון האמפירי המקבל הינו:

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

השימוש בגרסא האמפירית של פונקציית המחיר היא במקרים רבים בעייתית והיא גורמת בין היתר לתופעה המוכנה **overfitting** (התאמת יתר).

בשלב זה אנו נתעלם מבעיה זו ואנו נעסוק בה בהרחבה בהרצאה הבאה.

# (Empirical risk minimization (ERM

---

- רלוונטי למקרים בהם פונקציית המחיר מוגדרת כפונקציית סיכון.
- מנסה ישירות למצוא חזאי אשר ימזער את הסיכון האמפירי:

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

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

למה  $h_{\mathcal{D}}^*$  לא  $h^*$ ?

1. על מנת להדגיש את התלות של החזאי במדגם (לכל מדגם יהיה חזאי אופטימאלי אחר).

2. בכדי להבדיל את החזאי של ERM מהחזאי האופטימאלי.

נרצה להגביל את החזאי שלנו למשפחה מצומצמת של פונקציות. נעשה זאת על ידי שימוש במודל פרמטרי.

מוטיבציה:

1. הגבלת המודל למשפחה מצומצמת של מודלים מסייעת למזער את הoverfitting (בהרצאה הבאה).
2. יותר פרקטי לנסות לחפש פרמטרים של מודל מאשר חיפוש כללי של פונקציה במרחב הפונקציות.



# מודלים פרמטריים - המשך

---

מודל פרמטרי מגדיר את המבנה הכללי של הפונקציות במשפחה עד כדי מספר סופי של פרמטרים אשר חופשיים להשתנות.

לדוגמא:

$$h(\mathbf{x}; \boldsymbol{\theta}) = \frac{\theta_1^3 x_1 + x_4^{\theta_2}}{\log(\theta_3 x_2)}$$

# מודלים פרמטריים - דוגמאות

---

## 1. פונקציות לינאריות:

$$h(\mathbf{x}; \boldsymbol{\theta}) = \theta_1 x_1 + \theta_2 x_2 + \theta_2 x_2$$

## 2. פולינומים:

$$h(\mathbf{x}; \boldsymbol{\theta}) = \theta_1 + \theta_2 x_1 + \theta_3 x_1^2 + \theta_4 x_1^3$$

## 3. טור פוריה סופי:

$$h(x; \boldsymbol{\theta}) = \theta_1 \sin(\pi x) + \theta_2 \cos(\pi x) + \theta_3 \sin(2\pi x) + \theta_4 \cos(2\pi x)$$

## 4. רשתות נוירונים

# מודלים פרמטריים - מיפוי לוקטורים

- מודל פרמטרי ממפה כל פונקציה מהמשפחה הפרמטרית לוקטור.
  - יש לנו סט עשיר של כלים לעבודה עם וקטורים.
  - לדוגמא: שימוש ב **gradient descent** למציאת מינימום לוקאלי.
- ניתן כעת לרשום את בעיית האופטימיזציה כאופטימיזציה על הפרמטרים (במקום על  $h$ ):

$$\theta^* = \arg \min_{\theta} C(h(\cdot; \theta))$$

או במקרה של ERM:

$$\theta_D^* = \arg \min_{\theta} \frac{1}{N} \sum_{i=1}^N [l(h(\mathbf{x}_i; \theta), y_i)]$$

**מודל מהצורה:**

$$h(\mathbf{x}; \boldsymbol{\theta}) = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_D x_D$$

**או בצורה וקטורית:**

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

# איבר היסט (bias)

ניתן להוסיף למודל גם איבר bias:

$$h(\mathbf{x}; \boldsymbol{\theta}) = \theta_1 + \mathbf{x}^\top [\theta_2, \theta_3, \dots, \theta_{D+1}]^\top$$

בכדי לשמור על הכתיב הוקטורי נפריד את איבר ה bias משאר הפרמטרים:

$$h(\mathbf{x}; \boldsymbol{\theta}, \theta_0) = \theta_0 + \mathbf{x}^\top \boldsymbol{\theta}$$

לרוב נסמן אותו בעזרת  $b$  או  $\theta_0$ .

נראה בהמשך דרך נוחה יותר להוסיפת איבר ההיסט בעזרת שינוי של הוקטור  $\mathbf{x}$ .

# (Linear Least Squares (LLS

---

**:ERM + מודל לינארי + MSE**

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

**בעיית הLLS נפוצה מאד ומופיעה בתחומים רבים.**

**אחת התכונות הנחמדות ביותר של LLS הוא העובדה שניתן לפתור את הבעיית האופטימיזציה שלו באופן אנליטי.**

נגדיר את הוקטור והמטריצה הבאים:

• וקטור התגיות:

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

• מטריצת המדידות:

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

# Linear least squares - כתיב מטריצי

$$\mathbf{y} = [y^{(1)}, y^{(2)}, \dots, y^{(N)}]^\top \quad \mathbf{X} = \begin{bmatrix} - & \mathbf{x}^{(1)} & - \\ - & \mathbf{x}^{(2)} & - \\ & \vdots & \\ - & \mathbf{x}^{(N)} & - \end{bmatrix}$$

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

$$\begin{aligned} \boldsymbol{\theta}_D^* &= \arg \min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=0}^N (\mathbf{x}^{(i)\top} \boldsymbol{\theta} - y^{(i)})^2 \\ &= \arg \min_{\boldsymbol{\theta}} \frac{1}{N} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 \end{aligned}$$



# Linear least squares - פתרון סגור

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

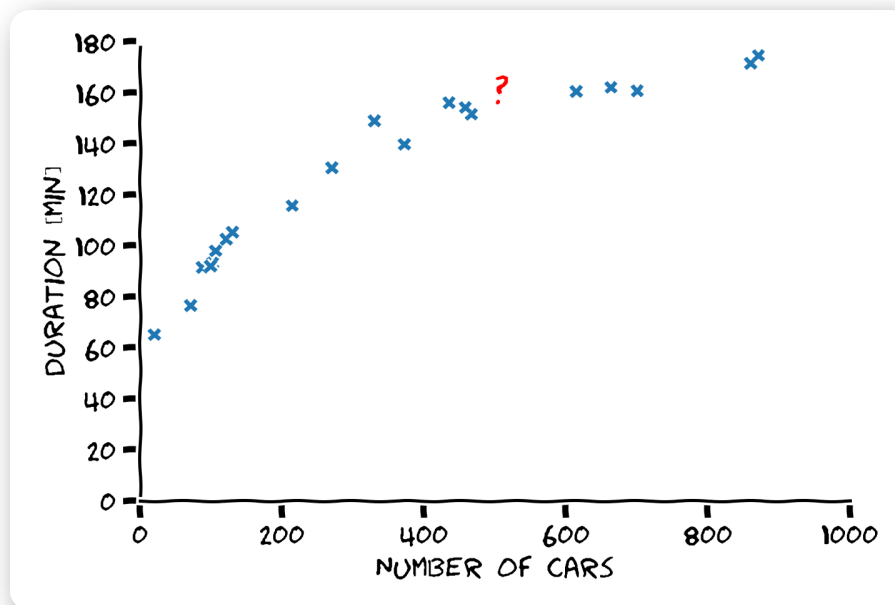
בבעיית האופטימיזציה הזו ניתן להגיע לפתרון סגור על ידי גזירה והשוואה ל-0:

$$\nabla_{\theta} \left( \frac{1}{N} \|X\theta - \mathbf{y}\|_2^2 \right) = 0$$

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

(את הפיתוח תראו בתרגול 3)

פתרון זה נכון רק כאשר המטריצה  $X^T X$  הפיכה. (בשבוע הבא נדבר על הנושא של רגולריזציה אשר יכול לעזור, בין היתר, במקרים שבהם המטריצה לא הפיכה).



נשתמש במודל:

$$h(x; \theta) = \theta x$$

ונפתור בעזרת LLS.

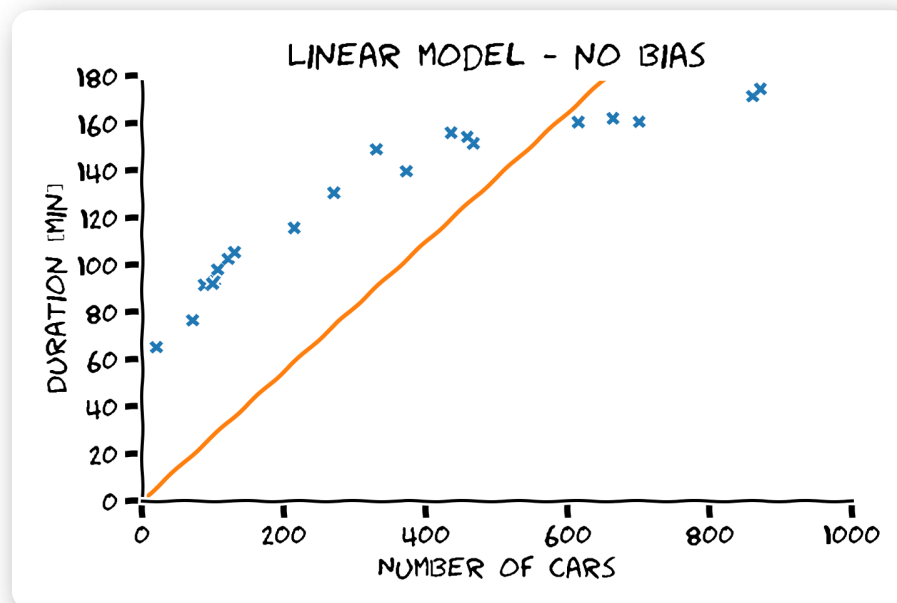
$$h(x; \theta) = \theta x$$

בעבור מקרה זה, נקבל ש:  $X = [x^{(1)}, x^{(2)}, \dots, x^{(N)}]^\top$

נחשב את  $\theta$  על ידי:

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

התוצאה המקבלת הינה:



**נרצה להשתמש במודל מהצורה:**

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

**בעיה: הפתרון הסגור של LLS לא מתייחס למודל זה (עם bias).**

**פתרון: ננסח מחדש את הבעיה כך שיתקבל מודל ללא איבר היסט.**

- אנו לא חייבים להשתמש בנתונים בצורתם הגולמית.

- מותר לנו לבצע עיבוד מקדים של הנתונים לפני שאנו מזינים אותם לחזאי.

- העיבוד המקדים יכול לפעול על כל וקטור המדידות  $\mathbf{x}$ , ולייצר וקטור חדש:

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

פעולת החיזוי תהיה:

$$\hat{y} = h(\Phi(\mathbf{x}); \theta)$$

את המוצא של הפונקציה  $\Phi$  מקובל לכנות וקטור המאפיינים **(features)**. השימוש במאפיינים מאפשר דברים כגון:

- הרחבת מודלים פשוטים למודלים מורכבים יותר.

- החלפת האופן שבו המידע מיוצג. לדוגמא:

  - החלפת יחידות.

  - הפיכת תמונת פנים לוקטור של מאפיינים כגון: המרחק בין העיניים, גוון העור, עד כמה הפנים אליפטיות וכו'

  - ניקוי רעשים בהקלטות audio.

- הפחתת overfitting (נראה לקראת סוף הקורס כשנדבר על הורדת מימד).

אנו נשתמש לפעמים בסימון הבא:

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

כאן  $\Phi$  הוא וקטור של פונקציות, כאשר כל פונקציה  $\varphi_i$  אחראית על ייצור של איבר אחד בוקטור  $\mathbf{x}_{\text{new}}$ :

$$x_{\text{new},i} = \varphi_i(\mathbf{x})$$

# מודלים לינאריים ומאפיינים

על ידי שילוב של מודל לינארי עם מאפיינים נוכל לקבל חזאים מהצורה:

$$\begin{aligned}\hat{y} &= h(\mathbf{x}; \theta) = h_{\text{linear}}(\Phi(\mathbf{x}); \theta) = \Phi(\mathbf{x})^\top \theta \\ &= \theta_1 \varphi_1(\mathbf{x}) + \theta_2 \varphi_2(\mathbf{x}) + \dots + \theta_M \varphi_M(\mathbf{x})\end{aligned}$$

זאת אומרת מודל שהוא קומבינציה לינארית של פונקציות של  $\mathbf{x}$ .

שימו לב: המודל נקרא מודל לינארי משום שהוא לינארי בפרמטרים שהם הנעלמים בבעיה (ולא ב  $\mathbf{x}$ )



# דוגמא: הוספה של איבר ההיסט

נוסיף כעת איבר היסט למודל שלנו לשיערוך זמן הנסיעה.  
נעשה זאת על ידי שימוש במאפיינים הבאים:

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

כל דגימה  $x$  תהפוך לוקטור  $x_{\text{new}} = [1, x]^T$  ומודל החיזוי שלנו יהיה:

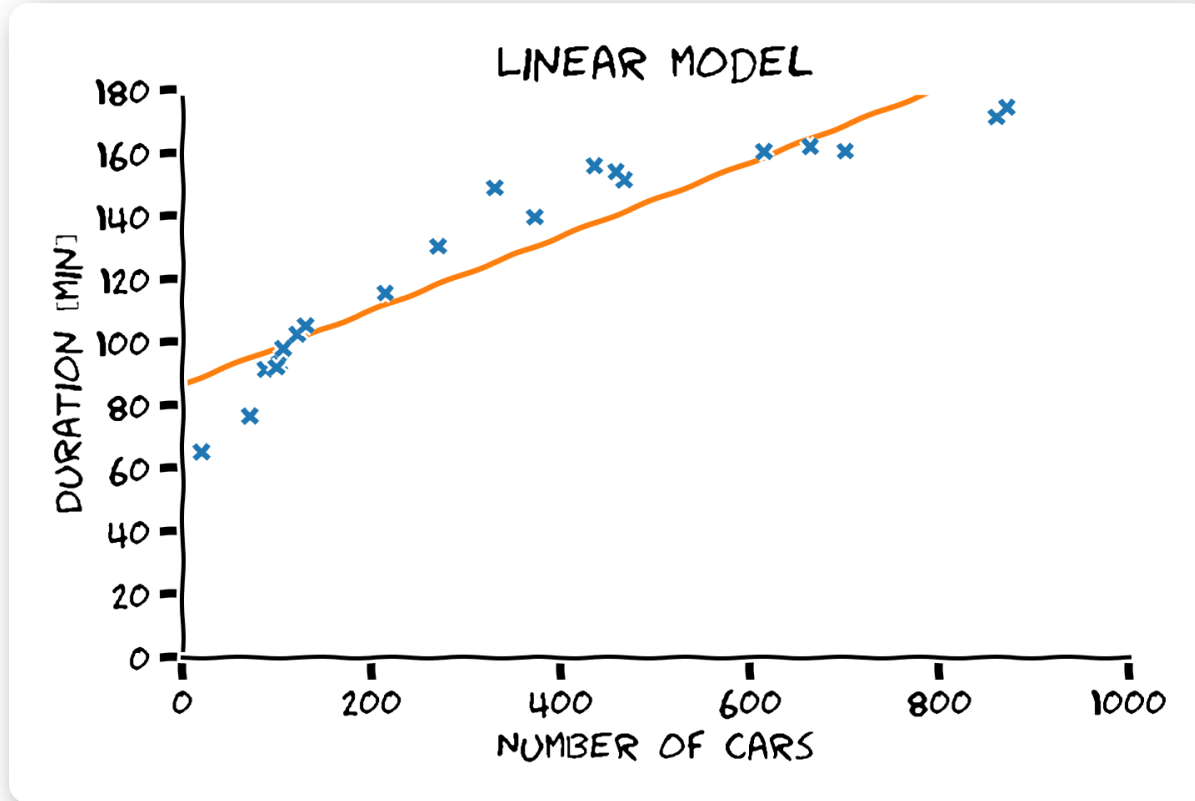
$$h(x; \theta) = \theta_1 + \theta_2 x$$

מטריצת המדידות  $X$  תהיה כעת:

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

# דוגמא: הוספה של איבר ההיסט

הצבה של מטריצה זו בנוסחא ל  $\theta_D^*$  נותנת את המודל הליניארי הבא:



# דוגמא נוספת - פולינומים

באותו אופן ניתן להשתמש במאפיינים בכדי לייצג מגוון רחב של פונקציות מורכבות יותר כגון פולינומים, כפי שיודגם בתרגול 3.

