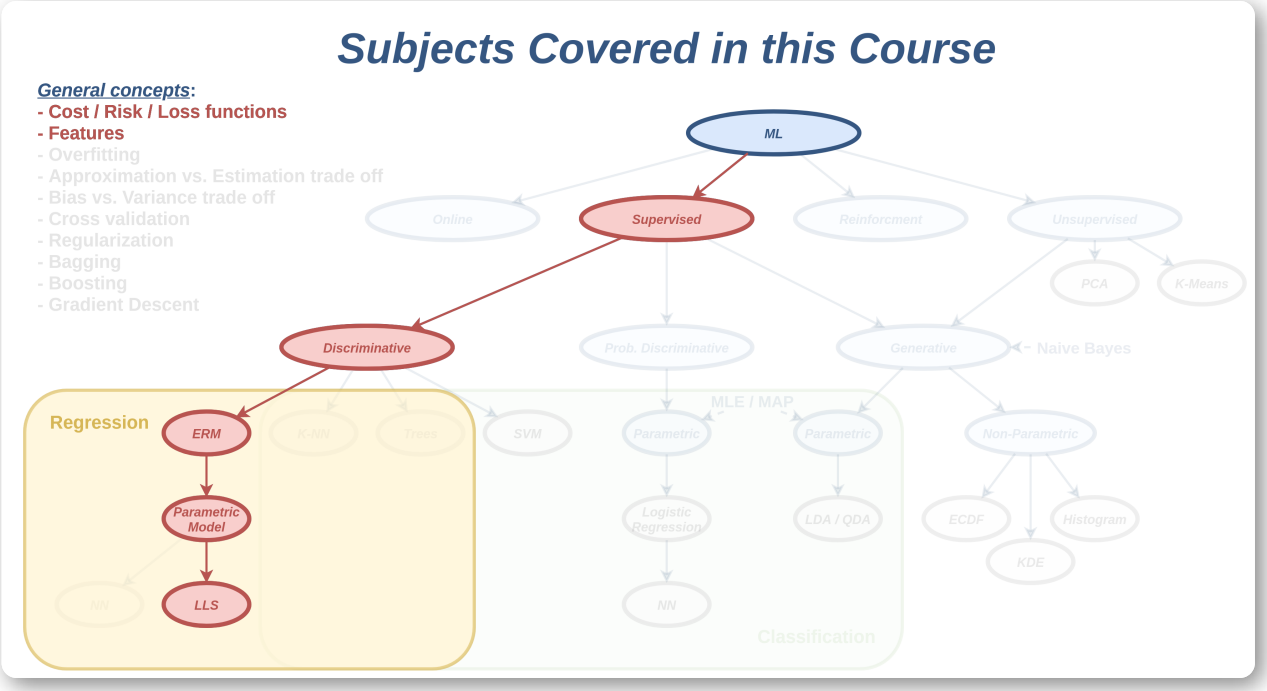
הרצאה 2 - רגרסיה לינארית

Slides PDF Code

מה נלמד היום



Supervised learing (למידה מונחית)

בהרצאה הקודמת הכרנו באופן כללי את סוגי הבעיות בהם עוסק התחום של מערכות לומדות והתחלנו לדבר על בעיות מסוג supervised learning (למידה מונחית). בעיות אלו הם הבסיסיות ביותר בתחום והבנה טובה של בעיות אלו היא הבסיס להבנה של כל שאר הבעיות במערכות לומדות. אנו נעסוק בבעיות מסוג זה לאורך רוב רובו של הסימסטר.

על מנת להבין ולהגדיר מה הם בעיות supervised learning עלינו ראשית לחזור על הנושא של בעיות חיזוי.

בעיית החיזוי

בתרגול 2 מופיעה תזכורת מורחבת יותר של התחום, אך לצורך הרצאה זו נסתפק בתיאור הקצר המופיע פה. בבעיית החיזוי 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. יתרה מזאת, בעיות חיזוי מלוות אותנו כמעט בכל פעולה יום יומית, לדוגמא:

> כאשר אנחנו מתלבטים האם לקחת איתנו מטריה ביציאה מהבית, אנו למעשה מנסים לבצע חיזוי של האם ירד גשם או לא על סמך פרטי מידע שיש בידינו, כגון התחזית ששמענו, כמות העננים בשמים, צבע העננים וכו'.

ביום יום אנחנו אולי לא מנסים לפתור את בעיות החיזוי באופן מתמטי, אך אנו כן מחזיקים בראש איזה שהוא מודל של הקשרים הסטטיסטים בין המשתנים השונים, ואנו מנסים לבצע את החיזוי על סמך אותו מודל באופן איכותי.

הקשר ל supervised learning

בבעיות חיזוי קלאסיות, אנו מניחים שהפילוג של כל המשתנים האקראיים ידוע, וכי האתגר הוא מציאת החזאי האופטימאלי על סמך הפילוג. לעומת זאת, בבעיות supervised learning (ובבעיות במערכות לומדות באופן כללי) אנו מניחים כי הפילוג אינו ידוע ובמקומו נתון לנו מדגם של דגימות מתוך אותו פילוג. את החזאי נאלץ כעת לבנות על סמך המדגם במקום על סמך הפילוג. במהלך הקורס אנו נדון נכיר שיטות שונות לבנות חזאים באופן זה ונדון בבעיות הקיימות בשיטות אלו.

סימונים ושמות

בקורס זה אנו נשתמש בסימונים והשמות הבאים:

- **Labels** (תויות / תגיות): y (או במקרה הוקטורי $\mathbf{y}$) - יהיה המשתנה / הוקטור האקראי שאותו אנו מנסים לחזות. בקורס זה ה y labels, יהיו כמעט תמיד סקלריים.
- **Observations \ measurements** (תצפיות או מדידות): x (או במקרה הוקטורי $\mathbf{x}$) - יהיה הוקטור האקראי אשר מכיל את המשתנים שלפיהם נרצה לבצע את החיזוי. במקרים מסויימים החיזוי יהיה על פי משתנה יחיד ואז x יהיה סקלר.
- $\hat{y}$ - תוצאת חיזוי כל שהיא.
- מרחב החזאים \ השערות $\mathcal{H}$. במרחב זה נמצאים כל החזאים האפשריים.
- $\hat{y} = h(\boldsymbol{x})$ - פונקציית החיזוי. נשים לב כי $h \in \mathcal{H}$.
- אנו נשתדל להשתמש ב D לסימון האורך של הוקטור $\boldsymbol{x}$

The dataset (המדגם)

כפי שציינו, את הבניה של החזאי אנו נעשה על פי מדגם מתוך הפילוג הלא ידוע. לרוב המדגם יהיה מורכב מזוגות של $\boldsymbol{x}$ ו y אשר יוצרו מתוך N דגימות **בלתי תלויות**:

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

נשתדל להשתמש תמיד ב N לסימון מספר הדגימות שבמדגם.

הנחת ה i.i.d

במערכות לומדות אנו תמיד נניח כי הדגימות במדגם נוצרו כולם מאותו הפילוג באופן בלתי תלוי אחת בשניה. זאת אומרת שזוג המשתנים $\{x^{(i)}, \mathrm{y}^{(i)}\}$ הינו בלתי תלוי סטטיסטית בזוג המשתנים $\{x^{(j)}, \mathrm{y}^{(j)}\}$ כאשר $i \neq j$.

מהו החזאי האופטימאלי

באופן כללי, כל פונקציה אשר ממפה מהמרחב של $\boldsymbol{x}$ למרחב של y היא פונקציית חיזוי חוקית. נשאלת אם כן השאלה מהי פונקציית החיזוי המוצלחת ביותר? באופן כללי היינו מעוניינים למצוא חזאי אשר לעולם לא טועה.בפועל, מכיוון שy הינו משתנה אקראי לא נוכל אף פעם לחזותו במדוייק (מלבד במקרים מיוחדים בהם y נקבע באופן חד ערכי על ידי $\mathbf{x}$).

מיכוון שבעבור כל חיזוי שנבחר אנו מצפים לשגיאה כל שהיא, אנו צריכים להגדיר דרך להשוות בין הטעויות שאותם מבצעים החזאים שונים. אנו צריכים להחליט לדוגמא איך נבחר בין חזאי שעושה כל הזמן שגיאות בינוניות לבין חזאי אשר רוב הזמן עושה שגיאות ממש קטנות אך פעם בכמה זמן עושה שגיאה מאד גדולה. הנושא הראשון שנעסוק בו בהרצאה יהיה הדרך שבה נרצה להשוות בין הביצועים של חזאים שונים.

Regression vs. Classification

מוקבל לחלק את הבעיות ב supervised learning לשני תתי תחומים:

- **בעיות regression (רגרסיה)** - בעיות בהם y הוא משתנה רציף.
- **בעיות classification (סיווג)** - בעיות בהם y הוא משתנה בדיד אשר יכול לקבל ערכים מתוך סט ערכים סופי (ולרוב קטן).

דוגמאות:

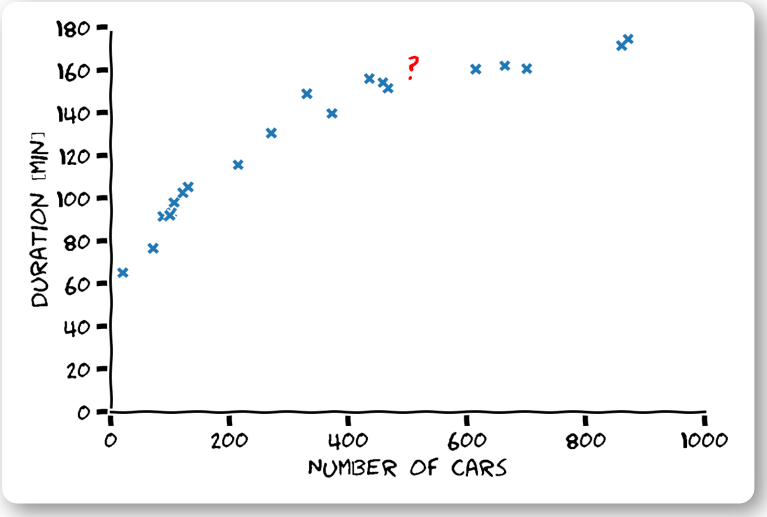
- **רגרסיה**: חיזוי זמן נסיעה בכביש החוף, חיזוי מרחקים לאובייקטים בתמונה, חיזוי מחירים של דירות וכו'.
- **סיווג**: חיזוי של המחלה בה חולה אדם מסויים על פי הסימפטומים שלו, חיזוי של האם דואר מסויים הוא spam או לא, חיזוי של האם עסקת אשראי מסויימים היא לגיטימית או הונאה וכו'/

(בעיקרון יכולים להיות גם בעיות בהם y בדיד ולא סופי. בבעיות מסוג זה לרוב פשוט מניחים שy רציף והופכים את הבעיה לבעיית רגרסיה ולבסוף מעגלים את התוצאה).

כפי שנראה בהמשך הקורס, אבחנה זו חשובה מכיוון שהאופי של y ישפיע על הדרך שבה ננסה לפתור את הבעיה.

בעיית רגרסיה לדוגמא

נסתכל על בעיית החיזוי של זמן הנסיעה בכביש החוף על סמך מספר המכוניות בכביש:



ננסח את הבעיה בצורה פורמאלית:

- y - Labels - המשתנה האקראי של זמן הנסיעה.
- x - Meassurments - 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 d על מנת למצוא חזאי $\hat{y} = h(x)$ אשר יהיה כמה שיותר מוצלח תחת קריטריון שאותו נצטרך להגדיר.

השלבים הכלליים לפתרון הבעיה

הרעיון מאחורי כמעט כל השיטות במערכות לומדות הוא זהה:

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)

נתחיל מהשלב הראשון של הגדרת הקריטריון של שלפיו נרצה לבחור את החזאי שלנו. הדרך המקובלת לעשות זאת הינה על ידי הגדרת פונקציית מחיר.

פונקציית המחיר $C(h)$ היא פונקציה אשר מעניקה לכל חזאי ציון. לרוב נהוג להגדיר את הפונקציה כך שככל שהציון נמוך יותר החזאי טוב יותר. אנו נגדיר את החזאי האופטימאלי כחזאי בעל הציון הנמוך ביותר מבין כל החזאים האפשריים. נסמן את החזאי האופטימאלי ב h^*:

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

ישנם דרכים רבות להגדיר את פונקציית המחיר, ואין דרך "נכונה" לעשות זאת. באופן כללי, הבחירה של פונקציית המחיר צריכה להתאים לבעיה שאותה רוצים לפתור בעזרת החזאי. באופן כללי פונקציית המחיר אמורה לשקף את המחיר אותו "נשלם" על שימוש בחזאי נתון כל שהוא, בפועל, במרבית המקרים משתמשים באחת מכמה פונקציות מחיר נפוצות.

נציג את אחת הדרכים הפופולריות להגדיר פונקציית מחיר, אשר עושה זאת על ידי שימוש בפונקציה הנקראת פונקציית הפסד (loss function).

Risk and loss functions (פונקציות סיכון והפסד)

פונקציית הסיכון הינה מקרה פרטי של פונקציית המחיר והיא מוגדרת באופן הבא:

בעוד שפונקציית המחיר מנסה לתת ציון ליכולת החיזוי הכללית של החזאי, פונקציית ההפסד l נותנת ציון לחיזוי בודד מסויים. שבהינתן חזאי h ודגימה אקראית עם $\boldsymbol{x}$ ו y מסויימים, ההפסד המשוייך לדגימה זו יהיה:

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

את פונקציית המחיר הכוללת ניתן כעת להגדיר כתוחלת של פונקציית ה loss על פני הפילוג של $\mathbf{x}$ ו y

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

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

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

פונקציות loss נפוצות

- פונקציית loss נפוצה לבעיות classification היא פונקציית ה **Zero-One loss** אשר מוגדרת באופן הבא:

  $$l(\hat{y}, y) = I\{\hat{y} \neq y\}$$

  לפונקציית ה risk אשר משתמשת ב loss הזה קוראים: **misclassification rate**.

- פונקציית loss נפוצה לבעיות regression היא פונקציית ה **l_2 loss** אשר מוגדרת באופן הבא:

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

  לפונקציית ה risk אשר משתמשת ב loss הזה קוראים: **mean squared errors (MSE)**.

- במקרים רבים נהוג להשתמש דווקא בשורש של ה MSE בכדי שפונקציית המחיר תחזיר ערכים באותם יחידות כמו y. במקרה זה קוראים לפונקציית המחיר **root mean squared errors (RMSE)**. התוספת של השורש לא משנה את בעיית האופטימיזציה (משום שהיא פונקציה מונוטונית עולה) ולכן הוא לא משפיע על החזאי המתקבל.

- פונקציית loss נפוצה נוספת לבעיות regression היא פונקציית ה **l_1 loss** אשר מוגדרת באופן הבא:

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

  לפונקציית ה risk אשר משתמשת ב loss הזה קוראים: **mean absolute errors (MAE)**.

בעיה: הפילוג של המשתנים האקראיים לא ידוע

הבעיה עם האופן שבו הגדרנו את פונקציית הסיכון הינה העובדה שהיא מוגדרת על ידי תוחלת על פני הפילוג של המשתנים האקראיים בבעיה שהוא כאמור לא ידוע. בעיה זו קיימת לא רק בפונקציות מחיר מסוג סיכון אלא גם בסוגים שונים של פונקציות מחיר אשר כמעט תמיד תלויות בפילוג של המשתנים האקראיים בבעיה.

שיערוך אמפירי של פונקציית המחיר / סיכון

בכדי לנסות ולהתמודד עם בעיה זו נוכל במקום לנסות ולחשב את ערכה של פונקציית המחיר באופן אנליטי, לנסות ולשערך את ערכה של פונקציית הסיכון על סמך אוסף של דוגמאות מתוך הפילוג (מדגם). שיעוך 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 $\mathcal{D}$.

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

ניתן להראות כי כאשר מספר הדגימות N הולך לאין סוף התוחלת האמפירית מתכנסת לתוחלת האמיתית במובן הסתברותי.

הסיכון האמפירי המקבל מהחלפה זו הינו:

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

התאמת יתר

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.

גישות לפתרון בעיות supervised learning

לפני שנעבור לשלב הבניה של החזאי אנו נציג שתי גישות שונות לבהם ניתן לגשת לבעיה. באופן כללי ניתן לחלק את השיטות לפתרון בעיות supervised learning לשתי הגישות הבאות:

גישה גנרטיבית (generative) - בגישה זו אנו ננסה להשתמש במדגם על מנת לנסות וללמוד את הפילוג הלא ידוע. שיטה זו נקראת גנרטיבית משום שהיא לומדת את הפילוג שמתוכו נוצרו (generated) הדגימות.

גישה דיסקרימינטיבית (discriminative) - בשיטה זו ננסה לבנות חזאי אופטימאלי על סמך הגרסא האמפירית של פונקציית המחיר בתקווה שהוא יקבל ציון טוב גם בגרסא הלא אמפירית של פונקציית המחיר (זאת אומרת, שהוא ידע להכליל בצורה טובה).

לכל אחד מהגישות יש את היתרונות והחסרונות שלה. במהלך הקורס נכיר אלגוריתמים משני הגישות ונעמוד על ההבדלים בניהם.

(Empirical risk minimization (ERM

בעבור מקרים בהם פונקציית המחיר מוגדרת כפונקציית סיכון, הגישה הדיסקרימינטיבית הבסיסית ביותר הינה לנסות לחפש באופן ישיר חזאי אשר ממזער את הסיכון האמפירי. הגישה הזו מכונה ERM (Empirical Rik Minimization) והיא מוגדרת על ידי בעיית האופטימיזציה הבאה:

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

שימו לב כי הוספנו את הכיתוב $\mathcal{D}$ מתחת לחזאי האופטימאלי של בעיית האופטימיזציה זו. עשינו זאת משתי סיבות:

1. על מנת להדגיש את התלות של החזאי במדגם (לכל מדגם יהיה חזאי אופטימאלי אחר).
2. בכדי להבדיל את החזאי המתקבל משיטת ה ERM מהחזאי האופטימאלי של הבעיה המקורית אשר באופן עקרוני יהיה שונה.

מודלים פרמטריים

לרוב אנו נרצה להגביל את החזאי שלנו למשפחה מצומצמת של פונקציות. לרוב אנו נרצה לעשות זאת על ידי בחירה של משפחה של פונקציות אשר מוגדרות על ידי מודל פרמטרי. ישנם שני סיבות עיקריות לכך:

1. כפי שנראה בהרצאה הבאה הגבלה זו חשובה בכדי לשפר את יכול ההכללה של החזאי ולסייע בהקטנת בעיית overfitting.
2. יותר פרקטי לנסות לחפש פרמטרים של מודל מאשר חיפוש כללי של פונקציה במרחב הפונקציות.

מודל פרמטרי מגדיר את המבנה הכללי של הפונקציות במשפחה עד כדי מספר סופי של פרמטרים אשר חופשיים להשתנות. את הפרמטרים של המודל נסמן בעזרת הוקטור $\boldsymbol{\theta}$. אנו נשתמש ב $h(\boldsymbol{x};\boldsymbol{\theta})$ בכדי לתאר חזאי מהמשפחה הפרמטרית עם פרמטרים $\boldsymbol{\theta}$. אין כל מגבלה על הצורה הכללית שלהמודל הפרמטרי, המודל הפרמטרי יכול להיות לדוגמא:

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

דוגמאות נוספות למודלים פרמטריים:

1. פונקציות לינאריות: $h(\boldsymbol{x};\boldsymbol{\theta}) = \theta_1 x_1 + \theta_2 x_2 + \theta_2 x_2$
2. פולינומים: $h(\boldsymbol{x};\boldsymbol{\theta}) = \theta_1 + \theta_2 x_1 + \theta_3 x_1^2 + \theta_4 x_1^3$
3. טור פוריה סופי: $h(\boldsymbol{x};\boldsymbol{\theta}) = \theta_1 \sin(\pi x) + \theta_2 \cos(\pi x) + \theta_3 \sin(2\pi x) + \theta_4 \cos(2\pi x)$
4. רשתות נוירונים

מודל פרמטרי למעשה ממפה כל פונקציה מהמשפחה הפרמטרית לוקטור. היתרון בעבודה עם וקטורים הינו שיש לנו סט עשיר של כלים בהם אנו יכולים להשתמש. לדוגמא, מכיוון שניתן לגזור לפי וקטורים, נוכל להשתמש ב gradient decent על מנת לחפש את המודל האופטימאלי במרחב המודלים. מיכוון שכל וקטור כעת מגדיר מודל מסויים (ולהיפך) ניתן לרשום את בעיית האופטימיזציה של מציאת המודל האופטימאלי כבעיית אופטימיזציה על וקטור הפרמטרים (במקום על h):

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

או במקרה של ERM:

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

מודל לינארי

המודל הפרמטרי הפשוט ביותר הינו המודל הלינארי. המודל הלינארי הוא בעל המבנה הבא::

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

דרך נוחה יותר לכתוב את המודל הזה היא בצורה וקטורית:

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

איבר היסט (bias)

ניתן להוסיף למודל גם איבר bias על מנת לקבל מודל מהצורה הבאה:

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

לשם הנוחות בכדי לשמור על הכתיב הוקטורי של המודל נפריד לרוב את איבר ה bias משאר הפרמטרים. לרוב נסמן אותו בעזרת b או θ_0:

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

אנו נראה מיד דרך נוחה יותר להוספת איבר ההיסט בעזרת שינוי של הוקטור $\boldsymbol{x}$ כך שהביטוי $\boldsymbol{x}^\top \boldsymbol{\theta}$ יכיל גם את איבר ההיסט.

(Linear Least Squares (LLS

מקרה מיוחד של בעיית ERM עם מודל לינארי, הוא המקרה שבו משתמשים בפונקציית MSE (פונקציית risk עם loss ריבועי (l_2)):

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

השימוש במודלים לינאריים וב MSE נפוץ מאד ולכן בעיית ה LLS מופיעה בתחומים רבים. בעיית האופטימיזציה המקבלת בעבור LLS הינה:

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

כתיב מטריצי

את בעיה זו ניתן לרשום גם בצורה קומפקטית על ידי הגדרת הוקטור והמטריצה הבאים:

- נדגיר את וקטור התגיות $\boldsymbol{y}$ כוקטור של כל התגיות במדגם:

$$\boldsymbol{y} = [y^{(1)}, y^{(2)}, \cdot, y^{(n)}]^\top$$

- נגדיר את המטריצת המדידות X כמטריצה של כל ה $\boldsymbol{x}$-ים במדגם:

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

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

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

(ניתן להראות זאת על ידי רישום הורמה כסכום ושימוש בעובדה ש $X\boldsymbol{\theta} = [\boldsymbol{x}^{(1)\top}\boldsymbol{\theta}, \boldsymbol{x}^{(2)\top}\boldsymbol{\theta}, \ldots, \boldsymbol{x}^{(N)\top}\boldsymbol{\theta}]^\top$)

פתרון סגור

מה שמיוחד בבעיית האופטימיזציה של LLS הינה העובדה שניתן להגיע לפתרון סגור לפרמטרים האופטימאליים על ידי גזירה והשוואה ל-0. הפתרון המתקבל הינו: (את החישוב עצמו אתם תראו בתרגול 3)

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

$$\Rightarrow \boldsymbol{\theta} = (X^\top X)^{-1} X^\top \boldsymbol{y}$$

פתרון זה נכון רק כאשר המטריצה $X^\top X$ הפיכה. בתרגול אנו נדון במשמעות של תנאי זה.

דוגמא

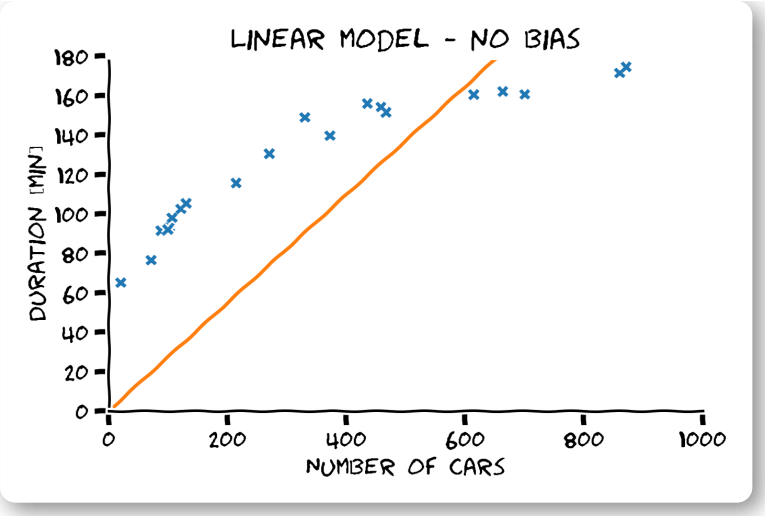
נשתמש במודל לינארי ובפתרון של בעיית ה LLS בכדי למצוא חזאי לבעיית שיערוך זמן הנסיעה. נתחיל במודל ללא היסט:

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

את הפרמטר θ האופטימאלי נוכל לחשב על ידי הצבה של הנקודות במדגם לתוך:

$$\theta^*_{\mathcal{D}} = (X^\top X)^{-1} X^\top \boldsymbol{y}$$

כאשר $X = [x^{(1)}, x^{(2)}, \ldots, x^{(N)}]^\top$. התוצאה המקבלת הינה



הוספת איבר היסט

בכדי להוסיף איבר היסט, עלינו להשתמש במודל מהצורה של:

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

הבעיה אם צורה זו הינה, שהפתרון הסגור שמצאו מתייחס למקרה שבו אין איבר היסט, לשם ככך ננסה לנסח מחדש את הבעיה כך שיתקבל מודל ללא איבר היסט.

עיבוד מקדים

נשים לב לעובדה שבבואנו לבצע משימת חיזוי אנו לא חייבים להשתמש בנתונים בצורת הגולמית ומותר לנו לבצע עיבוד מקדים של הנתונים לפני שאנו מזינים אותם לחזאי. נניח לדוגמא שאנו מעריכים שיהיה נוח יותר לבצע את החיזוי של זמן הנסיעה על פי הריבוע של כמות המכוניות על הכביש. במקרה שכזה נוכל פשוט להעלות המספר המכוניות בריבוע לפני שאנו מזינים אותם לחזאי. העיבוד המקדים יכול למעשה לפעול על כל וקטור המדידות $\boldsymbol{x}$, לעבד אותו ולייצר ממנו וקטור חדש. אנו נסמן ב Φ את הפונקציה אשר מקבלת את המידע הגולמי $\boldsymbol{x}$ מייצרת ממנו את $\boldsymbol{x}_{\text{new}}$:

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

פעולת החיזוי במקרים אלו תהיה:

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

את קלט החדש $\boldsymbol{x}_{\text{new}}$ מקובל לכנות וקטור ה**מאפיינים (features)**. השימוש במאפיינים מאפשר לנו מספרים דברים:

- הרחבת מודלים פשוטים, כגון המודל הלינארי למודלים מורכבים יותר (כפי שנראה כאן ובתרגול).
- שינוי האופן בו מיוצג המידע כך שיהיה לחזאי קל יותר לבצע את בעיית החיזוי. לדוגמא:
  - החלפת היחידות שבהם השתמשו לתיאור מדידה מסויימת
  - הפיכת תמונת פנים לוקטור של מאפיינים של פנים כגון: המרחק בין העיניים, גוון העור, עד כמה הפנים אליפטיות וכו'
  - ניקוי רעשים להקלטות audio.
- לקראת סוף הקורס נראה גם כיצד ניתן להשתמש במאפיינים על מנת להתמודד עם בעיית ה overfitting בעזרת שיטה המוכנה הורדת מימד.

במקרים רבים נרצה להתייחס לפונקציה אשר מייצרת איבר ספציפי ב $\boldsymbol{x}_{\text{new}}$, לשם כך נוח להתייחס לפונקציה Φ כוקטור של פונקציות $\varphi_i(\boldsymbol{x})$ אשר מייצרות כל אחת איבר אחד בוקטור $\boldsymbol{x}_{\text{new}}$:

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

סימון מתמטי מקובל בו נשתמש הינו הסימון הבא:

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

כאן Φ מוצגת כוקטור של פונקציות, כאשר הפעלה וקטור שכזה על $\boldsymbol{x}$ מייצרת את הוקטור של הפלטים של הפונקציות φ_i

דוגמא: הוספה של איבר ההיסט בעזרת מאפיינים

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

$$\hat{y} = h(\boldsymbol{x};\theta) = h_{\text{linear}}(\Phi(\boldsymbol{x});\theta) = \Phi(\boldsymbol{x})^\top \boldsymbol{\theta} = \theta_1\varphi_1(\boldsymbol{x}) + \theta_2\varphi_2(\boldsymbol{x}) + \cdots + \theta_M\varphi_M(\boldsymbol{x})$$

נחזור כעת לדוגמא של שיערוך זמן הנסיעה. נראה כעת כיצד ניתן להוסיף את איבר ההיסט על ידי שימוש במאפיינים. העבור הבחירה של במאפיינים הבאים:

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

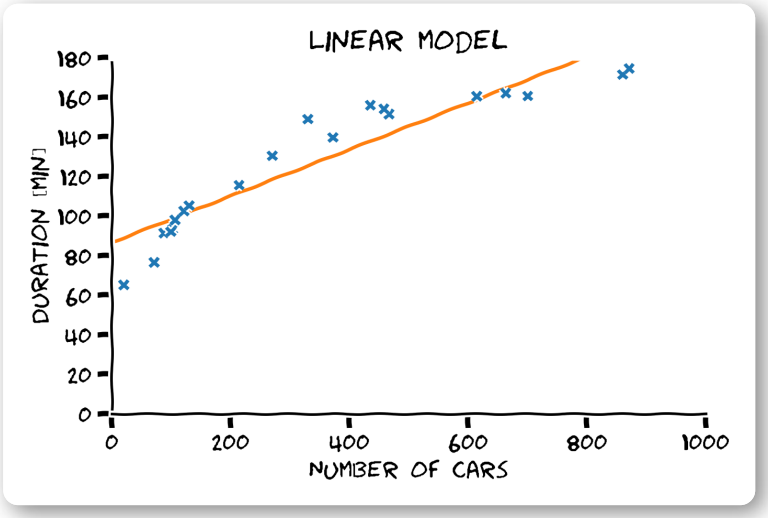
כל דגימה x תהפוך לוקטור $x_{\text{new}} = [1, x]^\top$ ומודל החיזוי שלנו יהיה:

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

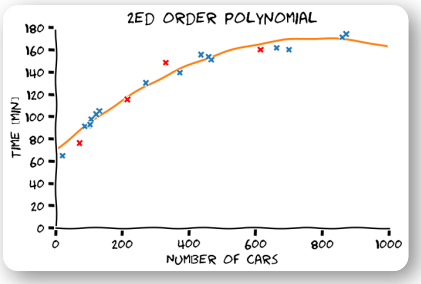
המטריצת המדידיות X תהיה כעת:

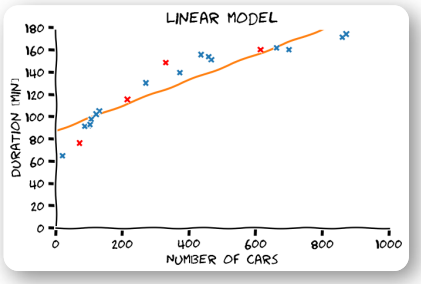
$$X = \begin{bmatrix} 1 & x^{(1)} \\ 1 & x^{(2)} \\ \vdots & \vdots \\ 1 & x^{(N)} \end{bmatrix}$$

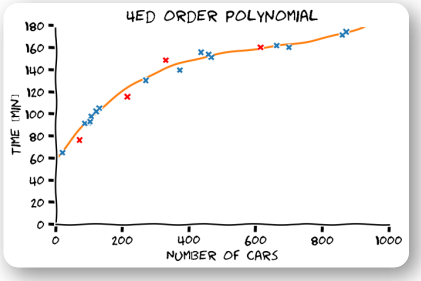
הצבה של מטריצה זו בנוחסא ל $\boldsymbol{\theta}^*_{\mathcal{D}}$ נותנת את המודל הלינארי הבא:

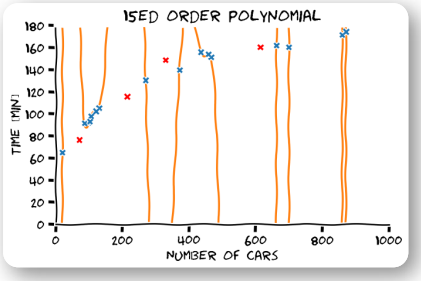


באותו אופן ניתן כמובן להשתמש במודל הלינארי בכדי לייצג מגוון רחב של פונקציות כגון פולינומים או קומבינציה של גאוסיאנים, כפי שיודגם בתרגול 3.






4ED ORDER POLYNOMIAL
TIME [MIN]
0
20
40
60
80
100
120
140
160
180
0
200
400
600
800
1000
NUMBER OF CARS


15ED ORDER POLYNOMIAL
TIME [MIN]
0
20
40
60
80
100
120
140
160
180
0
200
400
600
800
1000
NUMBER OF CARS