

תרגול 3 - Generalization & overfitting

[Slides](#)[PDF](#)[Code](#)

תקציר התיאוריה

מושגים

- **הכללה (generalization):** היכולת להסיק מן הפרט אל הכלל. היכולת של המודל להפיק תוצאות טובות גם על דגימות אשר לא הופיעו במדגם.
- **Overfitting (התאמת יתר):** התופעה שבה המודל לומד מאפיינים אשר מופיעים רק במדגם והם אינם מייצגים את התכונות של הפילוג האמיתי. Overfitting פוגע ביכולת ההכללה.
- **הערכת הביצועים / הציון של חזאי (יכולת הכללה):** הערכת המחיר (המתקבל מפונקציית המחיר) המתקבל בעבור חזאי נתון על הפילוג האמיתי.
- **יכולת הביטוי (expressiveness) של מודל פרמטרי:** היכולת של מודל פרמטרי לייצג (או לקרב) מגוון רחב של מודלים. לדוגמא לפולינום מסדר מאד גבוה יהיה יכולת ביטוי גבוהה בעוד שלמודל לינארי תהיה יכולת ביטוי נמוכה.
- **Hyper parameters** - הפרמטרים אשר משפיעים על המודל הפרמטרי או האלגוריתם, אך אינם חלק מהפרמטרים שעליהם אנו מבצעים את האופטימיזציה. דוגמאות:
 - סדר הפולינום שבו אנו משתמשים
 - הפרמטר η אשר קובע את גודל הצעד באלגוריתם ה gradient descent.
 - פרמטרים אשר קובעים את המבנה של רשת נוירונים.
- **סדר המודל:** כאשר ישנו hyper-parameter אשר שולט ביכולת הביטוי של המודל הפרמטרי (כגון המקרה של סדר של פולינום) אנו נכנה פרמטר זה לרוב הסדר של המודל.

הערכת ביצועים בעזרת test set (סט בחן)

במקרים בהם פונקציית המחיר מוגדרת בעזרת תוחלת (כמו במקרה הנפוץ של שימוש בפונקציות סיכון / הפסד) ניתן לשערך את ביצועיו של חזאי מסוים על ידי שימוש בתוחלת אמפירית ומדגם נוסף. לשם כך נפצל את המודל לשני תתי מדגמים:

- **Train set (סט אימון):** בו נשתמש לבנות את החזאי.
- **Test set (סט בחן):** בו נשתמש בכדי להעריך את ביצועי המערכת.

גודלו של ה test set

מצד אחד נרצה שסט הבחן יהיה גדול מספיק על מנת שיקרב בצורה טובה את ביצועיו האמיתיים של המודל אך מצד שני לא נרצה לגרוע יותר מידי דגימות מה training set. במקרים בהם המדגם מספיק גדול לא תהיה בעיה להפריש test set מספיק גדול מבלי לפגוע משמעותית בגודל המדגם, במקרים אחרים מקובל להשתמש בפיצול של 80% train ו-20% test.

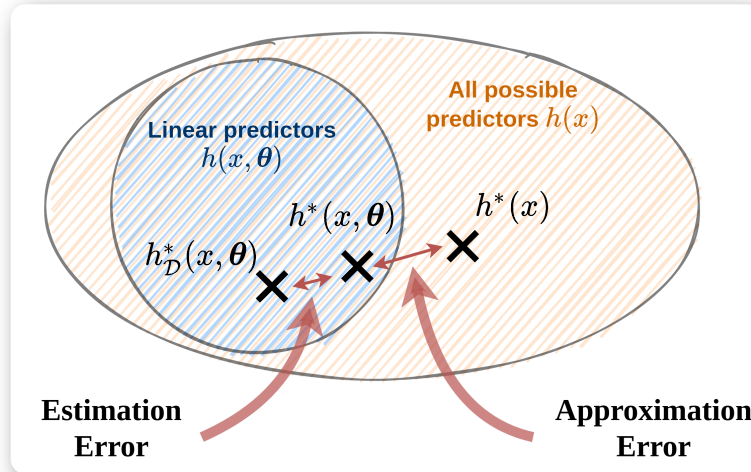
פירוק שגיאת החיזוי

בקורס זה נציג שני פירוקים נפוצים של שגיאת החיזוי בבעיות supervised learning.

Approximation-estimation decomposition

פירוק זה הוא רק רעיוני ולרוב לא ניתן לחשב אותו בפועל. בפירוק זה נתייחס לשלושת הגורמים הבאים בשגיאת החיזוי:

1. **Noise - ה"רעש" של התוויות:** השגיאה שהחזאי האופטימאלי צפוי לעשות. שגיאה זו נובעת מהאקראיות של התוויות y .
2. **Approximation error - שגיאת קירוב:** השגיאה עקב ההגבלה של המודל למשפחה מצומצמת של מודלים (לרוב מודל פרמטרי). שגיאה זו נובעת מההבדל בין המודל האופטימאלי h^* לבין המודל הפרמטרי האופטימאלי $h^*(\cdot, \theta)$.
3. **Estimation error - שגיאת השיערוך:** השגיאה הנובעת מהשימוש במדגם כתחליף לפילוג האמיתי וחוסר היכולת שלנו למצוא את המודל הפרמטרי האופטימאלי. שגיאה זו נובעת מההבדל בין המודל הפרמטרי האופטימאלי $h^*(\cdot, \theta)$ למודל הפרמטרי המשוערך על סמך המדגם $h_{\mathcal{D}}^*(\cdot, \theta)$.



Bias-variance decomposition

פירוק זה מתייחס למקרים שבהם פונקציית המחר הינה MSE (או RMSE).

המדגם \mathcal{D} שאיתו אנו עובדים הוא אקראי (משום שהוא אוסף של דגימות אקראיות) ולכן גם החזאי $h_{\mathcal{D}}$ שאותו ניצר על סמך המדגם הוא אקראי. נגדיר את החזאי הממוצע \bar{h} כחזאי המתקבל כאשר לוקחים תוחלת על החזאים המיוצרים על ידי אלגוריתם מסויים על פני כל המדגמים האפשריים.

$$\bar{h}(x) = \mathbb{E}_{\mathcal{D}} [h_{\mathcal{D}}(x)]$$

לשם הבהירות, אנו נשתמש בסימון $\mathbb{E}_{\mathcal{D}}$ בכדי לציין תוחלת על פני המדגמים האפשריים. (תוחלת ללא סימון \mathbb{E} תהיה לפי x ו y).

בעבור המקרה של MSE אנו יודעים כי החזאי האופטימאלי הינו: $h^*(x) = \mathbb{E}[y|x]$. על ידי שימוש בחזאי האופטימאלי והחזאי הממוצע ניתן לפרק את התוחלת על שגיאת ה MSE של אלגוריתם נתון באופן הבא:

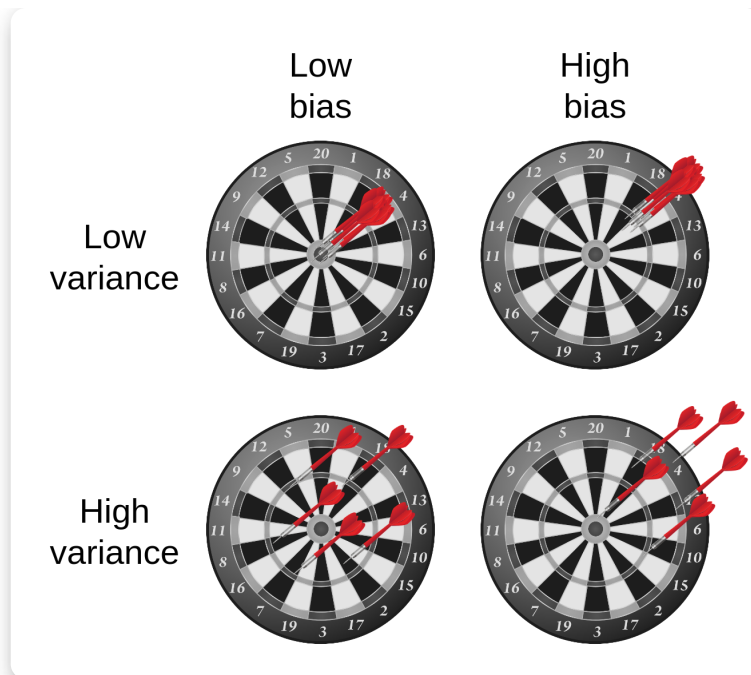
$$\mathbb{E}_{\mathcal{D}} [\mathbb{E} [(h_{\mathcal{D}}(x) - y)^2]] = \mathbb{E} \left[\underbrace{\mathbb{E}_{\mathcal{D}} [(h_{\mathcal{D}}(x) - \bar{h}(x))^2]}_{\text{Variance}} + \underbrace{(\bar{h}(x) - h^*(x))^2}_{\text{Bias}^2} + \underbrace{(h^*(x) - y)^2}_{\text{Noise}} \right]$$

בפירוק הזה:

- ה **variance** מודד את השונות של החזאים השונים המתקבלים ממדגמים שונים סביב החזאי הממוצע. זהו האיבר היחיד בפירוק אשר תלוי בפילוג של המדגם.
- ה **bias** מודד את ההפרש הריבועי בין החזיון של החזאי הממוצע לבין החזיון של החזאי האופטימאלי.
- ה **noise** (בודמה לפירוק הקודם) מודד את השגיאה הריבועית המתקבלת בעבור החזיון האופטימאלי (אשר נובעת מהאקראיות של y).

בתרגיל 3.1 נפתח את הפירוק הזה.

אילוסטרציה של bias ו variance:

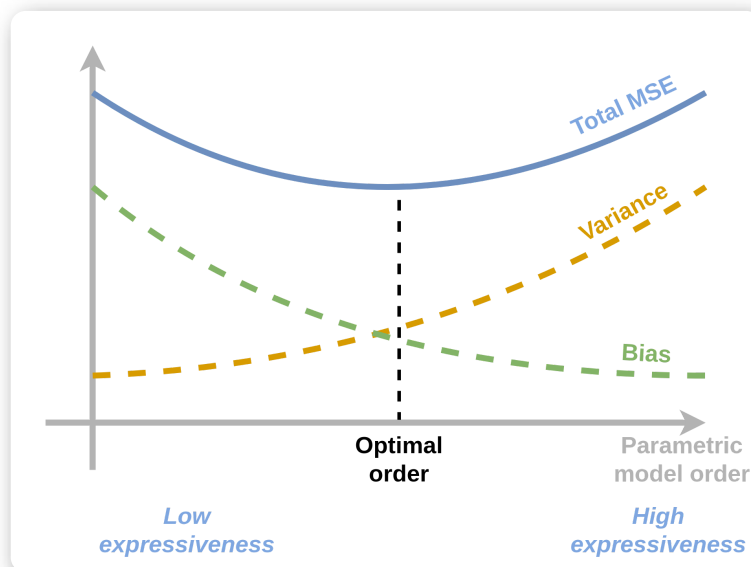


Tradeoffs

יכולת הביטוי של המודל הפרמטרי לרוב ישפיע מאד על גודלם של השגיאות אותם נקבל. לרוב התלות של השגיאות המתקבלות ביכולת הביטוי של המודל תקיים את הקשר הבא:

- מודל בעל יכולת ביטוי גבוהה לרוב יהיה בעל שגיאת קירוב / **bias** נמוך אך שגיאת שיערוך / **variance** גבוה.
- מודל בעל יכולת ביטוי נמוכה לרוב יהיה בעל שגיאת שיערוך / **variance** נמוך אך שגיאת קירוב / **bias** גבוה.

המודל עם הביצועים הטובים ביותר (יכולת הכללה טובה) ימצא באיזון שהיא נקודת ביניים בין שני הקצוות הנל, כפי שמתואר בשירטוט הסכימתי הבא:



שימוש ב validation set לקביעת hyper-parameters

מכיוון שה hyper-parameters אינם חלק מבעיית האופטימיזציה אנו צריכים דרך אחרת לקבוע אותם. לרוב נאלץ לקבוע את הפרמטרים האלו בשיטה של ניסוי וטעיה. זאת אומרת שיהיה עלינו פשוט נסות ערכים שונים ולבדוק את ביצועי המודל בעבור אותם ערכים.

מכיוון שאנו לא יכולים להשתמש ב test set בכדי לבנות את המודל שלנו או צריכים לייצר מדגם נפרד נוסף שעליו נוכל לבחון את ביצועי המודל בעבור ערכים שונים של hyper-parameters. אנו ניצר מדגם זה על ידי חלוקה נוספת של ה train set. למדגם הנוסף נקרא validation set.

במקרים רבים לאחר קביעת hyper-parameters אנו נאחד חזרה את ה validation set וה train set ונאמן מחדש את המודל על המדגם המאוחד (כל הדגימות מלבד ה test set).

רגולריזציה

דרך נוספת לנסות ולהקטין את ה overfitting של המודל על ידי הוספת איבר רגולריזציה לבעיית האופטימיזציה. מטרת איבר הרגולריזציה הינה להשתמש בידע מוקדם שיש לנו על אופי הבעיה לצורך בחירת המודל. הדרך שבה איבר הרגולריזציה עושה זאת הינה על ידי הוספת תיקון לבעיית האופטימיזציה כך שמודלים שלדעתנו פחות סבירים יקבלו ציון גבוה יותר. לרוב אנו נוסף לבעיית האופטימיזציה את איבר הרגולריזציה יחד עם קבוע כפלי נוסף λ אשר קובע את המשקל שאנו מעוניינים לתת לרגולריזציה.

בעיות אופטימיזציה עם רגולריזציה יהיו מהצורה הבאה:

$$\theta = \arg \min_{\theta} \underbrace{f(\theta)}_{\text{The regular objective function}} + \lambda \underbrace{g(\theta)}_{\text{The regularization term}}$$

שתי הרגולריזציות הנפוצות ביותר הינן:

- l_1 - אשר מוסיפה איבר רגולריזציה של $\|\theta\|_1$.
- Tikhonov regularization (l_2) - אשר מוסיפה איבר רגולריזציה של $\|\theta\|_2^2$.

רגולריזציות אלו מנסות לשמור את הפרמטרים כמה שיותר קטנים. המוטיבציה מאחורי הרצון לשמור את הפרמטרים קטנים הינה העובדה שבמרבית המודלים ככל שהפרמטרים קטנים יותר המודל הנלמד יהיה בעל נגזרות קטנות יותר ולכן הוא ישתנה לאט יותר ופחות "ישתולל".

המשקל אותו אנו נותנים לרגולריזציה λ הוא hyper-parameter של האלגוריתם שאותו יש לקבוע בעזרת ה validation set.

דוגמא: בעיות LLS עם רגולריזציה

Ridge regression: LLS + l_2 regularization

$$\theta = \arg \min_{\theta} \frac{1}{N} \sum_i (\mathbf{x}^{(i)\top} \theta - y^{(i)})^2 + \lambda \|\theta\|_2^2$$

גם לבעיה זו יש פתרון סגור והוא נתון על ידי:

$$\theta^* = (X^T X + \lambda I)^{-1} X^T \mathbf{y}$$

אנו נראה את הפתוח של פתרון זה בתרגיל 3.2.

LASSO: LLS + l_1 regularization

(LASSO = Linear Absolute Shrinkage and Selection Operator)

$$\theta = \arg \min_{\theta} \frac{1}{N} \sum_i (\mathbf{x}^{(i)\top} \theta - y^{(i)})^2 + \lambda \|\theta\|_1$$

לבעיה זו אין פתרון סגור ויש צורך להשתמש באלגוריתמים איטרטיביים כגון gradient descent.

תרגיל 3.1 - Bias-variance decomposition

1 הראו כי בעבור משתנה אקראי כל שהוא x וקבוע a ניתן לפרק את התחולת של המרחק הריבועי בין x לבין a באופן הבא:

$$\mathbb{E} [(x - a)^2] = \underbrace{\mathbb{E} [(x - \mathbb{E}[x])^2]}_{=\text{Var}(x)} + \underbrace{(\mathbb{E}[x] - a)^2}_{\text{bias}}$$

(2) הראו כי בעבור אלגוריתם אשר מייצר חזאיים h_D בהינתן מדגמים \mathcal{D} , ניתן לפרק את התוחלת (על פני מדגמים וחיזויים שונים) של שגיאת ה MSE באופן הבא:

$$\mathbb{E}_D [\mathbb{E} [(h_D(x) - y)^2]] = \mathbb{E} \left[\underbrace{\mathbb{E}_D [(h_D(x) - \bar{h}(x))^2]}_{\text{Variance}} + \underbrace{(\bar{h}(x) - h^*(x))^2}_{\text{Bias}^2} + \underbrace{(h^*(x) - y)^2}_{\text{Noise}} \right]$$

כאשר:

$$\bar{h}(x) = \mathbb{E}_D [h_D(x)]$$

1

$$h^*(x) = \mathbb{E} [y|x]$$

הדרכה:

1. הראו ראשית כי ניתן לפרק את שגיאת ה MSE בעבור מדגם נתון באופן הבא:

$$\mathbb{E} [(h_D(x) - y)^2] = \mathbb{E} [(h_D(x) - h^*(x))^2] + \mathbb{E} [(h^*(x) - y)^2]$$

לשם כך השתמשו בהחלקה על מנת להתנות את התוחלת ב x ולקבל תוחלת לפי y . הפעילו את הזהות מסעיף 1 על התוחלת של y .

2. הראו כי ניתן לפרק את התוחלת הזו באופן הבא:

$$\mathbb{E}_D [\mathbb{E} [(h_D(x) - h^*(x))^2]] = \mathbb{E} [\mathbb{E}_D [(h_D(x) - \bar{h}(x))^2] + (\bar{h}(x) - h^*(x))^2]$$

לשם כך החליפו את סדר התוחלות והשתמשו בזהות מסעיף 1 על התוחלת לפי \mathcal{D} .

3. השתמשו בשני הפירוקים הנ"ל על עמת להראות את פירוק ה bias-variance המלא.

(3) הניחו כי כאשר גודל המדגם הולך וגדל החזאים המתקבלים מהמודל מתכנסים (במובן הסטברותי) לחזאי ה"ממוצע": $h_D \rightarrow \bar{h}$. מה תוכלו לומר על התלות של איברי השגיאה בגודל המדגם?

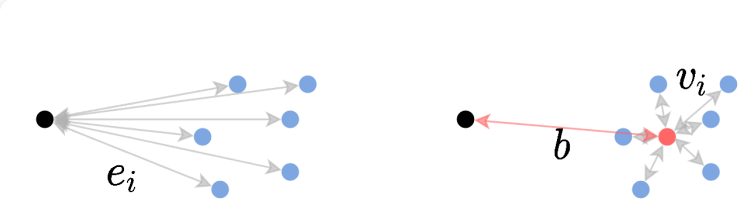
(ניתן להניח שכפי שקורה במרבית האלגוריתמים ש \bar{h} אינו תלוי בגודל המדגם)

(4) על פי תוצאת הסעיף הקודם, כיצד לדעתכם עשוי להשפיע גודל המדגם על סדר המודל שאותו נרצה לבחור?

פתרון 3.1

(1)

הזהות שאותה נתבקשנו להוכיח היא למעשה הכללה של הקשר הבא למשתנים אקראיים:



$$\frac{1}{N} \sum e_i^2 = b^2 + \frac{1}{N} \sum v_i^2$$

נוכיח את הזהות על ידי הוספה והחסרה של $\mathbb{E}[x]$ בתוך הסוגריים:

$$\begin{aligned}
\mathbb{E}[(x - a)^2] &= \mathbb{E}[(x - \mathbb{E}[x]) + (\mathbb{E}[x] - a)]^2 \\
&= \mathbb{E}[(x - \mathbb{E}[x])^2] - 2\mathbb{E}[(x - \mathbb{E}[x])(\mathbb{E}[x] - a)] + \mathbb{E}[(\mathbb{E}[x] - a)^2] \\
&= \mathbb{E}[(x - \mathbb{E}[x])^2] - 2\underbrace{(\mathbb{E}[x] - \mathbb{E}[x])}_{=0}(\mathbb{E}[x] - a) + \mathbb{E}[(\mathbb{E}[x] - a)^2] \\
&= \mathbb{E}[(x - \mathbb{E}[x])^2] + (\mathbb{E}[x] - a)^2
\end{aligned}$$

(2)

שלב ראשון

נפעל על פי ההדרכה. נחליק על ידי התניה ב x ונפעיל את הזהות מסעיף 1 על התוחלת הפנימית (לפי y):

$$\begin{aligned}
\mathbb{E}[(h_{\mathcal{D}}(x) - y)^2] &= \mathbb{E}[\mathbb{E}[(h_{\mathcal{D}}(x) - y)^2|x]] \\
&= \mathbb{E}[(h_{\mathcal{D}}(x) - \mathbb{E}[y|x])^2 + \mathbb{E}[(\mathbb{E}[y|x] - y)^2|x]]
\end{aligned}$$

נארגן מחדש את התוחלות:

$$\begin{aligned}
&= \mathbb{E}[(h_{\mathcal{D}}(x) - \mathbb{E}[y|x])^2] + \mathbb{E}[\mathbb{E}[(\mathbb{E}[y|x] - y)^2|x]] \\
&= \mathbb{E}[(h_{\mathcal{D}}(x) - \mathbb{E}[y|x])^2] + \mathbb{E}[(\mathbb{E}[y|x] - y)^2]
\end{aligned}$$

נשתמש כעת בעובדה ש $\mathbb{E}[y|x] = h^*(x)$ ונקבל:

$$= \mathbb{E}[(h_{\mathcal{D}}(x) - h^*(x))^2] + \mathbb{E}[(h^*(x) - y)^2]$$

בביטוי שקיבלנו האיבר הראשון הוא למעשה השגיאה הנובעת מההבדל בין החיזוי של מודל האידיאלי לבין החיזוי של מודל ספציפי ששנוצר ממדגם מסוים. נשים לב שהאיבר הראשון לא תלוי בכלל בפילוג של y . האיבר השני בביטוי שקיבלנו הוא השגיאה אותה עושה החזאי האופטימאלי והיא נובע מחוסר היכולת לחזות את y במדויק. נשים לב כי האיבר השני לא תלוי כלל במדגם.

שלב שני

על פי ההדרכה נפרק את התוחלת הבאה על ידי החלפת סדר התוחלות ושימוש בזיהות מסעיף 1 על התוחלת לפי \mathcal{D} :

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\mathbb{E}[(h_{\mathcal{D}}(x) - h^*(x))^2]] &= \mathbb{E}[\mathbb{E}_{\mathcal{D}}[(h_{\mathcal{D}}(x) - h^*(x))^2]] \\
&= \mathbb{E}[\mathbb{E}_{\mathcal{D}}[(h_{\mathcal{D}}(x) - \mathbb{E}_{\mathcal{D}}[h_{\mathcal{D}}(x)])^2] + (\mathbb{E}_{\mathcal{D}}[h_{\mathcal{D}}(x)] - h^*(x))^2]
\end{aligned}$$

נשתמש בסימון $\bar{h}(x) = \mathbb{E}_{\mathcal{D}}[h_{\mathcal{D}}(x)]$ ונקבל:

$$\mathbb{E}[\mathbb{E}_{\mathcal{D}}[(h_{\mathcal{D}}(x) - \bar{h}(x))^2] + (\bar{h}(x) - h^*(x))^2]$$

זהו הפירוק של השגיאה לרכיב ה variance של החזאי אשר מבטא את השגיאה הצפויה עקב ההשתנות של החזאי כתלות במדגם שאיתו נעבוד, ורכיב bias אשר מבטא את השגיאה אשר נובעת מההבדל בין החזאי ה"ממוצע" והחזאי האידיאלי.

נרכיב את הכל

נשתמש בפירוק הראשון על מנת לקבל:

$$\mathbb{E}_{\mathcal{D}}[\mathbb{E}[(h_{\mathcal{D}}(x) - y)^2]] = \mathbb{E}_{\mathcal{D}}[\mathbb{E}[(h_{\mathcal{D}}(x) - h^*(x))^2] + \mathbb{E}[(h^*(x) - y)^2]]$$

מכיוון שהאיבר השני לא תלוי ב \mathcal{D} נוכל להוציא אותו מהתוחלת על \mathcal{D} :

$$= \mathbb{E}_{\mathcal{D}}[\mathbb{E}[(h_{\mathcal{D}}(x) - h^*(x))^2]] + \mathbb{E}[(h^*(x) - y)^2]$$

נציב את הפירוק מהשלב השני ונקבל:

$$\begin{aligned}
&= \mathbb{E} \left[\mathbb{E}_{\mathcal{D}} \left[(h_{\mathcal{D}}(x) - \bar{h}(x))^2 \right] + (\bar{h}(x) - h^*(x))^2 \right] + \mathbb{E} \left[(h^*(x) - y)^2 \right] \\
&= \mathbb{E} \left[\underbrace{\mathbb{E}_{\mathcal{D}} \left[(h_{\mathcal{D}}(x) - \bar{h}(x))^2 \right]}_{\text{Variance}} + \underbrace{(\bar{h}(x) - h^*(x))^2}_{\text{Bias}^2} + \underbrace{(h^*(x) - y)^2}_{\text{Noise}} \right]
\end{aligned}$$

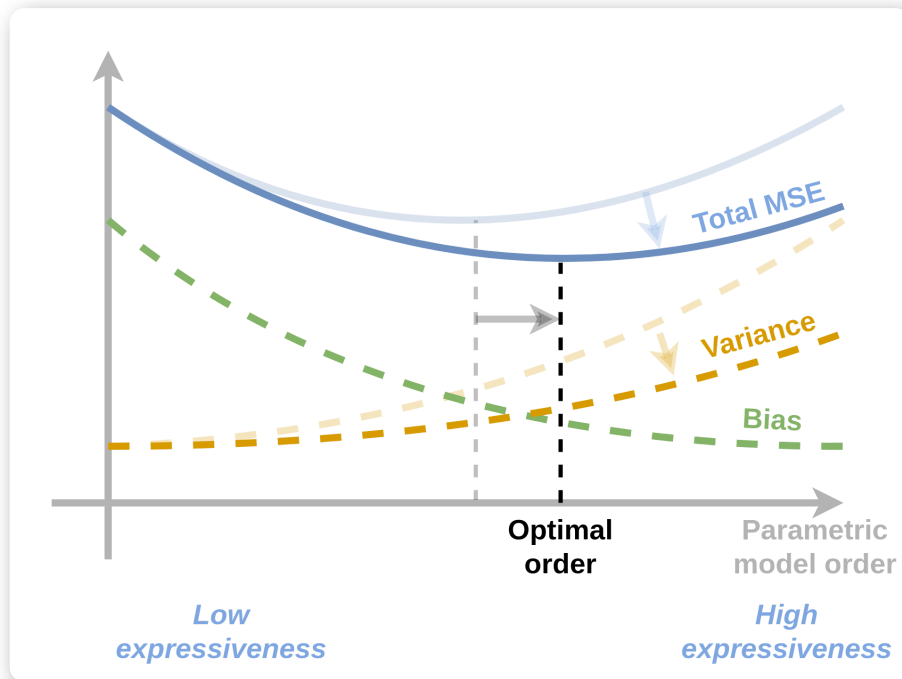
(3)

המשמעות של העובדה ש $h_{\mathcal{D}}$ מתכנס לממוצע שלו \bar{h} במובן הסתברותי הינה למעשה שה variance שלו קטן. זאת אומרת שרכיב ה variance בפירוק הני"ל יקטן. שאר האיברים במקרה זה לא יושפעו מהשינוי בגודלו של המדגם. (ישנם מקרים פחות נפוצים שבהם גם החזאי הממוצע \bar{h} תלוי בגודל המדגם ואז גם הוא יכול להשתנות).

(4)

השינוי של רכיב ה variance יכול כמובן להשפיע על סדר המודל האופטימאלי. באלגוריתמים טיפוסיים שגיאת ה variance תהיה זו שמושכת את משפחת המודלים להיות כמה שיותר מצומצמת (בעוד ששגיאת ה bias כן מושכת בכיוון ההפוך). אנו נצפה שכאשר ה variance יקטן תקטן גם ההשפעה שלו על השגיאה הכוללת ולרוב ניתן יהיה להוסיף ולהקטין את השגיאה על ידי הגדלת הסדר של המודל הפרמטרי.

ננסה להמחיש זאת גם בעזרת הגרף הסכימתי הבא:



כאשר הגרף של שגיאת ה variance ירד אנו מצפים כי נקודת המינימום של השגיאה הכוללת תזוז ימינה לכיוון מודלים מסדר גבוה יותר.

הערה: ניתוח זה כמובן נשען על התנהגות טיפוסית של אלגוריתמי supervised learning ואין הכרח שההתנהגות המתוארת בתשובה זו אכן תהיה ההתנהגות במציאות.

תרגיל 3.2 - רגולריזציה

(1) בעבור Ridge regression (המקרה של l_2 regularization של LLS) רשמו את בעיית האופטימיזציה ופתרו אותה על ידי גזירה והשוואה ל-0.

(2) נסתכל כעת על וריאציה של Ridge regression שבה אנו נותנים משקל שונה w_i לרגולריזציה של כל פרמטר. זאת אומרת, אנו נרצה להשתמש באיבר רגולריזציה מהצורה:

$$\sum_{i=1}^D w_i \theta_i^2$$

(כאן D הוא מספר הפרמטרים של המודל).

הדרכה: הגדירו את מטריצת המשקלים $W = \text{diag}(\{w_i\})$, רשמו את הבעיה בכתוב מטריוצי ופתרו אותה בדומה לסעיף הקודם.

(3) נסתכל כעת על אלגוריתם כללי שבו הפרמטרים של המודל נקבעים על פי בעיית האופטימיזציה הבאה

$$\theta^* = \arg \min_{\theta} f(\theta)$$

רשמו את הבעיות האופטימיזציה המתקבלות לאחר הוספה של איבר רגולריזציה מסוג l_1 ו l_2 .

(4) רשמו את כלל העדכון של אלגוריתם הגרדיאנט בעבור כל אחת משתי הרגולריזציות.

(5) על סמך ההבדל בין שני כללי העדכון הסבירו מה ההבדל בין האופן שבו שתי הרגולריזציות מנסות להקטין את הפרמטרים.

(6) על סמך שני כללי העדכון הסבירו מדוע רגולריזציות l_1 נוטה יותר לאפס פרמטרים מאשר רגולריזציות l_2 . (הניחו שצדדי העדכון קטנים מאד)

פתרון 3.2

(1)

תזכורת, בעיית ה LLS היא המקרה שבו אנו משתמשים ב

- MSE או RMSE כפונקציית המחיר / סיכון.
- ERM.
- מודל לינארי

בעיית האופטימיזציה של LLS הינה:

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

כאשר

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

כאשר נוסף לבעיית האופטימיזציה איבר של רגולריזציות l_2 נקבל:

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

נגזר ונשווה ל-0. נשתמש בנגזרת המוכרת $\nabla_{\mathbf{x}} \|\mathbf{x}\|_2^2 = 2\mathbf{x}$:

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

$$\Leftrightarrow \frac{2}{N} (X^T X \theta - X^T \mathbf{y}) + 2\lambda \theta = 0$$

$$\Leftrightarrow (X^T X + N\lambda I) \theta = X^T \mathbf{y}$$

$$\Leftrightarrow \theta = (X^T X + N\lambda I)^{-1} X^T \mathbf{y}$$

ניתן כמובן "לבלוע" את ה N בתוך הפרמטר λ , אך שינוי זה מצריך להתאים את הפרמטר λ לגודל המדגם ולעדכנו כאשר גודל המדגם משתנה (נגיד במקרה בו מפרישים חלק מהמדגם ל validation set).

(2)

בעיית האופטימיזציה כעת תהיה

$$\theta^* = \arg \min_{\theta} \frac{1}{N} \|X\theta - \mathbf{y}\|_2^2 + \lambda \sum_{i=1}^D w_i \theta_i^2$$

נפעל על פי ההדרכה. נגדיר את המטריצה:

$$W = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & & w_D \end{bmatrix}$$

בעזרת מטריצה זו ניתן לרשום את בעיית האופטימיזציה באופן הבא:

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

נגזור ונשווה ל-0:

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

(3)

בעיית האופטימיזציה בתוספת רגולריזציה l_2 תהיה:

$$\theta^* = \arg \min_{\theta} f(\theta) + \lambda \|\theta\|_2^2$$

בעיית האופטימיזציה בתוספת רגולריזציה l_1 תהיה:

$$\theta^* = \arg \min_{\theta} f(\theta) + \lambda \|\theta\|_1$$

(4)

כלל העדכון של gradient descent הינו:

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

כאשר $g(\theta)$ היא פונקציית המטרה של בעיית האופטימיזציה. בעבור רגולריזציה l_2 נקבל:

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_{\theta} f(\theta^{(t)}) - 2\eta \lambda \theta^{(t)}$$

בעבור רגולריזציה l_1 נקבל:

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

כאשר פונקציית ה sign פועלת איבר איבר.

(5)

שני האיברים שנוספו לכלל העדכון מנסים בכל צעד להקטין את וקטור הפרמטרים ולקרב אותו ל-0. ההבדל בין שני האיברים הינו שבעוד שהאיבר המתקבל ברגולריזציה l_2 הינו פורפוזיוני לגודל של האיברים בוקטור הפרמטרים האיבר של

רגולריזציית ה- l_1 הוא קבוע (עד כדי סימן). המשמעות של זה הינה שב- l_2 כלל שפרמטר מסויים גדול יותר כך הרגולריזציה תתאמץ יותר להקטין אותו ויחסית פחות תשפיע על הפרמטרים הקטנים. מנגד, רגולריזציית ה- l_1 תפעל להקטין את כל האיברים ללא קשר לגודלם.

6

כפי שצינו בסעיף הקודם, רגולריזציית ה- l_2 תשפיע באופן מועט יחסית על האיברים הקטנים ולא תתאמץ להקטין אותם ובעיקר תפעל להקטין את האיברים הגדולים. מנגד, רגולריזציית ה- l_1 תמשיך ולנסות להקטין את האיברים כל עוד הם שונים מ-0 ולכן בפועל היא תיטה לאפס יותר איברים.

הערה: בפועל בגלל שגודלו של איבר הרגולריזציה בגרדיאנט של l_1 קבוע הוא יקטין את האיברים לערכים קרובים ל-0 ואז יתחיל להתנדנד סביב ה-0.

K-fold cross validation

במקרים בהם גודלו של המדגם שנתון לנו הינו קטן לא נוכל להקצות כמות גדולה של דגימות לטובת ה- $validation\ set$. במקרים כאלה ה- $validation$ עלול להיות לא מאד מייצג ולפגוע בבחירה של ה- $hyper\ parameters$. במקרים כאלה נרצה למצוא דרך טובה יותר להעריך את ביצועי המודל בעבור כל בחירה של $hyper\ parameters$. שיטת ה- $K\text{-fold cross validation}$ מציעה שיטה לשפר את הדיוק על הערכת הביצועים על ידי מיצוע על כמה $validation\ sets$.

בשיטה זו נחלק את ה- $train\ set$ שלנו ל- K קבוצות ונבצע את הערכת הביצועים K פעמים באופן הבא:

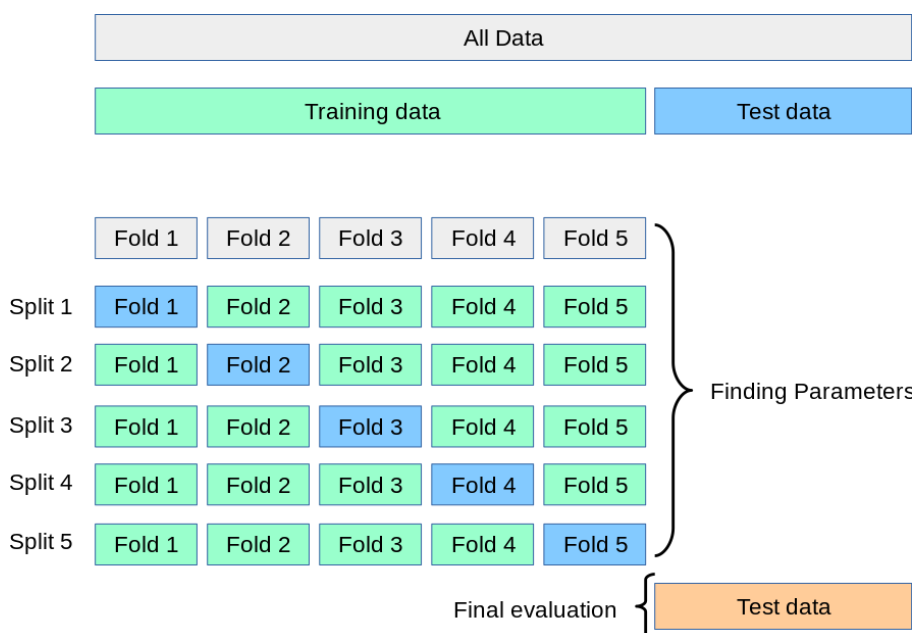
1. בכל פעם נבחר (על פי הסדר) את אחת הקבוצות לשמש כ- $validation\ set$ הנוכחי.
2. בניה של מודל על סמך ה- $K - 1$ קבוצות האחרות
3. חישוב הביצועים של המודל על סמך הקבוצה שנבחרה.

הביצועים הכוללים יהיו הממוצע של התוצאות אשר התקבלו ב- K החזרות.

גדולים אופייניים ל- K הינם בין 5 ל-10.

כמו תמיד, לאחר קביעת ה- $hyper\ parameters$ ניתן לאחד חזרה את כל הקבוצות ל- $train\ set$ אחד ולבנות בעזרתו את המודל תוך שימוש ב- $hyper\ parameters$ שנבחרו.

להלן סכימה של החלוקה של המדגם בעבור בחירה של $K = 5$ (שרטוט זה לקוח מתוך התיעוד של החבילה [scikit learn](#)):



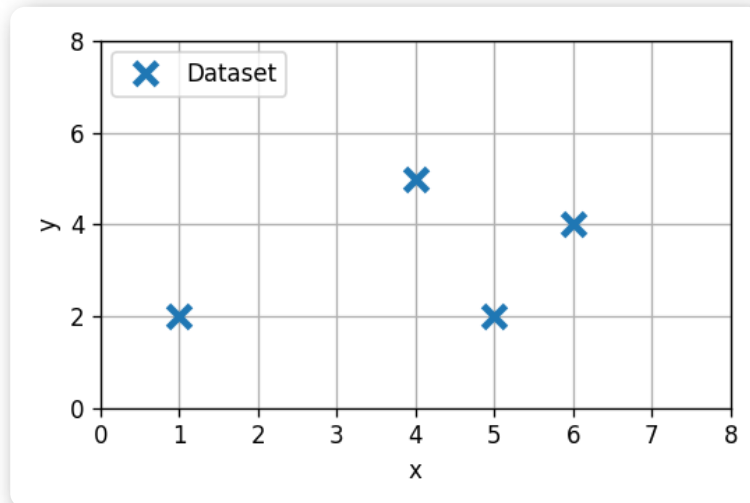
Leave-one-out cross validation

במקרים מסויימים (בעיקר כאשר ה train set מאד קטן) אנו נבחר לקחת את K להיות שווה למספר האיברים שב train set. במקרה זה גודלה של כל קבוצה יהיה 1. מקרה זה מכונה לרוב Leave-one-out cross validation.

תרגיל 3.3 - בחירת סדר המודל

נתון המדגם הבא:

$$\mathcal{D} = \{\{6, 4\}, \{1, 2\}, \{4, 5\}, \{5, 2\}\}$$



נרצה לנסות ולהתאים למדגם הנתון אחד משני מודלים: מודל לינארי מסדר 0 (פונקציה קבועה) או מסדר ראשון (פונקציה לינארית עם היסט). בתרגיל זה נבחן דרכים לקביעת סדר המודל.

נפצל את המדגם כך ששלושת הדגימות הראשונות יהיו ה train set והאחרונה תהיה ה test set.

1 השתמשו ב LLS על מנת להתאים כל אחד משני המודלים המוצעים ל train set. העריכו את ביצועי החזאי על פי שגיאת החיזוי המתקבל על הנקודה שב test set. מי מהמודלים נותן ביצועים טובים יותר?

2 מדוע לא נרצה לבחור את סדר המודל על סמך ההשוואה שעשינו על ה test set?

משלב זה והלאה נשכח שביצענו את הערכת הביצועים על ה test set וננסה לקבוע את סדר המודל על סמך validation set.

3 הפרישו מתוך ה train set את הדגימה השלישית על מנת שתשמש כ validation set. התאימו כעת את שני המודלים ל train set החדש והעריכו את ביצועיהם על ה validation set.

4 במקום להשתמש ב validation set קבוע, השתמשו ב leave-one-out על מנת לבחור מבין שני המודלים.

פתרון 3.3

(1)

נחלק את המדגם ל train set ו test set:

$$\mathcal{D}_{\text{train}} = \{\{6, 4\}, \{1, 2\}, \{4, 5\}\}$$

$$\mathcal{D}_{\text{test}} = \{\{5, 2\}\}$$

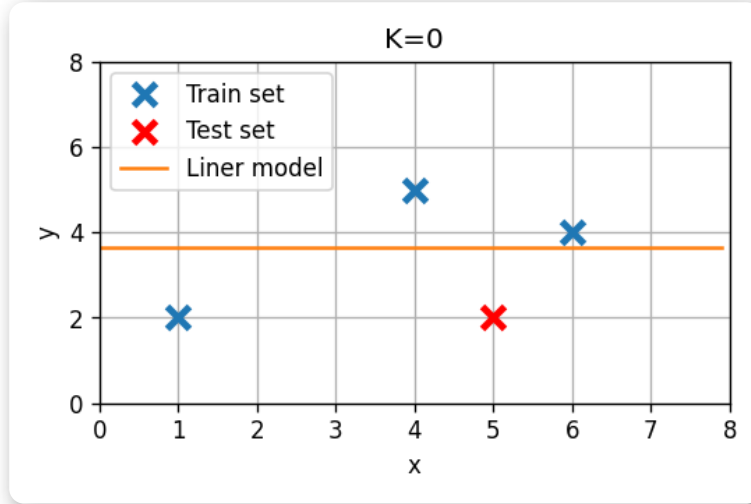
סדר 0

מודל מסדר 0 (פונקציה קבועה) הוא כמובן מקרה מנוון של מודל לינארי עם מאפיין יחיד של $\varphi(x) = 1$. במקרה זה אנו מצפים כי המודל אשר ימזער את השגיאה הריבועית יהיה פשוט פונקציה קבועה אשר שווה ל y הממוצע על ה train set. נראה כי זה אכן הפתרון המתקבל מתוך הפתרון הסגור. בעבור מודל זה המטריצה X והוקטור y יהיו:

$$X = [1, 1, 1]^T, \quad \mathbf{y} = [4, 2, 5]^T$$

הפרמטר האופטימאלי θ^* יהיה:

$$\theta^* = (X^T X)^{-1} X^T \mathbf{y} = \frac{\sum_{i=1}^N y^{(i)}}{N} = 3\frac{2}{3}$$



שגיאת החיזוי תהיה במקרה זה $|2 - 3\frac{2}{3}| = 1\frac{2}{3}$.

סדר ראשון

מודל זה הינו מודל לינארי עם המאפיינים:

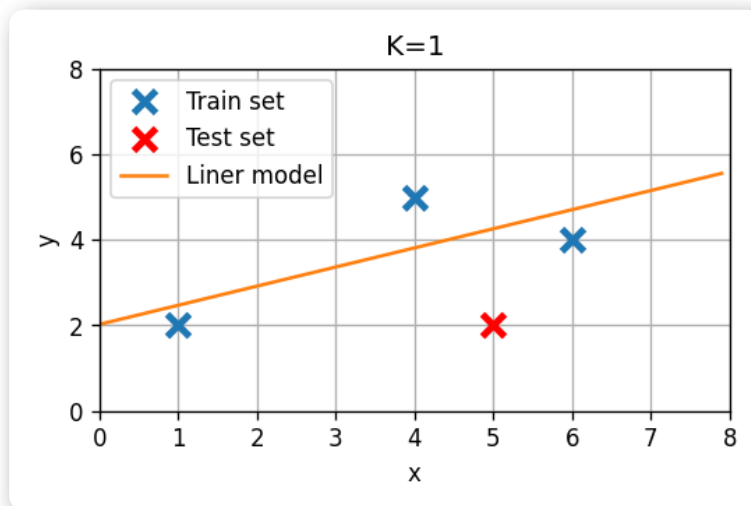
$$\varphi_1(x) = 1, \quad \varphi_2(x) = x$$

המטריצה X והוקטור \mathbf{y} יהיו:

$$X = \begin{bmatrix} 1 & 6 \\ 1 & 1 \\ 1 & 4 \end{bmatrix} \quad \mathbf{y} = [4, 2, 5]^T$$

הפרמטרים האופטימאליים θ^* יהיו:

$$\theta^* = (X^T X)^{-1} X^T \mathbf{y} = [77, 17]^T / 38$$



$$| \frac{77}{38} + \frac{17}{38} \cdot 5 - 2 | = 2.263$$

על סמך שיגאת החיזוי תהיה במקרה זה 2.263

(2)

תפקידו של ה test set הינו להעריך את ביצועי המודל הסופי לאחר שסיימנו את כל השלבים של בניית המודל כולל בחירת hyper parameters כגון סדר המודל. כאשר אנו מקבלים החלטה כל שהיא או קובעים פרמטר כל שהוא על סמך ה test set אנו למעשה גורמים למודל שלנו להתחיל לעשות overfitting ל test set הספציפי שבידינו ולכן לא נוכל להשתמש בו יותר על מנת לבצע הערכה בלתי מוטית של ביצועי המודל שלנו.

(3)

נקצה את הדגימה השלישית במדגם לטובת ה validation set:

$$\mathcal{D}_{\text{train}} = \{\{6, 4\}, \{1, 2\}\}$$

$$\mathcal{D}_{\text{validation}} = \{\{4, 5\}\}$$

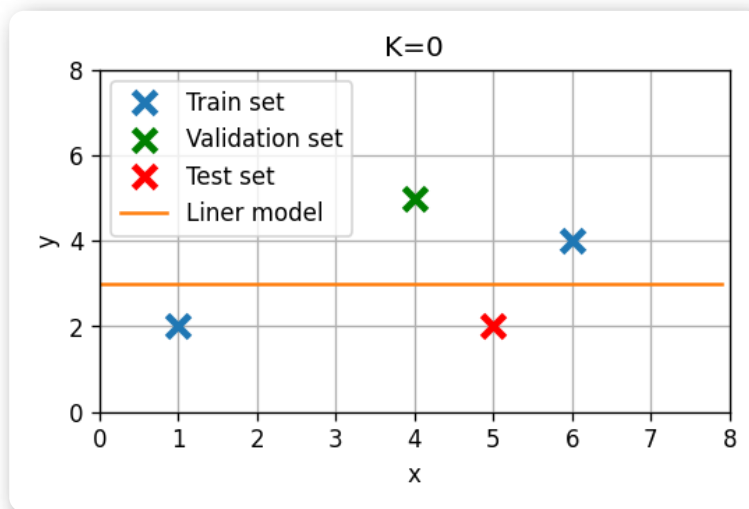
$$\mathcal{D}_{\text{test}} = \{\{5, 2\}\}$$

נתאים שוב את שני המודלים על סמך ה train set החדש ונעריך את שיגאת החיזוי על ה validation set:

סדר 0

$$X = [1, 1]^T, \quad y = [4, 2]^T$$

$$\theta^* = \frac{\sum_{i=1}^N y^{(i)}}{N} = 3$$



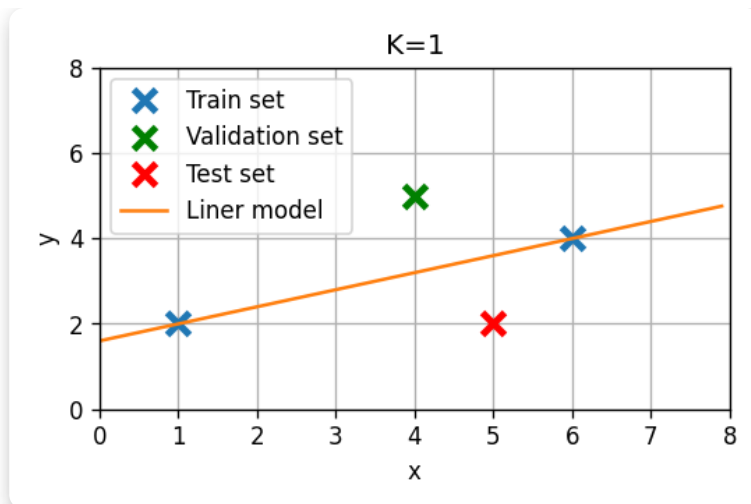
שיגאת החיזוי על ה validation set תהיה במקרה זה $|3 - 5| = 2$.

סדר ראשון

$$X = \begin{bmatrix} 1 & 6 \\ 1 & 1 \end{bmatrix} \quad y = [4, 2]^T$$

הפרמטרים האופטימאליים θ^* יהיו:

$$\theta^* = (X^T X)^{-1} X^T y = [8, 2]^T / 5$$



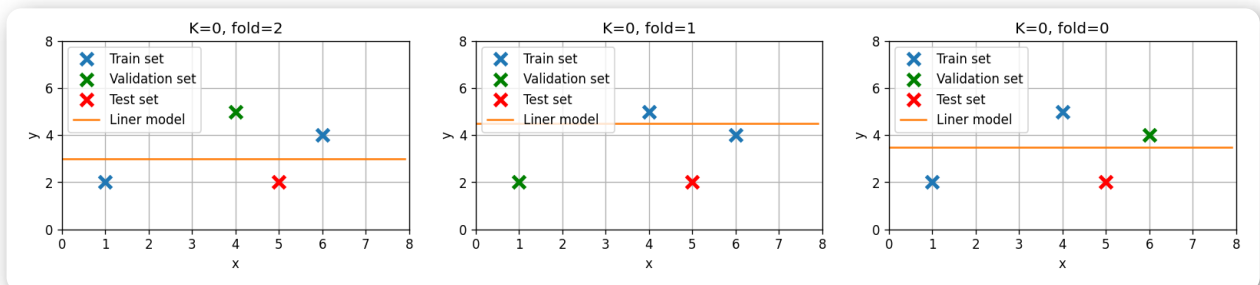
שיגאת החיזוי על ה validation set תהיה במקרה זה $|\frac{8}{5} + \frac{2}{5} \cdot 4 - 5| = 9/5$.

כעת נראה כי דווקא המודל מסדר ראשון הוא המודל העדיף. מכיוון ש ה validation set שלנו במקרה זה קטן מאד הוא לא מאד מייצג, ישנו סיכוי סביר שתוצאה זו התקבלה במקרה ושעל הפילוג האמיתי דווקא המודל מסדר 0 יכליל יותר טוב.

4

נחזור על הבחירה של סדר המודל בעזרת leave-one-out cross validation. במקרה זה אנו נחזור על החישוב של הסעיף הקודם 3 פעמים כשבכל פעם אנו בוחרים נקודה אחרת מה train set שתשמש כ validation set. את ביצועים של כל אחד מהמודלים נחשב בתור הממוצע על שלושת החזרות.

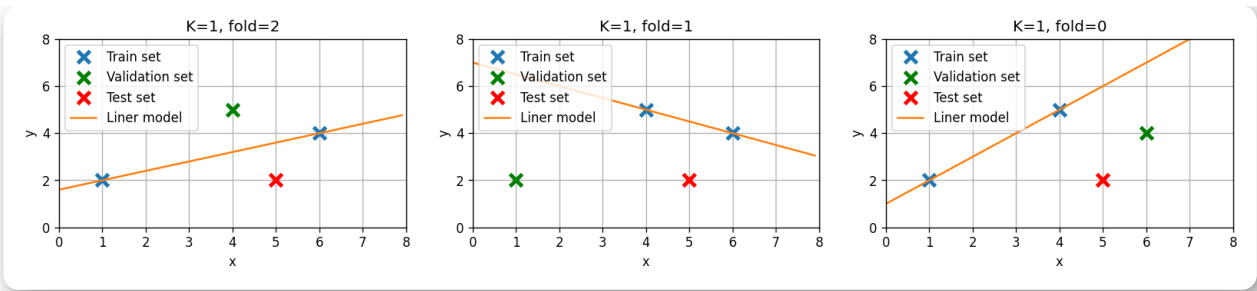
סדר 0



- **Fold 0:** $\theta^* = 3.5$. שיגאת חיזוי: 0.5
- **Fold 1:** $\theta^* = 4.5$. שיגאת חיזוי: 2.5
- **Fold 2:** $\theta^* = 3$. שיגאת חיזוי: 2

שיגאת חיזוי ממוצעת: $5/3$

סדר ראשון



- 3. שיגאת חיזוי: **Fold 0:** $\theta^* = [1, 1]^T$ •
- 4.5. שיגאת חיזוי: **Fold 1:** $\theta^* = [7, -0.5]^T$ •
- 1.8. שיגאת חיזוי: **Fold 2:** $\theta^* = [1.6, 0.4]^T$ •

שיגאת חיזוי ממוצעת: 3.1

על פי leave-one-out נראה שוב כי המודל מסדר 0 הוא העדיף. מכיוון ששיטה זו לא מסתמכת על נקודה אחת לקביעת סדר המודל ישנו סיכוי טוב יותר שה hyper-parameters אשר נבחרים בשיטה זו יניבו מודל אשר מכליל בצורה טובה יותר

דוגמא מעשית - חיזוי זמן נסיעה של מוניות בניו יורק

Code



נחזור לבעיה מהתרגול הקודם של חיזוי זמן הנסיעה של מוניות בניו יורק בעזרת המדגם הבא:

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

ay of ek	duration	dropoff northing	dropoff easting	pickup northing	pickup easting	tip amount	fare amount	payment type	trip distance	passenger count
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

בסוף התרגול הקודם השתמשנו במודל מהצורה של

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

אשר כולל תלות ליניארית במרחק האוירי שאותו צריכה המונית לעבור ותלות ריבועית בקואורדינטה של נקודת תחילת הנסיעה

הערכת הביצועים

נתחיל לחלק את המדגם ל 80% train set ול 20% test set. נשתמש ב train set על מנת לקבוע את הפרמטרים של המודל ונשערך את שגיאת ה RMSE על ה train set ועל ה test set. לאחר חישוב הפרמטרים והערכת הביצועים נקבל:

$$\text{RMSE}_{\text{train}} = 5.13 \text{ min}$$

$$\text{RMSE}_{\text{test}} = 5.16 \text{ min}$$

זאת אומרת שאנו צופים שנדע לחזות את זמן הנסיעה (על נסיעות שלא ראינו לפני) בדיוק של ± 5.16 דקות.

מודל פולינומיאלי

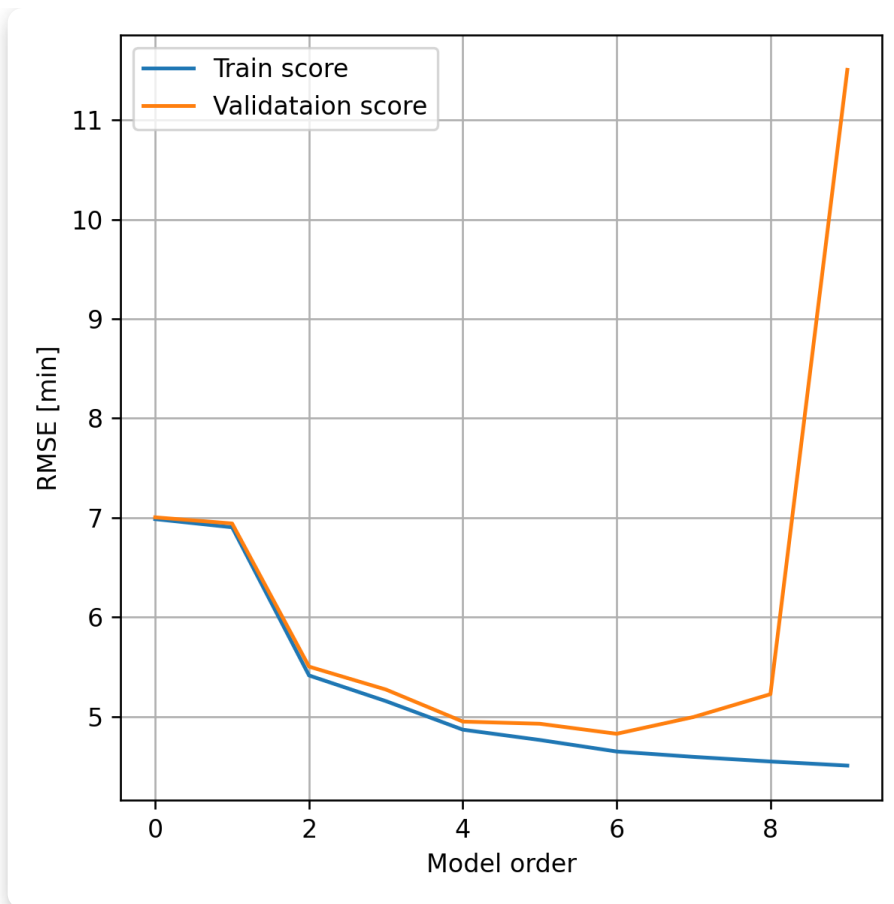
ננסה כעת להתאים מודל שהוא פולינום של קואורדינטת ההתחלה וקואורדינטת הסיום:

$$h(\mathbf{x}; \theta) = \theta_1 + \theta_2 x_{\text{pick east}} + \theta_3 x_{\text{pick north}} + \theta_4 x_{\text{drop east}} + \theta_5 x_{\text{drop north}} + \theta_6 x_{\text{pick east}}^2 + \theta_7 x_{\text{pick north}}^2 + \theta_8 x_{\text{drop east}}^2 + \theta_9 x_{\text{drop north}}^2 + \theta_{10} x_{\text{pick east}} x_{\text{pick north}} + \theta_{11} x_{\text{pick east}} x_{\text{drop east}} + \dots$$

קביעת סדר המודל

על מנת לקבוע את סדר המודל (החזקה המקסימאלית של הפולינום) נשתמש ב validation set. נפצל את ה train set ל 75% train set חדש ו 25% validation set.

נסרוק כעת את כל המודלים עד לסדר 9, בעבור כל סדר נאמן מודל על ה train set ונחשב את ביצועי המודל על ה validation set. חישוב זה נותן את התוצאה הבאה:



בגרף זה ניתן לראות את ה tradeoff בבחירת סדר המודל ואת תופעת ה overfitting. בקצה השמאלי של הגרף נמצאים המודלים ה"פשוטים" (פונקציה קבועה, פונקציה ליניארית וכו') באיזור זה השיגאה העיקרית היא שגיאת הקירוב (או ה bias). בקצה הימני נמצאים פולינומים בעלי מספר רב של מקדמים אשר מסוגלים לקרב מגוון רחב של מודלים. באיזור זה השיגאה העיקרית היא שגיאת השיערוך (או ה variance). אנו נחפש את סדר המודל האופטימאלי אשר נותן את שגיאת ההכללה הנמוכה ביותר. בקירוב זהו הסדר שנותן את הביצועים הטובים ביותר על ה validation set.

על פי תוצאה זו אנו נבחר את סדר המודל להיות 6.

אימון מחדש של המודל

כעת נאחד חזרה את ה validation set וה train set ונאמן מחדש את המודל וזה יהיה המודל הסופי בו נשתמש. נערך את ביצועי המודל הסופי בעזרת ה test set. חישוב זה נותן תוצאה של:

$$RMSE_{\text{train}} = 4.79 \text{ min}$$

$$RMSE_{\text{test}} = 4.81 \text{ min}$$

קיבלנו שיפור של כמעט 10% לעומת המודל שממנו התחלנו.

אופציה אלטרנטיבית - רגולריזציה

ניתן לחילופין לקבוע את סדר המודל להיות 9 ולהשתמש באיבר רגולריזציה על מנת למזער את ה overfitting. שימוש ב Ridge regression (רגולריזציה l_2) נותן את הביצועים הבאים:

$$RMSE_{\text{train}} = 4.82 \text{ min}$$

$$RMSE_{\text{test}} = 4.85 \text{ min}$$

אשר מאד קרובים לביצועים שקיבלנו בעבור מודל מסדר 6.