פרק 7: מבוא לבעיית הרגרסיה

מקור: HTF: 2.7-2.9, 7.2, 7.10

5.1 הגדרת בעיית הרגרסיה

בעיית הרגרסייה עוסקת בלימוד קשר פונקציונלי בין האלמנטים במרחב הקלט ($x \in X$) לאלמנטים במרחב הפלט ($y \in Y$).

כמו בעיית הסיווג, בעיית הרגרסיה שבה נעסוק בקורס זה הינה בעייה של למידה מודרכת: הקשר המבוקש בין הקלט לפלט נלמד מתוך אוסף דוגמאות $\{x_k, y_k\}$, כאשר y_k מציין את הערכת ה"מדריך" עבור הקלט המתאים לפלט x_k, דהיינו ההערכה הניתנת ללומד.

ההבדל בין בעיית הסיווג לבעיית הרגרסיה הינו באופי <u>מרחב הפלט</u>: בבעיית הסיווג מרחב זה הינו דיסקרטי וסופי, ולרוב חסר מבנה, ואילו בבעיית הרגרסיה הפלט הוא מספר – לרוב מספר ממשי – השייך למרחב מטרי בעל יחס סדר. בהתאם לכך, קיים הבדל במדדי הביצועים (קריטריוני השגיאה) המתאימים לשתי בעיות אלו.

<u>מרכיבים בסיסיים:</u> המרכיבים הבסיסיים של בעיית הרגרסיה הינם, אם כן:

X – מרחב הקלט. באופן טיפוסי X וקטור רב מימדי עם רכיבים ממשיים, דהיינו

$$x = (x_1, \ldots x_d), \quad x_i \in \mathbb{R}$$

Y – מרחב הפלט. באופן טיפוסי Y הוא המרחב הממשי $\mathbb{R}$ (או תת-קבוצה שלו).

$D = \{x_k, y_k\}_{k=1}^n$ – סדרת האימון (training), הכוללת דוגמאות קלט מתויגות.

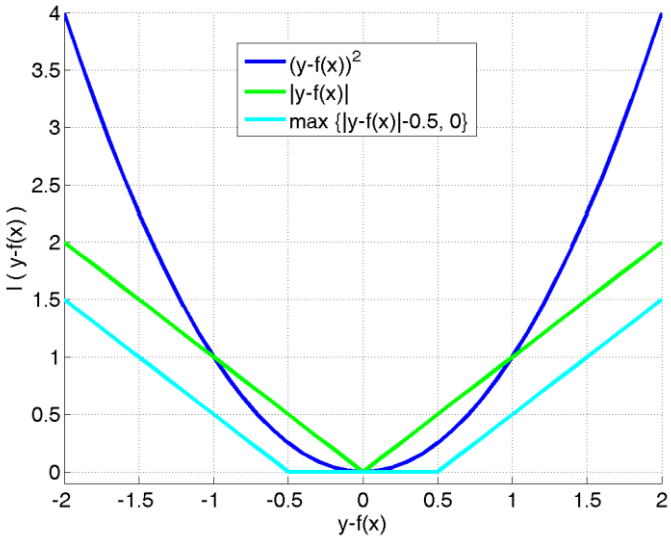
<u>מדידת השגיאה לדוגמא בודדת:</u> מטרת אלגוריתם הלמידה למצוא פונקציה $f(x)$ שערכיה יהיו קרובים ככל האפשר לערך הפלט y המתאים לקלט x . לצורך כך יש צורך להגדיר פונקציה דו-מקומית $l(y, f(x))$ המכמתת עד כמה שונה הפלט החזוי $f(x)$ מן הפלט הרצוי או הנכון y פונקציה זו מקבלת ערכים אי-שליליים בלבד וערכה שווה לאפס, כאשר מתקיים שוויון $f(x) = y$

לרוב, בבעיות ריגרסיה ערכה תלוי במשתנה יחיד $l(y, f(x)) = l(y - f(x))$

דוגמאות:

- שגיאה ריבועית: $l(y - f(x)) = \left(y - f(x)\right)^2$
- שגיאה אבסולוטית: $l(y - f(x)) = \left|y - f(x)\right|$
- שגיאה אבסולוטית, אדישה חלקית $l(y - f(x)) = \max\left\{\left|y - f(x)\right| - \varepsilon, 0\right\}$

שרטוט פונקציות השגיאה מופיע בתרשים הבא:



<u>מדידת השגיאה לכלל הדוגמאות</u>: קיים צורך להעריך איכות של פונקציות $f(x)$ עבור כלל הדוגמאות ולא כל אחת בנפרד. אפשרות אחת היא לקחת את הדוגמא הגרועה ביותר

$$l(f) = \sup_{(x,y)} l(y - f(x))$$

אלא שזו מתמקדת במקרה הגרוע ביותר, שלא תמיד הוא מאפיין ביחד לשאר הדוגמאות.

אפשרות שניה היא למצע, $l(f) = \int l(y - f(x))dP$. אפשרות זו לוקחת בחשבון את ההתנהגות האופיינית של השגיאה ביחס לכלל הדוגמאות, ולרוב קלה יותר לשימוש ואנליזה.

לרוב, המידה P אינה ידועה, ולכן מקרבים את הגודל האחרון על ידי מדגם סופי.

נניח עתה כי נתונה <u>סדרת בוחן</u> (test) $D = \{x_k, y_k\}_{k=1}^n$. סדרה זו אינה בהכרח זהה לסדרת הלימוד (כפי שנראה בהמשך). עבור פונקצית רגרסיה נתונה $f(x)$, נגדיר את <u>השגיאה הכוללת</u> (Sum of Errors) אופן הבא:

$$E_{SE}(f) = \sum_{k=1}^{n} l(y_k - f(x_k))$$

מדד זה נקרא גם <u>השגיאה האמפירית</u> ביחס לסדרת הבוחן D.

שני מקרים מיוחדים הם:

<u>השגיאה הריבועית הכוללת</u> (Sum of Square Errors) אופן הבא:

$$E_{SSE}(f) = \sum_{k=1}^{n} (y_k - f(x_k))^2$$

מדד זה נקרא גם <u>השגיאה האמפירית הריבועית</u> ביחס לסדרת הבוחן D.

מדד שגיאה מקובל נוסף הינו <u>סכום ערכי השגיאות</u> (Sum of Absolute Errors):

$$E_{SAE}(f) = \sum_{k=1}^{n} | y_k - f(x_k) |$$

מדד זה נקרא גם <u>השגיאה האמפירית המוחלטת</u> ביחס לסדרת הבוחן D. שם נוסף הינו שגיאת-ℓ_1, כיוון שניתן לראותו כנורמת-ℓ_1 של סדרת השגיאות.

ניתן עתה להגדיר את מטרת הלימוד כמציאת פונקצית רגרסיה $\hat{f}(x)$ שתביא למינימום את מדד השגיאה הנבחר.

5.2 הגישה הפרמטרית

הגישה הנפוצה ביותר לפתרון בעיית הרגרסיה הינה הגישה הפרמטרית. בגישה זו, אנו בוחרים ראשית משפחה פרמטרית של פונקציות, דהיינו משפחת פונקציות רגרסיה מהצורה $f(x,\theta)$, כאשר θ מציין וקטור פרמטרים (ממשיים) הניתנים לכיוונון. משפחה זו נקראת גם המודל. הלמידה מתבצעת על ידי קביעת וקטור הפרמטרים θ.

כיצד נבחר (נכוונן או נקבע) את הפרמטר θ? גישה פשוטה ומקובלת לכך הינה בחירה של θ המביא לערך קטן ככל האפשר של השגיאה האמפירית המתאימה ביחס לסדרת הלימוד. למשל:

$$E_{SSE}(\theta)=\sum_{k=1}^{n}(y_k-f(x_k,\theta))^2 \quad \rightarrow \quad \min \text{ over } \theta$$

כפי שנראה בהמשך, בגישה זו כוללת מספר בעיות שיש להתמודד עימן בזהירות.

מודלים מקובלים (עליהם נתעכב בהמשך) כוללים את הבאים.

א. מודלים ליניאריים בקלט. המודל הפשוט ביותר הינו המודל הבא

$$\hat{f}(x,\theta)=w_0+w_1x_1+\ldots+w_dx_d$$

כאשר $\theta=(w_0,w_1,\ldots,w_d)$ הוא וקטור הפרמטרים. מודל זה ליניארי הן בפרמטרים והן ברכיבי הקלט x. הפרמטרים w_i נקראים גם המשקלים, ובהתאם הסימון w.

מטרת תהליך הלימוד עבור מודל זה הינה מציאת הפונקציה הליניארית המתארת באופן הטוב ביותר את הקשר בין הקלט לפלט. יתרונו של המודל הליניארי הינה בפשטות החישוב: עבור פונקציית שגיאה ריבועית, ניתן לקבל נוסחה סגורה עבור הפרמטרים האופטימאליים.

ב. מודלים ליניאריים מוכללים: צרוף ליניארי של פונקציות בסיס. במקרים רבים, ההגבלה לפונקציות ליניאריות של הקלט הינה קשה מדי. למרבה המזל, ניתן להרחיב את יכולת התיאור של המודל בקלות יחסית מבלי לאבד את יתרונותיו החישוביים, וזאת על ידי שימוש בפונקציות בסיס. מודל זה מתואר על ידי:

$$\hat{f}(x,\theta) = w_1\phi_1(x) + \ldots + w_M\phi_M(x) = \sum_{m=1}^{M} w_m\phi_m(x)$$

הפונקציות $\phi_m(x)$ הינן פונקציות הנבחרות מראש, וקרויות פונקציות הבסיס. המודל $f(x,\theta)$ הינו, לפיכך, צרוף ליניארי של פונקציות בסיס.

בחירות אפשריות של פונקציות בסיס הינן פולינומים, פונקציות הרמוניות (טורי פורייה רב-ממדיים), פונקציות בסיס רדיאליות, ועוד (שיוזכרו בהמשך).

ג. מודלים לא- ליניאריים. הכוונה פה מודלים שאינם ליניאריים בפרמטרים. קיימות משפחות שונות של מודלים שהם בעלי תכונות רצויות לבעיות סיווג ורגרסיה. דוגמא חשובה למודל לא לינארי כזה הינו רשת ניירונים מלאכותית. במקרים אלה לא ניתן לחשב באופן אנליטי את הערך של θ המביא למינימום את השגיאה האמפירית, ויש להיעזר באלגוריתמי אופטימיזציה נומריים, כגון אלגוריתם הגרדיאנט.

5.3 שיטות גרעין ורגרסיה מקומית

שיטות מבוססות גרעין הינן שיטות א-פרמטריות אשר מבוססות על מתן הערך $\hat{f}(x_0)$ בנקודה רצויה x על ידי מיצוע ערכי סדרת הלימוד בנקודות סמוכות לנקודה המבוקשת. ניתן לראות גישות אלו כהכללה או התאמה של שיטות Nearest Neighbor למקרה של רגרסיה.

מסווג השכן הקרוב: תהי $\{x_k, y_k\}_{k=1}^{n}$ סדרת הלימוד, אותה אנו שומרים בזיכרון. בהינתן קלט חדש x, נמצא את תבנית הקלט x_k הקרובה ביותר ל- x, ונסווג את x בהתאם לפלט של x_k:

$$f_{NN}(x) = y_{k(x)}$$

כאשר

$$k(x) = \arg\min_{k=1,\ldots,n} d(x, x_k)$$

ואילו $d(x, x_k)$ הוא המרחק בין x ל- x_k.

ניתן כמובן להכליל גישה זאת ולקחת ממוצע של יותר משכן אחד, או אפילו באמצעות קרוב פונקציה באמצעות גרעין היא הממוצע המשוקלל הבא (ממוצע Nadaraya-Watson):

$$\hat{f}(x_0) = \frac{\sum_{k=1}^{n} K(x_0, x_k) y_k}{\sum_{k=1}^{n} K(x_0, x_k)}$$

פונקציית הגרעין $K(x_0,x)$ היא פונקצייה המרוכזת סביב x_0 ודועכת כאשר x רחוק מ- x_0.
הגרעין הגאוסי הבא הוא בין פונקציות הגרעין הנפוצות ביותר:

$$K(x_0,x)=\frac{1}{\lambda}\exp\left[-\frac{\|x-x_0\|^2}{2\lambda}\right]$$

הפרמטר λ שולט על על רוחב הגאוסיאן, וכפי שכבר ראינו פרמטר זה הוא בעל השפעה מכרעת על טיב הקרוב המתקבל.

ניתן לראות כי גישת הפתרון פה היא של Lazy Learning .

<u>רגרסיה מקומית</u>: שיטה נוספת המשלבת בין התכונות המקומית של פונקציות הגרעין לגישה הפרמטרית היא שיטת הרגרסיה המקומית. בגישה זו המשערך של $f_0(x_0)$ נבחר באופן פרמטרי באמצעות מודל מקומי $f(x,\theta)$, דהיינו $\hat{f}(x_0)=f(x_0,\hat{\theta})$ כאשר הפרמטר θ מביא למינימום את השגיאה הריבועית המקומית:

$$E(\theta,x_0)=\sum_{k=1}^{n}K(x_0,x_i)(y_k-f(x_k,\theta))^2 \quad\rightarrow\quad \min \text{ over } \theta$$

$f(x,\theta)$ היא באופן טיפוסי פולינום מסדר נמוך ב- x. נציין כי:

- עבור $f(x,\theta)=\theta$ מתקבל משערך הגרעין הבסיסי (Nadaraya-Watson).
- עבור $f(x,\theta)=\theta_0+\theta^T x$, דהיינו פונקציה לינארית, מתקבלת שיטה נפוצה הידועה בשם Locally Linear Regression).

5.4 ניסוח סטטיסטי של בעיית הרגרסיה

<u>מודל סטטיסטי:</u> כחלק מהגדרת ופתרון הבעייה נוח לעיתים להניח מודל מסוים עבור הקשר בין הקלט x והפלט y. המודל הסטטיסטי הפשוט ביותר הינו מודל הרעש האדיטיבי:

$$y=f_0(x)+e$$

כאשר $f_0(x)$ פונקציה דטרמיניסטית (אך לא ידועה), ואילו e מציין שגיאה אקראית, בלתי תלויה ב- x, ובעלת ממוצע 0: $E(e)=0$. לפיכך $E(y|x)=f_0(x)$. מודל זה קרוי מודל

רגרסיה. בבעיה הלומדת, מטרתנו ללמוד פונקציה $\hat{f}(x)$ אשר מהווה קרוב טוב ככל האפשר לפונקציה $f_0(x)$. לימוד זה יתבצע מתוך סדרת הלימוד D, ובסיוע ידע מוקדם לגבי הבעיה.

הפונקציה הנלמדת $\hat{f}(x)$ תיקרא גם פונקצית הרגרסיה, או המודל הנלמד. הערך $\hat{f}(x_0)$ בנקודה נתונה x_0 ייקרא המקרב או המשערך של $f_0(x_0)$ בנקודה זו.

כפי שכבר הזכרנו, המסגרת הבסיסית של בעיית הרגרסיה היא המודל ההסתברותי הבא:

$$y = f_0(x) + e$$

כאשר $f_0(x)$ פונקציה לא ידועה, ו- e הינו רעש אדיטיבי.

כאשר $f_0(x)$ נתונה בצורה פרמטרית והסטטיסטיקה של e ידועה, הבעיה הופכת לבעייה סטנדרטית של שערוך פרמטרים סטטיסטי. נדגים זאת בקצרה על ידי חישוב משערך סבירות המירבית למקרה זה (גישה לא בייסיאנית). ניתן כמובן להשתמש פה גם בשיטות בייסיאניות.

נתבונן במודל הרגרסיה הבסיסי, בצורתו הפרמטרית:

$$y = f(x,\theta) + e$$

כאשר x וקטור הפלט, y משתנה הפלט, $f(\cdot,\theta)$ משפחת פונקציות כלשהי (בפרמטר θ), θ פרמטר לא ידוע, ו- e רעש בעל פילוג נתון. ביתר פירוט, נניח כי סדרת הדוגמאות $\{x_k, y_k\}_{k=1}^n$ מקיימת את הקשר

$$y_k = f(x_k,\theta) + e_k$$

כאשר $\{e_k\}$ סדרה של משתנים בלתי תלויים, בעלי פילוג שולי זהה $p_e(\cdot)$ (ובלתי תלוי ב- x).
משערך הסבירות המירבית יוגדר פה באופן הבא:

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} L(\theta)$$

כאשר פונקצית הסבירות $L(\theta)$ מוגדרת על ידי:

$$L(\theta) = p\left(\{y_k\}_1^n \mid \{x_k\}_1^n, \theta\right) = \prod_{k=1}^n p(y_k \mid x_k,\theta) = \prod_{k=1}^n p_e\left(y_k - f(x_k,\theta)\right)$$

לדוגמא, אם $p_e(\cdot)$ היא בעלת פילוג נורמלי, כלומר $e_k \sim N(\mu_e,\sigma_e^2)$, נקבל כי:

$$p_e\left(y_k - f(x_k,\theta)\right) = \frac{1}{\sqrt{2\pi}\sigma_e}\exp(-\frac{1}{2\sigma_e^2}(y_k - f(x_k,\theta))^2)$$

ולכן

$$L(\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma_e}\right)^n \exp\left\{-\frac{1}{2\sigma_e^2}\sum_{k=1}^{n}(y_k - f(x_k,\theta))^2\right\}$$

מכאן נובע כי מקסימיזציה של $L(\theta)$ שקולה למינימיזציה של הסכום בחזקה, כלומר

$$\hat{\theta}_{MLE} = \arg\min_{\theta\in\Theta}\sum_{k=1}^{n}(y_k - f(x_k,\theta))^2$$

נזכור (מסעיף 3.1) כי הסכום האחרון הינו למעשה השגיאה הריבועית הכוללת $E_{SSE}(\theta)$. קיבלנו כי <u>במקרה של שגיאה גאוסית, משערך הסבירות המירבית מתקבל ע"י מינימיזציה של השגיאה הריבועית הכוללת</u>, $E_{SSE}(\theta)$. אבחנה זו נותנת מוטיבציה (נוספת) לשימוש בקריטריון השגיאה הריבועית להערכת הפרמטרים.

5.5 בעיות יסוד בלמידה מודרכת

נתאר עתה מספר בעיות בסיסיות הקשורות ללמידה מודרכת, דהיינו למידה אינדוקטיבית מתוך סדרת דוגמאות. לשם קונקרטיזציה נתייחס בעיקר לבעיית הרגרסיה הפרמטרית, אולם הבעיות המוזכרות אינן ייחודיות למקרה זה.

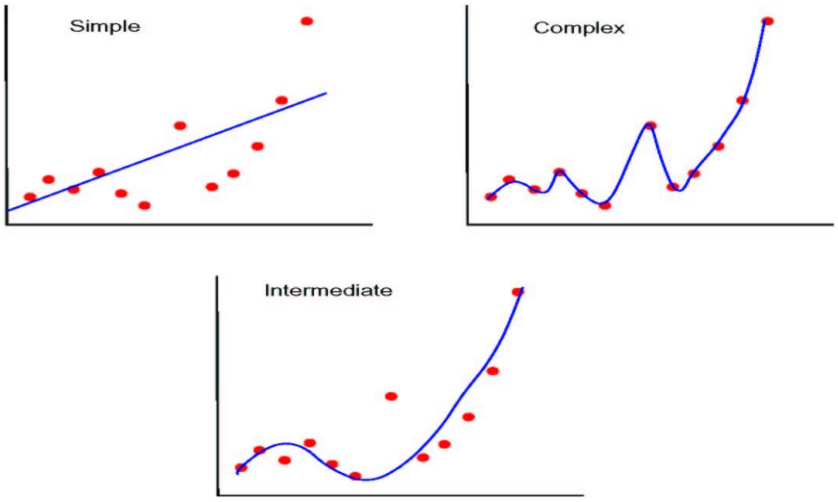
א. בחירת המודל

בחירת אופי המודל (רשת ניירונים, מודל ליניארי, מודל א-פרמטרי) המתאים לבעיה מסוימת הינה בעיה ללא פתרון חד-משמעי. בחירת המודל תסתמך בעיקר על ניסיון קודם עם בעיות דומות, השוואת ביצועי מודלים וגישות שונות לבעיה הקיימת, שיקולים חישוביים, והעדפה אישית.

ב. בחירת סדר המודל

סדר המודל (מספר הפרמטרים) קובע, כעיקרון, את הסיבוכיות האפשרית של משפחת הפונקציות אותן המודל עשוי לתאר. הדילמה הבסיסית בבחירת סדר המודל (או ביתר דיוק סיבוכיות המודל) היא כלהלן:

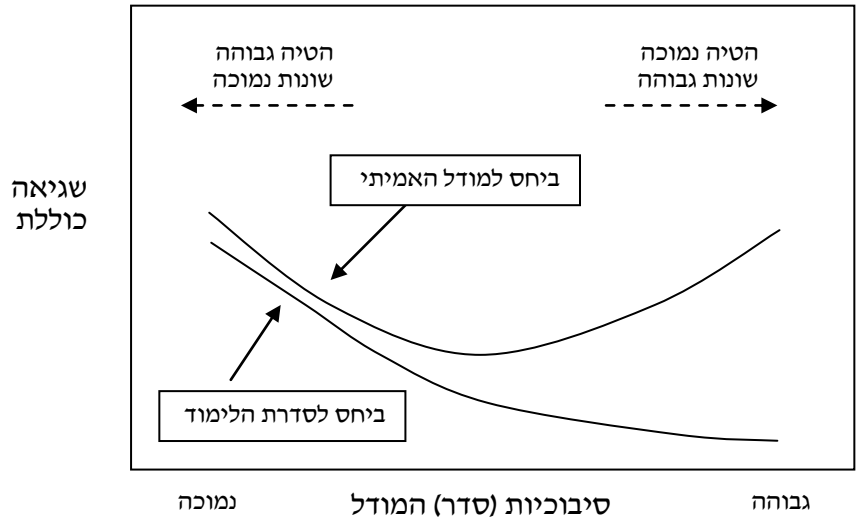
- מודל פשוט מדי (בעל סדר נמוך) עלול לא לאפשר תיאור מדויק של הקשר ה"אמיתי" בין הקלט לפלט. למשל: מודל ליניארי ייתן התאמה גרועה עבור פונקציה ריבועית.
- מודל מסובך מדי (בעל סדר גבוה) עלול לדרוש מספר רב מאוד של דוגמאות (או לחילופין זמן לימוד ארוך כאשר מספר הדוגמאות אכן גדול) על מנת לבצע הכללה סבירה. לדוגמא: להתאמת מודל ריבועי לפונקציה ריבועית (סקלרית) נדרשות 3 דוגמאות. התאמה טובה של מודל פולינומיאלי מסדר 700 יידרש מספר גדול בהרבה.



הניגוד בין שני קטבים אלה קרוי הדילמה של הטיה לעומת שונות **(Bias-Variance Tradeoff)**:

- ההטיה (bias) מציינת את המרחק המינימלי האפשרי בין פונקציה כלשהי מתוך המודל לפונקציה הנלמדת האמיתית (המרחק יימדד על ידי נורמה מתאימה, למשל כאינטגרל על השגיאה הריבועית הנקודתית). ההטיה תקטן ככל שסדר המודל גדל.
- השונות (variance) מציינת את המרחק בין הפונקציה האופטימאלית מתוך המודל – זו שנותנת את ההתאמה הטובה ביותר לפונקציה האמיתית – לבין זו שהתקבלה בפועל לאחר התאמה לסדרת הלימוד. השונות תגדל ככל שסדר המודל גבוה יותר.

המרחק בין המודל שנלמד לפונקציה האמיתית מתקבל, לפחות עקרונית, על ידי סיכום מרכיבים אלה. אופי התלות של מרחק זה בסדר המודל מודגם באיור הבא (לפי ציור 2.11 בספר HTF:2001), עבור סדרת לימוד בעלת אורך נתון.



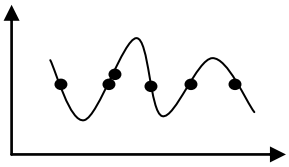
סדר המודל האופטימלי לבעיה זו הוא בנקודת המינימום של עקום ההתאמה למודל האמיתי. יש לציין כי השונות קטנה, עקרונית, ביחס הפוך למספר הדוגמאות בסדרת הלימוד, בעוד ההטיה (לפי הגדרתה) אינה תלויה במספר זה. לכן ככל שמספר הדוגמאות גדל, נקודת האופטימום תתקבל עבור סדר מודל גבוה יותר.

<u>הערה</u>: יש לציין כי הקטנת סדר המודל (מספר הפרמטרים) אינה הדרך היחידה להקטנת הסיבוכיות האפקטיבית של המודל. דרך נוספת, ובעלת חשיבות רבה, הינה הטלת הגבלות על גודל ותחום השינוי של הפרמטרים, או על תכונות שונות של הפונקציה עצמה כגון "חלקות" (גודל הנגזרת הראשונה או השניה). בפרט, בגישת <u>הרגולריזציה הפרמטרית</u> מוסיפים לשגיאה האמפירית איבר נוסף אשר מטיל "קנס" על גודל הפרמטרים, כך שהוא גורם להטיה לכיוון של ערכי פרמטרים נמוכים. בגישה זו מתאפשרת שליטה רציפה בפשרה (tradeoff) שבין גודל הפרמטרים להתאמה לסדרת הלימוד.

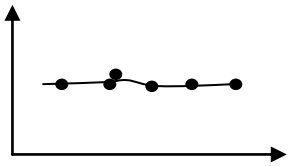
ג. התאמת יתר

אבחנה נוספת הקשורה במודלים בעלי סדר גבוה הינה התאמת היתר. התאמת יתר פרושה התאמה מדויקת של המודל לסדרת הלימוד, אשר מובילה להתאמה גרועה ל"פונקציה האמיתית" (דהיינו: לדוגמאות חדשות). תופעה זו תתבטא בשונות גבוהה עבור מודלים בעלי סדר גבוה, כפי שכבר צוין לעיל.

תופעת התאמת היתר בולטת במיוחד כאשר תהליך הלימוד הוא הדרגתי (איטרטיבי) כך שהתאמת המודל לסדרת הלימוד משתפרת עם זמן הלימוד. התופעה מודגמת בציור הבא:

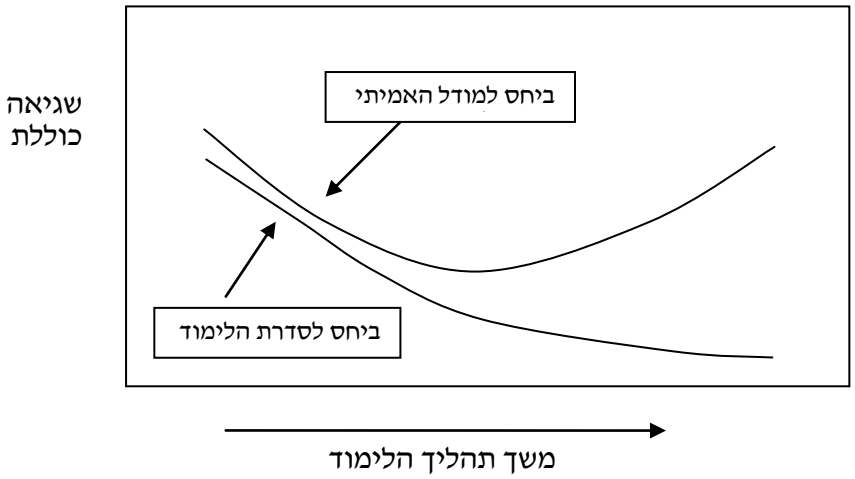


התאמה מושלמת (סיום תהליך הלימוד)
לקו ישר "רועש"



התאמה חלקית (בתחילת תהליך הלימוד)
לקו ישר "רועש"

את תופעת התאמת היתר כתלות בזמן הלימוד ניתן לתאר איכותית בעזרת האיור הבא:



5.6 הערכת ביצועים באמצעות קבוצות בוחן

השאלות המתבקשות מהדיון בסעיף הקודם הן: כיצד לבחור מודל מבין מספר אפשרויות; כיצד לבחור את סדר המודל; וכיצד להימנע מתופעת התאמת היתר. הבעיה כמובן שהמודל האמיתי אינו ידוע, כך שלא ניתן למדוד את השגיאה ביחס למודל זה ולבחור בסדר המודל האופטימאלי, או לעצור את תהליך הלימוד לאחר משך הזמן האופטימאלי.

שאלות אלו, ובפרט נושא בחירת סדר המודל, נחקרו באופן עמוק המסגרות תיאורטית שונות. עם זאת, הגישה הפשוטה ביותר ליישום והבנה היא אמפירית באופייה, ומתבססת על הערכת השגיאה ביחס ל"מודל האמיתי" בעזרת סדרת בוחן.

<u>קבוצות אימות ובוחן</u>: הגישה הפשוטה ביותר לבדיקת ביצועים היא לחלק את קבוצת הדוגמאות הקיימת למספר תת-קבוצות, כאשר הראשונה משמשת ללימוד (training set), והאחרות לבחירת וכיוונון המודל ובדיקת ביצועים סופית. חלוקה מקובלת היא כלהלן:

א. סדרת הלימוד (Training set) – משמשת לכיוונון הפרמטרים (ע"י מזעור השגיאה האמפירית וכו').

ב. סידרת האימות (Validation set) – משמשת לבחירת מודל וסדר מודל

ג. סידרת הבוחן (Test set) – משמשת להערכת ביצועים סופית.

חשוב להדגיש כי אין "לערבב" בין קבוצות אלו. שימוש בסדרת האימות, למשל, לצורך כיוונון הפרמטרים עשוי להוביל להערכה אופטימית של השגיאה, ובהתאם לבחירה בסדר מודל גבוה והתאמת יתר. שימוש בסדרת הבוחן באחד משלבי הלימוד יוביל להערכה אופטימית של ביצועי המערכת.

חלוקת הדוגמאות הקיימות לקבוצות השונות תתבצע באופן אקראי, כדי למנוע חוסר איזון באופי הדוגמאות. כלל אצבע לבחירת גודל הקבוצות הוא: 50% (לימוד), 25% (אימות), 25% (בוחן). יחסים אלה עשויים להשתנות בהתאם למידת הרעש הצפויה במידע, ובהתאם למספר הכולל של הדוגמאות הזמינות.

גישה זו פשוטה למימוש וחסכונית בזמן. אולם כאשר מספר הדוגמאות הינו קטן, הויתור על 50% מתוכן בשלב הלימוד הופך להיות משמעותי. במקרה זה ניתן להשתמש בגישה הבאה.

<u>אימות צולב (Cross Validation)</u>: בגישה זו קבוצת הדוגמאות מחולקת ל-K קבוצות זרות ושוות גודל (בקירוב). עבור כל אחת מקבוצות אלו ($i = 1,\ldots,K$) מתבצע התהליך הבא:

1. שלב הלימוד מתבצע על סט הדוגמאות הכולל, <u>פרט</u> לקבוצה ה- i.

2. השגיאה האמפירית של המסווג שהתקבל מחושבת על פני הקבוצה ה- i.

בסיום התהליך מחושבת השגיאה הכוללת כסכום K השגיאות שהתקבלו. זו השגיאה בה נשתמש לבחירת המודל וסדר המודל.

בחירה אופיינית עבור K הינה 5 עד 10. במקרה הקיצוני שבו $K = n$ (כמספר הדוגמאות הכולל), כל קבוצה תכלול דוגמא אחת בלבד. מקרה זה מכונה:

. Leave-one-out cross-validation