**בעיות רגרסיה (Regression)**

בעיות רגרסיה עוסקות בחיזוי ערכים מספריים רציפים על בסיס קשרים בין משתני קלט (Features) למשתנה פלט Target)). כמו למשל חיזוי מחיר טיסה.



משוואת הרגרסיה הלינארית:

מקדמי הפיצרים (משקלים -W) - משקלים הם הפרמטרים שהמודל לומד במהלך תהליך האימון. כל משקל מייצג את מידת ההשפעה של תכונה ספציפית על התוצאה הסופית של המודל. בפשטות - הם "השיפועים" של הקו בכל ממד (תכונה).

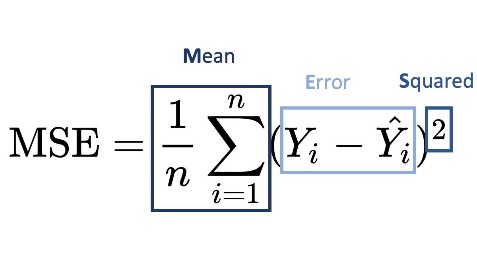
X – משתני הקלט

B- ביאס (Bias) מייצג את השאריות (Residuals) - השגיאה בין הערך החזוי לערך האמיתי.

הביאס הוא הערך הקבוע שמתווסף למודל כדי להזיז את הקו (או המישור) למעלה או למטה.

הביאס מאפשר למודל להתאים נתונים שאינם עוברים דרך הראשית (0,0).

y – במודל לינארי, הניבוי מחושב על ידי הסכום המשוקלל של התכונות בתוספת הביאס. כלומר - הערך אותו אנו חוזים.

מדדים להערכת ביצועי מודלים ברגרסיה:

* MSE (Mean Squared Error)

ממוצע ריבועי ההפרשים בין התחזיות לערכים האמיתיים.

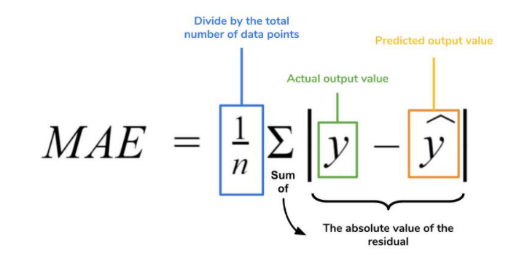
ככל שהMSE נמוך יותר – כך המודל נחשב ליותר מדויק.

הפרש גדול בין MSE של נתוני האימון לבין נתוני בדיקה מצביע על כך שהמודל לא סובל מoverfitting.

מטרת המדד לשים דגש על טעויות גדולות בזכות ריבוע ההפרשים.

חיסרון המדד הוא שהוא רגיש מאוד לערכים חריגים (outliers) - מעניק משקל גדול יותר לטעויות גדולות.

במצבים בהם יש הרבה ערכים חריגים שימוש בMSE עשוי להטות את השגיאה ולכן נעדיף להשתמש בMAE.



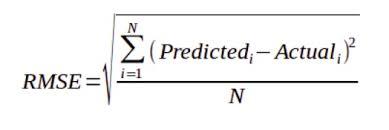
* MAE (Mean Absolute Error)

ממוצע ההפרשים המוחלטים בין הערכים החזויים לערכים האמיתיים.

מטרת המדד למדוד את ממוצע גודל הטעויות בין התחזיות לערכים בפועל, מבלי להתחשב בכיוון הטעויות.

חיסרון: אינו מעניש טעויות גדולות כמו MSE.

ערך MAE נמוך מצביע על כך שהמודל מנבא בצורה מדויקת, לפחות בממוצע, את הערכים האמיתיים.

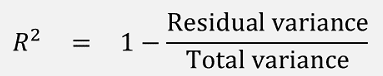
* RMSE (Root Mean Squared Error)

שורש ריבועי הממוצע של ה-MSE.

מודד כמה תחזיות המודל קרובות לערכים האמיתיים בממוצע ביחידת מידה של הנתונים המקוריים, מה שמקל על פרשנותו.

ערך RMSE קטן מעיד על התאמה טובה יותר של המודל לנתונים.

ערך RMSE נמוך גם על נתוני אימון וגם על בדיקה -> מעיד על מודל מדויק ועמיד בפני Overfitting

* R² (R-Squared)

המדד נקרא גם מדד פירסון - coefficient of determination

מדד המסביר את אחוז השונות בנתונים שהמודל מצליח להסביר בטווח שנע בין 0 ל-1.

ערך קרוב ל-1 מעיד על התאמה גבוהה.

* פונקציית העלות (Cost Function) - פונקציה שמטרתה למדוד את רמת הטעויות של המודל. לדוגמא - ברגרסיה לינארית: פונקציית העלות הנפוצה היא MSE.
* ירידת גרדיאנט (Gradient Descent):

אלגוריתם אופטימיזציה המשמש בתהליך הלמידה של מודלים כדי למזער את פונקציית העלות (Cost Function). פונקציית העלות מודדת את השגיאה בין חיזויי המודל לערכים בפועל. האלגוריתם פועל באופן איטרטיבי כדי לעדכן את מקדמי המודל (Coefficients) כך שהשגיאה תתקרב למינימום. בכל איטרציה, האלגוריתם מתקדם בכיוון השיפוע השלילי של פונקציית העלות עד שהשגיאה מתכנסת למינימום.

* קצב למידה (Learning Rate): קצב הלמידה (α) הוא פרמטר מרכזי המשפיע על עדכון מקדמי המודל בירידת גרדיאנט. הוא קובע את גודל הצעד שהאלגוריתם עושה בכל איטרציה. מספר ממשי (בדרך כלל בין 0 ל-1):

קצב למידה קטן מדי: תהליך הלמידה איטי מאוד.

קצב למידה גדול מדי: עלול לגרום לאי-יציבות ולאי-התכנסות.

ברגרסיה לינארית, ירידת גרדיאנט משמשת למזער את פונקציית העלות, שמודדת את השגיאה (למשל, שגיאה ריבועית ממוצעת - MSE). קצב הלמידה משפיע על אופן התקדמות האלגוריתם לעבר הפתרון האופטימלי (המינימום של פונקציית העלות). שימוש נכון בקצב למידה הוא חיוני כדי שהמודל יתכנס לשגיאה מינימלית באופן מהיר ויציב.

* Residual Plot - גרף השאריות

כלי המאפשר להעריך את התאמת המודל רגרסיה לנתונים.

בעזרת הגרף נוכל לראות את השאריות (Residuals) של מודל רגרסיה. השאריות הן ההפרשים בין הערכים הנצפים (observed values) לערכים החזויים (predicted values) על ידי המודל.

מודל טוב -> השאריות מפוזרות בצורה אקראית סביב הציר האופקי .

מודל בעייתי -> תבניות ברורות בשאריות עשויות להצביע על חוסר לינאריות או Overfitting

|  |  |
| --- | --- |
| חישוב השאריות - הפרש בין הערך האמיתי לערך החזוי | train\_residuals = y\_train - y\_train\_pred  test\_residuals = y\_test - y\_test\_pred |
| יצירת הגרף השאריות – הנקודות עבור האימון ועבור המבחן בנפרד (המבחן באדום). הנקודות עם שקיפות כדי שלא יעלו זו על זו. | plt.figure(figsize=(10, 5))  plt.scatter(y\_train\_pred, train\_residuals, alpha=0.7, label="Train Residuals")  plt.scatter(y\_test\_pred, test\_residuals, alpha=0.7, label="Test Residuals", color="red")  plt.axhline(y=0, color='black', linestyle='--', linewidth=1) |

בחירת הפיצ'רים (Features) לרגרסיה – קריטריונים:

1. שונות גבוהה – פיצ'רים עם שונות וסטיית תקן גבוהות מספקים יותר מידע, מכיוון שהם מבטאים שונות בין הדגימות. פיצ'רים עם שונות נמוכה או ערכים כמעט קבועים אינם תורמים למודל ולכן עדיף להסירם.
2. מתאם עם המשתנה המוסבר – פיצ'רים בעלי קורלציה גבוהה עם המטרה (Target) יכולים לשמש כמנבאים טובים. פיצ'ר ללא שונות לא יכול להיות במתאם עם המטרה ולכן אינו רלוונטי.