

# אנליסט אקוויטי, על רגרסיה לינארית כבר שמעת?

רגרסיה לינארית היא כלי לחיזוי ערך רציף על סמך הנחת קשר לינארי בין ה-Target (היעד) ל-Features (המאפיינים). יצאתי לראיין את רועי פולניצר בנושא

ברצון. למידה בהשגחה (Supervised Learning) עוסקת בחיזוי ערכו של

יעד (Target, משתנה מוסבר) אחד או יותר, כאשר היעד הוא משתנה מסוים שאותו אנו מבקשים לחזות. לפני שנתחיל אני רוצה לעשות אבחנה בין מאפיין (Feature, משתנה מסביר) לבין תווית (Label, ערכו בפועל של המשתנה המוסבר). בעוד שמאפיין הוא משתנה מסוים המשמש באלגוריתם של למידת מכונה, הרי שתווית היא ערכו של היעד. למשל אם אני רוצה לחזות שכר כיעד על סמך גיל כמאפיין אז משכורת של 50 אלף ש"ח בחודש זה תווית.



תודה על ההבהרה, רוב הכלכלנים והסטטיסטיקאים לא מכירים את הטרמינולוגיה החדשה הזו שנולדה בשנים האחרונות בערוגות של מדעי המחשב והנדסת תוכנה. הם ואני מכירים את הקונספטים הללו מתחום האקונומטריקה או

מדען הנתונים רועי פולניצר מכהן כמנכ"ל האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA). רועי בעל תואר M.B.A. במנהל עסקים (עם התמחות בבניית מודלים מתמטיים וסטטיסטיים) ותואר B.A. בכלכלה (עם התמחות בכלים ושיטות לאנליזה ותחקור מידע), שניהם מאוניברסיטת בן-גוריון בנגב ובהצטיינות. רועי מחזיק בכמה הסמכות מקצועיות רלוונטיות למדע נתונים ולמידת מכונה, ביניהן הסמכה מקצועית "מנהל סיכונים מוסמך" (CRM - Certified Risk Manager) מטעם האיגוד הישראלי למנהלי סיכונים (IARM - Israeli Association of Risk Managers) המעידה על כך שהמחזיק בה בקיא בפיתוח, יישום ותיקוף מודלים סטטיסטיים ואלגוריתמים מתמטיים כגון DT, NB ו-PCA לניהול סיכונים תפעוליים/ביטוחיים. בשנה האחרונה רועי ייסד את האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA - Professional Data Scientists' Israel Association) ולכן יצאתי לראיין אותו בנושא.

רגרסיה לינארית (Linear Regression) משמשת סטטיסטיקאים מזה שנים רבות. המתמטיקאי המפורסם קרל פרידריך גאוס הוא זה שהציע לראשונה, בערך בשנת 1800, את גישת הריבועים הפחותים (OLS - Ordinary Least Squared) העומדת בבסיס הרגרסיה הלינארית. בלמידת מכונה אין צורך להניח קשר לינארי היות ומרבית הכלים בלמידת מכונה מובילים למודלים לא-לינאריים מורכבים. למרות זאת, הרגרסיה הלינארית נותרה כלי חשוב בלמידת מכונה. זהו למעשה זהו אחד הכלים הראשונים המשמשים אנליסטים בלמידה בהשגחה. רגרסיה לינארית פשוטה עוסקת במזעור השגיאה הריבועית הממוצעת (MSE) כאשר ערכו של היעד (המשתנה המוסבר) נחזה על סמך מאפיין (משתנה מסביר) אחד או יותר. בראיין זה רועי פולניצר יסביר כיצד ניתן להכניס מאפיינים קטגוריאליים (Categorical Features, מאפיינים שאינם נומריים) לתוך הרגרסיה לינארית על מנת שאלו יישמשו לחיזוי וידבר על רגרסיות מסוג Ridge, Lasso ו-Elastic Net אשר שימושיות במיוחד כאשר החיזוי מתבצע על סמך מספר רב של מאפיינים.

הזכרתי בפתיח שלי את המונח למידה בהשגחה, האם תרצה להסביר אותו לקוראים?

ההתקבלה בשלב השני. ובשלב הרביעי התוכנה אומדת את ה- Standard Error של הרגרסיה כשורש הריבועי של התוצאה שהתקבלה בשלב השלישי.

### ולא יוצאת אותה תוצאה בשתי השיטות?

לא. היות והתוכנת הללו מורידות 1 ממספר התצפיות לפני שהן מנכות את מספר המאפיינים. לפיכך ה- Standard Error בפלטי הרגרסיה שמפיקות התוכנות הללו לעולם יהיה גבוה יותר מה- RMSE.

### האם תוכל להסביר את פלט הרגרסיה של Eviews שבתמונה?

Dependent Variable: DPRICE						
Method: Least Squares						
Date: 19/11/19 Time: 19:39						
Sample: 1 25						
Included observations: 25						
Variable	Coefficient	Std. Error	t-Statistic	Prob.		
C	1. -44.38708	3. 3.428232	5. 12.94751	7. 0.00000		
MPRICE	2. 0.584302	4. 0.014709	6. 39.72358	8. 0.00000		
R-squared	9. 0.985634	Mean dependent var	14. 173.6440			
Adjusted R-squared	10. 0.985009	S.D. dependent var	15. 44.07479			
S.E. of regression	11. 5.396413	Akaike info criterion	6.285964			
Sum squared resid	12. 669.7893	Schwarz criterion	6.383474			
Log likelihood	-76.57455	F-statistic	16. 1577.963			
Durbin-Watson stat	13. 0.590787	Prob(F-statistic)	17. 0.000000			

1. ברצון. נתתי מספרים מ-1 עד 17 לתוצאות פלט הרגרסיה.
2. ההטייה (Bias).
3. המשקולות (Weight) של המאפיין (Feature).
4. שגיאת התקן של האומד להטייה.
5. שגיאת התקן של האומד למשקולות של המאפיין.
6. t סטטיסטי לבדיקת ההשערה שההטייה שווה לאפס.
7. t סטטיסטי לבדיקת ההשערה שהמשקולות של המאפיין שווה לאפס.
8. רמת המובהקות המינימלית לדחיית ההשערה שבסעיף 5.
9. רמת המובהקות המינימלית לדחיית ההשערה שבסעיף 6.
10.  $R^2$ .
11.  $R^2$  מתוקן.
12. שגיאת התקן הנאמדת של ריבועי הפרשים שבין התווית של היעד בפועל והתווית של היעד לפי המודל. שגיאת התקן מחושבת באמצעות ארבעת השלבים שפירטתי מוקדם יותר.
13. סכום ריבועי הפרשים הנאמדים (RSS).
14. סטטיסטי לבדיקת קיומו של מתאם סדרתי.
15. ממוצע ערכי היעדים במדגם.
16. שגיאת התקן של הנאמדת של ריבועי הפרשים שבין התווית של היעד בפועל לבין הערך שבסעיף 15.
17. F סטטיסטי לבדיקת ההשערה שכל המשקולות פרט להטייה שוות ל-0.
18. רמת המובהקות המינימלית לדחיית ההשערה שבסעיף 16.

### אוקיי, מהי פונקציית המטרה ברגרסיה לינארית ונילה?

פונקציית המטרה ברגרסיה לינארית ונילה (Plain Vanilla) היא למזער (קרי, להביא למינימום) את השגיאה הריבועית הממוצעת (MSE).

### אני שומע כל הזמן את הביטוי רגולריזציה בהקשר של רגרסיה לינארית בקרב מדעני נתונים. האם תוכל בבקשה להסביר את המונח רגולריזציה?

רגולריזציה (Regularization) הינה הליך של פישוט המודל שמטרתו להימנע מהתאמת יתר (Over-Fitting) על ידי הקטנת המשקולות

### המידול הסטטיסטי, בהתאמה. מה ההבדל בין סט נתוני אימון לסט נתוני תיקוף?

סט נתוני האימון (Training Data Set) הינו סט הנתונים המשמש לאמידת הפרמטרים עבור המודל הנבדק, בעוד שסט נתוני התיקוף (Validation Data Set) הינו סט הנתונים המשמש לקביעת טיב פעולת המודל, שפותח על סמך סט נתוני האימון, על נתונים אחרים.

### אז מה זה סט נתוני בדיקה?

סט נתוני הבדיקה (Test Data Set) הינו סט הנתונים המשמש לקביעת רמת הדיוק של המודל שנבחר לבסוף, מבין כל המודלים שפיתחנו על סמך סט נתוני האימון ותיקפנו על סמך סט נתוני התיקוף.

### מה מאפיין את הרגרסיה הלינארית?

זוהי למעשה רגרסיה שבה הקשר שבין היעד למאפיינים מונח כלינארי. רגרסיה לינארית בנויה מהטייה (Bias, חותך/קבוע הרגרסיה) וממשקולות (Weights, שיפועים/פרמטרים/מקדמי הרגרסיה). אבחנה נוספת שאני רוצה לעשות היא בין ההטייה למשקולות. בעוד שהטייה הינה איבר קבוע (גודל אוטונומי), הרי שמשקולות הינה מקדם/פרמטר של ערך המאפיין.

### מהו כיוול מאפיינים? ולמה צריך אותו?

כיוול מאפיינים (Feature Scaling) הינו שלב הכרחי הן בלמידה ללא השגחה והן בלמידה בהשגחה. מדובר בהליך מסוים שנועד להבטיח שהמאפיינים נמדדים על בסיס אותו קנה מידה.

### אני זוכר שכשאני למדתי רגרסיה לינארית במסגרת אקונומטריקה בתואר הראשון שלי אז עסקנו הרבה מאוד בבדיקת ה-t הסטטיסטי וה-P-value של מקדמי הרגרסיה וכמוכן בבדיקת ה-R<sup>2</sup> של הרגרסיה, האם גם בלמידת מכונה מדעני נתונים מתעסקים באינדיקטורים הללו?

טוב. ראשית בו נסביר לקוראים את שלושת המושגים הללו. t סטטיסטי (t-statistic) הוא הערך של פרמטר מסוים מחולק בשגיאת התקן שלו ברגרסיה הלינארית, בעוד ש-P-value הוא ההסתברות ברגרסיה לינארית לקבל t סטטיסטי גבוה כמו זה שהיה נצפה אילו למאפיין הנבדק לא הייתה יכולת הסבר.  $R^2$  (R-squared statistic) הוא הפרופורציה של השונות ביעד המוסברת על ידי המאפיינים ברגרסיה לינארית. לדוגמא,  $R^2$  של 66% פירושו ש-66% מההשתנות של היעד מוסברים על ידי מאפייני הרגרסיה. לאחר שהסברתי מהם שלושת המושגים הללו, אני אוכל לומר לך שלאינדיקטורים הללו אין מקום בלמידת מכונה. הם לא מעניינים את מדעני הנתונים. את מדעני הנתונים מעניין רק דבר אחד בלבד והוא שורש השגיאה הריבועית הממוצעת (RMSE) או סטיית התקן של השגיאות.

### אתה מתכוון למה שבפלט הרגרסיה ב-Excel מכוונה Standard Error של הרגרסיה או למה שפלט של תוכנת Eviews מכוונה S.E. of regression?

בגדול כן. שורש השגיאה הריבועית הממוצעת (RMSE- Root Mean Squared Error) מחושב כסטיית התקן של ההפרשים שבין התווית של היעד בפועל לבין התווית של היעד שהתקבלה מהרגרסיה הלינארית. האומדנים שאתה מדבר עליהם מחושבים בצורה טיפה שונה. תוכנת Excel וגם חבילת התוכנה Eviews מחשבות תחילה את סכום ריבועי הפרשים שבין התווית של היעד בפועל לבין התווית של היעד שהתקבלה מהרגרסיה הלינארית. בשלב השני, התוכנה מחשבת את מספר התצפיות פחות 1 ומנכה מהתוצאה את מספר המאפיינים ברגרסיה (לא כולל ההטייה). בשלב השלישי, התוכנה מחלקת את התוצאה שהתקבלה בשלב הראשון בתוצאה

לחילופין ערך של 1 אם יש שיפוע עדין, לחילופין חילופין ערך של 2 אם יש שיפוע בינוני, או לחילופין-חילופין חילופין ערך של 3 אם יש שיפוע רציני.

**לצורך חיזוי מחירי בתים פרטיים, כיצד היית מטפל במאפיין קטגוריאל שזמחה את השכונה שבה ממוקם הבית?**

אני הייתי יוצר משתנה דמה עבור כל אחת מהשכונות. משתנה הדמה עבור שכונה מסוימת מקבל ערך של 1 אם הבית ממוקם באותה שכונה או לחילופין ערך של 0 אם הבית לא ממוקם באותה שכונה.

**האם תוכל להציג דוגמא לחישוב ה- RMSE שלדברך הוא כל כך חשוב בלמידת מכונה?**

וודאי. ניקח את מדגם הנתונים הבא (10 תצפיות) של גילאים מול שכר שנתי של בעלי מקצוע במקצוע מסוים באזור מסוים.

#	Age	Salary
1	25	ש 135,000
2	55	ש 260,000
3	27	ש 105,000
4	35	ש 220,000
5	60	ש 240,000
6	65	ש 265,000
7	45	ש 270,000
8	40	ש 300,000
9	50	ש 265,000
10	30	ש 105,000

אם נרצה להריץ רגרסיה ליניארית שבה היעד הוא השכר השנתי והמאפיין הוא הגיל, אז פלט הרגרסיה שנקבל בתוכנת ה- Excel יראה כך:

Regression output for linear model						
Regression Statistics						
Multiple R	0.73558279					
R Square	0.54108204					
Adjusted R Squar	0.4837173					
Standard Error	52747.8296					
Observations	10					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	26243831770	2.62E+10	9.43231	0.01531984	
Residual	8	22258668230	2.78E+09			
Total	9	48502500000				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	51160.4153	56360.28669	0.907739	0.39054	-78806.6389	181127.469
Age	3827.3052	1246.189443	3.071207	0.01532	953.587194	6701.02321

ואם נחשב את ה- RMSE ואת ה- Standard Error של הרגרסיה באקסל נקבל את התוצאות הבאות:

Regression output for linear model				
Age	Salary	Estimate	Error	Error <sup>2</sup>
25	135,000	146,843	-11,843	140,257,723
55	260,000	261,662	-1,662	2,762,913
27	105,000	154,498	-49,498	2,450,017,922
35	220,000	185,116	34,884	1,216,886,665
60	240,000	280,799	-40,799	1,664,536,157
65	265,000	299,935	-34,935	1,220,471,930
45	270,000	223,389	46,611	2,172,571,397
40	300,000	204,253	95,747	9,167,560,135
50	265,000	242,526	22,474	505,095,267
30	105,000	165,980	-60,980	3,718,508,120
		<b>RMSE</b>	<b>49,731.1307</b>	
		<b>Standard Error</b>	<b>52,747.8296</b>	

ברגרסיה. כאשר מדברים על רגולריזציה לרוב מדברים על 3 סוגים של רגרסיות ליניאריות: Ridge, Lasso, ו-Elastic Net.

**מהי רגרסיה מסוג Ridge?**

רגרסיה מסוג Ridge פירושה רגולריזציה באמצעות הוספת סכום ריבועי המשקולות לפונקציית המטרה ברגרסיה ליניארית ונילה.

**מהי פונקציית המטרה עבור רגרסיה מסוג Ridge?**

במקרה של רגרסיה מסוג Ridge פונקציית המטרה היא להביא למינימום את סך הצברם של ה- MSE בתוספת מכפלת קבוע מסוים (Constant) בסכום ריבועי המשקולות.

**מהו יתרון המרכזי של רגרסיה מסוג Ridge?**

רגרסיה מסוג Ridge מקטינה את גודל המשקולות כאשר הקורלציה בין המאפיינים היא גבוהה.

**מהי רגרסיה מסוג Lasso?**

רגרסיה מסוג Lasso פירושה רגולריזציה באמצעות הוספת סכום הערכים המוחלטים של הפונקציית המטרה ברגרסיה ליניארית ונילה.

**מהי פונקציית המטרה עבור רגרסיה מסוג Lasso?**

במקרה של רגרסיה מסוג Lasso פונקציית המטרה פונקציית המטרה היא להביא למינימום את סך הצברם של ה- MSE בתוספת מכפלת קבוע מסוים בסכום הערכים המוחלטים של המשקולות.

**מהו יתרון המרכזי של רגרסיה מסוג Lasso?**

רגרסיה מסוג Lasso מעמידה על אפס את ערכי המשקולות של המאפיינים, שלהם השפעה מועטה על תוצאות החיזוי.

**אז ניתן לומר שבעוד שרגרסיה מסוג Ridge "מייבשת" את המאפיינים הלא רלוונטיים, הרי שרגרסיה מסוג Lasso מאפסת אותם. יפה. מהי אם כך רגרסיה מסוג Elastic Net?**

רגרסיה מסוג Elastic Net היא שילוב של רגרסיה מסוג Ridge ורגרסיה מסוג Lasso.

**מהי פונקציית המטרה עבור רגרסיה מסוג Elastic Net?**

במקרה של רגרסיה מסוג Elastic Net פונקציית המטרה פונקציית המטרה היא להביא למינימום את סך הצברם של ה- MSE בתוספת מכפלת קבוע מסוים בסכום ריבועי המשקולות בתוספת מכפלת קבוע שונה בסכום הערכים המוחלטים של המשקולות.

**לצורך חיזוי מחירי בתים פרטיים, כיצד היית מטפל במאפיין קטגוריאל שמקבל "כן" אם בבית יש מזגן ו- "לא" אם בבית אין מזגן?**

אני הייתי משתמש במשתנה דמה (Dummy Variable) בודד שמקבל ערך של 1 אם יש בבית מזגן וערך של 0 אם אין בבית מזגן.

**לצורך חיזוי מחירי בתים פרטיים, כיצד היית מטפל במאפיין קטגוריאל שמתאר את המגרש שעליו יושב הבית כ- "ללא שיפוע", "בעל שיפוע עדין", "בעל שיפוע בינוני" ו- "בעל שיפוע רציני"?**

אני הייתי משתמש במשתנה דמה בודד שמקבל ערך של 0 אם אין שיפוע,