נתנאל חסון | צילום: עמי ארליך ||

אקטואר, על רגרסיה לוגיסטית כבר שמעת?

רגרסיה לוגיסטית היא כלי לסיווג תצפיות לאחת משתי קטגוריות. מדובר בכלי חזק לסיווג אשראי קמעונאי. יצאתי לראיין את רועי פולניצר בנושא

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 למד אקטואריה (יישום טכניקות סטטיסטיות וכריית נתונים לבעיות ניתוח סדרות עתיות, צמצום מימדים, אופטימיזציה וסימולציה) בכמה מקומות וביניהם בתוכנית ללימודי תעודה באוניברסיטת חיפה. בשנה האחרונה רועי ייסד את האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA- Professional Data Scientists' Israel Association) ולכן יצאתי לראיין אותו בנושא.

רגרסיה לוגיסטית (Logistic Regression) משמשת אקטוארים מזה שנים רבות. למרות שהמתמטיקאי המפורסם קרל פרידריך גאוס הוא זה שהציע לראשונה, בערך בשנת 1816, את גישת הנראות המירבית (MLE) העומדת בבסיס הרגרסיה הלוגיסטית. הרגרסיה הלוגיסטית עושה שימוש בפונקציית סיגמואיד ולכן היא מתאימה למצבים בינאריים (חדלות פירעון או לא חדלות פירעון). אקטוארים משתמשים ברגרסיה לוגיסטית לאמידת ההסתברות שמאורע מסוים יתממש, כמו למשל שלווה קמעונאי לא יוכל לעמוד בהתחייבות פיננסית להחזר חוב או הלוואה בנקאית, כלומר יגיע למצב של פשיטת רגל. רגרסיה לוגיסטית כרוכה במקסום הנראות הכוללת (Total Likelihood) על ידי שינוי ההטייה והמשקולות. בראיון זה רועי פולניצר יסביר על הרגרסיה הלוגיסטית.

מה הקשר בין למידה בהשגחה לבין רגרסיה לוגיסטית?

קיימים שני סוגים של מודלים של למידה בהשגחה: כאלו המשמשים לפתרון בעיית חיזוי של משתנה נומרי וכאלה המשמשים לפתרון בעיית סיווג, כלומר, חיזוי לאיזו מבין שתי קטגוריות תצפיות חדשות ישתייכו. רגרסיה לוגיסטית (Logistic Regression) היא אחד הכלים המשמש לפתרון בעיית סיווג. נניח שיש לנו מספר מאפיינים (משתנים מסבירים) שחלקם בכלל משתני דמה שנוצרו מתוך מאפיינים קטגוריאליים. עוד נניח כי ישנן שתי קטגוריות שאליהן התצפיות יכולות להשתייך. לקטגוריה אחת נתייחס כאל תוצאה חיובית (על פי רוב זה יהיה הדבר שאותו אנו רוצים לחזות), בעוד שלקטגוריה השנייה כאל תוצאה שלילית.

דוגמא קלאסית לסיווג היא זיהוי של דואר זבל על סמך מילים הנמצאות במייל. דואר זבל יסווג כתוצאה חיובית ודואר שאינו זבל יסווג כתוצאה שלילית. למעשה אנו משתמשים בתווית של 1 עבור תוצאה חיובית ובתווית של 0 עבור תוצאה שלילית. נשאלת השאלה מה כל זה קשור לרגרסיה לוגיסטית? למעשה ניתן להשתמש ברגרסיה לוגיסטית לחישוב ההסתברות לקבל תוצאה חיובית. הרגרסיה הלוגיסטית עושה זאת באמצעות פונקציית סיגמואיד.

מהי פונקציית הסיגמואיד (Sigmoid)?

פונקציית סיגמואיד היא פונקציה מתמטית בעלת עקומה בצורת "S" שערכיה נעים בין 0 ל- 1. ברגיל, פונקציית סיגמואיד מתייחסת למקרה פרטי של פונקציה לוגיסטית, כאשר פונקציית סיגמואיד מוגדרת על ידי מספרים ממשיים (מספרים שיכולים להיות שלמים או שברים, חיוביים או שליליים) והיא עולה בצורה מונוטונית. פונקציית סיגמואיד נראית כך:

$$f(y) = \frac{1}{1 + e^{-Y}}$$

מה מייצג ה- Y בפונקציית הסיגמואיד?

כאשר Y גבוה מאוד ושלילי, אז e^{-Y} גבוה מאוד ולכן פונקציית ה- Q קרובה לאפס. כאשר Y גבוה מאוד וחיובי, אז e^{-Y} נמוך מאוד ולכן פונקציית ה- Q קרובה לאחד. לשאלתך אנו קובעים את Y כשווה להטייה בתוספת צירוף לינארי של המאפיינים:

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_mX_m$$

כך שההסתברות לכך שהתצפיות 'יפלו' בתוך הקטגוריה הראשונה היא:

$$Q = \frac{1}{1 + exp\left(-a - \sum_{j=1}^{m} b_jX_j\right)}$$

מהי פונקציית המטרה ברגרסיה לוגיסטית?

רגרסיה לוגיסטית לא משתמשת בשיטת הריבועים הפחותים (OLS) שעומדת בבסיס רעיון הרגרסיה הלינארית, כי אם בשיטת הנראות המירבית (MLE- Maximum Likelihood Estimation). בסטטיסטיקה, שיטת ה- MLE היא שיטה כללית לבחירת פרמטרים (משקולות במקרה דנן שלפנינו) מתוך סט של תצפיות באופן כזה שימקסם את הסיכוי שהתצפיות יתרחשו. פונקציית המטרה ברגרסיה לוגיסטית נראית כך:

$$\sum_{\substack{\text{תוצאות}\\\text{חיוביות}}} ln(Q) + \sum_{\substack{\text{תוצאות}\\\text{שליליות}}} ln(1 - Q)$$

כאשר Q היא ההסתברות החזויה לקבלת תוצאה חיובית. למעשה המקסום של הפונקציה הזו מביא לבחירת ה- a וה- b_j. הסכימה הראשונה היא על כל התצפיות המביאות לתוצאות חיוביות והסכימה השנייה היא על כל התצפיות המביאות לתוצאות שליליות. שיטת מעלה הגרדיאנט (Gradient Ascent) משמשת לצורך קבלת פתרון נומרי מאחר ואין פתרון אנאליטי לבעיית הנראות המירבית של רגרסיה לוגיסטית.

כאשר רגרסיה לינארית משמשת לחיזוי משתנה נומרי מסוים, רמת הדיוק של הרגרסיה נמדדת באמצעות שורש השגיאה הריבועית הממוצעת (RMSE). האם גם הנכונות (Accuracy) של הרגרסיה הלוגיסטית נמדדת באמצעות שורש השגיאה הריבועית הממוצעת (RMSE)?

ראשית, נסביר לקוראים שנכונות פירושה אחוז התצפיות שסווגו נכונה. שנית, שורש השגיאה הריבועית הממוצעת (RMSE- Root Mean Squared Error) מחושב כסטיית התקן של ההפרשים שבין התווית (Label) של היעד (Target) בפועל לבין התווית של היעד שהתקבלה מהרגרסיה הלינארית. שלישית, מאחר שרגרסיה לוגיסטית מבוססת על גישת הנראות המירבית ולא על גישת הריבועים הפחותים, או אז מדד ה- RMSE לא מהווה מדד לנכונות הרגרסיה הלוגיסטית.

האם ניתן למדוד את נכונות הרגרסיה הלוגיסטית באמצעות אחוז התצפיות שסווגו נכונה על ידי הרגרסיה?

בגדול כן, אבל זה לא תמיד עובד. לדוגמא אם אנחנו אנו מנסים לזהות הונאה בכרטיסי אשראי מתוך מאפיינים כמו מספר החיובים היומי, סוג הרכישות, וכך הלאה. אז אם למשל רק 1% מעסקאות הן עסקאות הונאה אז אנו יכולים לקבל 99% דיוק פשוט מעצם זה שנסווג את כל העסקאות כעסקאות טובות. הבעיה היא שיש חוסר איזון בין גודל הקטגוריות. כי למעשה יש לנו בדוגמא הזו שתי קטגוריות: עסקאות טובות ועסקאות הונאה והקטגוריה הראשונה היא הרבה יותר גדולה מהשנייה. אם הקטגוריות היו זהות בגודלן (או דומות בקירוב בגודל) הרי שמדד הדיוק שהצעת היה עובד מצויין. לצערי, במרבית המקרים אין לנו קטגוריות שוות גודל.

כיצד אם כך ניתן לפתור את הבעיה הזו?

דרך אחת לטפל בבעיה הזו מכונה Under-Sampling והיא גורסת שעל מנת ליצור סט אימון מאוזן יש להשמיט תצפיות מקטגוריית הרוב. לדוגמא בסיטואציה שהרגע הצגתי, מדען הנתונים יכול ליצור סט אימון באמצעות איסוף נתונים על 100,000 עסקאות הונאה ולהצמיד אותן למדגם מקרי של 100,000 עסקאות טובות.

אם ניתן להשמיט תצפיות מקטגוריית הרוב, אז בטח ניתן ליצור בדרך סטטיסטית כזו או אחרת תצפיות נוספות ולהוסיף אותן לקטגוריה הקטנה?

בדיוק כך. גישה אחרת לטיפול בבעיה היא Over-Sampling והיא גורסת שעל מנת ליצור סט אימון מאוזן יש להוסיף תצפיות לקטגוריית המיעוט על ידי יצירת תצפיות סינטטיות. שיטה זו מכונה SMOTE (Synthetic Minority Over-sampling Technique).

אז איזון גודל הקטגוריות הכרחי רק לרגרסיה לוגיסטית?

ממש לא. איזון גודל הקטגוריות הוא לא רק הכרחי עבור רגרסיה לוגיסטית אלא גם עבור שיטות נוספות של למידת מכונה כמו מכונת וקטורים תומכים (SVM) ורשתות נוירונים (Neural Networks), מאחר והוא גורם להן לעבוד הרבה יותר טוב. למעשה איזון הקטגוריות מאפשר ל"אחוז התצפיות שסווגו נכונה" לשמש כפונקציית המטרה. כלומר- באמצעות שימוש בקריטריון זה אנו עשויים לסווג תצפית מסוימת לקטגוריה החיובית אם ההסתברות, Q, להשתייך לקטגוריה זו הינה גבוהה יותר מ- 0.5 ולהשתייך לקטגוריה השלילית אחרת.

מה קורה כאשר עלות הסיווג של תצפית חדשה כחיובית כאשר היא בעצם שלילית שונה מעלות הסיווג של אותה תצפית כשלילית כאשר היא בעצם חיובית?

במצב שכזה הרי שחברה מסוימת לא תרצה לבצע סיווג בהתבסס על האם ההסתברות, Q, גבוהה יותר או נמוכה יותר מ- 0.5. אם רוצים לקבל החלטה, אז יכול להיות שימושי להציג למקבל ההחלטה טווח של קריטריונים אלטרנטיביים המכונה עקומת ה- ROC (עקומה אופיינית למסווג, Receiver Operating Characteristic).

האם תוכל לתת איזשהי דוגמא ליישום של הרגרסיה הלוגיסטית עבור החלטות אשראי?

ניקח לדוגמא סט נתונים של חברת Prosper Marketplace לגבי החלטות האשראי שלה. Prosper Marketplace היא מלווה מסוג peer-to-peer (הלוואות חברתיות) המאפשר למשקיעים להלוות כסף ללווים ללא תיווך. Prosper Marketplace משתמשת בלמידת מכונה ומפרסמת נתונים על הלוואותיה. ניסיתי לשפר את הקריטריון של Prosper Marketplace באמצעות למידת מכונה. מצ"ב תמצית מהנתונים שבהם השתמשתי.

סט נתוני האימון ששימש לחיזוי הלוואות שהגיעו לחדלות פירעון

בעלות על בית, X_1, בעלות = 1, שכירות = 0	הכנסה (באלפי דולר), X_2	יחס החוב להכנסה, X_3	ציון אשראי (FICO), X_4	תוצאת ההלוואה, טובה = 1, חדלת פירעון = 0
1	43.304	18.47	690	0
1	136.000	20.63	670	1
0	38.500	33.73	660	0
1	88.000	5.32	660	1
.....	.....	.....	.....	.....
.....	.....	.....	.....	.....

השתמשתי בסט אימון וסט בדיקה. סט האימון מורכב מ- 8,695 תצפיות שמתוכן 1,499 הן הלוואות שהגיעו למצב של חדלות פירעון ויתר ה- 7,196 הן הלוואות שהוכחו כטובות. סט הבדיקה מורכב מ- 5,916 תצפיות שמתוכן 1,058 הן הלוואות שהגיעו למצב של חדלות פירעון ויתר ה- 4,858 הן הלוואות שהיו טובות. השתמשתי בארבעה מאפיינים (אחד מהם, בעלות על בית, היה קטגוריאלי ולכן היה עליי לקדד אותו באמצעות משתנה דמה שמקבל 0 או 1). המשקולות שנאמדו עבור סט האימון מוצגות להלן:

משקולות אופטימליות (באמצעות אקסל או שפת R)

מאפיין	סימן	משקולת, b_i
בעלות על בית	X_1	0.1395
הכנסה (באלפי דולר)	X_2	0.0041
יחס החוב להכנסה	X_3	-0.0011
ציון אשראי	X_4	0.0113

ההטייה (החותך) נאמדה כ- 6.5645- והרגרסיה כולה נראית כך:

$$Y = -6.5645 + 0.1395X_1 + 0.0041X_2 - 0.0011X_3 + 0.0113X_4$$

כעת על מדען הנתונים להחליט על קריטריון לסיווג האם הלוואה מסוימת הינה טובה או לא טובה. ההחלטה על קריטריון כרוכה בהגדרת רמת סף, Z, עבור הערך של Q כך שאם $Q > Z$ או אז ההלוואה מסווגת כטובה ואם $Q \leq Z$ או אז ההלוואה מסווגת כלא טובה. ניתן לסכם את התוצאות, המתקבלות מיישום ערך מסוים של Z על סט הבדיקה, באמצעות מה שמכונה מטריצת טעות (Confusion Matrix). מטריצת הטעות מציגה את הקשר שבין הסיווגים לבין התוצאות בפועל.

מטריצת הטעות עבור סט הבדיקה כאשר Z = 0.75

	סיווג כלא חדלת פירעון	סיווג כחדלת פירעון
תוצאה חיובית (אין חדלות פירעון)	77.59%	4.53%
תוצאה שלילית (חדלות פירעון)	16.26%	1.62%

מטריצת הטעות עבור סט הבדיקה כאשר Z =0.8

	סיווג כלא חדלת פירעון	סיווג כחדלת פירעון
תוצאה חיובית (אין חדלות פירעון)	55.34%	26.77%
תוצאה שלילית (חדלות פירעון)	9.75%	8.13%

מטריצת הטעות עבור סט הבדיקה כאשר Z =0.85

	סיווג כלא חדלת פירעון	סיווג כחדלת פירעון
תוצאה חיובית (אין חדלות פירעון)	28.65%	53.47%
תוצאה שלילית (חדלות פירעון)	3.74%	14.15%

מה אומרות התוצאות במטריצות הטעויות שאתה מציג כאן?

על מנת להבין את התוצאות של מטריצת הטעות ראשית אני אגדיר את ארבעת האלמנטים של מטריצת הטעות כדלקמן:

- חיובי אמיתי (TP- True Positive): כאשר הן הסיווג והן התוצאה הם חיוביים.
- שלילי כוזב (FN- False Negative): כאשר הסיווג הוא שלילי אך התוצאה היא חיובית.
- חיובי כוזב (FP- False Positive): כאשר הסיווג הוא חיובי אך התוצאה היא שלילית.
- שלילי אמיתי (TN- True Negative): כאשר הן הסיווג והן התוצאה הם שליליים.

האם אתה יכול בבקשה לסדר את ההגדרות הללו במטריצת הטעות על מנת שנבין יותר טוב כיצד הן משתלבות?

	סיווג כתוצאה חיובית	סיווג כתוצאה שלילית
תוצאה חיובית	TP	FN
תוצאה שלילית	FP	TN

להלן היחסים הנגזרים ממטריצת הטעות:

$$\text{The False Positive Rate} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

$$\text{The False Negative Rate} = \frac{\text{FN}}{\text{TP} + \text{FN}}$$

$$\text{True Positive Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{The True Negative rate} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

סידרתי את תוצאות היחסים הללו במטריצה הבאה:

	Z = 0.75	Z = 0.80	Z = 0.85
שיעור החיוביים הכוזבים (FPR)	90.93%	54.54%	20.89%
שיעור השליליים הכוזבים (FNR)	5.52%	32.60%	65.11%
שיעור החיוביים האמיתיים (TPR)	94.48%	67.39%	34.89%
שיעור השליליים האמיתיים (TNR)	9.07%	45.46%	79.11%
נכונות (Accuracy)	79.21%	63.47%	42.80%
דיוק (Precision)	82.67%	85.02%	88.47%

נכונות פירושה אחוז התצפיות שסווגו נכונה. כפי שציינתי מקסום הנכונות לא בהכרח יוצר את המודל הטוב ביותר. אכן, בדוגמא שלי הנכונות ממוקסמת ל- 82.12% באמצעות סיווג פשוט של כל התצפיות כחיוביות (קרי, סיווג כלא חדלת פירעון).

מה פירוש שיעור החיוביים האמיתיים (True Positive Rate)?

זוהי ההסתברות המותנה לסווג תצפית מסוימת כחיובית מותנה בכך שידוע שהתוצאה היא חיובית. בסטטיסטיקה מדד זה מכונה רמת ביטחון או רמת סמך. לדוגמא, ההסתברות לסווג הלוואה מסוימת כלא חדלת פירעון אם ידוע שלא ארעה חדלות פירעון.

מה פירוש שיעור החיוביים הכוזבים (False Positive Rate)?

זוהי ההסתברות המותנה לסווג תצפית מסוימת כחיובית מותנה בכך שידוע שהתוצאה היא שלילית. בסטטיסטיקה מדד זה מכונה רמת המובהקות או אלפא. לדוגמא, ההסתברות לסווג הלוואה מסוימת כלא חדלת פירעון אם ידוע שארעה חדלות פירעון.

מהו מדד שיעור השליליים האמיתיים (True Negative Rate)?

זוהי ההסתברות המותנה לסווג תצפית מסוימת כשלילית מותנה בכך שידוע שהתוצאה היא שלילית. בסטטיסטיקה מדד זה מכונה עוצמת המבחן. לדוגמא, ההסתברות לסווג הלוואה מסוימת כחדלת פירעון אם ידוע שארעה חדלות פירעון.

מהו מדד שיעור השליליים הכוזבים (False Negative Rate)?

לאמור- זוהי ההסתברות המותנה לסווג תצפית מסוימת כשלילית מותנה בכך שידוע שהתוצאה היא חיובית. בסטטיסטיקה מדד זה מכונה טעות מסוג 2 או ביתא. לדוגמא, ההסתברות לסווג הלוואה מסוימת כחדלת פירעון אם ידוע שלא ארעה חדלות פירעון.

מהו דיוק (Precision)?

דיוק הוא אחוז הסיווגים החיוביים אשר היו נכונים. לאמור- זוהי ההסתברות המותנה שתוצאה מסוימת היא חיובית מותנה בכך שידוע שהתצפית סווגה כחיובית. בסטטיסטיקה מדד זה מכונה ערך ניבוי חיובי. לדוגמא, ההסתברות שהלוואה מסוימת לא תגיע לחדלות פירעון אם ידוע שהיא סווגה כלא חדלת פירעון.

אז ישנם למעשה מספר יחסי תחלופה.

בדיוק כך. אני יכול להגדיל את שיעור השליליים האמיתיים (קרי, לזהות אחוז גבוה יותר של הלוואות שיגיעו לחדלות פירעון) רק אם אני אזהה אחוז נמוך יותר של הלוואות שהוכחו כטובות. כמובן שהנכונות יורדת ככל ששיעור השליליים האמיתיים עולה.

לפני שהצגת את הדוגמא הזכרת את עקומת ה- ROC. האם תרצה להרחיב עליה?

אני אציג תרשים שמתאר את שיעור החיוביים האמיתיים (TPR) מול שיעור החיוביים הכוזבים (FPR). תרשים זה מכונה עקומת ה- ROC.

השטח שמתחת לעקומה (AUC) הינו דרך פופולארית לסכם את יכולת הסיווג של המודל. אם ה- AUC הוא 1.0, אז המודל הוא מושלם היות ושיעור החיובים האמיתיים (TPR) הוא 100% ושיעור האמיתיים הכוזבים (FPR) הוא 0%. הקו המנוקד שבתרשים לעיל מתאים ל- AUC של 0.5, מה שמתאים למודל ללא יכולת סיווג. למשל, למודל שמבצע סיווג מקרי יש AUC של 0.5.

אז מודלים עם AUC שקטן מ- 0.5 הם למעשה גרועים יותר ממודלים שמבצעים סיווג מקרי. מה לגבי המודל שלך?

עבור הנתונים שבחנתי למודל יש יכולת סיווג נמוכה. בהינתן שחברת Prosper Marketplace כבר משתמשת בלמידת מכונה לצורך קבלת החלטות ההלוואה שלה ושאני משתמש רק בארבעה מאפיינים, אז אין זה מפתיע שה- AUC של המודל שלי הוא רק מעט מעל ל- 0.5. הנקודה החשובה היא שאין לצפות שמודל שכזה יבצע סיווג מושלם.

אז מהו המבחן העיקרי עבור מודל סיווג?

המבחן העיקרי הוא האם המודל מסוגל לקבל החלטות שטובות לפחות כמו ההחלטות שהיה מקבל בנאדם. רוצה לומר שבעת ההחלטה על הערך הראוי של Z (קרי, מיקום על עקומת ה- ROC) על הלווה לשקול הן את הרווח הממוצע מהלוואות שלא מגיעות לחדלות פירעון והן את ההפסד הממוצע מההלוואות שמגיעות לחדלות פירעון.

רגע, אז מה קורה אם הרווח מהלוואה שלא מגיעה לחדלות פירעון הוא X, בזמן שההפסד מהלוואה שמגיעה לחדלות פירעון הוא 4X?

אז הרווח של המלווה הוא הגבוה ביותר כאשר הוא ממקסם את הפונקציה הבאה:

$$X \times TP - 4X \times FP$$

עבור האלטרנטיבות שבדקנו לעיל (Z של 0.75, 0.8 ו- 0.85) הפונקציה הבאה שווה 12.55X, 16.34X ו- 13.69X, בהתאמה. זה מצביע על כך שמשלושת ערכי ה- Z האלטרנטיביים, Z = 0.8 הוא הרווחי ביותר. נקודה נוספת היא שרגרסיות מסוג Ridge, Lasso ו- Elastic Net יכולות לשמש בשילוב עם רגרסיה לוגיסטית ממש כמו בשילוב עם רגרסיה רגילה. כמובן שזה שימושי מבחינה פוטנציאלית כאשר ישנם מאפיינים רבים. אני רק אעיר שזה נעשה אך ורק עבור אמידת הפרמטרים ולא עבור הסיווג משעה שהפרמטרים כבר נאמדו.