

סיכום רגרסיה לינארית

רגרסיה לינארית היא בוודאי איננה טכניקה חדשה של למידת מכונה. למעשה רגרסיה לינארית משחקת תפקיד מרכזי במחקרים אמפיריים מזה שנים רבות. מדעני נתונים (Data Scientists) אימצו את הרגרסיה הלינארית ככלי לניבוי/חיזוי.

ליישומים של למידת מכונה לעיתים קרובות ישנם מאפיינים (Features, משתנים מסבירים) רבים, אשר חלקם מתואמים מאוד (Highly Correlated) האחד עם השני. במקרה שכזה רגרסיה לינארית מהימנה תייצר תוצאה שתספק משקולת (Coefficient, מקדם) חיובית גבוהה לערכים עבור מאפיין מתואם אחד ומשקולת שלילית גבוהה לערכים עבור מאפיין מתואם אחר.

אחת הגישות להקטנת המשקולות של המאפיינים במודל הרגרסיה נקראת רגרסיה מסוג Ridge. גישה אחרת נקראת רגרסיה מסוג Lasso. רגרסיה מסוג Lasso מקטינה לאפס את המשקולות של המאפיינים הלא חשובים. גישה נוספת נקראת רגרסיה מסוג Elastic Net והיא עושה שימוש ברעיונות שעומדים בבסיסן של הרגרסיות מסוג Ridge ו-Lasso. למעשה, רגרסיה מסוג Elastic Net יכולה לשמש להשגת היתרונות של שתי הרגרסיות הללו (קרי, קבלת משקולות שקטנות יותר בגודלן והשמטת המאפיינים לא חשובים).

ניתן להכניס משתנים קטגוריאליים (Categorical Variables) לרגרסיה לינארית באמצעות יצירת משתנה דמה (Dummy Variable), אחד לכל קטגוריה. משתנה הדמה עבור תצפית מסוימת נקבע כשווה ל-1 אם התצפית 'נופלת' באותה קטגוריה ו-0 אם לאו.

רגרסיה לוגיסטית (Regression Logistic), ממש כמו רגרסיה לינארית, משמשת במחקר האמפירי מזה שנים רבות. כעת הרגרסיה הלוגיסטית הופכת לכלי סיווג חשוב עבור מדעני נתונים. באופן טבעי קיימות שתי קבוצות. האחת מכונה "חיובית"; האחרת מכונה "שלילית". פונקציית הסיגמואיד (Function Sigmoid) בצורת S משמשת להגדרת ההסתברות של תצפית מסוימת 'ליפול' בקבוצה החיובית.

הליך חיפוש איטרטיבי משמש למציאת הפונקציה הלינארית של ערכי המאפיינים שכאשר מכניסים אותה לתוך פונקציית הסיגמואיד היא עושה את העבודה הטובה ביותר בהקצאת הסתברות גבוהה לתוצאות חיוביות והסתברות נמוכה לתוצאות שליליות. מקובל לסכם את תוצאות השימוש ברגרסיה לוגיסטית על נתוני סט הבדיקה (Test Set) באמצעות שימוש במטריצת הטעות (Matrix Confusion).

משעה שסיימנו לבנות את הרגרסיה הלוגיסטית, עלינו להחליט באיזה אופן ישמשו התוצאות. כך למשל, אם הרגרסיה הלוגיסטית משמשת להחלטת מתן אשראי, הרי שעל מקבל ההחלטה להגדיר ערך Z מסוים. כאשר ההסתברות לקבלת תוצאה חיובית מההלוואה נאמדת כגבוהה יותר מ- Z , או אז הבקשה להלוואה מאושרת. כאשר ההסתברות לקבלת תוצאה חיובית מההלוואה נמוכה יותר מ- Z , או אז הבקשה להלוואה נדחית.

קיים יחס תחלופה בין ההצלחה לזהות הלוואות טובות לבין הצלחה לזהות הלוואות שתגענה למצב של חדלות פירעון (Default). באופן טבעי שיפור ההצלחה לזהות הלוואות שתגענה למצב של חדלות פירעון מביא להרעה בהצלחה לזהות הלוואות טובות, ולהיפך. ניתן לסכם את יחס תחלופה הזה באמצעות עקומת ה-ROC (עקומה אופיינית למסווג, Receiver Operating Characteristic Curve) המקשרת בין שיעור החיובים האמיתיים (TPR, כלומר, רמת הביטחון או רמת הסמך ופירושו האחוז מהזמן שבו תוצאה חיובית מסווגת כחיובית) לבין שיעור החיוביים הכוזבים (FPR, רמת המובהקות ופירושו האחוז מהזמן שבו תוצאה חיובית מסווגת כשלילית).



פרטים אודות כותב המאמר: מדען הנתונים רועי פולניצר, PDS

- מייסד ומנכ"ל האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA), מייסד ויו"ר לשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (IAVFA) ובעלים של פירמת הייעוץ וההדרכה שווי פנימי.
- מחזיק בתואר M.B.A. במנהל עסקים עם התמחות בניהול סיכונים ואקטואריה ותואר B.A. בכלכלה עם התמחות במימון שניהם בהצטיינות מאוניברסיטת בן-גוריון בנגב, דיפלומה בניהול סיכונים פיננסיים (FRM) מאוניברסיטת אריאל, תואר Financial Risk Manage מארגון בינ"ל GARP, תואר Certified Risk Manage מארגון ישראלי IARM, תואר Fellow Actuary מארגון ישראלי IAVFA ותואר Professional Data Scientist מארגון ישראל PDSIA.
- בעל ניסיון אינטנסיבי של מעל עשור וחצי שנים בתחום מדע הנתונים ולמידת המכונה, הכולל ביצוע מחקרי מידע מעמיקים לשם הפקת תובנות עסקיות, ניקוי, טיוב וסידור של המידע המשמש למחקרים השונים, הפעלת אלגוריתמים שונים של מידול, כריית נתונים ו-Machine Learning על המידע ובניית תהליכי הכנת המידע והאופטימיזציה של האלגוריתמים השונים.
- מרצה לתכנות בשפות R ו-Python, לניהול סיכונים, הערכות שווי ואקטואריה והנדסה פיננסית.