

מודלים סטטיסטיים ויישומיהם 52518 תשע"ח – תרגיל 9

להגשה עד 1.1.18 בשעה 23:55

1. בשעור ראינו כי עבור רגרסיה בינארית, אלגוריתם ניוטון-רפסון יכול להרשם בצורה

$$\beta^{(m+1)} = (X^T W(\beta^{(m)}) X)^{-1} X^T W(\beta^{(m)}) Z(\beta^{(m)})$$

כאשר $W(\beta)$ היא מטריצה אלכסונית עם הערכים

$$w_i(\beta) = \Omega'(\beta^T X_i) (Y_i - g(\beta^T X_i)) - \Omega(\beta^T X_i) g'(\beta^T X_i)$$

באלכסון וכן מתקיים

$$Z(\beta^{(m)}) = X\beta^{(m)} - W(\beta^{(m)})^{-1} R(\beta^{(m)})$$

אם נשנה את ההגדרה של w_i להיות

$$w_i(\beta) = -\Omega'(\beta^T X_i) (Y_i - g(\beta^T X_i)) + \Omega(\beta^T X_i) g'(\beta^T X_i)$$

(כלומר ההגדרה הקודמת כפול -1), כמו ברשימות אשר הועלו לאתר, נקבל

$$Z(\beta^{(m)}) = X\beta^{(m)} + W(\beta^{(m)})^{-1} R(\beta^{(m)})$$

שהוא הניסוח המקובל יותר. זכרו כי עבור המודל הלוגיסטי מתקיים $\Omega(u) = 1, \Omega'(u) = 0$. $g(u) = \frac{e^u}{1+e^u} \rightarrow$

א. כתבו ב-R פונקציה המבצעת את האלגוריתם הנ"ל עבור המקרה של רגרסיה לוגיסטית.

ב. הריצו את הפונקציה שכתבתם על הקובץ `mrfit.csv`.

ג. השוו את התוצאות מהסעיף הקודם לתוצאות הרצת רגרסיה לוגיסטית מובנית ב-R, באמצעות הפקודה

`glm(data, formula, family = binomial(link = "logit"))`

2. עבור נתוני הקובץ `mrfit.csv`,

א. חשבו ידנית אומד לרווח הסמך (ברמת בטחון של 95%) למנת יחס הסיכויים (Odds Ratio) הנובעת משינוי

ערך הכולסטרול ב-30.

ב. חשבו אומד ורווח סמך ל- $p(x)$ עבור פרט עם הנתונים הבאים:

$$Age = 54, \quad Map = 85, \quad Diab = 0, \quad Chol = 214, \quad Smoke = 1$$

3. הקובץ `spam.train.csv` מכיל 57 משתנים מסבירים עבור 3681 הודעות דוא"ל ועוד משתנה תוצאה (`spam`), האם

המייל קוטלג כדואר זבל או לא.

א. הריצו מודל רגרסיה לוגיסטית על הנתונים, מהם המשתנים המובהקים יותר?

ב. עבור נתוני הקובץ `spam.test.csv`, חשבו את ההסתברות של כ"א מהדגימות להיות מקוטלגת כספאם. סמנו

$$\hat{y}_i = 1 \text{ אם ההסתברות גדולה מחצי ואפס אחרת.}$$

ג. מהו אחוז הדיוק של המודל אותו בנית? (דהיינו, אחוז המקרים בהם $\hat{y}_i = y_i$)

ד. כעת, בנו מודל חדש לנתוני הקובץ `spam.train.csv`, המבוסס רק על המשתנים המובהקים אותם מצאתם קודם.

ה. בדומה לסעיף ג' – מהו אחוז הדיוק בניבוי של המודל המצומצם על נתוני הקובץ `spam.test`?

4. מחקר בדק את האינדיקציה ללימודים אקדמיים ($y = 1$ אם הנבדקת למדה\לומדת) כתלות בהכנסת ההורים (x)

ובהשכלתם (m, f אם האמא\אבא למדו). המודל שהתקבל הוא $\text{logit}(\hat{p}) = -1.9 + 0.02x + 0.82m + 1.33f$.

בהנחה שהמשתנים y, m, f בינאריים ושטווח הערכים של x הוא 0.5 עד 130 בקפיצות של 0.5.

א. שרטטו ב-R גרף של כל ה- \hat{p}_i האפשריים כתלות ב- x_i וצבעו את התוצאות לפי הערכים האפשריים של m, f

(כלומר, ארבעה גרפים על אותה מערכת צירים – אחד לכל ערך אפשרי של הזוג (m, f)). האם ניתן להבחין

בהבדל משמעותי בין הקטגוריות?

ב. עבור כל אחת מהאפשרויות לערכי m, f , מהי רמת ההכנסה המינימלית עבורה מתקיים $P(y = 1) =$

$$P(y = 0)$$