

מודלים סטטיסטיים ויישומיהם 52518 5 תשע"ח – תרגיל 9

להגשה עד 1.1.18 בשעה 23:55

1. בשעור ראיינו כי עבור רגרסיה בינהרית, אלגוריתם ניוטון-רפסון יכול להרשם בצורה

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

באשר $Z(\beta)$ היא מטריצה אלכסונית עם הערכבים

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

באלבוסון וכן מתקיים

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

אם נשנה את ההגדירה של w_i להיות

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

(בלומר ההגדירה הקודמת בפול 1-), במו ברשימות אשר הועלו לאחר, נקבל

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

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

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

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

ג. השוו את התוצאות מה壽יף הקודם לתוצאות הרצת רגרסיה לוגיסטיבית מובנית ב-R, באמצעות הפקודה

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

2. עבור נתוני הקובץ `mrfit.csv`

א. חשבו ידנית אומד לרווח הסמרק (ברמת בטחון של 95%) למנת יחס הסיבוכים (Odds Ratio) הנובעת משינוי ערק הכולסטרול ב-30.

ב. חשבו אומד ורווח סמרק ל-(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 דואר אלקטרוני ועוד משתנה תוצאה (`spam`), האם המייל קופולג בדואר זבל או לא.

א. הריצו מודל רגרסיה לוגיסטיבית על הנתונים, מהם המשתנים המובאים יותר?

ב. עבור נתוני הקובץ `spam.test.csv`, חשבו את ההסתברות של ב"א מהדցימות להיות מקוטלתgas בספאם. סמן $\hat{y}_i = 1$ אם ההסתברות גדולה מhalb וAppending אפס אחרת.

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

ד. בעת, בנו מודל חדש לנatoi הקובץ `spam.train`, המבוסס רק על המשתנים המובאים אותם מצאתם קודם.

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

4. מחקר בדק את האינדייקציה לסטודנטים אקדמיים ($y = 1$ אם הנבדקת למדה\לומדת) בתלות בהכנתה ההורורים (x)

ובהשכלתם ($f, m = 1$ אם האמא\אבא למד). המודל שהתקבל הוא $\hat{y} = -1.9 + 0.02x + 0.82m + 1.33f$. בהנחה שהמשתנים f, m, y בינהרים ושטוחה הערכיהם של x הוא 0.5 עד 0.30 בקפיצות של 0.5.

א. שרטטו ב-R גרף של ה- \hat{y} האפשרים בתלות ב- x וקבעו את התוצאות לפי הערכיהם האפשריים של f, m (בלומר, ארבעה גרפים על אותה מערכת ציריים – אחד לכל ערך אפשרי של הזוג f, m). האם ניתן להבחין

בהבדל משמעותי בין הקטגוריות?

ב. עבור כל אחת מהאפשרויות לערכי f, m , מהי רמת ההכנה המינימלית עבורה מתקיים $P(y = 1) = ?P(y = 0)$