שיטות מחקר בקוגניציה (6132)

תרגיל 3

14 ביוני 2022

נהלי עבודה והגשה

יש להגיש את התרגיל בתיבת ההגשה הייעודית במודל, **עד לתאריך 25.6** בשעה 23:59.
יש להעלות לתיבת ההגשה קובץ בפורמט PDF ובו תשובות מילוליות **לכל** השאלות בתרגיל, בצירוף הגרפים הנדרשים. בנוסף, יש להעלות לתיבת ההגשה קובץ R שמכיל את כל הקוד הנדרש בתרגיל. על קובץ הקוד לרוץ ללא שגיאות ולהפיק את כל התוצרים (לרבות הגרפים) שמוצגים במסמך ה-PDF. הקוד ייבדק ידנית, אנא הקפידו על תיעוד ועל קריאות.
שאלות על התרגיל ניתן לפרסם, גם באנונימיות, בפורום יחידה 3. להוראות נוספות, ראו מסמך נהלי הקורס.
בהצלחה!

טעינת הנתונים

כל השאלות בתרגיל מתייחסות לנתונים המצורפים בקובץ wine_quality.csv. הקובץ מכיל נתונים על מאפייניהם הכימיים של מספר רב של יינות, כמו גם ציון איכות לכל יין שנקבע כממוצע דירוגים שניתנו ע״י שופטים מומחים בתחום. בנוסף, לכל יין מצוין תחת המשתנה type האם מדובר ביין לבן או אדום. מקורם של הנתונים, לפני התאמות שנעשו בהם לטובת התרגיל, במאמר הזה.
טענו את הנתונים ל-R והמירו את המשתנה הקטגוריאלי type לטיפוס factor.

שאלה 1

בשאלה זו נראה כיצד ניתן להשתמש במסגרת העבודה של רגרסיה לינארית מרובה על-מנת לבצע רגרסיה עם מודל שאינו לינארי, נעסוק בתופעה של התאמת-יתר (overfitting) שעשויה להתרחש במודלי רגרסיה מרובה ונכיר את שיטת leave-one-out (LOO) להתמודדות איתה.

א. נתמקד בקשר בין רמת החומצה הציטרית ביין לבין איכותו. כדי להמחיש טוב יותר את התופעה, נעבוד עם מדגם חלקי של הנתונים. סננו את הנתונים כך שיישארו רק רמות החומצה הציטרית וציוני האיכות של **25** היינות **הראשונים** בקובץ. רמות החומצה הציטרית נתונות במשתנה **citric.acid**. ציוני האיכות נתונים במשתנה **quality**.

ב. נסמן את המשתנה citric.acid באות x. הוסיפו לנתונים מסעיף א' ארבעה משתנים חדשים: x^2, x^3, x^4, x^5.

ג. צרו חמישה מודלי רגרסיה לינארית:

- מודל רגרסיה פשוטה עם המנבא x והמנובא quality.
- מודל רגרסיה מרובה עם המנבאים x, x^2 והמנובא quality.
- מודל רגרסיה מרובה עם המנבאים x, x^2, x^3 והמנובא quality.
- מודל רגרסיה מרובה עם המנבאים x, x^2, x^3, x^4 והמנובא quality.
- מודל רגרסיה מרובה עם המנבאים x, x^2, x^3, x^4, x^5 והמנובא quality.

לכל אחד מהמודלים, דווחו את **p-value** למובהקות המודל כולו, ואת **גודל האפקט** (R^2).

ד. לכל אחד מהמודלים מסעיף ג', חשבו **SSE**. איזה מודל השיג את ערך ה-SSE **הנמוך ביותר**?

תזכורת:

$$\text{SSE} = \sum_{i=1}^{25} (\hat{y}_i - y_i)$$

כאשר $\hat{y}_i$ הוא ערך ה-quality המנובא של התצפית ה-i ו-y_i הוא ערך ה-quality האמיתי של התצפית ה-i.

רמז: כדי לקבל את ערכי הניבויים על הנתונים שאליהם המודל הותאם, עבור מודל השמור במשתנה בשם m, השתמשו בפקודה `predict(m)`.

ה. לכל $d \in \{1, 2, 3, 4, 5\}$, הציגו גרף פיזור המתאר את הקשר בין x ל-quality והוסיפו לגרף קו רגרסיה המתאים לפולינום ממעלה d, בעזרת geom_smooth באופן הבא:

```
geom_smooth(method = "lm", formula = y ~ poly(x, d), se = F)
```

הבהרה: בסעיף זה יש להגיש חמישה גרפים, בכולם פיזור נקודות זהה, והם נבדלים רק בקו הרגרסיה המוצג בהם.

ו. על סמך הגרפים, מהי מעלת הפולינום (החזקה הכי גבוהה של x) שהכי מתאימה לנתונים בעיניכן? האם האינטואיציה הזאת עולה בקנה אחד עם מעלת הפולינום של המודל שהשיג ערך SSE מינימלי בסעיף ד'?

התופעה שחזינו בה כעת נקראת **התאמת-יתר** (overfitting): הוספת מנבאים למודל תמיד מקטינה את גורם הטעות שלו, אבל החל משלב מסוים הוספת המנבאים רק מאפשרת למודל לתפוס את ה"רעש", כך שהוא מתאר היטב את הנתונים שאליהם הותאם, אך יכולת ההכללה שלו לאוכלוסייה כולה פוחתת. אחת הגישות הנפוצות להערכת שגיאת ההכללה של המודל נקראת **cross validation**.

הסעיפים הנותרים בשאלה זו הם סעיפי רשות, ואינם מהווים חלק מהציון. בסעיפים אלו נכיר ונממש שיטת cross validation שנקראת leave-one-out (LOO):

ז. עבור כל אחת מהתצפיות במדגם מסעיף ב':

- הסירו אותה מהמדגם והתאימו את חמשת המודלים מסעיף ג' ל-24 התצפיות הנותרות.
- השתמשו במודלים אלו כדי לנבא את ערך התצפית שלא השתתפה בהתאמתם. אם מדובר בתצפית בשורה ה-i של data frame בשם data, והמודל שהותאם על שאר התצפיות שמור במשתנה בשם m, ניתן לקבל את הניבוי המבוקש באמצעות הפקודה הבאה:

```
predict(m, newdata = data[i,])
```

ניבויים שהתקבלו בצורה כזאת ייקראו "ניבויי LOO". ניבוי LOO עבור התצפית ה-i יסומן $\hat{y}_i^{\text{LOO}}$.

בסיום התהליך תקבלו ניבויי LOO לכל אחת מ-25 התצפיות עבור כל אחד מהמודלים.

ח. לכל אחד מהמודלים, חשבו את הערך הבא:

$$\text{SSE}_{\text{LOO}} = \sum_{i=1}^{25} \left(\hat{y}_i^{\text{LOO}} - y_i\right)$$

הציגו גרף של SSE_{LOO} כפונקציה של מעלת הפולינום של המודל (החזקה הכי גבוהה של x שמהווה מנבא במודל).

ט. בשיטת LOO, המודל שנעדיף הוא זה שהשיג את ערך **SSE_{LOO} המינימלי**. האם המודל שהשיג את ערך SSE_{LOO} המינימלי הוא זה ששיערתן בסעיף ו'?

שאלה 2

הבהרה: משאלה זו והלאה יש לעבוד עם סט הנתונים המלא ($N = 6497$) שנטען בחלק "טעינת הנתונים", לא עם הסט החלקי ששימש בשאלה 1.

בהרצאה ובתרגול ראינו כיצד ליצור מדגם בוטסטראפ לסטטיסטי המחושב ממדגם יחיד. בשאלה זו נלמד ליצור מדגם בוטסטראפ לסטטיסטיים המחושבים ממדגמים מזווגים, ונשתמש בשיטת הבוטסטראפ כדי לחשב רווחי סמך עבורם.

נתחיל מביצוע מבחן ספירמן:

א. לכל סוג יין (אדום / לבן), הציגו **גרף פיזור** של הקשר בין ציון איכות היין לבין אחוז האלכוהול בנפח היין. אחוזי האלכוהול נתונים במשתנה **alcohol** וציוני האיכות נתונים במשתנה **quality**.

ב. לכל סוג יין, בצעו **מבחן מתאם ספירמן** בין אחוז האלכוהול בנפח היין לבין איכות היין. דווחו את תוצאות שני המבחנים (**מקדם מתאם** ו-p**-value** לכל אחד מסוגי היין). ענו בנפרד לכל סוג יין: האם נמצא קשר מובהק בין אחוז אלכוהול בנפח לבין איכות היין? האם עוצמת הקשר גבוהה?

נשים לב שהפונקציה המובנית למבחן ספירמן ב-R לא מספקת רווח סמך עבור מקדם המתאם. נחשב את רווחי הסמך בעצמנו, באמצעות שיטת הבוטסטראפ:

ג. לכל סוג יין, צרו **מדגם בוטסטראפ של זוגות** ערכים מהמשתנים alcohol ו-quality. על גודל מדגם הבוטסטראפ להיות כגודל המדגם המקורי. לסיום, חשבו **במדגם הבוטסטראפ** את מקדם מתאם ספירמן בין alcohol ו-quality. **שימו לב:** יש לדגום בכל פעם **זוג** ערכים. כלומר, זוגות הערכים במדגם הבוטסטראפ חייבים להיות זוגות שהופיעו במדגם המקורי. אחת הדרכים להשיג זאת היא ליצור תחילה מדגם בוטסטראפ של אינדקסים, ולהשתמש בווקטור האינדקסים כדי לקבל את הערכים המתאימים מכל אחד מהמשתנים.

ד. חזרו על סעיף ג' 10000 פעמים ושמרו את התוצאות בנפרד לכל סוג יין. בסיום התהליך תקבלו שני וקטורים באורך 10000 כל אחד. כל וקטור כזה מכיל **התפלגות בוטסטראפ** של מקדם מתאם ספירמן באחד מסוגי היין.

ה. לכל סוג יין, חשבו את אחוזונים 2.5 ו-97.5 של התפלגות הבוטסטראפ שנוצרה בסעיף ד'. אלו **גבולות רווח הסמך ברמת בטחון 95% עבור מקדם מתאם ספירמן**.

נציג את רווחי הסמך בצורה גרפית:

ו. לכל סוג יין, הציגו היסטוגרמה של התפלגות הבוטסטראפ מסעיף ד', יחד עם קווים אנכיים שמציינים את גבולות רווח הסמך שחושב בסעיף ה'.

ז. עבור איזה מסוגי היין התקבל רווח סמך רחב יותר? כיצד ניתן להסביר זאת על-סמך הנתונים?

ח. **בונוס:** הציגו גרף נוסף, בו רווחי הסמך יבוטאו כ-**error bars** מסביב לערכי מקדמי מתאם ספירמן **מהמדגם המקורי**. הקפידו על עקרונות **ויזואליזציה נכונה** שנלמדו בקורס, ובפרט שימו לב לציין על-גבי הגרף את משמעות ה-error bars.

שאלה 3

בשאלה זו נתרגל שימוש במבחן מאן-וויטני ונכיר את אחד מגדלי האפקט המתאימים לו.

א. הציגו גרף של הקשר בין התפלגות ציוני האיכות של היין לבין סוג היין. סוגי היינות נתונים במשתנה **type** וציוני האיכות נתונים במשתנה **quality**. ניתן להציג לכל סוג יין boxplot או violin plot, או שילוב של אחד מהם עם jitter plot.

ב. השתמשו **במבחן מאן-וויטני** כדי לבחון את ההשערה שיש הבדל בין התפלגות ציוני האיכות של יינות לבנים לבין התפלגות ציוני האיכות של יינות אדומים. דווחו את תוצאות המבחן (***U***, המסומן בפלט המבחן כ-W, ו-***p*-value**).

ג. חשבו את ערכי U_1, U_2 עבור מבחן מאן-וויטני מסעיף ב'. אילו מהם הופיע בפלט שהתקבל בסעיף ב'?

ג. גודל האפקט במבחן מאן-וויטני נקרא **rank-biserial correlation**, ובמבחן **דו-זנבי** הוא נתון ע"י:

$$\text{RBC} = \left|\frac{U_1 - U_2}{n_1 n_2}\right|$$

בהתאם לפרשנות U_1, U_2 שהוצגה בתרגול, $\frac{U_1}{n_1 n_1}$ הוא שיעור זוגות התצפיות שבהם התצפית ממדגם 1 הייתה גדולה מהתצפית ממדגם 2, מתוך כל זוגות התצפיות האפשריים. $\frac{U_2}{n_1 n_1}$ הוא שיעור זוגות התצפיות שבהם התצפית ממדגם 2 הייתה גדולה יותר מהתצפית ממדגם 1, מתוך כל זוגות התצפיות האפשריים. גודל האפקט במבחן הדו-זנבי הוא ערכו המוחלט של ההפרש בין שיעורים אלה. נזכור שמתקיים $U_1 + U_2 = n_1 n_2$ ולכן:

$$\text{RBC} = \left|\frac{U_1 - U_2}{n_1 n_2}\right| = \left|\frac{U_1 - (n_1 n_2 - U_1)}{n_1 n_2}\right| = \left|\frac{2U_1}{n_1 n_2} - 1\right|$$

ובאותו אופן גם:

$$\text{RBC} = \left|1 - \frac{2U_2}{n_1 n_2}\right| = \left|\frac{2U_2}{n_1 n_2} - 1\right|$$

בחרו באחד מהביטויים של rank-biserial correlation, חשבו אותו ודווחו את גודל האפקט עבור המבחן שבוצע בסעיף ב'.