מטלת פרויקט 5

מטרת משימה 5 בפרויקט היא לתרגל את הגישה הבייסיאנית והתמודדות עם נתונים חסרים.

במשימה זו ניתן להשתמש בכל פונקציה בפייתון. הפעילו שיקול דעת ובחרו בעצמכם כיצד להציג את התוצאות של הסעיפים השונים

חלק ראשון: הגישה הבייסיאנית

בחלק זה השתמשו בשאלת מבחן בה השתמשתם במשימה 2: האם ההתפלגות של משתנה רציף X בקטגוריה אחת שונה מההתפלגות של X בקטגוריה השנייה, כאשר הקטגוריות נקבעות על ידי משתנה בינארי Y. אם יש לכם יותר משתי קטגוריות, הגבילו את עצמכם לשתי קטגוריות. כל רווחי הסמך או המהימנות יבוצעו ברמה של 95%.

1. בחרו באופן אקראי תת-מדגם בגודל 200. אלו הנתונים שאיתם נעבוד בחלק זה. נתייחס לנתונים האלו כנתונים הנצפים. בנוסף בחרו באקראי תת-מדגם בגודל 1000 שאינו מכיל נקודות מהנתונים הנצפים, ואליו נתייחס כנתוני העבר.
2. נגדיר משתנה חדש בינארי Z המבוסס על המשתנה X באופן הבא: נבחר ערך סף τ כך ש $Z = \begin{cases} 1 & X > \tau \\ 0 & X \leq \tau \end{cases}$. תהליך כזה מכונה דיכוטומיזציה. הסף τ יכול להיקבע על פי ערך שנראה לכם מתאים מהנתונים או כחציון (או אחוזון אחר) של המשתנה X.
   נגדיר את ההסתברות
   $$P(Z = 1|Y = j) = p_j \qquad j = 1,2$$
   אנו מעוניינים לאמוד את לוג יחס הסיכויים $\psi = \eta(p_1) - \eta(p_2)$, כאשר
   $\eta(p) = \log \frac{p}{1-p}$

   א. אמדו את ψ וחשבו רווח סמך מבוסס בוטסטראפ.

   ב. השתמשו בפריור יוניפורמי סטנדרטי עבור כל p_j $(j = 1,2)$, אמדו את ψ וחשבו רווח מהימנות.

   ג. השתמשו בפריור של ג'פרי עבור כל p_j $(j = 1,2)$, אמדו את ψ וחשבו רווח מהימנות.

   ד. השתמשו בנתוני העבר כדי לחשב פריור ל-p_j $(j = 1,2)$. אתם יכולים להניח שהפריור הוא ממשפחת $Beta$ (כלומר, אמדו את הפרמטרים של ההתפלגות מן נתוני העבר, אפשר גם באמצעות פונקציות ספריה). חשבו את ההתפלגות האפוסטריורית, אמדו את ψ וחשבו רווח המהימנות.

   ה. השוו בין האומדים השונים ל-ψ. מהי מסקנתכם?

   הדרכה לסעיפים ב'-ה': ניתן להיעזר בסימולציות כדי לחשב רווחי מהימנות (ראו בדוגמה 11.4 בספר).

חלק שני: נתונים חסרים

בחלק זה נרצה להשוות בין השיטות השונות לטיפול בנתונים חסרים. בחלק זה, בחרו משתנה מסביר רציף שנסמנו ב-X ומשתנה מוסבר אחד שהוא רציף שנסמנו ב-Y. אנחנו נייצר באופן מלאכותי נתונים חסרים במשתנה המוסבר ונבחן את ההצלחה של השיטות השונות.

1. בחרו באופן אקראי תת-מדגם בגודל 1000 ללא נתונים חסרים.
2. נרצה לאמוד את Y בעזרת רגרסיה לינארית על המשתנה המסביר. אמדו את מקדמי הרגרסיה β_0, β_1 כאשר אין נתונים חסרים וחשבו להם רווחי סמך (בעזרת מטריצת השונות).
3. נרצה למחוק כ-500 מהערכים של Y כך שככל ש-Y יותר גדול, הסיכוי שלו להימחק יותר גדול.

   הדרכה: ניתן לעשות זאת במספר דרכים. דרך אחת היא: סדרו את ערכי ה-Y מקטן לגדול. עבור $i = 1, \dots, 1000$ הגרילו משתנה מקרי ברנולי עם הסתברות להצלחה p_i התלויה ב-i כך שלמשל

   $$p_1 = \frac{1}{5} < \cdots < p_{500} = \frac{1}{2} < \cdots < p_{1000} = \frac{4}{5}$$

   ומחקו את כל הנקודות שמשתנה ברנולי שלהם יצא אחד.
4. נרצה לחזור על שאלה 2 בחלק זה כאשר ישנם נתונים חסרים. השתמשו במאגר הנתונים שקיבלתם בסעיף הקודם.
   א. אמדו את מקדמי הרגרסיה על בסיס הנתונים השלמים בלבד, ללא שורות בהם יש נתונים חסרים. חשבו להם רווחי סמך (בעזרת מטריצת השונות).
   ב. השלימו את הנתונים החסרים בעזרת הממוצע של ערכי ה-Y. השלמה כזו נקראת mean imputation. אמדו את מקדמי הרגרסיה וחשבו להם רווחי סמך (בעזרת מטריצת השונות). האם התוצאה שקיבלתם שונה מהסעיף הקודם?
   ג. השוו את התוצאות שקיבלתם בשני הסעיפים האלו לשאלה 2.
5. נרצה להשתמש בשיטת הנראות לנתונים החסרים.
   א. הסתכלו על ההסיטוגרמה של ערכי המשתנה המסביר X. הציעו מודל להתפלגות של ה-Xים (למשל, ה-Xים מתפלגים נורמלית, או אחיד).
   ב. נרצה לאמוד את התוחלת של המשתנה המוסבר Y באמצעות שיטת הנראות. שימו לב שבמקרה זה, ניתן להעזר במשוואה

   $$\hat{\mu}_Y = \hat{\beta}_0 + \hat{\beta}_1 \hat{E}[X]$$

   הסבירו מדוע זו המשוואה המתקבלת מהפעלת שיטת הנראות, וציינו תחת אילו הנחות ניתן להשתמש בשיטה.
   ג. עבור כל אחד ממודלי הרגרסיה שקיבלתם בשאלה 2 ובסעיפים 4א' ו4ב' (כזכור, בכל אחד מהסעיפים קיבלתם אומדים למקדמי הרגרסיה β_0, β_1) אמדו את התוחלת של המשתנה המוסבר Y בעזרת המשוואה מהסעיף הקודם. השוו בין התוצאות שקיבלתם.