מתרגל: ניר קרת
סמסטר ב' תשפ"א

תרגיל 8

- יש לענות על כל השאלות.
- הגשת התרגילים היא **בזוגות קבועים**.
- שאלות המסומנות ב-** הן שאלות רשות.
- את התרגילים יש להגיש לתיבת ההגשה במודל.
- יש להגיש הכל בקובץ PDF אחד, למשל על ידי שימוש ב Rmarkdown.

1) שאלה זו מספקת מבט אחר על ה *Jackknife*, שיכול לתת תובנות מעניינות. נתון לנו מדגם $\vec{x}$ של n תצפיות *iid*. כמו כן, נניח והאומד שלנו הינו הפעלת הפונקצייה s על המדגם, כלומר $s(\vec{x}) = \hat{\theta}$. נגדיר את $\hat{\theta}_{(-i)}$ להיות הפעלת הפונקצייה s על המדגם, ללא התצפית ה-i.

כעת נגדיר "*pseudo-values*": $\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{(-i)}$.

א. הראו, שאם האומד שלנו הינו לינארי, אז מתקבל שה*pseudo-values* הם *iid*. אומד לינארי הינו מהצורה $s(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$ עבור פונקצייה f כלשהי.

ב. הראו שאם ניקח את ממוצע ה*pseudo-values* , אז נקבל את האומד ל-θ המתוקן להטייה לפי *JK*:

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{\theta}_i = \hat{\theta}_{JK} = \hat{\theta} - \widehat{bias}_{JK}(\hat{\theta})$$

ג. הראו שאם נתייחס אל ה pseudo-values כתצפיות iid ונחשב את השונות המדגמית שלהם (הבלתי מוטה), נקבל את אומד JK לשונות.

ד. מה ניתן להסיק מכך על "ההנחות" ששיטת JK מניחה על הפונקצייה s כדי שאומדי הJK יהיו נכונים בקירוב? האם זה עולה בקנה אחד עם הדוגמא של החציון שראיתם בכיתה, בה JK "מפשל"?

2) במודל ישנו קובץ נתונים בשם *ex8data1* המכיל 4 עמודות. הניחו כי הנתונים מגיעים מהתפלגות רב-מימדית כלשהי.

א. הפרמטר אותו אנו מעוניינים לאמוד הינו הערך העצמי הגדול ביותר במטריצת השונויות של ההתפלגות ממנה מגיעים הנתונים. מצאו אומד נקודתי עבור פרמטר זה, ומצאו רווח סמך בר"ס 95% בעזרת בוטסטרפ כאשר *B=1000*. (ניתן להשתמש בפונקצייה *eigen* ב-*R*). את רווח הסמך עליכם לחשב בשיטות 2 ו-3 שלמדתם בכיתה: על ידי *pivot* ועל ידי אחוזונים.

ב. בעזרת מדגמי הבוטסטרפ שיצרתם בסעיף הקודם, מצאו אומד להטייה של האומד הנקודתי מהסעיף הקודם ואומד לשונות שלו.
ג. מצאו אומד להטייה ולשונות של האומד שלכם באמצעות *jackknife* (כל פעם עליכם להסיר שורה מהמדגם). השוו את התוצאות שקיבלתם לתוצאות של ה*bootstrap*.

3) פתרו את שאלת שיעורי הבית שבשקופית 11 אשר במצגת 6.

4) במודל ישנו קובץ נתונים בשם *ex8data2* ובו שתי עמודות *X* ו –*Y*. נניח ו-*Y* הוא המשתנה התלוי ו-*X* הינו המשתנה המסביר, וברצוננו למדל את הקשר $E(Y|X=x)$ בשיטה א-פרמטרית של *k-NN (k Nearest Neighbours)*. לפי שיטה זו, נגדיר את *k* ("מספר השכנים"), ובהינתן ערך x_{new} מסויים נמצא את ה-"שכונה" ($N_{x_{new}}$) של *k* התצפיות שערכי ה-*X* שלהן הם הקרובים ביותר לערך x_{new} זה. התחזית שלנו עבור אותו x_{new} תהיה:

$$\hat{Y}_{new} = \frac{1}{k} \sum_{i \in N_{x_{new}}} Y_i$$

ברצוננו למצוא ערך *k* אופטימלי. נניח ואנו מתלבטים בין 3 ערכי *k* אפשריים:

K=3,15,100.

א. על ידי *leave-one-out CV*, מצאו מהו הערך *k* הממזער את מדד ה *PRESS-* (*Predicted Residual Error Sum of Squares*)

$$\mathrm{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2$$

ב. הגרילו 1000 ערכים חדשים באופן אחיד בטווח הנצפה של ערכי *x*, וחשבו עבורם תחזית לפי *k-NN*, באמצעות מדגם הלמידה המקורי. עשו זאת לכל אחד משלושת הערכים המוצעים עבור *k*. שרטטו את קווי התחזית עבור התצפיות החדשות, לפי כל אחד מהערכים של *k*, תארו את ההבדלים, ונמקו באיזה הייתם מעדיפים להשתמש. האם זה עולה בקנה אחד עם הסעיף הקודם?

מתרגל: ניר קרת
סמסטר ב' תשפ"א

(שאלת רשות) (שאלה *16.15* מהספר *An introduction to the bootstrap* של *Efron, Tibshirani*)**

5) נסתכל על בעיית ה *one-sample*, כאשר נתון לנו מדגם בודד $\vec{X}$ וברצוננו לבדוק את ההשערה

$$H_0: E(X) = \mu_0$$
$$H_1: E(X) \neq \mu_0$$

מכיוון שלא בטוח שהנתונים מגיעים מההתפלגות תחת H_0 ראיתם בכיתה אפשרות לתקן את הנתונים על ידי כך שנגדיר

$\widetilde{x_k} = x_k - \bar{x} + \mu_0$ *k=1,…,n*

וכעת נגריל מדגמי בוטסטרפ מ $\vec{\tilde{x}}$.

הצעה חלופית לתקן את המדגם ככה שיקיים את השערת האפס בעולם הבוטסטרפ היא ההצעה הבאה:

במקום לתת משקל של 1/n לכל תצפית במדגם, אנו ניתן משקל אחר p_i לכל תצפית, כך שיתקיים $\sum_{i=1}^{n} p_i x_i = \mu_0$. בנוסף, מכיוון שאנו רוצים לשמור על F^* (ההתפלגות המתוקנת) קרובה ככל האפשר ל $\hat{F}$, נרצה למזער את ה"מרחק" בין ההתפלגויות. ישנן דרכים שונות לחשב מרחק בין התפלגויות, אם כי אחת הנפוצות היא "מרחק" ה Kullback–Leibler (למעשה המילה "מרחק" אינה מדוייקת, כי מדובר במדד שאינו סימטרי), המוגדר באופן הבא (עבור התפלגויות דיסקרטיות):

$$d(F_1, F_2) = \sum_{x \in S} p(x) \ln\left(\frac{p(x)}{q(x)}\right)$$

כאשר F_1 ו-F_2 מוגדרות על אותו מרחב הסתברות S, וכן מתקיים

$Supp(F_1) \subseteq Supp(F_2)$, ו–$p(x), q(x)$ הן ההסתברויות לקבל את הערך x תחת F_1 (p) ו-F_2 (q) בהתאמה. ה"מרחק" הוא מהתפלגות F_2 להתפלגות F_1, כלומר F_2 תהיה במקרה שלנו $\hat{F}$.

א. השתמשו בכופלי לגרנז', והראו שהפתרון האופטימלי הממזער את המרחק הנ"ל תחת האילוצים $\sum p_i = 1$,$\sum p_i x_i = \mu_0$ הינו מן הצורה:

$$p_i = \frac{e^{tx_i}}{\sum_{j=1}^{n} e^{tx_j}}$$

כאשר t הינו קבוע אשר נבחר כך שיתקיים האילוץ $\sum p_i x_i = \mu_0$. (שימו לב שאין אפשרות למצוא אותו בצורה סגורה, אלא בשביל למצוא אותו יש לפתור משוואה בצורה נומרית).

ב. עבור הנתונים *ex8data3* השתמשו בשתי השיטות המתוארות לעיל, וכן במבחן *t* רגיל על מנת לבדוק את ההשערה:

$$\mu_0 = 5$$
$$\mu_0 \neq 5$$

והשוו בין התוצאות.

הערה: בשאלה זו אתם רשאים להשתמש בפונקצייה *uniroot* ב-*R* אשר מוצאת שורש לפונקצייה.