שיטות כריית נתונים ובינה עסקית – 096411

מבחן סיום – מועד א'

מרצים: דר' דוד עזריאל, דר' תמיר חזן

מתרגלים: ליאון ענבי, מיכל מלמד

25 ביולי 2017

ת.ז.:_____

הוראות – נא לקרוא בעיון רב

- משך הבחינה 3 שעות.
- חומר עזר מותר לבחינה הינו מחשבון וכל חומר כתוב.
- אין להפריד אף דף מטופס הבחינה.
- במבחן זה ארבע שאלות ובכל שאלה מספר סעיפים. יש לבחור שלוש מתוכן ולענות עליהן במלואן. משקלה של כל שאלה הינו 33 נקודות כאשר לציון הסופי תתווסף נקודה אחת נוספת.
- שימי לב: תיבדקנה רק שלוש השאלות הראשונות לפי סדר הופעתן במחברת הבחינה. אין טעם לפתור יותר משלוש שאלות מכיוון שהשאלה הרביעית לא תיבדק.
- המבחן מנוסח בלשון נקבה אך מתייחס לשני המינים
- בסיום המבחן יש למסור את טופס הבחינה.
- בהצלחה!!!

שאלה 1 (33 נק')

אומד רשת אלסטית (*Elastic Net*) מוגדר להיות אומד הרגרסיה הבא:

$$\widehat{\boldsymbol{\beta}}^{\boldsymbol{EN}} = \operatorname{argmin}_{\widetilde{\boldsymbol{\beta}}} \left\|\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}\right\|^2 + \lambda_1 \sum_{j=1}^{p} \left|\widetilde{\beta}_j\right| + \lambda_2 \left\|\widetilde{\boldsymbol{\beta}}\right\|^2$$

עבור $\lambda_1, \lambda_2 \geq 0$. כאשר $\boldsymbol{X}$ $(n \times p)$ הינה מטריצת התצפיות (ללא עמודת אחד).

א. (8) הראי כי אומד הרשת האלסטית הינו הכללה של *Lasso* וגם של *Ridge*.

ב. (15) נניח שמודל הרגרסיה הינו $\boldsymbol{Y} = \boldsymbol{\beta} + \epsilon$. כלומר, אין חותך, $n = p$ ומתקיים $\boldsymbol{X} = \boldsymbol{I}$.

השלימי את אומד הרשת האלסטית והוכיחי את התוצאה:

$$\hat{\beta}_i^{\,EN} = \begin{cases} \dfrac{Y_i - \frac{\lambda_1}{2}}{1+\lambda_2} & Y_i \geq ? \\ 0 & ? \leq Y_i \leq ? \\ \dfrac{Y_i + \frac{\lambda_1}{2}}{1+\lambda_2} & Y_i \leq ? \end{cases}$$

ג. (10) עבור λ_1 קבוע, האם ההטיה של האומד שחישבת בסעיף ב' עולה או יורדת עם הגדלת λ_2? באופן דומה, האם שונות האומד עולה או יורדת עם הגדלת λ_2? **בססי את תשובתך באופן מתמטי והסבירי <u>בקצרה</u> את התוצאה.**

שאלה 2 (33 נק')

תהי D התפלגות ויהי $S = (z_1, \dots, z_m)$ כאשר $z_i = (x_i, y_i)$ סט תצפיות אימון בלתי תלויות ושוות הסתברות (iid) ותהי z' תצפית בלתי תלויה ושוות הסתברות נוספת. תהי $U(m)$ התפלגות אחידה על פני $[m]$. יהי A אלגוריתם למידה ונסמן $A(S)$ את פלט האלגוריתם על מדגם S.

נגדיר $S^{(i)} = (z_1, \dots, z_{i-1}, z', z_{i+1}, \dots, z_m)$

א. (6) הסבירי את משמעות הביטוי $E_{S\sim D^m}\left[L_D\left(A(S)\right)\right]$ והראי כי לכל ערך של $1 \le i \le m$ מתקיים:

$$E_{S\sim D^m}\left[L_D\left(A(S)\right)\right] = E_{(S,z')\sim D^{m+1}}\left[l\left(A(S^{(i)}, z_i\right)\right]$$

ב. (6) הסבירי את משמעות הביטוי $E_{S\sim D^m}\left[L_S\left(A(S)\right)\right]$ ומדוע מתקיים:

$$E_{S\sim D^m}\left[L_S\left(A(S)\right)\right] = E_{(S,z')\sim D^{m+1}, i\sim U(m)}[l(A(S, z_i)]$$

ג. (7) הוכיחי כי

$$E_{S\sim D^m}\left[L_D\left(A(S)\right) - L_S\left(A(S)\right)\right] = E_{(S,z')\sim D^{m+1}, i\sim U(m)}\left[l\left(A(S^{(i)}, z_i\right) - l(A(S, z_i)\right]$$

ד. (7) הסבירי את משמעות השוויון שהתקבל וכיצד ניתן להשתמש בו על מנת להבטיח שלא תתקיים התאמת יתר תוך שימוש באלגוריתם מסוים.

ה. (7) כעת נגדיר באמצעות $I = \{i_1, \dots, i_{|I|}\}$ תת קבוצה של תצפיות מתוך S

תהיינה $Z = \{z'_1, \dots, z'_{|I|}\}$ תצפיות iid נוספות ונגדיר את

$$S^{(I)} = \left(z_1, \dots, z_{i_1-1}, z'_1, z_{i_1+1}, \dots, z_{i_2-1}, z'_2, z_{i_2+1}, \dots, z_{i_{|I|}-1}, z'_{|I|}, z_{i_{|I|}+1}, \dots, z_m\right)$$

נסחי את השוויון מסעיף ג' כך שיתאים לקבוצת התצפיות I ולמדגם $S^{(I)}$. הסבירי את תשובתך, תארי את ההבדלים וצייני יתרונות וחסרונות לכל גישה.

רמז: נגדיר את $\binom{m}{|I|}$ תתי הקבוצות בגודל $|I|$ מתוך אברי S באמצעות $\left[\binom{m}{|I|}\right]$ ונגדיר את $U_{|I|}$ להיות ההתפלגות האחידה על פני $\left[\binom{m}{|I|}\right]$ לצורך בחירת תת-הקבוצה.

שאלה 3 (33 נק')

נתונות m תצפיות $x_1, \dots, x_m$ ממימד $p > 2$ אשר מקיימות

$$\frac{1}{m}\sum_{i=1}^{m} x_i = 0$$

נסמן ב-X את המטריצה שמכילה את התצפיות בעמודות ואת מטריצת השונויות של התצפיות ב-$XX^T = \Sigma$.

עבור כל וקטור u נגדיר את $y = u^T x$ להיות הטרנספורמציה של x באמצעות u.

נחפש את וקטור היחידה u $(u^T u = 1)$ אשר ממקסם את השונות של התצפיות $y_1, \dots y_m$:

$$u_1 = argmax_u \left\{ \sum_{i=1}^{m} (y_i - \bar{y})^2 \right\}, \qquad s.t. \;\; u^T u = 1$$

ראינו כי u_1 הינו הוקטור העצמי של Σ אשר מתאים לערך העצמי המקסימלי λ_1

א. (5 נק') הסבירי **באופן מתמטי** לשם מה נחוץ האילוץ ש-u הינו וקטור יחידה.

ב. בבואנו להוסיף את הכיוון השני המשמר את מירב השונות (Principal Component 2), אנחנו ראשית אוכפים את האילוץ כי המשתנים $u_1^T x_i$ ו- $u_2^T x_i$ יהיו בלתי מתואמים: $\sum_{i=1}^{m} (u_1^T x_i)(u_2^T x_i) = 0$.

הסבירי בקצרה את הדרישה **והוכיחי** כי היא גוררת את האילוץ שu_1 ו-u_2 יהיו ניצבים זה לזה $u_1^T u_2 = 0$.

רמז: זכרי את הגדרת Σ ואת העובדה ש-u_1 הינו וקטור עצמי של Σ.

ג. ניתן לתאר את בעיית האופטימיזציה מסעיף ב' באופן הבא:

$$u_2 = argmax_{u:\, u^T u_1 = 0, \|u\| = 1} \left\{ \sum_{i=1}^{m} (u^T x_i)^2 \right\}$$

$$u_2 = argmax_{u,\lambda,\delta} \left\{ \sum_{i=1}^{m} (u^T x_i)^2 - \lambda(u^T u - 1) - \delta(u^T u_1) \right\}$$

כאשר הוקטור האופטימלי, u_2, הינו הוקטור העצמי של Σ אשר מתאים לערך העצמי השני בגודלו λ_2.

רמז: זכרי את האילוץ מסעיף ב' והיעזרי בו על מנת למצוא את ערך δ.

שאלה 4 (33 נק')

בידינו מדגם ובו שלוש תצפיות $\boldsymbol{Z} = X_1, X_2, X_3 \sim^{iid} F$, ונגדיר את הפונקציה $\phi(\boldsymbol{Z}) = \frac{X_1+X_2}{2}$.

ברצוננו לחשב אומד bootstrap עבור $Var\big(\phi(\boldsymbol{Z})\big)$.

א. (10) יהי $\boldsymbol{Z}^* = X_1^*, X_2^*, X_3^* \sim^{iid} \hat{F}_3$. הראי כי

$$\phi(\boldsymbol{Z}^*) = \begin{cases} \dfrac{X_1+X_2}{2} & with\ probability\ {}^2/_9 \\ \dfrac{X_1+X_3}{2} & with\ probability\ {}^2/_9 \\ \dfrac{X_2+X_3}{2} & with\ probability\ {}^2/_9 \\ X_1 & with\ probability\ {}^1/_9 \\ X_2 & with\ probability\ {}^1/_9 \\ X_3 & with\ probability\ {}^1/_9 \end{cases}$$

ב. (13) תארי דרך אנליטית לחישוב אומד bootstrap עבור $Var\big(\phi(\boldsymbol{Z})\big)$.

אין צורך להציג נוסחה מפורשת אך עלייך להציג <u>באופן מפורט וברור</u> דרך (אלגוריתם) לחישוב האומד באמצעות סעיף א' וללא שימוש במדגמי bootstrap.

ג. (10) כעת הניחי כי בידייך התצפיות הבאות במדגם $\boldsymbol{Z} = (X_1 = 0, X_2 = 0, X_3 = 1)$. והרצת 1000 מדגמי bootstrap בתוצאות הבאות:

0,0,0	326 פעמים
0,0,1	130 פעמים
0,1,0	141 פעמים
0,1,1	84 פעמים
1,0,0	149 פעמים
1,0,1	68 פעמים
1,1,0	73 פעמים
1,1,1	29 פעמים

מהו ערך אומד bootstrap עבור $Var\big(\phi(\boldsymbol{Z})\big)$ בהתבסס על 1000 המדגמים הנ"ל?