שיטות כריית נתונים ובינה עסקית – 096411

מבחן סיום – מועד א'

מרצים: דר' דוד עזריאל, דר' תמיר חזן

מתרגל: ליאון ענבי

25 ביולי 2016

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')

נתונות n תצפיות חד מימדיות $X_1, \dots, X_n$ אשר מגיעות מעירוב של שתי קבוצות. התצפיות בכל כל קבוצה מתפלגות מעריכית. כל תצפית מגיעה מהקבוצה הראשונה בהסתברות π.

תזכורת: פונקציית הצפיפות של משתנה מקרי מעריכי עם פרמטר θ היא $\theta e^{-\theta x}$.

א. (10 נק') הראי כי פונקציית *log-likelihood* של הנתונים הינה:

$$l(X|\theta_1, \theta_2, \pi) = \sum_{i=1}^{n} log\left(\pi\theta_1 e^{-\theta_1 x_i} + (1-\pi)\theta_2 e^{-\theta_2 x_i}\right)$$

ב. (15 נק') תארי אלגוריתם *EM* לאמידת הפרמטרים π, θ_1, θ_2.

תזכורת: עבור $Y_1, \dots, Y_m \sim \exp(\theta)$ אומד נראות מרבית ל- θ הוא $\hat{\theta} = \frac{1}{\frac{1}{m}\sum_{i=1}^{m} Y_i}$

נניח כי התקבלו בסעיפים הקודמים האומדים הבאים: $\hat{\pi} = 0.6, \hat{\theta}_1 = \frac{1}{3}, \hat{\theta}_2 = \frac{1}{4}$.

ג. (8 נק') מאיזו קבוצה סביר יותר לקבל תצפית שערכה $X = 5$? הסבירי תשובתך.

שאלה 2 (33 נק')

נתון מודל רגרסיה פשוטה: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \;\; i = 1, \dots, n$ כאשר מתקיים $\sum_{i=1}^{n} X_i = 0$

א. (11 נק') הראי שאומד *RIDGE* ל-$\hat{\beta}_1$ עבור פרמטר λ מסוים הוא:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2 + \lambda}$$

ב. (11 נק') הראי שהתוחלת והשונות של האומד מסעיף א' הינן:

$$E(\hat{\beta}_1) = \beta_1 \frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i^2 + \lambda} \qquad Var(\hat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^{n} X_i^2}{\left(\sum_{i=1}^{n} X_i^2 + \lambda\right)^2}$$

ג. (11 נק') האם ההטיה בריבוע של האומד מסעיף א', $\left(E\left[\hat{\beta}_1 - \beta_1\right]\right)^2$ עולה או יורדת ב-λ? האם השונות של האומד עולה או יורדת ב-λ? הסבירי את התוצאה תוך התייחסות למטרה של השימוש ב-*RIDGE*.

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-$\Sigma = XX^T$.

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 $y_1, \dots y_m$:

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

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

באופן שקול ניתן לבטא את האילוץ הנ"ל ע"י שימוש בכופלי לגרנז' וניסוח בעיית האופטימיזציה הבאה:

$$u_1 = argmax_u \left\{ \sum_{i=1}^{m} (y_i - \bar{y})^2 - \lambda(u^T u - 1) \right\}$$

ב. (15 נק') הראי כי u_1 הינו וקטור עצמי של מטריצת השונויות Σ

תזכורת: וקטור עצמי v למטריצה A עם ערך עצמי λ שמתאים לו מקיימים:

$$Av = \lambda v$$

ג. (5 נק') הראי כי מבין הוקטורים העצמיים של Σ, הוקטור העצמי u_1 אשר ממקסם את השונות של Y יהיה הוקטור העצמי אשר מתאים לערך העצמי המקסימלי.

רמז: זכרי כי u_1 הינו: (a) וקטור עצמי של Σ, (b) וקטור יחידה.

ד. (8 נק') נתונות התצפיות הבאות: $(-1,1), (0,0), (1,-1)$

מצאי את הוקטור u_1 עבורן, הציגי את ערכי y_1, y_2, y_3 ואת שגיאת השחזור של התצפיות.

שאלה 4 (33 נק')

בידינו m תצפיות $(x_1, y_1), \dots, (x_m, y_m)$ כאשר $y \in \{1, \dots, k\}$, לכל (x, y) נגדיר וקטור $\phi(x, y)$ ונתייחס לפונקציית ההסתברות הבאה:

$$p(y|x, w) = \frac{e^{w^T \phi(x,y)}}{\sum_{y'=1}^{k} e^{w^T \phi(x,y')}}$$

במקרה הבינארי $y \in \{-1, +1\}$ הגדרנו:

$$p(y = +1|w, x) = \frac{e^{w^T x}}{1 + e^{w^T x}}$$

א. (5 נק') הראי כי המקרה הבינארי הינו מקרה פרטי של המקרה ובו k מחלקות. כלומר, הראי כי קיים $\phi(x, y)$ עבורו פונקציות ההסתברות שוות.

אמד נראות מרבית יהיה:

$$w^* = argmax_w \left\{ \sum_{i=1}^{m} \log\big(p(y_i|x_i, w)\big) \right\}$$

ב. (13 נק') עבור המקרה הבינארי, הראי כי מתקיים:

$$\sum_{i=1}^{m} p(y = +1|x_i, w^*) x_i = \sum_{i; y_i = +1} x_i$$

שימי לב כי x, w שניהם וקטורים.

כעת נניח כי שלושה סטודנטים מימשו שלושה מסווגים f_1, f_2, f_3 כאשר כל מסווג הינו פונקציה $f\big((X_{train} Y_{train}), x_{new}\big) \in \{1, \dots, m\}$ אשר מקבלת תצפיות אימון $(X_{train} Y_{train})$ ומחזירה סיווג (מחלקה) עבור התצפית החדשה x_{new}.

על מנת להשוות בין המסווגים ולקבוע מי מהם בעל ביצועים טובים יותר חילקנו את m התצפיות לשתי קבוצות G_1, G_2 וחישבנו את הערכים הבאים לכל מסווג:

$$E_{1,1} = \frac{100}{|G_1|} \sum_{(x_i, y_i) \in G_1} I\{f(G_1, x_i) = y_i\}$$

$$E_{1,2} = \frac{100}{|G_2|} \sum_{(x_i, y_i) \in G_2} I\{f(G_1, x_i) = y_i\}$$

ג. (5 נק') מהם שני הערכים $E_{1,1}, E_{1,2}$? הסבירי תשובתך בקצרה.

ד. (10 נק') התקבלה התוצאה הבאה עבור שלושת המסווגים:

	f_1	f_2	f_3
$E_{1,1}$	*10*	10	*12*
$E_{1,2}$	*15*	50	*12*

האם ניתן לומר כי המסווג f_3 עדיף על סמך התוצאות? הציעי לפחות בדיקה נוספת אחת שתשפר את יכולת ההחלטה לגבי מי המסווג העדיף מבין השלושה ונמקי את תשובתך.