למידה סטטיסטית מבוססת נתונים– 096411

מועד א'

מרצים: דר' דוד עזריאל, דר' תמיר חזן

מתרגלים: עומר בן פורת, מיכל מלמד

16 ביולי 2018

ת.ז.:_____

הוראות – נא לקרוא בעיון

- משך הבחינה 3 שעות.
- חומר עזר מותר לבחינה הינו מחשבון כיס וחמישה דפי A4.
- אין להפריד אף דף מטופס הבחינה.
- במבחן זה שני חלקים:
  - **חלק פתוח**: כולל ארבע שאלות, ובכל שאלה מספר סעיפים. יש לבחור שלוש מתוכן ולענות עליהן במלואן. משקלה של כל שאלה הינה 30 נקודות. שימו לב: תיבדקנה רק שלוש השאלות הראשונות לפי סדר הופעתן במחברת הבחינה. אין טעם לפתור יותר משלוש שאלות מכיוון שהשאלה הרביעית לא תיבדק.
  - **חלק סגור**: כולל חמש שאלות נכון\לא נכון בשווי ארבע נקודות כל אחת, יש לענות על כל השאלות בחלק זה, כאשר הציון המרבי אותו ניתן לקבל עבור חלק זה הוא 10 נקודות. כל שאלה מכילה טענה, ויש להחליט אם הטענה נכונה או לא. את התשובה יש לענות בטופס הבחינה במקום המתאים לכך.
- טופס הבחינה כולל 6 דפים.
- בסיום המבחן יש למסור את טופס הבחינה ומחברת הטיוטה.
- בהצלחה!!!!

חלק פתוח (90 נקודות)

בחלק זה יש לבחור שלוש שאלות בלבד. כל שאלה שווה 30 נקודות.

שאלה 1:

(רק תשובה הכוללת נימוקים מתמטיים ומילוליים מלאים תקבל את מלוא הנקודות)

נתון מודל רגרסיה לינארית פשוטה עם שתי תצפיות:

$$Y_1 = w_1 + w_2 X_1 + \epsilon_1, \;\; Y_2 = w_1 + w_2 X_2 + \epsilon_2$$

נתון כי התצפיות ממורכזות, כלומר $X_1 + X_2 = Y_1 + Y_2 = 0$.

א. (10 נק') הראו כי אמד ריבועים פחותים הינו $\widehat{w}_1 = 0$ וכן $\widehat{w}_2 = \frac{Y_1}{X_1}$.

ב. (10 נק') הראו כי אמד Ridge הינו $\widehat{w}_1^R = 0$, וכן $\widehat{w}_2^R = \frac{Y_1 X_1}{X_1^2 + \frac{\lambda}{2}}$.

ג. (5 נק') הראו כי עבור Lasso מתקיים $\widehat{w}_1^L = 0$, וכן $\widehat{w}_2^L$ פותר את בעיית המינימיזציה הבאה:

$$\underset{\widetilde{w}}{\operatorname{argmin}}\, L(\widetilde{w}) \text{ , where } L(\widetilde{w}) = 2(Y_1 - \widetilde{w} X_1)^2 + \lambda |\widetilde{w}|$$

ד. (5 נק') הראו כי אם בנוסף $4|X_1 Y_1| < \lambda$, מתקיים $\widehat{w}_2^L = 0$

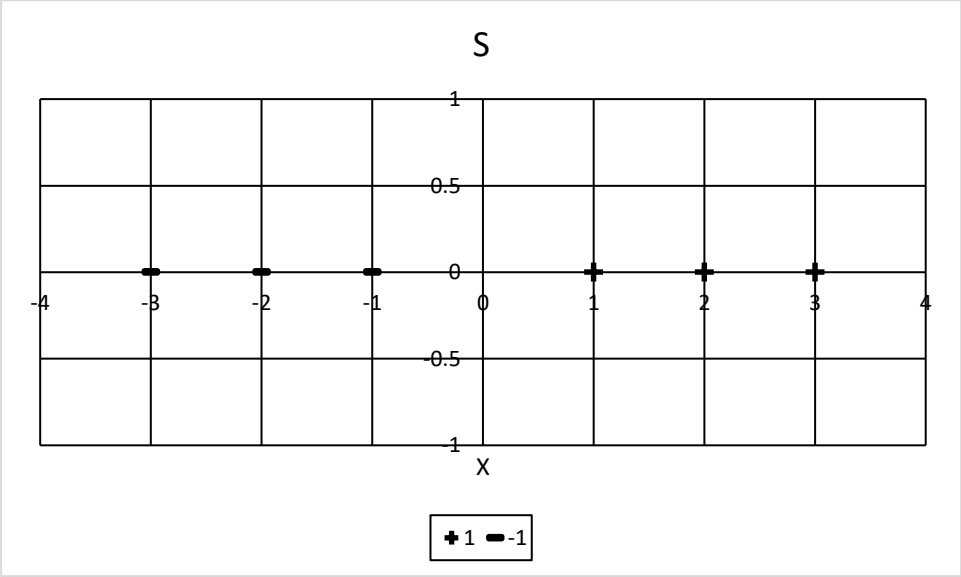
(רמז: הראו כי $L(0) = 0$ וכן שכאשר $4|X_1 Y_1| < \lambda$ אז לכל $\widetilde{w}$ מתקיים $L(\widetilde{w}) \geq 0$).

שאלה 2:

נתון מדגם עם 6 תצפיות $S = \{(x_1, y_1), \dots (x_6, y_6)\}$ כאשר כל x_i הוא מספר ממשי וכל $y_i = sign(x_i)$

באופן מפורש, המדגם מכיל את התצפיות: $S = \{(-3,-1), (-2,-1), (-1,-1), (1,1), (2,1), (3,1)\}$

ובאופן גרפי:



א. (5 נק') רשמו את אלגוריתם ה *perceptron*.
כמה מסווגים שונים האלגוריתם יכול למצוא עבור מדגם האימון הנתון?
מה מספר הצעדים המקסימלי שהאלגוריתם יבצע עד לעצירה עבור מדגם האימון הנתון?
רשמו את אחד המסווגים האופטימליים.

ב. (10 נק') רשמו את התיאור המתמטי של *hard-SVM*.
מהו המסווג האופטימלי עבור מדגם האימון הנתון?

ג. (10 נק') מהו התיאור המתמטי של *soft-SVM* עם מקדם רגולריזציה λ? הנח כי החותך $b = 0$
ניתן להשתמש בפונקציית ההפסד $\max\{0, 1 - y\langle w, x\rangle\}$, שנקראת *hinge-loss*.
מהו ה- w האופטימאלי עבור λ כלשהו?
עבור אילו ערכים של λ פונקציית ההפסד היא 0 עבור כל נקודות המדגם?

ד. (5 נק') שגיאת הסיווג היא מספר הפעמים שבהם תווית האימון שונה מתווית הפרדיקציה, והיא נקראת *zero-one loss*. אם יש מסווג שעבורו ה *hinge-loss* שווה לאפס על מדגם האימון, מה אפשר להגיד על שגיאת הסיווג על מדגם האימון. נמקו את תשובתכם.

שאלה 3:

אלגוריתם ה *AdaBoost* מבצע שלושה צעדים:

i) $h_t = WL(D^{(t)}, S)$

ii) $\epsilon_t = \sum_{i=1}^{m} D_i^{(t)} 1_{[h_t(x_i) \neq y_i]}$ וגם $w_t = \log\left(\frac{1}{\epsilon_t} - 1\right)/2$

iii) $D_i^{(t+1)} \propto D_i^{(t)} e^{-w_t y_i h_t(x_i)}$

א. (5 נק') הסבירו את ההנחה על $WL()$ ואת כל אחד מהצעדים באלגוריתם.

ב. (10 נק') הראו ש $2\sqrt{\epsilon_t(1-\epsilon_t)} = \sum_{i=1}^{m} D_i^{(t)} e^{-w_t y_i h_t(x_i)}$.

ג. (10 נק') הראו ש $1/2 = \sum_{i=1}^{m} D_i^{(t+1)} 1_{[h_t(x_i) \neq y_i]}$.

ד. (5 נק') מהו המספר המינימלי של צעדים הנדרש באלגוריתם ה *AdaBoost* בכדי שפונקציית ההפסד שלו תהיה 0 על מדגם האימון?

שאלה 4:

נתון מדגם $X_1, \dots X_m$ כאשר כל תצפית בו הינה מספר ממשי חד מימדי, אשר נדגמה באופן בלתי תלוי מהתפלגות $\mathcal{D}$. עוד ידוע כי $\mathcal{D}$ הינה עירוב (mixture) של שני אשכולות (clusters), כך שהתצפיות באשכול אחד מתפלגות לפי התפלגות נורמלית (עם תוחלת ושונות לא ידועים) והתצפיות באשכול השני מתפלגות לפי התפלגות מעריכית עם פרמטר $\lambda = 1$. נזכיר כי פונקציות הצפיפות של ההתפלגויות הנ"ל נתונות ע"י

$$f_{exp}(t) = \lambda e^{-\lambda t} 1_{t \geq 0}, \quad f_{normal}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

א. (5 נק') הגדירו במפורש את הפרמטרים הלא ידועים ורשמו את לוג-הנראות של התצפיות.

ב. (20 נק') נסחו אלגוריתם *EM* להערכת הפרמטרים הלא ידועים. רמת הפירוט צריכה להיות כזאת שמאפשרת מימוש של האלגוריתם. בפרט, הסבירו מהו שלב *E* ומהו שלב *M*. יש לכתוב נוסחאות מפורשות לעדכון כל אחד מן האמדים.

ג. (5 נק') עבור תצפית שלילית (כלומר תצפית i כך ש $X_i < 0$), מה תהיה ההסתברות שהיא שייכת לאשכול הנורמלי ולא למעריכי? הסבירו את תשובתכם.

חלק סגור (10 נקודות)

בחלק זה חמש שאלות נכון\לא נכון, בשווי ארבע נקודות כל אחת. יש לענות על כל השאלות ומספר הנקודות המרבי לחלק הוא 10. **יש לסמן X במקום המתאים בטבלה המופיעה בסוף החלק הסגור.**

שאלה 5:

יש ברשותנו קופסא שחורה ALG אשר, בהינתן קבוצה של וקטורים $\mathcal{S}$ ומספר k, מריצה את אלגוריתם *k-means* שנלמד בכיתה. הקריאה $ALG(\mathcal{S},k)$ מחזירה את ערך פונקציית המטרה המתאים.

טענה: "אם $k_1 < k_2$, בהכרח מתקיים ש $ALG(\mathcal{S},k_1) \geq ALG(\mathcal{S},k_2)$".

שאלה 6:

טענה: "אלגוריתם *AdaBoost* מייצר סדרה של מסווגים חלשים בלתי תלויים, ובכך מגדיל את ההסתברות להצלחה בתיוג דוגמה חדשה".

שאלה 7:

בתרגול 13 לקחנו את נתוני *USArrests*, הכוללים 50 תצפיות וארבעה משתנים מסבירים. לאחר נירמול ביחס לממוצע ולסטיית התקן, הפעלנו אלגוריתם *PCA* על המדגם. המודל שחזר הינו

Perform PCA

```
pr.out=prcomp(USArrests,center=TRUE, scale=TRUE)
pr.out
```

```
## Standard deviations (1, .., p=4):
## [1] 1.5748783 0.9948694 0.5971291 0.4164494
##
## Rotation (n x k) = (4 x 4):
##                 PC1        PC2        PC3         PC4
## Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
## Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
## UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
## Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

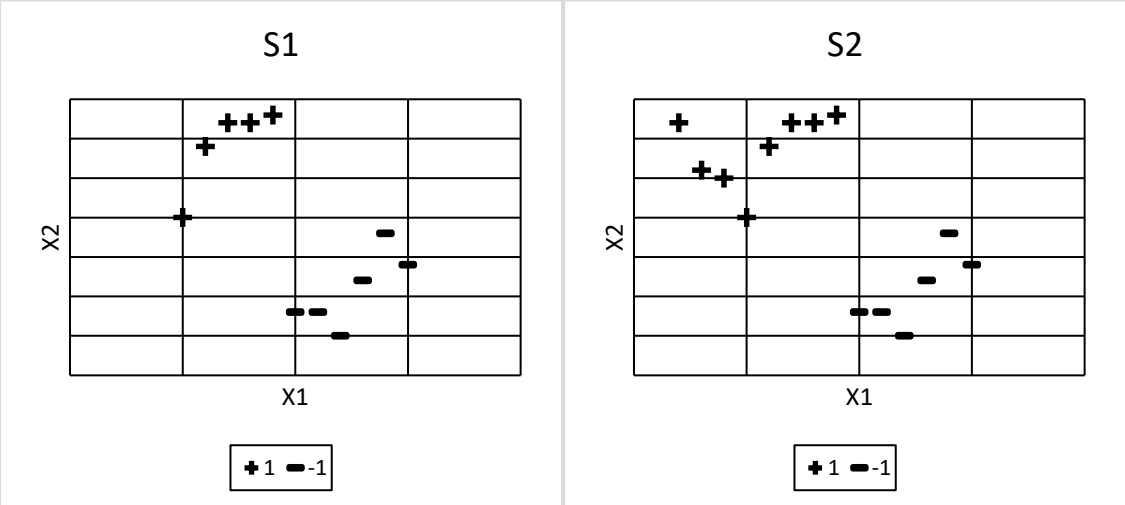
טענה: "שימוש בשלושת ה *PC* (*principal components*) הראשונים יביא לשחזור של מעל 90% מן השונות".

שאלה 8:

טענה: "סיבוכיות זמן תיוג דוגמה חדשה ב *soft-SVM* עם קרנל לינארי זהה לתיוג עם קרנל פולינומיאלי".

שאלה 9:

הגרפים הבאים מציגים שני מדגמים, S1 ו- S2, כאשר S2 מכיל את כל התצפיות מתוך S1 וכן שלוש תצפיות נוספות (השייכות למחלקה 1+).



הפעילו את אלגוריתם *hard-SVM* על כל אחד מהמדגמים, וקיבלו מישור מפריד ביחס לכל מדגם.

טענה: "המישור המפריד המתאים למדגם S1 שונה מזה המתאים למדגם S2".

יש לסמן את תשובותיכם בטבלה הבאה:

	הטענה נכונה	הטענה אינה נכונה
שאלה 5		
שאלה 6		
שאלה 7		
שאלה 8		
שאלה 9		