Notebook No:_____:מס' מחברת I.D number:_____:מס' ת.ז

בית ספר "אפי ארזי" למדעי המחשב המרכז הבינתחומי
The Efi Arazi school of computer science
The Interdisciplinary Center

סמסטר ב' תשע"ז
Spring 2017

מבחן מועד א בלמידה ממוכנת
Machine Learning Exam A

Lecturer : Prof Ariel Shamir,
Dr. Zohar Yakhini
Time limit : 3 hours
Additional material or calculators are not allowed in use!

Answer 5 out of 6 from the following questions (each one is 20 points)
Good Luck!

מרצים: פרופ אריאל שמיר,
ד"ר זהר יכיני

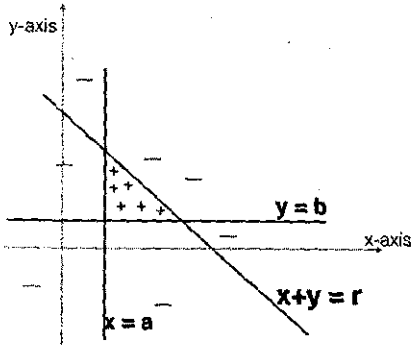
משך המבחן: 3 שעות
אין להשתמש בחומר עזר ואין להשתמש במחשבונים!

יש לענות על 5 מתוך 6 השאלות הבאות
לכל השאלות משקל שווה (20 נקודות)
בהצלחה!

יש לכתוב בצד השמאלי של הדף בלבד

שאלה 1 (5 סעיפים)

ב-$\mathbb{R}^2$, נתון מרחב היפותזות של "משולשים ישרי זווית ושווי שוקיים שפונים ימינה" – שהיא קבוצת כל המשולשים שווי השוקיים כאשר הצלעות השוות הן מקבילות לצירים והצלע השלישית נמצאת מימין למעלה (ראי ציור למטה).



משולשים כאלו נוצרים ע"י חיתוך של הקווים הישרים x=a, y=b ו- x+y=r, כאשר r>a ו- r>b.

בצורה פורמלית אפשר לייצג את מרחב ההיפותזות כך:

$$H_\Delta = \{h_{a,b,r} : a, b, r \in \mathbb{R}, r > a, r > b\}$$

כאשר חלוקת המרחב מוגדרת לחיובי ושלילי:

$$h_{a,b,r} = \begin{cases} 1 & if\ x \geq a\ and\ y \geq b\ and\ x + y \leq r \\ -1 & otherwise \end{cases}$$

כלומר, $h_{a,b,r}$ יסווג 1+ לכל דוגמא שתהיה בתוך המשולש (כולל אם היא נופלת על צלעותיו) ו 1- אחרת.

להלן שרטוט של היפותזה $h_{a,b,r}$ מתוך מרחב זה:

נתונים שלושת החסמים, שלמדנו בכיתה, על מס' דוגמאות האימון הנדרשות:

$$m \geq \frac{1}{\varepsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$

$$m \geq \frac{1}{2\varepsilon^2}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$

$$m \geq \frac{1}{\varepsilon}\left(8 \cdot VC(H)\log_2\frac{13}{\epsilon} + 4\log_2\frac{2}{\delta}\right)$$

א. בהינתן $1 \leq a = b \leq n$ וגם $1 \leq r \leq 2n$ ובנוסף ש- a, b, r הם מספרים שלמים, חשבי את גודל מרחב ההיפותזות.

ב. בהינתן ש n=100:

1. הגדירי אלגוריתם למציאת היפותזה קונסיסטנטית $h_c \in H_\Delta$ בהנחה שאין רעש (קבוצת האימון נוצרה ע"י קונספט $c \in H_\Delta$).
2. הראי שמרחב זה הוא PAC learnable ע"י האלגוריתם שהגדרת בסעיף הקודם (אין צורך להראות חסם הדוק).

ג. בהינתן $a = b = 0$ ו-r יכול להיות כל מס' בהתאם להגדרת המרחב (לא חייב להיות מס' שלם), חשבי את ה-VC dimension.

ד. נתון אלגוריתם למידה במרחב ההיפותזות מסעיף ב (המרחב מסעיף א כאשר n=100) המבטיח טעות אימון 0. כמה דוגמאות אימון צריך הלימוד במרחב ההיפותזות זה כדי להבטיח בהסתברות לפחות 90% היפותזה עם טעות לכל היותר של 0.05? הראי את החישוב בלבד (אין צורך לקבל תוצאה סופית).

ה. האם התשובה לסעיף ד תשתנה כאשר מרחב ההיפותזות יוגדר ע"פ סעיף ג? אם לא, מדוע? אם כן, הסברי כיצד תחושב כמות דוגמאות האימון הנדרשת? (אין צורך לקבל תוצאה סופית).

שאלה 2 (7 סעיפים)

נתון אוסף של m מופעים שמוגדרים על ידי תכונה x כלשהי (כלומר הערך x_i הוא ערך התכונה של מופע i ונתונים ערכי פונקציית מטרה y המוגדרים על כל מופע כ-y_i. נתונה הנוסחה למדידת קורלציה Pearson Correlation Coefficient בין תכונה x לערכי פונקציית המטרה y:

$$\rho = \frac{\sum_{i=1}^{m}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{m}(x_i - \mu_x)^2 \sum_{i=1}^{m}(y_i - \mu_y)^2}}$$

א. הסבירי מה מודדת קורלציה, מה משמעות המונה ומה משמעות המכנה בנוסחה הנתונה ומה הערכים האפשריים למדד הקורלציה?

ב. הסבירי מה הקשר בין קורלציה לבין תלות בין x ל- y (בתור משתנים מקריים)? הסבירי מה המשמעות בכל אחד מהמקרים הבאים עבור הקשר והתלות בין x (המסביר) ל- y(המוסבר):
- הקורלציה היא 1
- הקורלציה היא 0.8-
- הקורלציה היא 0

ג. הסבירי מה הקשר בין מדד הקורלציה לרגרסיה לינארית המנסה לשערך (להסביר) את y בעזרת x. באיזה מהמקרים בסעיף הקודם כדאי להשתמש ברגרסיה לינארית ובאיזה לא? הסבירי מדוע והאם תהיה טעות בשערוך בכל אחד מהמקרים.

עתה נניח כי למופעים n תכונות (features) שונות ולא רק תכונה אחת .נסמן כל תכונה באינדקס תחתון כלומר x_i תהיה תכונה i וכדי לסמן מופע כלשהו מהאוסף נשתמש באינדקס עליון. כלומר x_i^d יסמן את התכונה ה-i של המופע ה-d מתוך m המופעים באוסף.

ד. הגדירי (כולל נוסחה) מהי מטריצת covariance (או scatter) של התכונות של אוסף המופעים והסבירי מה הקשר בין האיברים שבה לבין מדד הקורלציה מסעיף א?

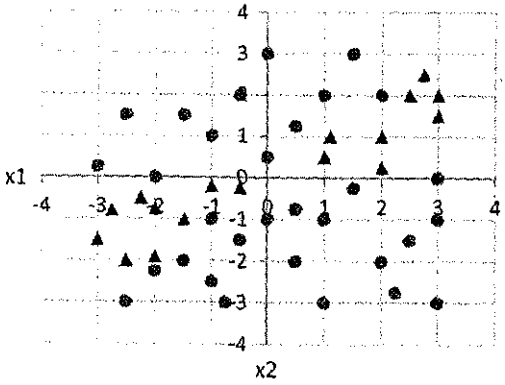
ה. הסבירי מה מטרת אלגוריתם Principal Components Analysis (PCA).

ו. הסבירי מה מטרת אלגוריתם Linear Discriminant Analysis (LDA).

ז. הסבירי מה ההבדל בין מטריצת ה-scatter שבשימוש ב-PCA לבין זו שבשימוש באלגוריתם LDA?

שאלה 3 (6 סעיפים)

נתונה קבוצת האימון ע"פ הציור הבא:



א. האם יצליח אלגוריתם הפרספטרון למצוא מפריד טוב לקבוצה זו? אם כן, מהן משקולות המפריד? אם לא, הסבירי מדוע לא?

ב. הפעלנו את אלגוריתם SVM על אותה קבוצת אימון, וניסינו קרנלים שונים. לבסוף קיבלנו את כלל ההחלטה הבא:

$$t(x) = \mathrm{sgn}(g(x)) = \mathrm{sgn}\left(\sum_{i \in SV} \alpha_i t_i (x_i \cdot x)^2\right)$$

כאשר

SV – היא קבוצת הווקטורים התומכים (support vectors).

α_i – הוא המשקל של ווקטור התמיכה ה-i.

t_i – ה-class של וקטור התמיכה ה-i.

הסבירי מהו הקרנל $K(x,y)$ שנבחר?

ג. מצאי $\varphi: \mathbb{R}^2 \to \mathbb{R}^3$, כך שהקרנל מהסעיף הקודם ייתן $K(x,y) = \varphi(x) \cdot \varphi(y)$.

ד. מצאי את המשקולות של המפריד הלינארי ב-$\mathbb{R}^3$, ששקול ל $x_2(x_2 - x_1) = 0$.

ה. נניח שהקרנל שנבחר הינו RBF – radial basis function. כלומר:

$$K(x,y) = \varphi(x) \cdot \varphi(y) = exp\left(-\frac{1}{2}\|x-y\|^2\right)$$

הוכיחי שלכל שתי נקודות x, y ב-$\mathbb{R}^2$ מתקיים:

$$\|\varphi(x) - \varphi(y)\|^2 = 2 - 2exp\left(-\frac{1}{2}\|x-y\|^2\right)$$

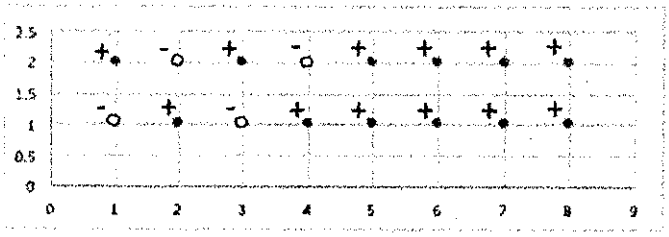
ו. לכל אחת מהטענות הבאות, החליטי נכון / לא נכון ונמקי

1. לאחר המיפוי למרחב הגבוה בעזרת הקרנל RBF, ייתכן ואלגוריתם הפרספטרון ישיג תוצאות טובות יותר ביחס לתוצאות במרחב המקורי (הנמוך).
2. לאחר המיפוי למרחב הגבוה בעזרת הקרנל RBF, ייתכן ואלגוריתם 1-NN (1 Nearest Neighbor), המשתמש במרחק אוקלידי ללא משקלים, ישיג תוצאות טובות יותר ביחס לתוצאות במרחב המקורי (הנמוך).
   רמז: השתמשי בזהות מסעיף ה'.
3. התשובה לסעיף ו2' תישאר זהה גם אם משתמשים ב-k-NN המשתמש במרחק אוקלידי עם משקלים (שכן שקרוב יותר מקבל משקל גבוה יותר), כאשר $k > 1$.

שאלה 4 (7 סעיפים)

נתונה הטבלה והגרף של קבוצת נתונים בעלי שתי תכונות x1, x2 משתי מחלקות + ו-"-" אשר משמשת כקבוצת אימון ללמידה של עצי החלטה.
אנו יוצרים שני עצי החלטה להלן:
עץ TOver הוא עץ אשר גדל עד הסוף ללא הגבלה וללא pruning
עץ TUnder הוא עץ בעל צומת בודדת שבה נכללים כל הנתונים.

instance	x1	x2	Value
1	1	2	+
2	2	1	+
3	3	2	+
4	4	1	+
5	5	1	+
6	5	2	+
7	6	1	+
8	6	2	+
9	7	1	+
10	7	2	+
11	8	1	+
12	8	2	+
13	1	1	-
14	2	2	-
15	3	1	-
16	4	2	-



א. הסבירי מהו Goodness of Split שקובע ע"פ איזה תכונה נפצל צומת בעץ החלטה והראי מה הנוסחה שלו.

ב. נניח כי אנו משתמשים במדד GiniIndex בתור הפונקציה המודדת עד כמה קבוצה הומוגנית כדי לבנות את העץ TOver :

$$GiniIndex(S) = 1 - \sum_{i=1}^{c}(p_i)^2 = 1 - \sum_{i=1}^{c}\left(\frac{|S_i|}{|S|}\right)^2$$

הסבירי (כולל נוסחה) כיצד משתמשים במדד זה כדי לקבוע את Goodness of Split.

ג. הסבירי כיצד נקבע את הפיצול הראשון בעץ TOver? כמה חישובים של Goodness of Split יש לבצע? (אין צורך להציב מספרים ולחשב מספר סופי או להגיע לתוצאה אלא רק להסביר כמה חישובים ואיזה חישובים יש לבצע כדי למצוא את הפיצול הראשון).

ד. נניח כי הפיצול הראשון בעץ TOver מוגדר על ידי הנוסחה (x1<4.5) כמה עלים יהיו בעץ בסוף הלמידה?

ה. האם יתכן מצב בו מפסיקים את בניית העץ לפני שכל העלים הומוגניים (שיש בהם דוגמאות ממחלקה אחת בלבד). הסבירי מדוע וכיצד נקבע הסיווג של מופע חדש שמגיע לעלה לא הומוגני כזה בעץ ההחלטה אחרי שהסתיימה הלמידה.

ו. מה תהיה הטעות הכוללת בשיטת leave one out בעץ TOver? הסבירי!

ז. מה תהיה הטעות הכוללת בשיטת leave one out בעץ TUnder? הסבירי!

שאלה 5 (6 סעיפים)

נתונה קבוצת מופעים S שאנו רוצים לחלק ל-k קבוצות (כלומר לבצע clustering). להלן אלגוריתם אשר נקרא k-medoids ודומה ל:k-means

Initialize $c_1, \ldots, c_k$ by randomly selecting k elements from S
Loop:
  Assign all n samples to their closest c_i and create k clusters $S_1, \ldots, S_k$
  For each cluster S_i $(1 \le i \le k)$ define a new c_i:
    choose $c_i \in S_i$ whose distance to all other members in S_i is the smallest
Until no change in $c_1, \ldots, c_k$
Return $c_1, \ldots, c_k$

instance	x	y
p_1	2	6
p_2	4	7
p_3	5	8
p_4	6	1
p_5	6	4
p_6	7	3
p_7	5	6

א. נניח כי הקבוצה S מונה 7 מופעים בעלי 2 תכונות כנתון בטבלה משמאל. הריצי את אלגוריתם k-medoids לחלוקה לשתי קבוצות (כלומר k=2) כאשר מאתחלים את הריצה עם p_1-וp_5 בתור המרכזים הראשונים כלומר בשלב הראשון $c_1=p_1$ ו- $c_2=p_5$. (רמז: ראשית ציירו את המופעים על מישור דו ממדי). בכל שלב צייני מי המרכזים ומה חלוקת המופעים לכל קבוצה - אין צורך להראות את כל החישובים בכל שלב.

ב. מה ההבדל העיקרי בין אלגוריתם k-means ו- k-medoids?

ג. הסבירי בנוסחה איזה פונקציה מביא אלגוריתם k-means למינימום?

ד. כיצד היית משנה את הפונקציה מסעיף ג להתאים אותה לאלגוריתם k-medoids?

ה. נניח כי אנו מריצים את שני האלגוריתמים על אותה הקבוצה האם אנו מצפים שערך הפונקציה שאותה מביא k-medoids למינימום יהיה קטן, גדול או שווה לערך הפונקציה שאותה מביא k-means למינימום? הסבירי מדוע!

ו. איזה בעיה חמורה ניתן למצא באלגוריתם k- medoids כפי שהוא מנוסח למעלה?

שאלה 6 (6 סעיפים)

נתון אבחון המורכב משתי תכונות כמותיות x1 ו-x2.
ידוע, בהתבסס על היסטורית מדידות ארוכת טווח, שההתפלגות המותנית של הערכים של תכונות אלו בכל אחת מהמחלקות (classes) נתונה ע"י (D ו-H הן המחלקות, כאשר D= disease ו- H= healthy):
לתכונה הראשונה

$$f(x_1|D) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

$$f(x_1|H) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-3)^2}{2}\right)$$

ולתכונה השנייה

$$f(x_2|D) = \begin{cases} 1 & 0 \le x_2 \le 1 \\ 0 & \text{Otherwise} \end{cases}$$

$$f(x_2|H) = \begin{cases} e & 1-\frac{1}{e} \le x_2 \le 1 \\ 0 & \text{Otherwise} \end{cases}$$

מומלץ (אבל לא חובה) שתשרטטי לעצמך את ההתפלגויות של הערכים בכל אחת מהתכונות, לטובת הבנה טובה יותר.

א. הסבירי מה ההבדל בין פרדיקציה לפי ML (maximum likelihood) ופרדיקציה לפי MAP (maximum aposteriori).

ב. מה תהיה הפרדיקציה לפי ML במקרים (הנפרדים) הבאים:
1. לפציינט מסוים התקבל במדידה הערך $x_2 = 0.25$.
2. לפציינט אחר התקבל במדידה הערך $x_1 = 1$.

ג. בהינתן הסתברות prior כלשהי P(H), הגדירי את הנוסחה לחישוב MAP באבחון זה. מה הערך המינימלי של P(H), כדי שהפרדיקציה לפי MAP, ע"פ הנתונים בסעיף ב2', תהיה H?

ד. הניחי שבנוסף למדידה בסעיף ב2' נמדד גם הערך $x_2 = 0.95$. מה הערך המינמלי של P(H), כדי שהפרדיקציה לפי MAP, במקרה זה, תהיה H?

ה. במקרה נוסף, נמדדו הערכים $x_1 = 8$, $x_2 = 0.95\left(1-\frac{1}{e}\right)$. בנוסף נתון ש- P(H)=0.9. מה תהיה הפרדיקציה לפי MAP במקרה זה? האם את חושבת שתוצאה זו מייצגת הטיה או חסרון של גישת הסיווג הנ"ל? איך תתגברי על בעיה זו?

בהצלחה!