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

סמסטר ב' תשע"ז
Spring 2018

מבחן מועד א בלמידה ממוכנת
Machine Learning Exam A

Lecturer: Prof Zohar Yakhini
Time limit: 3 hours
Additional material or calculators are not allowed in use!

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

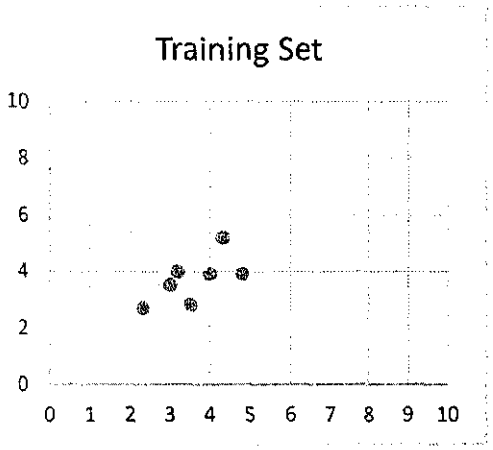
מרצה: פרופ זהר יכיני
משך המבחן: 3 שעות
אין להשתמש בחומר עזר ואין להשתמש במחשבונים!

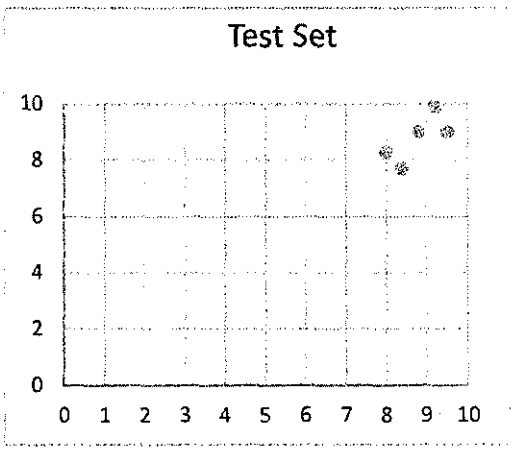
יש לענות על 5 מתוך 6 השאלות הבאות
לכל השאלות משקל שווה (20 נקודות)
בהצלחה!

שאלה 1 (5 סעיפים)

א. כתבי את הנוסחה של MSE (Mean Squared Error) בהקשר של פרדיקציה של פונקציה $y = f(x)$ (רגרסיה).

ב. נתון ה-training set ו-test set הבאים:





את משתמשת באחד האלגוריתמים הבאים להעריך את $y = f(x)$ בעזרת ה-training set:

1. Linear Regression
2. Regression 2-NN (רגרסיה עם 2-nearest-neighbor)

למי מהאלגוריתמים תהיה טעות קטנה יותר בחישוב MSE על ה- **test set**?

ג. איך היית משנה את ההגדרה של MSE loss, שהגדרת בסעיף א', כך שתוכלי לתת לכל instance משקל w_i שונה בחישוב ה-loss?

ד. כתבי את הפסאודו קוד של רגרסיה לינארית (linear regression), כולל צעד העדכון, העושה שימוש ב-gradient descent ב- stochastic mode שנלמד בכיתה.

איך היית משנה את צעד העדכון כך שימזער את הפונקציה שכתבת בסעיף ג'.

ה. בכיתה ניסחנו את בעיית הרגרסיה הלינארית בעזרת המשוואה הבאה:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \ ||X\theta - y||_2^2$$

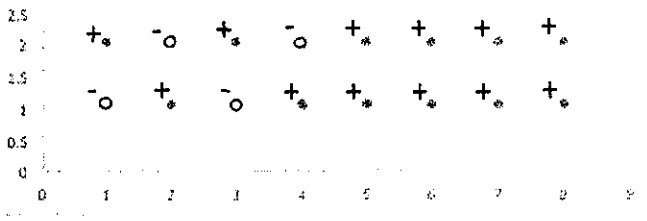
מצאי מטריצה W שבעזרתה תוכלי לעדכן את משוואת הרגרסיה הלינארית ולהגדיר פתרון פסאדו אינווֹרס (pseudo inverse) עבור הפונקציה שהגדרת בסעיף ג'.

הסבירי את כל הצעדים. בתשובתך ציייני מהי המטריצה ואת המשוואה המעודכנת.

שאלה 2 (4 סעיפים)

נתונה הטבלה והגרף של קבוצת נתונים בעלי שתי תכונות רציפות x1, x2 משתי מחלקות "+" ו-"-".
אנחנו משתמשים בקבוצה זו כקבוצת אימון ללמידה של עצי החלטה.
עץ T_N הינו עץ בינארי העושה שימוש ב-Goodness of Split לביצוע פיצולים בעץ וגדל עד שמגיע לגובה N *או* שלא ניתן יותר לפיצול (המוקדם מבניהם).
לדוגמא:
T_1 יהיה עץ בעל פיצול אחד (שורש ושני בנים).
T_2 יהיה עץ בעל פיצול אחד בשורש ולכל היותר עוד פיצול אחד בכל אחד משני הבנים.
* שימי לב שהעץ אינו חייב להיות סימטרי.

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 מהי פונקציית *impurity* ואיך *Goodness of Split עושה שימוש* בפונקציית impurity φ ע"מ לקבוע ע"פ איזה תכונה נפצל בצומת בעץ החלטה. בהסבר יש לכלול את הנוסחה של Goodness of Split.

ב. האם ייתכן מצב שעץ שנבנה ע"י שימוש ב- Goodness of Split יהיה גבוה יותר מאשר עץ שנבנה בצורה אחרת ומגיע גם לעלים טהורים? אם כן, תני דוגמא שמראה *זאת*. אם לא, הסבירי למה לא.

ג. האם הפיצול הראשון בעץ T_1 הנלמד על קבוצת אימון יהיה שונה מהפיצול הראשון בעץ T_3 הנלמד על אותה קבוצת אימון? הסבירי!

ד. בבניית עצי החלטה מסוג T_1 ו T_2 בעזרת ה-training data הנתון מעלה, מה הטעות המתקבלת בכל אחד משני המקרים בנפרד בשיטת leave one out? מי משתי השיטות מביאה לתוצאה טובה יותר ע"פ מה שנמדד בשיטת leave one out?

שאלה 3 (5 סעיפים)

א. מצאי את המינימום והמקסימום של הפונקציה $x + 4y$ עם האילוץ $x^2 + 9y^2 = 1$.

ב. בהינתן ה dataset הבא (פונקציית ה-XOR):

X1	X2	Y
+1	+1	-1
+1	-1	+1
-1	+1	+1
-1	-1	-1

השתמשי בלמה הבאה להראות שה-dataset הנ"ל אינו ניתן להפרדה לינארית. אין צורך להוכיח את הלמה.

למה:

נניח שמפריד לינארי חוזה פרדיקציה $y \in \{-1,+1\}$ על 2 נקודת $z, z' \in \mathbb{R}^2$ (כלומר, $h(z) = h(z') = y$). אזי, המפריד ייתן את אותה פרדיקציה על כל נקודות ביניים, כלומר:

$$\forall \alpha \in [0,1] \quad h\big((1-\alpha)z + \alpha z'\big) = y$$

ג. מצאי מיפוי φ למרחב עם ממד לבחירתך, אשר ממפה את ה-dataset מסעיף ב', כך שבממד החדש הוא יהיה ניתן להפרדה לינארית ומצאי את המפריד הלינארי בממד החדש.

ד. בהינתן המיפוי $\varphi(x) = (1, x, x^2, x^3, \dots, x^N)$ עבור $x \in (-1,+1)$ (הקטע הפתוח בין 1- ל 1+) עם מספר טבעי N כלשהו. הראי שלמיפוי φ קיימת פונקציית קרנל $K(x,y)$.

ה. הראי שהפונקציה $K(x,y) = \frac{1}{1-xy}$ עבור $x, y \in (-1,+1)$ (הקטע הפתוח בין 1- ל 1+), היא פונקציית קרנל עם מיפוי מתאים הממפה למרחב אינסופי.

שאלה 4 (4 סעיפים)

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

Initialize $c_1, \ldots, c_k$ randomly
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:
    b_i = the center of the cluster (average point)
    if $|S_i| > 2$ (if the number of samples in S_i is larger than 2):
      x = the sample with the highest distance from b_i
      c_i = the center of the cluster without x
    else:
      $c_i = b_i$
Until no change in $c_1, \ldots, c_k$
Return $c_1, \ldots, c_k$

instance	x_1	x_2
p_1	1	0
p_2	0	1
p_3	2	1
p_4	1	2
p_5	6	6

א. איזה פונקציה שואף אלגוריתם k-means הסטנדרטי להביא למינימום? כתבי נוסחה.

ב. נניח כי הקבוצה S מונה 5 מופעים בעלי 2 תכונות כנתון בטבלה. הריצי את האלגוריתמים k-means-outlier ואת k-means לחלוקה לשתי קבוצות (כלומר k=2) כאשר מאתחלים את הריצה עם $c_1=(0, 0)$ ו- $c_2=(2,2)$ בתור קבוצת המרכזים הראשונית (המלצה: ראשית ציירי את המופעים על מישור דו ממדי). בכל שלב צייני מי המרכזים ומה חלוקת המופעים לכל קבוצה. **אין צורך להראות את כל החישובים בכל שלב.**

ג. האם אלגוריתם k-means-outlier מתכנס? אם כן, הוכיחי שהאלגוריתם מתכנס. אם לא, הראי דוגמה שבה האלגוריתם אינו מתכנס.

ד. להלן פסאדו קוד לגרסה של Fuzzy K-Means (הידוע גם בכינויו soft clustering):

Initialize $c_1, \ldots, c_k$ randomly
Loop:
  Calculate a distance vector for each sample with distances to each cluster
  For each sample j, convert the distance vector to probability vector
    $w_j = (w_{j1}, \ldots, w_{jk})$
  For each cluster S_i ($1 \le i \le k$) define a new c_i:
    $c_i = \frac{\sum_{j=1}^{n} w_{ji} x_j}{\sum_{j=1}^{n} w_{ji}}$
Until no change in $c_1, \ldots, c_k$
Return $c_1, \ldots, c_k$

כתבי הצעה לפסאדו קוד לגרסה של אלגוריתם Fuzzy K-Means-Outlier.

שאלה 5 (4 סעיפים)

א. תני דוגמא למרחב דוגמאות X ומרחב H של היפותזות $h: X \to \{-1,+1\}$ (היפותזות בינאריות), כך ש:

$$VC(H) = 2018$$

ב. להלן 3 הנוסחאות ל- sample complexity שנלמדו בכיתה:

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

נתון מרחב הדוגמאות $X = [0,1] \times [0,1]$. יהי $N \in \mathbb{N}$ כאשר $N \geq 2$, נגדיר קבוצה $A = \{\frac{1}{N}, \frac{2}{N}, \dots, 1\}$ ואת מרחבי ההיפותזות הבאים:

- $H_1 = \{h: h(x_1,x_2) = +1 \Leftrightarrow x_1 \in [0,a] \wedge x_2 \in [0,a],\ a \in A\}$
- $H_2 = \{h: h(x_1,x_2) = +1 \Leftrightarrow x_1 \in [0,a] \wedge x_2 \in [0,b],\ a,b \in A\}$
- $H_3 = \{h: h(x_1,x_2) = +1 \Leftrightarrow x_1 \in [0,a] \wedge x_2 \in [0,b],\ a,b \in [0,1]\}$

כל היפותזה היא מלבן המקביל לצירים בעל קודקוד בראשית הצירים (instance יסווג כחיובי אם ורק אם הוא נופל בתוך מלבן בעל הקודקודים [(0,0), (a,0), (a,b), (0,b)]).

לכל אחד מהמקרים הבאים, השתמשי באחת מהנוסחאות לחישוב כמות ה-instances הנדרשת להבטיח טעות של 0.1 בהסתברות של לפחות 95%:

1. כאשר מנסים ללמוד קונספט, c, שנמצא במרחב H_3 בעזרת מרחב ההיפותזות H_2.
2. כאשר מנסים ללמוד קונספט, c, שנמצא במרחב H_1 בעזרת מרחב ההיפותזות H_2.
3. כאשר מנסים ללמוד קונספט, c, שנמצא במרחב H_3 בעזרת מרחב ההיפותזות H_3.

שאלה 6 (4 סעיפים)

נתון בדיקת איכות 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 (G ו-B הן המחלקות, כאשר G = good ו- B = bad):
לתכונה הראשונה

$$f(x_1|G) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-2)^2}{2}\right)$$
$$f(x_1|B) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-5)^2}{2}\right)$$

ולתכונה השנייה

$$f(x_2|G) = \begin{cases} \frac{1}{3} & 0 \le x_2 \le 1 \\ \frac{2}{3} & 2 \le x_2 \le 3 \\ 0 & otherwise \end{cases}$$

$$f(x_2|B) = \begin{cases} e^2 & 3 - \frac{1}{e^2} \le x_2 \le 3 \\ 0 & otherwise \end{cases}$$

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

א. מה תהיה הפרדיקציה לפי ML במקרים (הנפרדים) הבאים:
   1. למוצר מסוים התקבל במדידה הערך $x_2 = 2.9$.
   2. למוצר אחר התקבל במדידה הערך $x_1 = 3$.

ב. בהינתן הסתברות prior כלשהי P(B), הגדירי את הנוסחה לחישוב Naïve Bayes MAP באבחון זה.
מה הערך המינימלי של P(B), כדי שהפרדיקציה לפי Naïve Bayes MAP, ע"פ הנתונים בסעיף א2', תהיה B?

ג. הניחי שבנוסף למדידה בסעיף א2' נמדד גם הערך $x_2 = 2.99$.
מה הערך המינימלי של P(B), כדי שהפרדיקציה לפי Naïve Bayes MAP, במקרה זה, תהיה B?

ד. במקרה נוסף, נמדדו הערכים $x_1 = 6$, $x_2 = 0.975\left(3 - \frac{1}{e^2}\right)$. בנוסף נתון ש- P(B)=0.9. מה תהיה הפרדיקציה לפי Naïve Bayes MAP במקרה זה? האם את חושבת שתוצאה זו מייצגת הטיה או חסרון של גישת הסיווג הנ"ל? איך תתגברי על בעיה זו?

בהצלחה!

Standard formula sheet – IDC 2018

1. Distributions:

Normal $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Bernoulli trials - $B(n,p)$ $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$

Poisson $P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$

Geometric $P(X=k) = (1-p)^{k-1}p$

2. Decision Trees:

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

Entropy $Entropy(S) = -\sum_{i=1}^{c}\frac{|S_i|}{|S|}\log\frac{|S_i|}{|S|}$

3. Gradient descent and update steps:

Linear regression $\theta_j := \theta_j - \alpha\frac{1}{m}\sum_{d\in D}\left(h_\theta\left(x^{(d)}\right) - y^{(d)}\right)\cdot x_j^{(d)}$

Perceptron $w_j := w_j - \eta\sum_{d\in D}(o^{(d)} - t^{(d)})x_j^{(d)}$

Dual perceptron If $o^{(d)}\cdot t^{(d)} < 0$ then:

$\alpha_j = \alpha_j + \eta$

4. Logistic regression:

$P(h(x)=1) = \frac{1}{1+e^{-w^Tx}}$

5. SVM:

Primal objective function $\frac{1}{2}\|w\|^2 + \gamma\sum_d \xi_d - \sum_d \alpha_d(t_d(w^Tx_d + w_0) - 1 + \xi_d) - \sum_d \mu_d\xi_d$

s.t $\alpha_d \geq 0 \quad \mu_d \geq 0$

Dual objective function $\sum_d \alpha_d - 1/2\sum_d\sum_e \alpha_d\alpha_e t_d t_e x_d^T x_e$

s.t $\sum_d \alpha_d t_d = 0,\ 0 \leq \alpha_d \leq \gamma$

6. EM (for Bernoulli distributions):

$New\ w_{A_j} = \frac{1}{N}\sum_{i=1}^{N} r(x_i, A_j)$

$p_{A_j} = \frac{1}{(New\ w_{A_j})N}\sum_{i=1}^{N} r(x_i, A_j)v(i)$