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

סמסטר ב' תשפ"א
Spring 2021

מבחן מועד ב בלמידה ממוכנת
Machine Learning Exam B

Lecturer: Prof Zohar Yakhini
Time limit: 3 hours

Answer 4 out of 5 from the following questions. Each question is 25 points.

Good Luck!

מרצה: פרופ זהר יכיני
משך המבחן: 3 שעות

יש לענות על 4 מתוך 5 השאלות הבאות.
לכל השאלות משקל שווה (25 נקודות)

בהצלחה!

ניתן להשתמש בדפי העזר המצורפים, מחשבון ומילון בלבד. כל חומר עזר אחר אסור.

יש להסביר\להוכיח את כל התשובות.

You can use the attached formula sheet, a calculator and a dictionary. All other material should not be used.

Prove/explain all your answers.

שאלה 1 – תיאוריה (25 נקודות)

נתון $X = \mathbb{R}^2$. תהי $C = H$ קבוצת כל המלבנים עבורם הקודקוד השמאלי התחתון נמצא בראשית הצירים והקודקוד הימני העליון נמצא ברביע הראשון. נקודות שנמצאות בתוך המבלנים מסווגות כחיוביות ונקודות מחוץ למלבנים מסווגות כשליליות.

פורמלית:

לכל שני מספרים $t, r \in \mathbb{R}_+$

נגדיר $h(t,r) = \{(x,y) \mid \ 0 \le x \le t \wedge 0 \le y \le r\}$.

ובעת נגדיר את מרחב ההיפותזות:

$$H = \{h(t,r) \mid t, r \in \mathbb{R}_+\}$$

1. (5 נק') חשב את מימד ה-VC של H. הוכיחו את תשובתכם.

2. (6 נק') הצע אלגוריתם עקבי (consistent) L שמקבל בקלט נקודות מתויגות $D^m = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\} \subseteq \mathbb{R}^2$ ומחזיר $h \in H$.

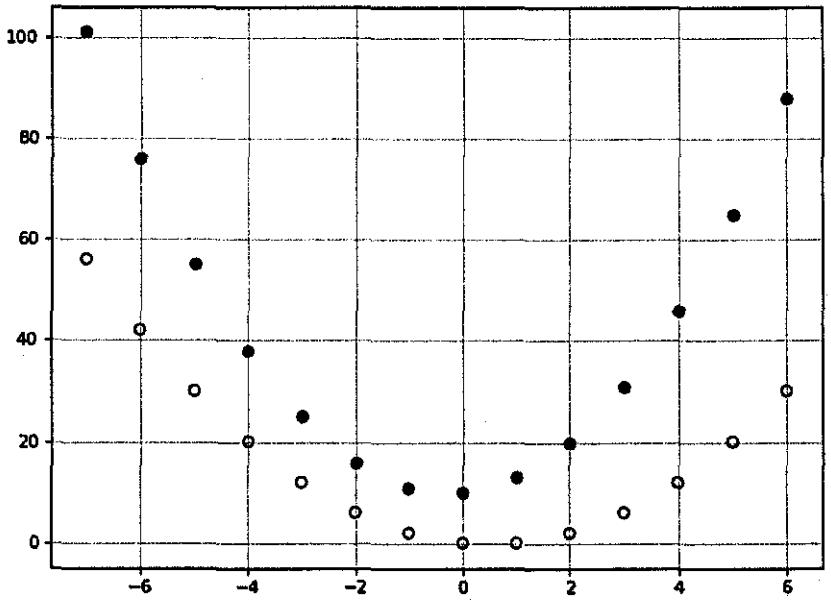
3. (7 נק') חשב חסם מספק על מורכבות המדגם (sample complexity) בלמידה של C ע"י H באמצעות האלגוריתם L שהצעת.

4. (7 נק') החסם שחישבת בסעיף הקודם אינו הדוק. הסבר מדוע.

שאלה 2 – SVM (25 נקודות)

נתון דאטה D במתואר באיור באשר העיגולים הריקים מייצגים את הקלאס השלילי והעיגולים המלאים את הקלאס החיובי. הדאטה מכיל את הנקודות הבאות:

x	y	$class$
-6	76	+
2	20	+
5	65	+
-2	6	-
-4	20	-
5	20	-



1. משוואת הפרבולה מוגדרת על ידי המשוואה הבאה:

$$y = ax^2 + bx + c$$

a. (3 נק') מצא את משוואת הפרבולה העליונה (עיגולים מלאים).
b. (3 נק') מצא את משוואת הפרבולה התחתונה (עיגולים ריקים).

2. נתונה הטענה הבאה:
הנח שמסווג ליניארי חוזה את אותו ערך $y \in \{-1,1\}$ עבור שתי נקודות $h(z) = -z, z' \in \mathbb{R}^n$
$h(z')$, אזי הוא חוזה את אותו ערך פרדיקציה עבור כל נקודת ביניים באופן הבא:

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

a. (8 נק') הוכח את הטענה.
b. (5 נק') השתמש בטענה על מנת להוכיח שהדאטה לעיל אינו ניתן להפרדה ליניארית.

3. (6 נק') מצא פונקציית מיפוי $\Phi(x,y)$ ל-$\mathbb{R}^4$ כך ש-$\Phi(D)$ ניתן להפרדה ליניארית והוכח את נכונות המיפוי באמצעות ווקטור w כך שנקודה (x,y) תסווג למחלקה החיובית אם ורק אם $w \cdot \Phi(x,y) \geq 0$.

שאלה 3 – Clustering (25 נקודות)

1. (5 נק') נתון הדאטה הבא: $\{0, 4, 5, 20, 25, 39, 43, 44\}$ ואלגוריתם ה- Hierarchical Clustering בפי שנלמד בביתה. באמצעות שתי פונקציות מרחק – Single linkage ו- Complete linkage, מהם האלמנטים בשני הקלאסטרים האחרונים אשר יאוחדו?

תזכורת: **Single linkage** משתמש במרחק המינימלי בין כל הנקודות הנמצאות בסטים שונים ו-**Complete linkage** משתמש במרחק המקסימלי בין כל הנקודות בסטים השונים. כמו כן, השיטות הללו מייצגות דרכים שונות לאיחוד קלאסטרים.

2. (5 נק') נתון אלגוריתם ה-Naïve Cluster Growing כפי שנלמד בביתה, בעל ערך Threshold שנקבע מראש, $T = 1$. יש לייצר שני דוגמאות לדאטה-סטים באופן הבא:
   a. דאטה-סט A מכיל 5 נקודות ואחרי ריצת Naïve Cluster Growing מתקבלים 5 קלאסטרים.
   b. דאטה-סט B מכיל 5 נקודות ואחרי ריצת Naïve Cluster Growing מתקבל קלאסטר יחיד.

   השתמשו במרחק אוקלידי ובבל מימד שתבחרו עבור הדאטהסט. עבור כל דאטה-סט, רשמו את כל הקואורדינטות של הנקודות בו.

3. (6 נק') עבור כל טענה, בחרו האם היא נכונה או שגויה. נמקו את תשובותיכם.
   a. יש לקבוע את מספר הקלאסטרים מראש כפרמטר עבור k-means ו- Hierarchical Clustering.
   b. אלגוריתם k-means משתמש באתחול רנדומלי בעוד ש- Hierarchical Clustering אינו משתמש באתחול כזה.

4. (9 נק') נתונה וריאצה של אלגוריתם ה-k-means בשם k-means-outlier-L1 לפי ה-Pseudocode הבא:

1. Initialize k centers $c_1, \dots, c_k$ randomly unless centers are given.
2. Loop until there is no change in $c_1, \dots, c_k$:
   a. Assign all n samples to their closest center c_i and create k clusters, $S_1, \dots, S_k$.
   b. For each cluster define a new center:
      If $|S_i| > 2$:
         let x_i be the point in S_i with the largest L_1 distance from c_i.
         Calculate the new center c_i using $S_i \backslash \{x_i\}$.
      otherwise:
         Calculate the new center c_i using S_i.

#	x_1	x_2
1	1	0
2	0	1
3	1	5
4	3	1
5	6	5
6	6	6

הריצו את אלגוריתם k-means-outlier-L1 על הדאטה בטבלה באשר $k = 2$ באמצעות המרכזים ההתחלתיים הבאים: $c_1 = (0,0), c_2 = (4,1)$.

שאלה 4 – Logistic Regression (25 נקודות)

1. (6 נק') פונקציית ה-Loss עבור Logistic Regression מוגדרת באופן הבא:

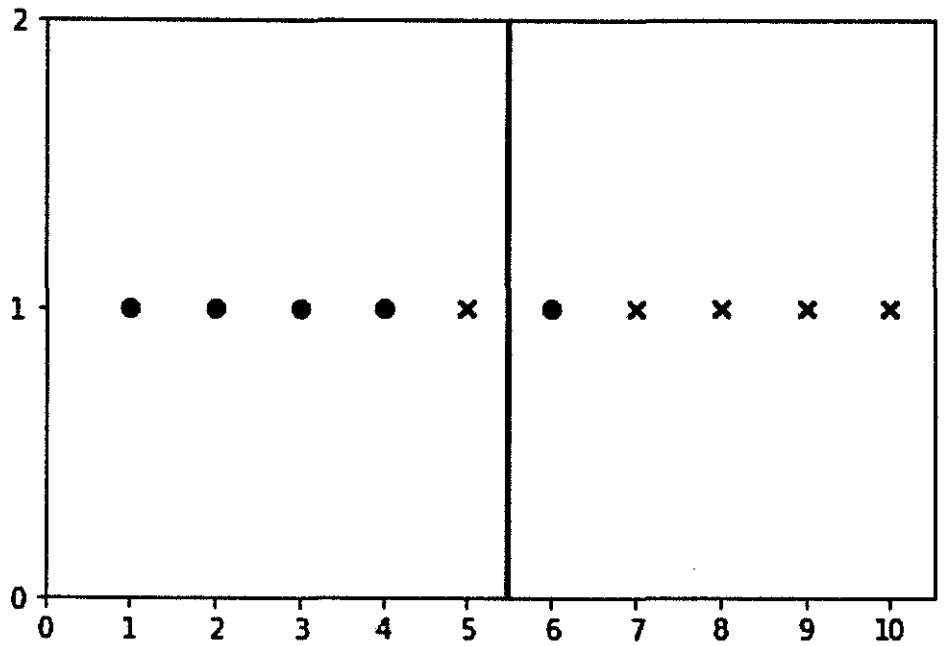
$$J = -\sum_{d=1}^{m} y^{(d)} \ln\left(h_\theta\left(x^{(d)}\right)\right) + \left(1 - y^{(d)}\right) \ln\left(1 - h_\theta\left(x^{(d)}\right)\right)$$

$$h_\theta\left(x^{(d)}\right) = \frac{1}{1 + e^{-\theta^T x^{(d)}}}$$

בהינתן דאטה הניתן להפרדה ליניארית, האם מובטח שאלגוריתם ה-Logistic Regression תמיד ימצא מפריד ליניארי מושלם (שגיאת אימון = 0)?

2. (9 נק') ענו על השאלות הבאות ונמקו בקצרה:
   a. באיזו שיטה נעשה שימוש על מנת לבצע אופטימיזציה ל-Cost function של Logistic regression (מציאת הפרמטרים האופטימליים - θ)?
   b. האם ניתן להשתמש בשיטת ה-Pseudo Inverse על מנת למצוא את הפרמטרים האופטימליים ב-Logistic Regression?
   c. האם ניתן להשתמש ב-Logistic Regression עבור סיווג שאינו ביניארי (יותר מ-2 מחלקות)? אם כן, הסבירו בקצרה ביצד. אם לא, נמקו בקצרה מדוע?

3. (10 נק') נתון באיור לפניכם דאטה בעל שתי מחלקות (המחלקה השלילית מסומנת בעיגול, המחלקה החיובית מסומנת ב-x) ואת הקו המפריד שנמצא באמצעות Logistic regression (הקו הישר במרכז האיור). בנו את ה-ROC Curve עבור מסווג זה באמצעות ציור 11 נקודות של ה-ROC Curve (יש לצייר במחברת – **אין לרשום על גבי טופס הבחינה**).



שאלה 5 – עצי החלטה (25 נקודות)

נתונה טבלה של נתונים בעלי שתי תכונות רציפות, x_1, x_2 ושתי מחלקות, $+, -$. אנו משתמשים בנתונים הללו כקבוצת אימון לטובת אימון עץ החלטה.

נגדיר T_N להיות עץ החלטה בינארי העושה שימוש ב-Goodness of split לביצוע פיצולים בעץ וגדל עד שמגיע לגובה N או עד שלא ניתן יותר לבצע פיצולים (המוקדם מביניהם). לדוגמה:

לעץ T_1 יהיה פיצול אחד (שורש ושני בנים). לעץ T_2 יהיה פיצול אחד בשורש ולכל היותר פיצול אחד בכל אחד משני הבנים. שימו לב שהעץ אינו חייב להיות סימטרי.

Instance	X1	X2	Value
1	2	3	+
2	1	3	-
3	1	4.5	-
4	1	2	+
5	1	5	-
6	2	5	-
7	1	6	-
8	2	6	-
9	1	7	-
10	2	7	-
11	1	8	-
12	2	8	-
13	2	2	+

1. (5 נק') הסבירו מהי Impurity Function φ וביצד Goodness of split משתמש בה על מנת לבצע פיצול בעץ החלטה. על הסבר לכלול נוסחה ברורה.

2. (10 נק') האם ניתן לבנות עץ באמצעות Goodness of split שיהיה גבוה יותר מעץ שנבנה בשיטה אחרת, כאשר שניהם יגיעו בסופו של דבר לעלים טהורים? אם כן, ספקו דוגמה. אם לא, הסבירו מדוע.

3. (5 נק') האם הפיצול השני ב-T_2 שנלמד באמצעות סט אימון כלשהו יהיה שונה מהפיצול השני שהתחרש ב-T_3 שנלמד באמצעות אותו סט אימון? הסבירו.

4. (5 נק') בבניית עצים מסוג T_1 ו-T_2 באמצעות סט האימון הנתון מעלה, מה תהיה השגיאה המתקבלת בכל אחד משני המקרים בנפרד בשיטת Leave one out? מי משתי השיטות מביאה לתוצאה טובה יותר על פי הנמדד בשיטת Leave one out?

בהצלחה!

Standard formula sheet – IDC TASHPA

1. Distributions:

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

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

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:

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

5. SVM:

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

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):

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

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

7. Linear Regression (closed form): $\theta^* = \underset{\theta}{\operatorname{argmin}}\|y - X\cdot\theta\|_2^2 = (X^T X)^{-1}X^T y$