250200.5780 - **למידה חישובית/כריית מידע**
מבחן מועד א'

- למבחן 11 שאלות "אמריקאיות" ובוחרים תשובה **אחת** נכונה
- שאלה 12 היא בכתב (תשובות קצרות)
- כל התשובות יימלאו בטבלת תשובות (קובץ WORD מצורף)
- את קובץ התשובות יש להעלות לתיבת הגשה במודל וגם לשלוח במייל (solewicz@g.jct.ac.il)
- חומר סגור
- משך המבחן שעה וחצי

בהצלחה!

1) בבעיה של חיזוי מחירי בתים השתמשנו ב-Linear Regression.
נניח שאחרי ה-gradient descent מצאנו פרמטרים שהובילו ל-Loss=0 ב-training set.
לפניך מספר טענות:

1. ניתן לחזות מחיר עבור בתים נוספים (שלא בקבוצת האימון) בוודאות.
2. כל המחירים שבקבוצת האימון עומדים על אותו קו ישר.
3. נניח שבשורות קובץ ה-data הבתים מסודרים לפי מחיר (הבית הכי זול בשורה הראשונה והבית הכי יקר באחרונה). בשלב חלוקת קובץ ה-data ל-train ו-test נכון לערבב (shuffle) את שורות לפני החלוקה ולא מספיק פשוט לחתוך את ה-data לשני חלקים לפי הסדר המקורי.

אלו טענות נכונות? (8 נק')

A. כולם נכונות
B. כולם לא נכונות
C. 1,2 נכונות
D. 1,3 נכונות
E. 2,3 נכונות

---

2) נניח שיש לך X dataset של 50 דוגמאות וכל אחד עם 200000 features. רוצים להשתמש ב-linear regression כמודל חיזוי (***y=bX***). עדיף לאמן ע"י gradient descent או ע"י הנוסחה הסגורה שלמדנו?
(8 נק')
תזכורת לנוסחה:

$$b=(X^TX)^{-1}X^Ty$$

A) ע"י הנוסחה כי היא מדויקת ומהירה יותר לחישוב
B) ע"י gradient descent כי הוא מהיר יותר לחישוב
C) שניהם אותו דבר
D) ע"י הנוסחה מפני שה- gradient descent עלול להיתקע במינימום לוקאלי
E) אי אפשר לדעת מראש מה יותר טוב

---

3) אתה מאמן מודל logistic regression. לפניך מספר טענות:

1. הוספת הרבה features חדשים למודל ימנע overfitting
2. הוספת feature חדש למודל לא תוריד ביצועים ב-training set.
3. אם תכניס regularization למודל לא תוריד ביצועים עבור דוגמאות שב-training set.
4. אם תכניס regularization למודל לא תוריד ביצועים עבור דוגמאות שלא ב-training set.

אלו טענות נכונות? (8 נק')

A. כולם נכונות
B. כולם לא נכונות
C. 1,2,3 נכונות
D. 2,4 נכונות
E. 2 נכונה

---

4) מימשת Regularized Logistic Regression ע"מ לסווג בין תמונות של כלבים ושל חתולים. למרות שה-ERROR היה מאוד נמוך בשלב האימון, הוא מאוד גבוה בתמונות ה-test. לפניך מספר אפשרויות ע"מ לשפר את הביצועים.

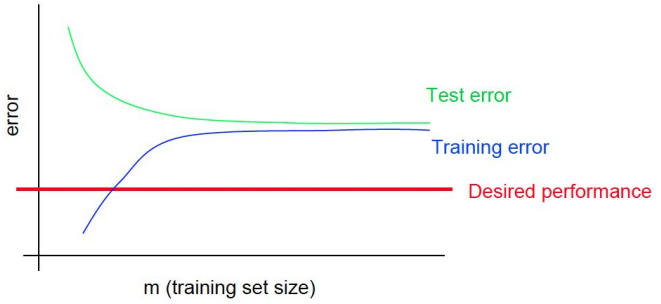
1. להגדיל מספר ה-features (למשל ע"י הרחבה של polynomial features).
2. להקטין את מספר הדוגמאות באימון.
3. להגדיל את מספר הדוגמאות באימון.
4. לנסות לצמצם את מימד ה-features (למשל ע"י PCA).

מה עליך לנסות ? (8 נק')

A) 1
B) 2
C) 1,2
D) 3,4
E) אף אחת מהאפשרויות לא סבירות

---

5) למסווג מסוים יש עקומת למידה כזו:



מה הטענה הנכונה? (8 נק')

A) הוא סובל מ-bias גבוה ולכן אנסה להרחיב את מימד ה-features
B) הוא סובל מ-bias גבוה ולכן אנסה לצמצם את מימד ה-features
C) הוא סובל מ-variance גבוה ולכן אנסה להרחיב את מימד ה-features
D) הוא סובל מ-variance גבוה ולכן אנסה לצמצם את מימד ה-features
E) הוא סובל מ-bias גבוה ולכן אנסה להוסיף regularization
F) הוא סובל מ-variance גבוה ולכן אנסה להוסיף regularization

---

6) בתהליך אימון מסוים, השתמשת ב-cross-validation ע"מ לקבל ערך אופטימלי לפרמטר "**Z**". לפי ה-**T**raining **E**rror ו-**E**rror **V**alidation שקיבלת עבור כל partition, איזה ערך ל-**Z** היית בוחר? (8 נק')

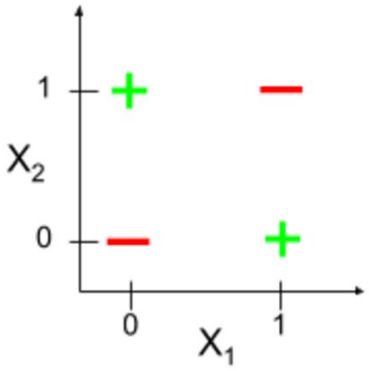
Z	TE	VE
1	105	90
2	200	85
3	250	96
4	105	85
5	300	100

A) Z=1
B) Z=2
C) Z=3
D) Z=4
E) Z=5
F) אי-אפשר לקבוע Z ע"פ התוצאות של ה- Cross-Validation שהתקבלו

---

7) סימנו את כל המסווגים שיגיעו ל- training error = 0 בבעיית ה-XOR?

(8 נק')



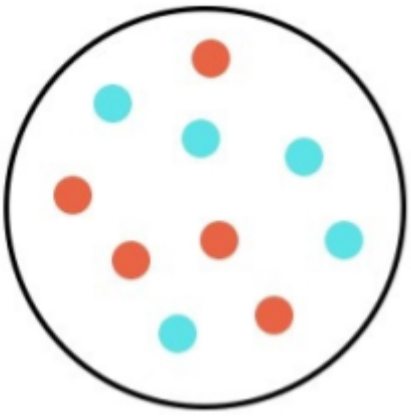
1. Logistic regression
2. SVM with quadratic (degree=2) kernel
3. Depth-2 decision tree
4. 3-NN classifier

A. כולם נכונות
B. כולם לא נכונות
C. 1,2 נכונות
D. 1,3 נכונות
E. 1,4 נכונות
F. 2,3 נכונות
G. 2,4 נכונות

---

8) תזכורת: אפשר לחשב את האנטרופיה (H) של ערך בינארי כמו בציור (x נקודות אדומות ו-y נקודות כחולות) ע"י הנוסחה:

$H(x,y) = -x/(x+y)* log_2 (x/(x+y)) -y/(x+y)* log_2 (y/(x+y))$



שאלות: רוצים לסווג את ה-DATASET הבא ע"י עץ החלטה.

Weight	Height	Gender	Eye color	output
L	H	M	D	1
L	L	M	D	1
H	H	F	D	1
H	L	M	D	1
H	H	M	D	1
L	L	M	D	0
L	H	F	D	0
L	L	F	D	0
H	H	F	D	0
H	L	F	D	0

a. מה האנטרופיה של ה-OUTPUT (לפני החלוקה)? (4 נק')

A. H(5,10)
B. 1
C. H(10,10)
D. H(5,0)
E. H(0,5)
F. אף תשובה לא נכונה

b. איזה FEATURE יתחלק ראשון לפי יכולתו להוריד יותר את האנטרופיה של ה-output? (כזכור, ה-SPLIT הראשון נעשה ע"י ה-FEATURE יותר חשוב ע"פ ירידה באנטרופיה.)
(4 נק')

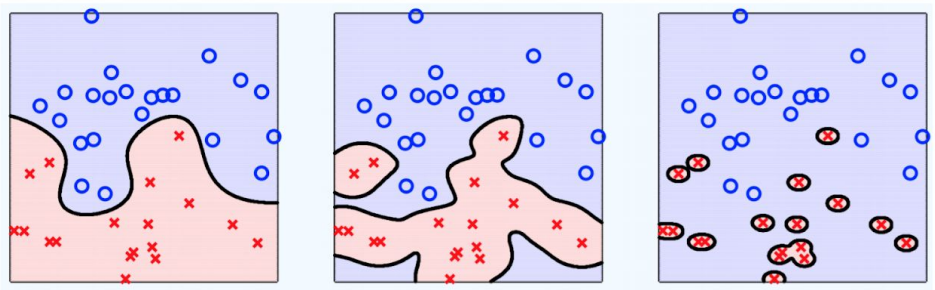
A. WEIGHT
B. HEIGHT
C. GENDER
D. EYE COLOR
E. לא ניתן לקבוע מראש

---

9) בהינתן ה-KERNEL הבא:

$$K(\mathbf{a},\mathbf{b}) = \exp\left(-\gamma\|\mathbf{a}-\mathbf{b}\|^2\right)$$

אתה מאמן Kernel-SVM לסווג בין "**X**" לבין "**O**" כמו בציור. מה הערכים המתאימים ל- γ עבור המקרים I,II,III ?
(5 נק')



I II III

	III	II	I
A	γ =100	γ =10	γ = 1
B	γ =1	γ =10	γ =100
C	γ =1	γ =100	γ =10
D	אין דרך לדעת		

---

10) בבעיה עסקית מסוימת התבקשת לבנות מודל רווחי שנדרש לעמוד בתנאים הבאים:

- **F**alse **P**ositives עולים פי-5 יותר יקר מ-**F**alse **N**egatives ללקוח.
- recall>0.8

בנית מסווג בינארי כלשהו וחישבת עבורו **C**onfusion **M**atrices לפי thresholds שונים. איזה **CM** מספקת את הדרישות? (8 נק')
תזכורת:

$$\textbf{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$\textbf{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

TN	FP	FN	TP	CM
91	9	22	78	A
99	1	21	79	B
96	4	10	90	C
98	2	18	82	D
שום תשובה לא נכונה				E

---

11) קיבלת dataset של N דוגמאות עם M features כל אחד. לפניך מספר טענות:

1. PCA עוזר לצמצם מימדים (M)
2. k-means עוזר לצמצם כמויות של דוגמאות שונות (N)
3. מציאת מספר ה-clusters האופטימלי הוא חלק מובנה באלגוריתם k-means
4. מציאת מספר ה-principal components האופטימלי הוא חלק מובנה באלגוריתם PCA

מה נכון? (8 נק')

A. כולם נכונות
B. כולם לא נכונות
C. 1,2 נכונות
D. 1,3 נכונות
E. 1,4 נכונות
F. 2,3 נכונות
G. 2,4 נכונות

---

12) ענה במשפט אחד על כל אחת משלושת השאלות:

a. בבעיית סיווג מסוימת עם כמות גדולה של נתונים לאימון, אתה נדרש לעמוד בזמנים מאוד קצרים בזמן ריצה בשטח. במה היית משתמש ולמה: k-NN או Logistic Regression ? (אל תתייחס להבדלים ב-Accuracy.)

(5 נק')

b. נניח ש- k=1. האם ה-training error של k-NN יכול להיות גדול יותר מזה של ה-logistic regression?

(5 נק')

c. יש לך אפשרות לבחור בין מימוש Boosting או Random Forest, איזה מהם יותר מתאים למימוש ממוחשב מקבילי? (נניח שלשניהם אותו Accuracy.)

(5 נק')