התחיל ב:	21/01/2022, 09:01
מצב	הסתיים
הושלם ב-	21/01/2022, 10:16
הזמן שלקח	1 שעה 15 דקות

שאלה **1**
הושלם
ניקוד השאלה: 10.00

לפניכם פרמטרים של אלגוריתמי למידה, המשפיעים על הרגלוריזציה למניעת overfitting.

א. עומק העץ, בעץ החלטה

ב. מקדם L2 ברגרסיית Ridge

ג. K ב-KNN

באיזה מהם, **הגדלת** ערך הפרמטר תוביל לרגולריזציה **גבוהה** יותר?

יש לבחור תשובה אחת:

- א, ג
- א, ב
- א, ב, ג
- ב, ג

התשובה הנכונה: ב, ג

שאלה **2**
הושלם
ניקוד השאלה: 10.00

לפניכם כמה טענות לגבי BAGGING. סמנו את הנכונה.

יש לבחור תשובה אחת:

- בשיטה זו, ככל שמגדילים את כמות הלומדים החלשים, משפרים את הביצועים אבל מאריכים את זמן החישוב
- המטרה העיקרית של שיטה זו היא להפחית את ה-bias של אלגוריתמי הלמידה
- בשיטה זו דוגמים עם חזרה מתוך הנתונים, כדי ליצור תתי-קבוצות שונים של הנתונים עבור הלומדים החלשים
- שיטה זו לא יעילה עם רגרסיה לוגיסטית, מכיוון שכל הלומדים החלשים לומדים בדיוק את אותן משקלות

התשובה הנכונה: בשיטה זו דוגמים עם חזרה מתוך הנתונים, כדי ליצור תתי-קבוצות שונים של הנתונים עבור הלומדים החלשים

שאלה **3**

הושלם

ניקוד השאלה: 10.00

נניח שרוצים ללמוד מסווג עבור טקסט של דוא"ל, כדי להחליט אם מדובר ב-SPAM או NOT_SPAM. ההשערה היא שמספר המופעים של האות X בטקסט יכולה לשמש כנתון יחיד לאלגוריתם הלמידה, ועל פיו לקבל החלטה.

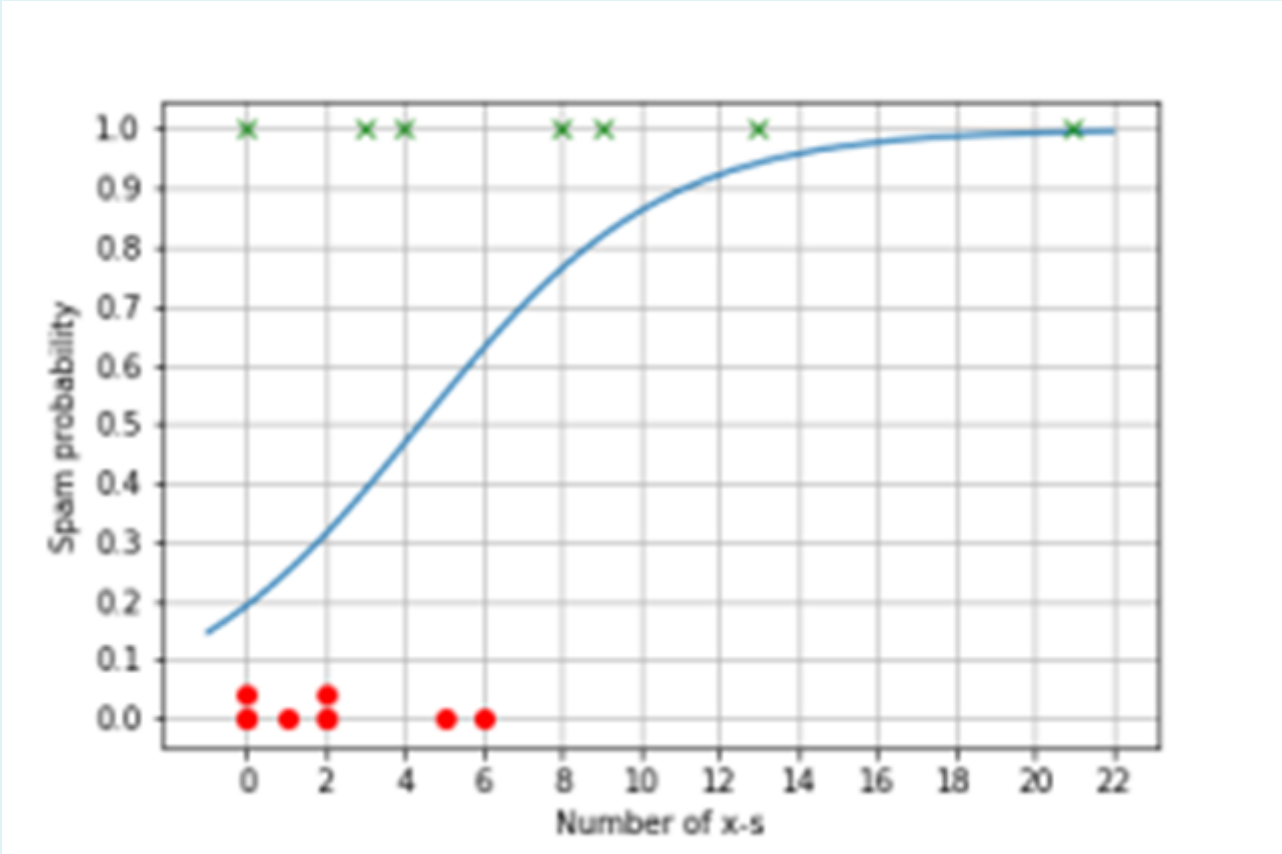
לצורך הלמידה, נלקחו שבע דוגמאות חיוביות (SPAM=1) ושבע דוגמאות שליליות (NOT_SPAM=0), נספרו מופעי האות X בכל אחת מהדוגמאות, והתקבלו הנתונים הבאים:

[SPAM = [0, 3, 4, 8, 9, 13, 21

[NOT_SPAM = [0, 0, 1, 2, 2, 5, 6

(כל מספר מייצג דוא"ל אחד).

הריצו אלגוריתם של רגרסיה לוגיסטית, והתקבל הגרף הבא:



מקראה:

x- SPAM

o- NOT_SPAM

בהנחה שישנה דרישה קריטית שאף דוא"ל שהוא NO_SPAM יסווג בטעות בתור SPAM, ביחרו את האפשרות שמנסחת במונחים של מדדי למידה את הדרישה הזו, ואת ערך הסף שיספק דרישה זו:

יש לבחור תשובה אחת:

- precision נמוך עבור SPAM, סף 0.7
- precision גבוה עבור SPAM, סף 0.7
- precision נמוך עבור SPAM, סף 0.2
- precision גבוה עבור SPAM, סף 0.2

התשובה הנכונה: precision גבוה עבור SPAM, סף 0.7

שאלה **4**
הושלם
ניקוד השאלה: 10.00

נניח שהריצו gradient descent לפתרון בעיית רגרסיה ליניארית, והגיעו למצב בו אין התכנסות של האלגוריתם, גם לאחר 10000 איטרציות.

נתונות האפשרויות הבאות:

א. קצב הלמידה קטן מדי, כדאי להעלות אותו

ב. התכונות (features) הן בקנה מידה לא אחיד, כדאי לנרמל את הנתונים

ג. האלגוריתם נתקע במינימום מקומי, כדאי לאתחל מנקודה אחרת

איזו מהן כדאי לנסות?

יש לבחור תשובה אחת:

- ב, ג
- א, ב
- א, ג
- א, ב, ג

התשובה הנכונה: א, ב

שאלה **5**
הושלם
ניקוד השאלה: 10.00

נתונה מטריצת הנתונים הבאה לאימון של עץ החלטה, עם תיוגים (label) בעמודה y:

y	רוחב	משקל
1	6	3
0	2	7
1	9	6
0	4	2

איזה מה-splits הבאים ייתן את ה-information gain הגבוה ביותר?

יש לבחור תשובה אחת:

- רוחב > 2
- משקל > 3
- משקל > 6
- רוחב > 4

התשובה הנכונה: רוחב > 4

שאלה **6**
הושלם
ניקוד השאלה: 10.00

נתון dataset עם מיליון דוגמאות לא מתוייגות, ממימד 5, כל המימדים מטיפוס float. רוצים לנתח את הנתונים, אך בגלל מגבלות של שטח זכרון וכח חישוב, רוצים להשתמש רק ב-10 דוגמאות מייצגות.

איזו מבין השיטות הבאות יכולה לשמש לבחירת 10 דוגמאות מייצגות:

יש לבחור תשובה אחת:

- Polynomial Feature Expansion
- K-means
- PCA
- Logistic Regression with LASSO regularization

התשובה הנכונה: K-means

שאלה **7**
הושלם
ניקוד השאלה: 10.00

נניח שהריצו רגרסיה ליניארית על נתונים, והבחינו בתופעה הבאה: כאשר ה-training set גדל, מנקודה מסויימת היתה עליה גדולה של שגיאת ה-test לעומת שגיאת ה-train, שירדה.

מבין האפשרויות הבאות, איזו מהווה סיבה עיקרית להתנהגות זו.

א. variance גבוה

ב. BIAS גבוה

ג. variance נמוך

יש לבחור תשובה אחת:

- ג
- א, ב
- א
- ב

התשובה הנכונה: א

שאלה **8**
הושלם
ניקוד השאלה: 10.00

באילו מהשיטות הבאות אפשר להשתמש עבור dimensionality reduction:

א. PCA

ב. Random Trees

ג. Kernels

יש לבחור תשובה אחת:

- א, ג
- א, ב, ג
- א, ב
- ב, ג

התשובה הנכונה: א, ב

שאלה **9**
הושלם
ניקוד השאלה: 10.00

השלימו את המשפט על ידי האפשרות הסבירה ביותר:

הוספת פונקציות בסיס נוספות במודל ליניארי ע"י BOOSTING...

א. מקטינה את ה-bias של המודל

ב. מקטינה את ה-variance של המודל

יש לבחור תשובה אחת:

- א
- ב
- שתי האפשרויות נכונות
- אף אפשרות אינה נכונה

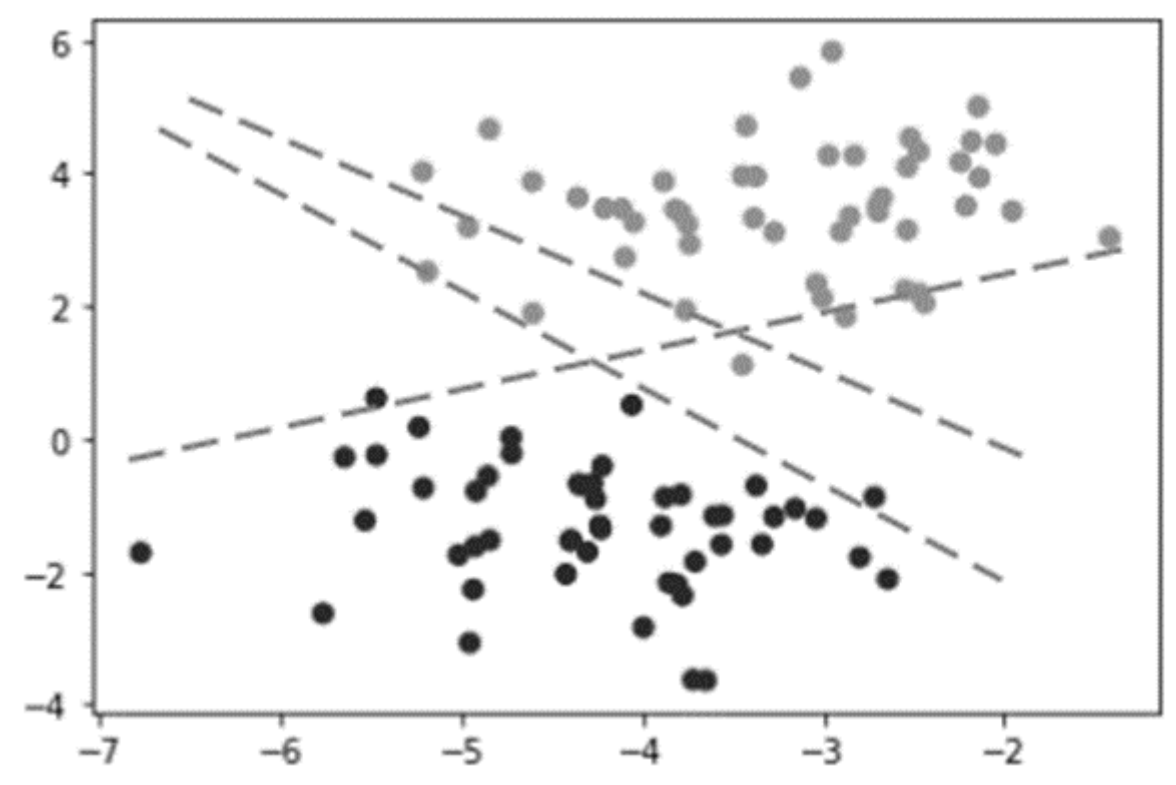
התשובה הנכונה: א

שאלה **10**

הושלם

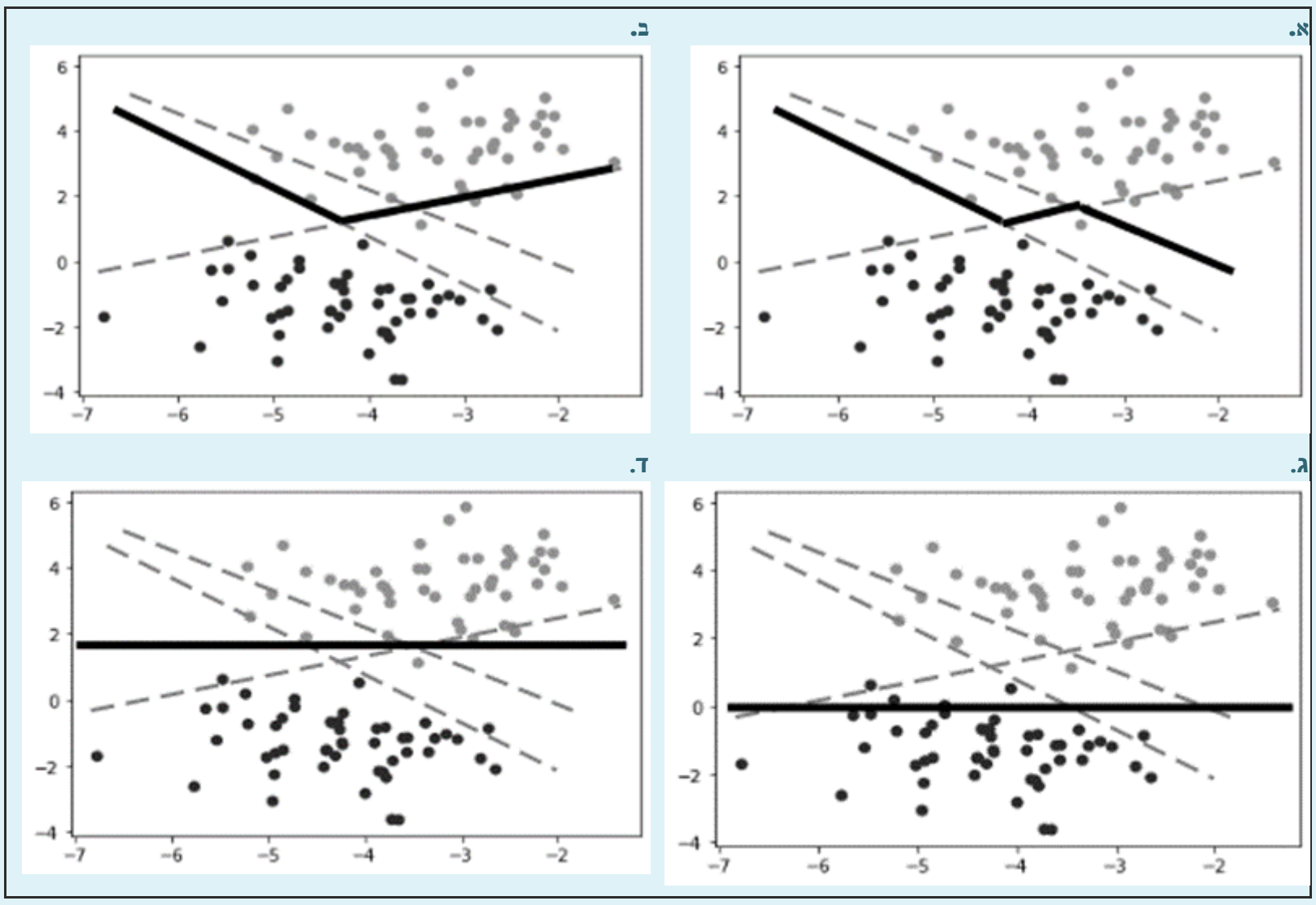
ניקוד השאלה: 10.00

באיור שלפניכם, הקווים המקווקווים מייצגים את גבולות ההחלטה (decision boundary) עבור שלושה מסווגים על נתוני האימון. כל הקווים הללו מסווגים את הנקודות שמעליהם כמחלקה 1 (אפור) ואת הנקודות שמתחתם כמחלקה 0 (שחור).



עליכם לבחור מבין האיורים הבאים את האיור בו מסומן גבול ההחלטה (הקו העבה) של מסווג הצבעות הרוב (majority voting ensemble classifier) שמופעל על תוצאות שלושת מסווגי הבסיס (הקווים המקווקווים).

(לכל איור יש אות).



יש לבחור תשובה אחת:

- ג
- א
- ד
- ב

התשובה הנכונה: א