

ניקוד שאלה 1

1	ציור עקומי אפס של X	A
1	שיוך נכון של עקומי אפס לא ולץ	B
1	מציאת נקודות שבת (אפשר גם גרפית). צריך להיות ברור לחלוטין מהן כל נקודות השבת	C
2	נוסחא כללית יעקוביאן	D
1	ערכים עצמיים נכונים ליד A	E
1	סיווג נכון ליד A	F
1	ערכים עצמיים נכונים ליד B	G
1	סיווג נכון ליד B	H
2	כיוון מסלולים נכון בהתחלה של ארבעת המסלולים	I
2	כיוון מסלולים נכון ברוב תאי השטח וחציות עקומי אפס	J
1	כיוון מסלולים נכון בכל תאי השטח וחציות עקומי אפס	K
1	התייחסות לספירלה	L
1	התייחסות לאוכף עם מסלולים BC	M

ניקוד שאלה 2

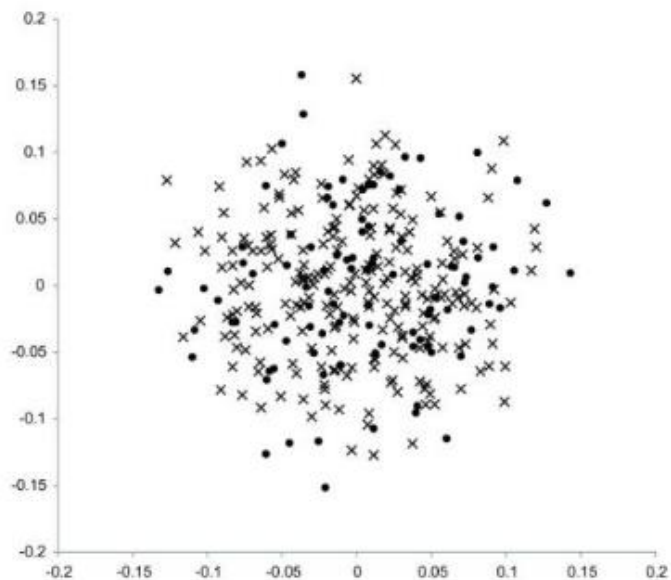
1	נוסחאות נכונות $x(2)$	A
1	נוסחאות נכונות $x(4)$	B
1	נוסחאות נכונות $x(5)$	C
1	ערכים מספריים נכונים $x(4), x(5)$	D
1	גרף תואם את זמנים $0 < t < 5$	E
1	$x(6) > x(3)$	F
1	$x(6) < 3$	G
1	דעיכה החל מזמן $t=6$	H

1	$F(0) > 0$	I
2	נקודת שבת יציבה ב $x=1$	J
2	נקודת שבת יציבה ב $x=3$	K
1	אין נקודות יציבות נוספות בין 1 ל 3	L
2	נקודת המינימום של F בין $x=1$ לבין נקודת השבת הלא יציבה מקיימת: $-3 < f(a) < -1$	M

שאלה 3 – PCA

1. ה-PCA המצורף מתאר gene expression של דוגמאות מחולים (מסומנים ב-X) ודוגמאות מבריאים (מסומנים בנקודות). כפי שניתן לראות, אין הפרדה ברורה בין הקבוצות. האם ניתן לחפש הבדלים בין החולים באופן אחר באמצעות ה-PCA? אם לא – הסבר. אם כן – תאר כיצד ומה יתרונות וחסרונות השיטה.

פתרון: (2 נק', 1 נק' אם הציעו לנרמל את הנתונים. $\frac{1}{2}$ נק' אם הציעו להשתמש בשיטות מקובלות שאינן PCA) – ניתן לחפש סיגל ב-PC3 והלאה. היתרון הוא שיתכן שנמצא סיגל שלא נמצאים בשני ה-PCs הראשונים. החסרון הוא שנתפוס פחות מהשונות בנתונים של ה-data ב-PCs נמוכים יותר.



2. הסבר כיצד נתונים קיצוניים במדידות (outliers) ישפיעו על hierarchical clustering לעומת PCA. על מי מהם הם ישפיעו יותר ולמה?

פתרון: (2 נק', ניקוד חלקי אם אין הצדקה או הצדקה לא נכונה) ישפיעו יותר על PCA מכיוון שהשונות בנתונים קובעת מי יהיה ה-PC הראשון וכן הלאה. בעוד שבקלאסטרינג ערך קיצוני ככל הנראה יתקבץ לקבוצה "צדדית" ולא ישפיע על המבנה הכללי שלי הדנדוגרמה.

3. (4 נק' סה"כ, 1 נקודה לסעיף) עבור כל משפט כתוב – נכון או לא נכון.

- a. ה-loadings של PC1 יהיו הגדולים ביותר. (לא נכון)
- b. PCA מביאה למקסימום את מספר ה-PCs. (לא נכון)
- c. PCA היא שיטה Unsupervised. (נכון)
- d. PCA היא שיטה שיכולה לשמש למזעור מספר המימדים בנתונים וגם למציאת הפקטורים שמשפיעים על השונות בנתונים. (נכון)

שאלה 4 – מרחקים ו-clustering

1. (2 נקודות, יתקבלו בן עם עשו את החישוב לאיטרציה הנוכחית או לזו שבאה אחריה, במקרה האחרון, ניקוד חלקי, עד הורדת נקודה, אם עשו טעות חישובית מינורית אך ברור שהבינו את האלגוריתם) ניח שברצונכם לקלסטר 7 דוגמאות ל-3 מקבצים באמצעות K-means. לאחר איטרציה אחת המקבצים C1, C2 ו-C3 הם אלה:

C1: {(2,2), (4,4), (6,6)}

C2: {(0,4), (4,0)}

C3: {(5,5), (9,9)}

מה יהיו הצנטרואידים במידה ותרצו להמשיך לאיטרציה נוספת?

פתרון: עבור C1: (4,4), עבור C2: (2,2), עבור C3: (7,7)

$$C1 = ((2+4+6)/3, (2+4+6)/3) = (4, 4)$$

$$C2 = ((0+4)/2, (4+0)/2) = (2, 2)$$

$$C3 = ((5+9)/2, (5+9)/2) = (7, 7)$$

2. בשתי ריצות שונות של אלגוריתם K-means לאותו סט של נתונים – האם בהכרח נקבל את אותן תוצאות clustering? – הסבר.

פתרון: (1 נקודה, ניקוד חלקי אם לא הסבירו) לא. K-means היא יוריסטיקה, כלומר פתרון מקורב, שעלול להגיע למקסימה מקומית. בהנתן נתוני התחלה שונים (צנטרואידים התחלתיים שונים), יתכן ונקבל תוצאות שונות.

3. קלאסטרינג היררכי בוצע על 4 נקודות. נתונה מטריצת המרחקים בין כל זוג נקודות:

	P1	P2	P3	P4
P1	0	13	28	39
P2	13	0	14	25
P3	28	14	0	11
P4	39	25	11	0

הסבירו את שני צעדי המיזוג הראשונים בקלאסטרינג היררכי אם שיטת המיזוג היא single linkage?

פתרון (5 נקודות, 3- אם לא זיהו את הזוגות הנכונים למיזוג, 2- אם טעות נגררת, 1/2- לכל חישוב מרחק לא נכון): ראשית יחוברו הנקודות P3 ו-P4 מכיוון שביניהם המרחק הקטן ביותר.

ואז יש לנו קלאסטר חדש {P3,P4}. המרחק בינו לבין כל שאר הנקודות מוגדר כמינימום של כל זוג נקודות מתוך המקבץ ומחוצה לו:

$$D(P1, \{P3,P4\}) = \min(D(P1,P4), D(P1,P3)) = D(P1,P3) = 28$$

$$D(P2, \{P3,P4\}) = \min(D(P2,P4), D(P2,P3)) = D(P2,P3) = 14$$

	P1	P2	{P3,P4}
P1	0	13	28
P2	13	0	14
{P3,P4}	28	14	0

לפי מטריצת המרחקים החדשה שהתקבלה כעת נקבץ את הנקודות P1 ו-P2 כי ביניהם המרחק הקטן ביותר.