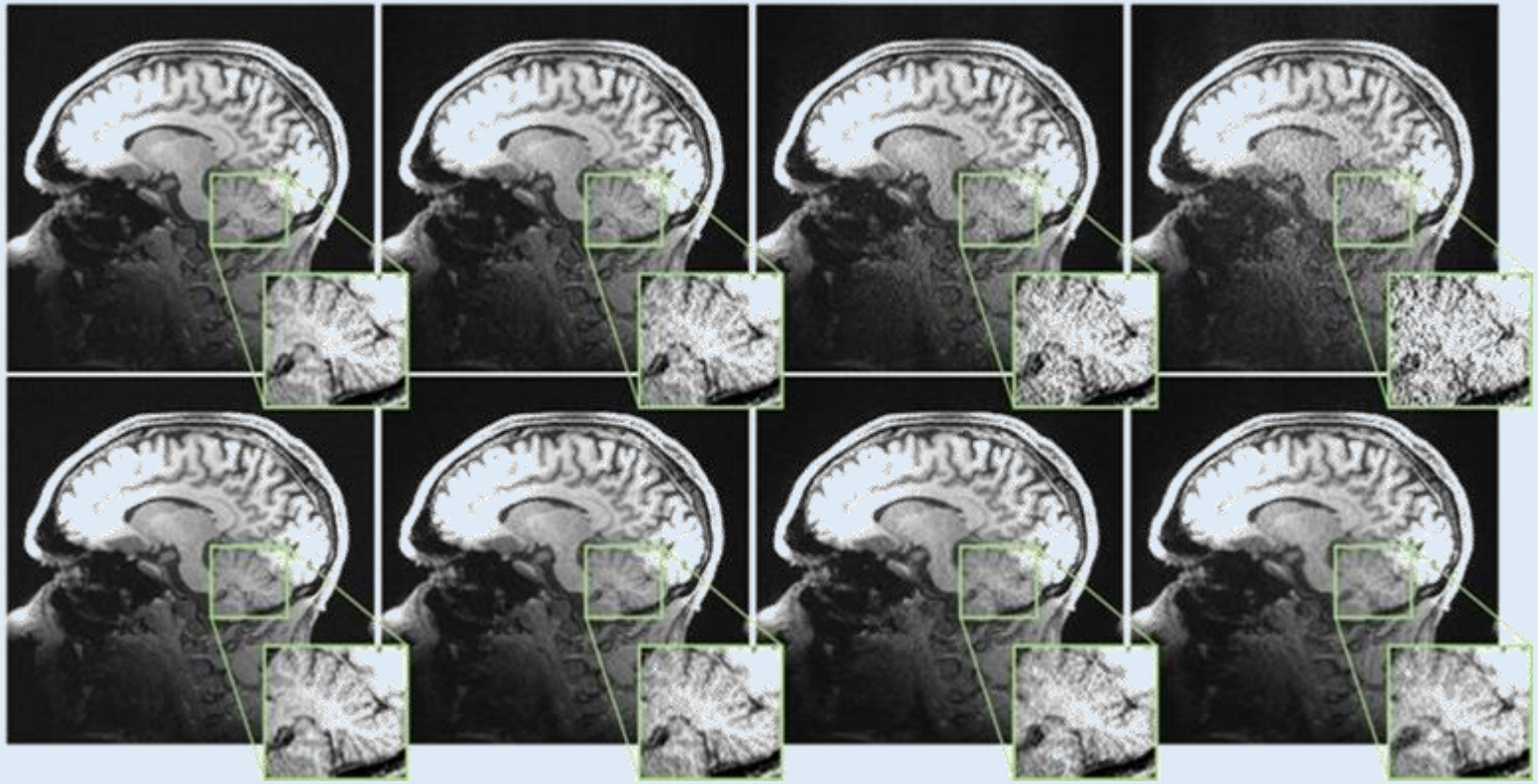
אִשכּוֹל (Clustering) רב-מימדי

<u>המטרה: לנתח מידע ברב מימד בצורה יעילה ומדוייקת ככל הניתן</u>

דוגמא:
שימוש ב-Machine learning כדי לחזות על פי סריקת MRI למי יש בעיות



סוגי למידה

איך עושים את זה?

Patient	Age	Gender	Diabetes	Cholesterol
1001	54	M	Yes	170
1002	45	M	No	135
1003	34	F	Yes	185
1004	61	M	Yes	200
1005	58	F	No	140

נחלק את סוגי הנתונים שאנחנו יכולים לקבל לשניים:

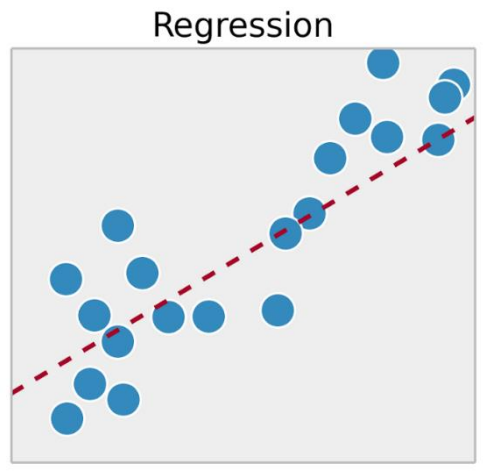
- כאשר יש לנו גם את נתוני ההתחלה (רמות כולסטרול) וגם את מה שהיינו רוצים לגלות (למי יש סוכרת ולמי לא). הלמידה הזו נקראת **למידה מונחית (supervised learning)** והמטרה שלה היא למצוא מודל שיכול לשייך דוגמא לאחת הקבוצות.

Patient	Age	Gender	Cholesterol
1001	54	M	170
1002	45	M	135
1003	34	F	185
1004	61	M	200
1005	58	F	140

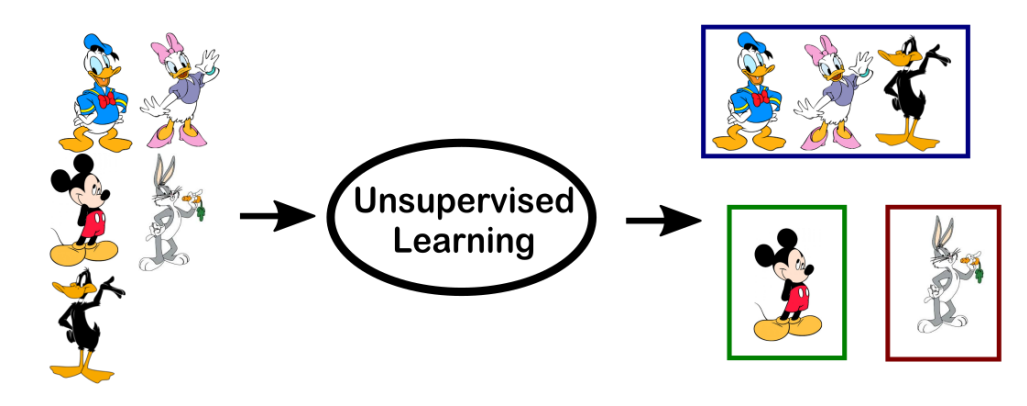
- כאשר יש לנו רק את נתוני ההתחלה (רמות כולסטרול) ונרצה להבין מהו המבנה של הנתונים שלנו (למשל, להבין האם רמות כולסטרול מחלקות אנשים לקבוצות). הלמידה הזו נקראת **למידה שאינה מונחית (unsupervised learning)** והמטרה שלנו תהיה לאתר תבניות/קבוצות בנתונים על סמך דמיון בין הדוגמאות.

סוגי למידה

לכל אחד מסוגי הלמידה יש מודלים סטטיסטיים בהם ניתן להשתמש



- **למידה מונחית (supervised learning)** – נרצה לשייך input מסויים ל-output מסויים. אפשר להשתמש למשל ברגרסיה.



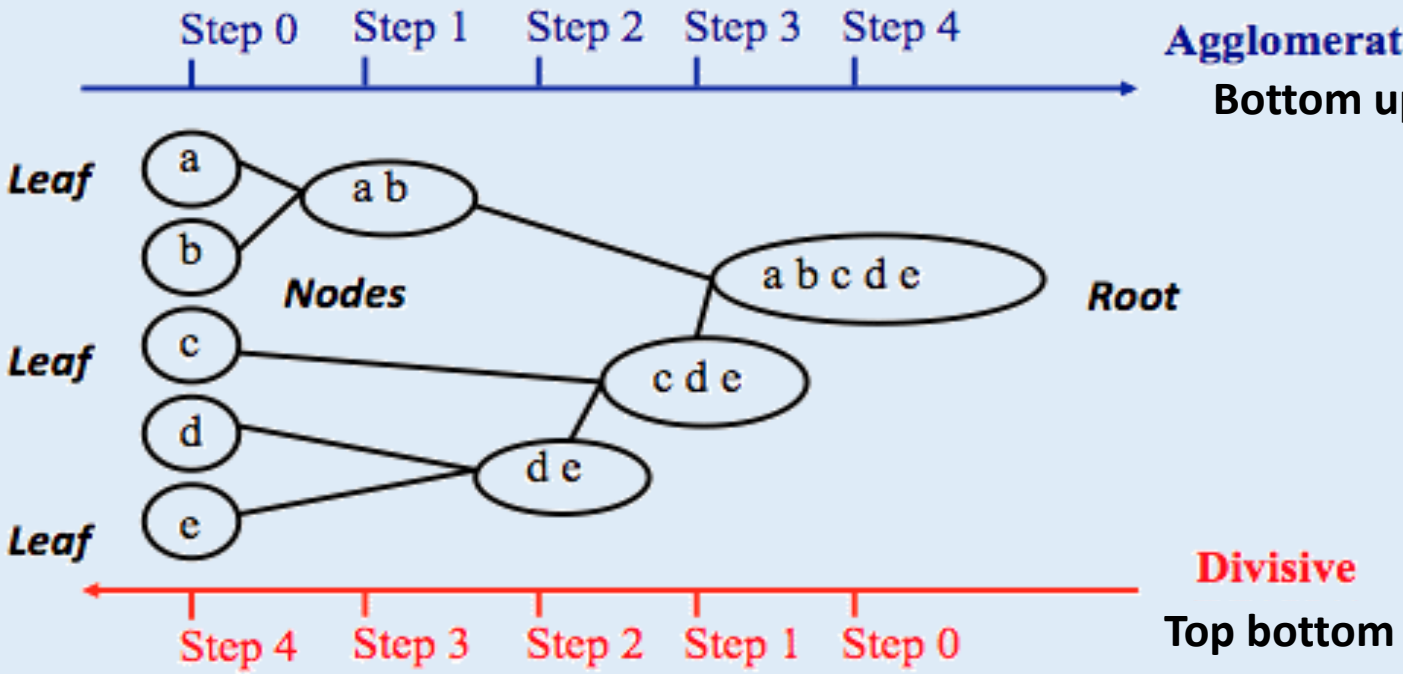
- **למידה שאינה מונחית (unsupervised learning)** – נרצה ללמוד על המבנה הטבעי של הנתונים מבלי להשתמש בידע מוקדם. משתמשים בעיקר בשיטות של אִשכּוּל (clustering) כגון k-means, hierarchical clustering, PCA וכו'.

Unsupervised Learning

מטרת שיטות האשכול השונות היא לחלק את הדוגמאות לקבוצות במרחב רב-מימדי, כך שהמרחק של דוגמאות בתוך אותו אשכול קטן יחסית למרחקן מדוגמאות ששייכות לאשכול אחר.

בכל שיטת אשכול צריכים תחילה למדוד את המרחקים בין הדוגמאות, ולשם כך צריך לבחור שיטה למדידת מרחק.

Hierarchical Clustering



נשתמש בשיטת bottom-up כדי לעשות

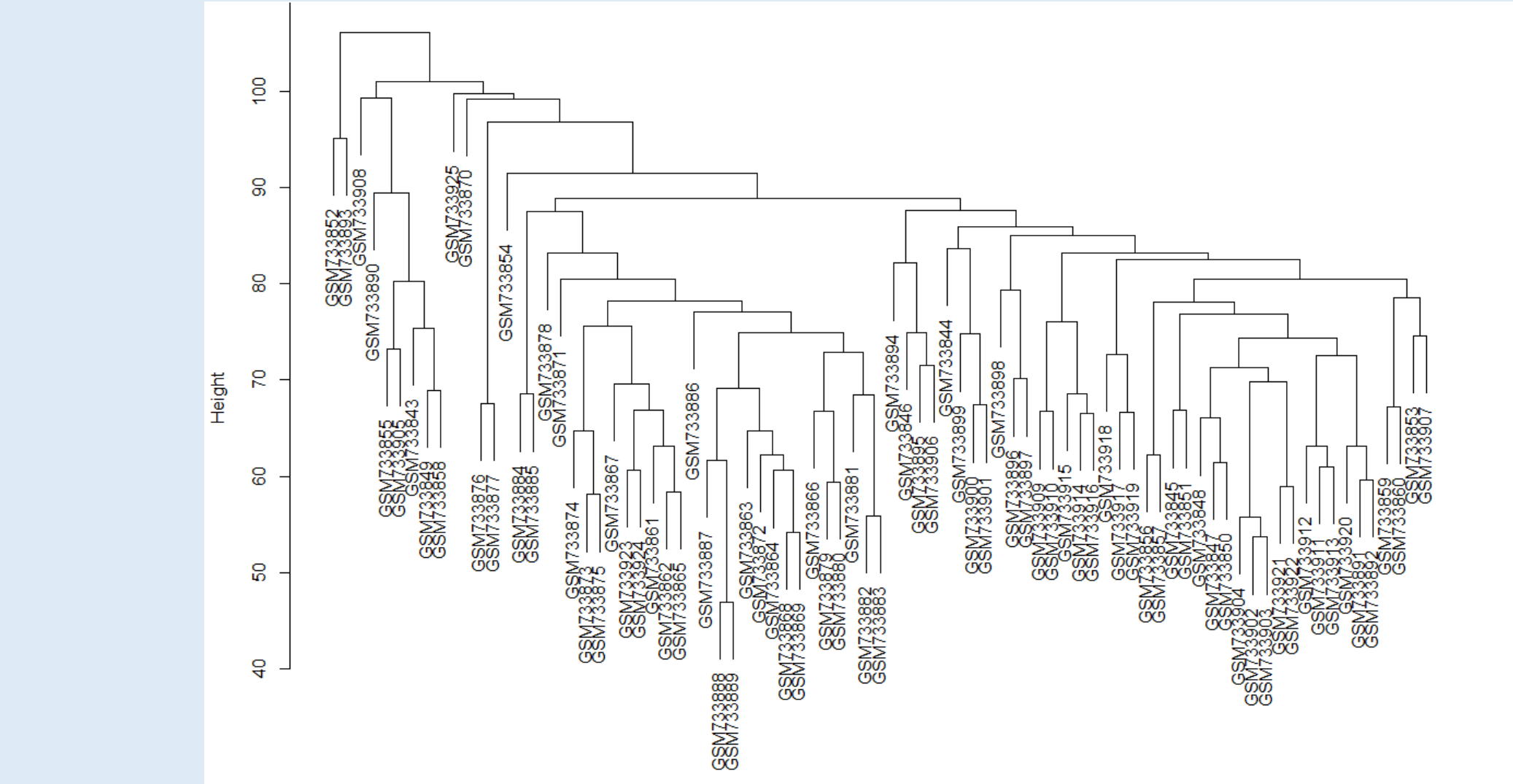
קלאסטרינג היררכי:

מתחילים לאחד קבוצות קטנות לפי מרחק זו מזו.

0. נבחר מרחק בו נרצה להשתמש
1. נחשב את המרחק בין כל שתי נקודות
2. נאחד קלאסטרים* לפי המרחק עד שלא נשאר מה לאחד
3. נאחד** את שתי הקבוצות הקרובות ביותר

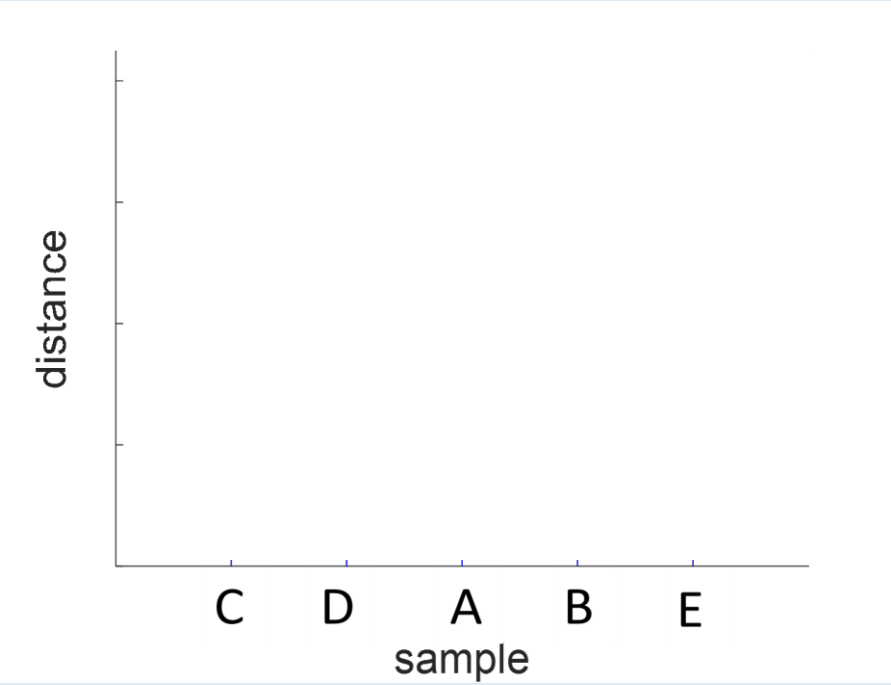
* גם נקודות בודדות הן קלאסטרים
** איחוד יעשה באחת משיטות ה-linkage

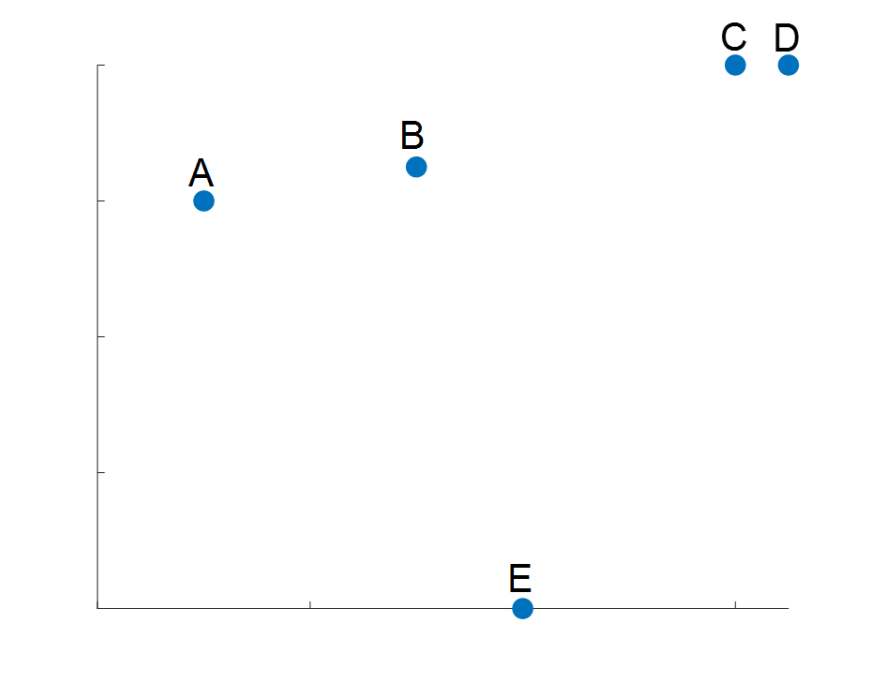
דנדרוגרמה


Height
40
50
60
70
80
90
100
GSM733852
GSM733893
GSM733908
GSM733890
GSM733855
GSM733905
GSM733843
GSM733849
GSM733858
GSM733925
GSM733870
GSM733854
GSM733876
GSM733877
GSM733884
GSM733885
GSM733878
GSM733871
GSM733874
GSM733875
GSM733867
GSM733923
GSM733924
GSM733862
GSM733861
GSM733865
GSM733886
GSM733887
GSM733888
GSM733889
GSM733863
GSM733872
GSM733864
GSM733868
GSM733869
GSM733866
GSM733879
GSM733880
GSM733881
GSM733882
GSM733883
GSM733894
GSM733846
GSM733895
GSM733906
GSM733844
GSM733899
GSM733900
GSM733901
GSM733898
GSM733896
GSM733897
GSM733909
GSM733910
GSM733915
GSM733914
GSM733916
GSM733918
GSM733917
GSM733919
GSM733856
GSM733857
GSM733845
GSM733851
GSM733848
GSM733847
GSM733850
GSM733904
GSM733902
GSM733903
GSM733921
GSM733922
GSM733912
GSM733911
GSM733913
GSM733920
GSM733891
GSM733892
GSM733859
GSM733860
GSM733853
GSM733907

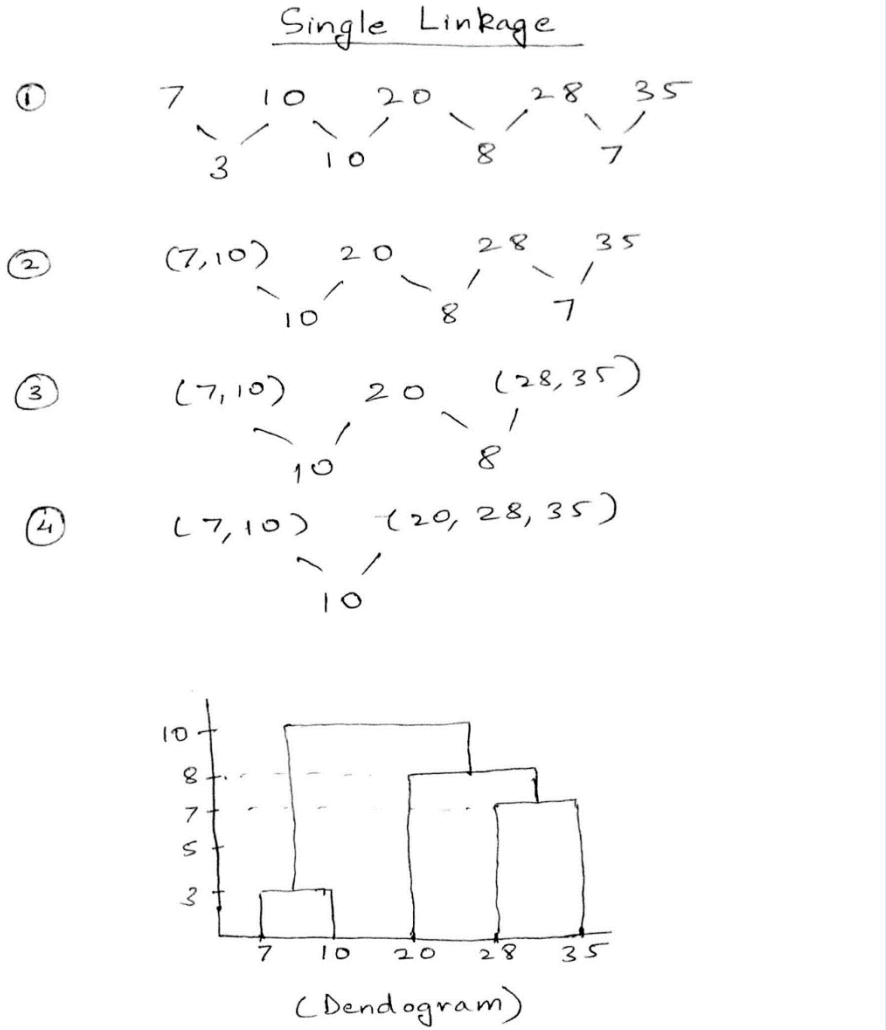
שאלה – ציור דנדרוגרמה

צייר דנדרוגרמה סכמטית עבור הנקודות הבאות.





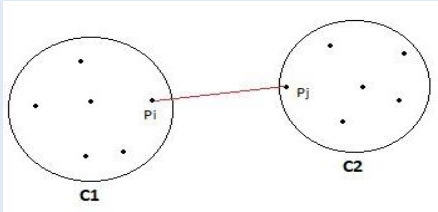
מרחק בין קלאסטרים

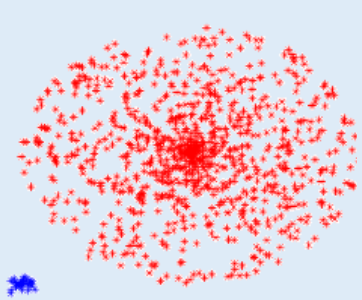


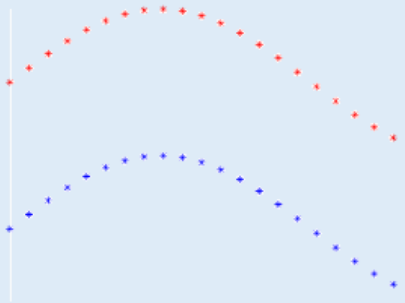
המרחק שלפיו עושים את האיחוד

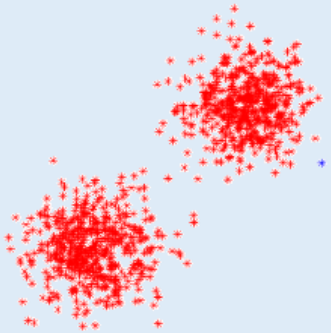
Single linkage

המרחק בין שני קלאסטרים שווה למרחק המינימלי בין הנקודות בקבוצה האחת לקבוצה האחרת.

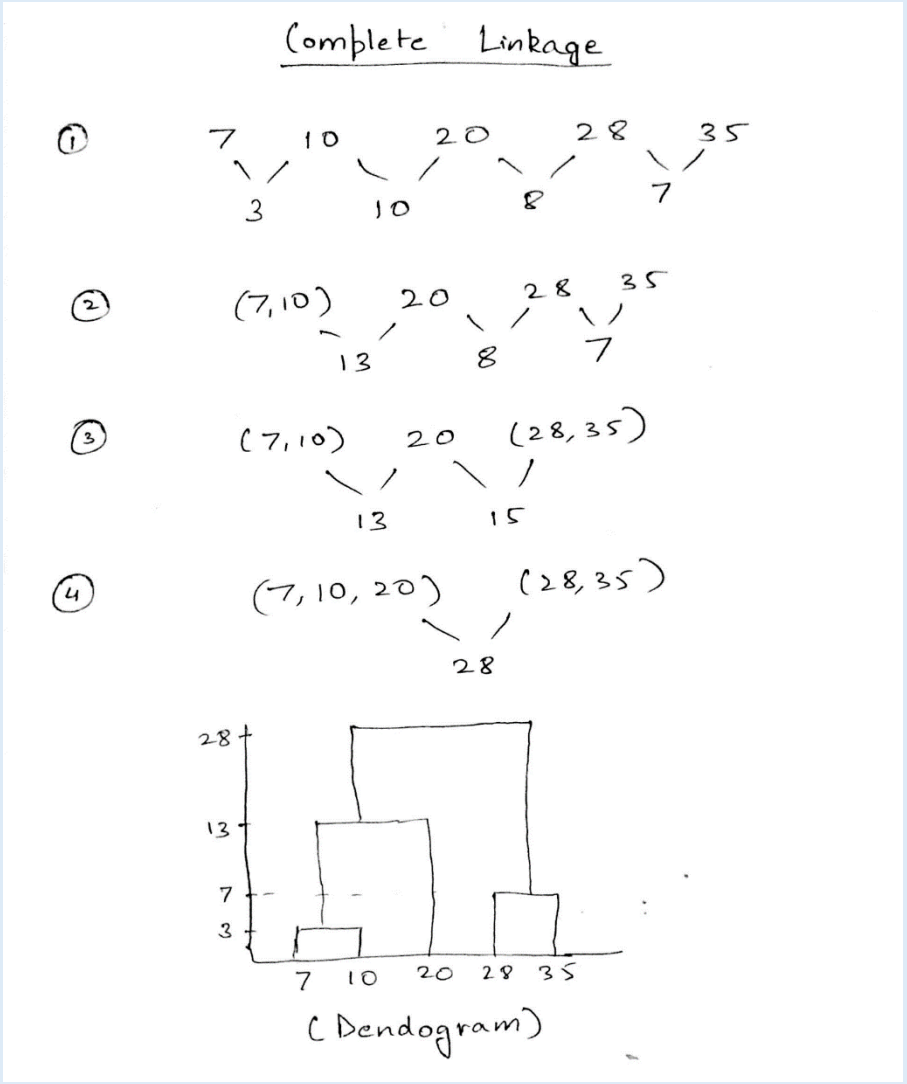








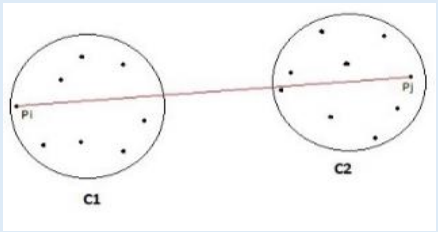
מרחק בין קלאסטרים

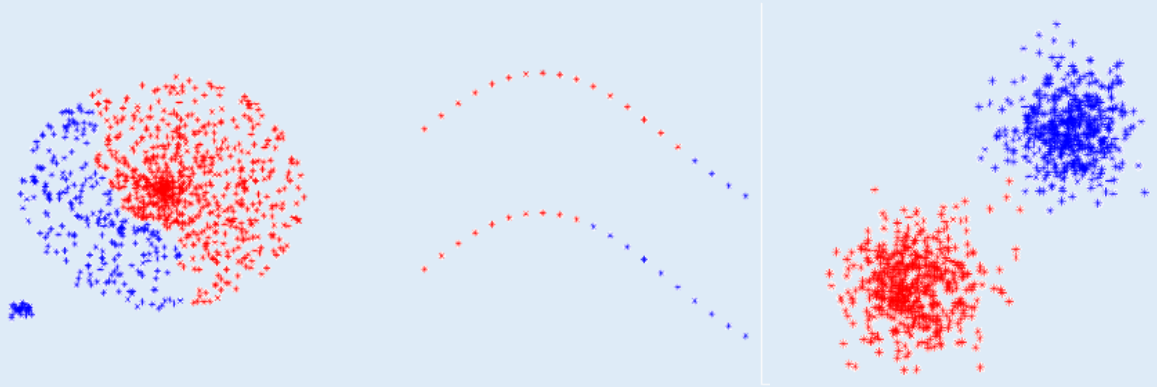


המרחק שלפיו עושים את האיחוד

Complete linkage

המרחק בין שני קלאסטרים שווה למרחק המקסימלי בין הנקודות בקבוצה האחת לקבוצה האחרת.



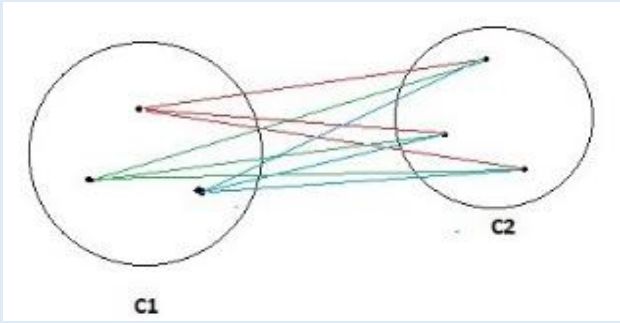


מרחק בין קלאסטרים

המרחק שלפיו עושים את האיחוד

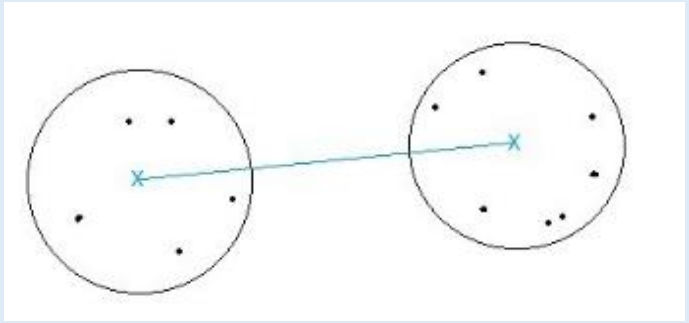
Average linkage

המרחק בין שני קלאסטרים שווה לממוצע המרחק בין הנקודות בקבוצה האחת לקבוצה האחרת.



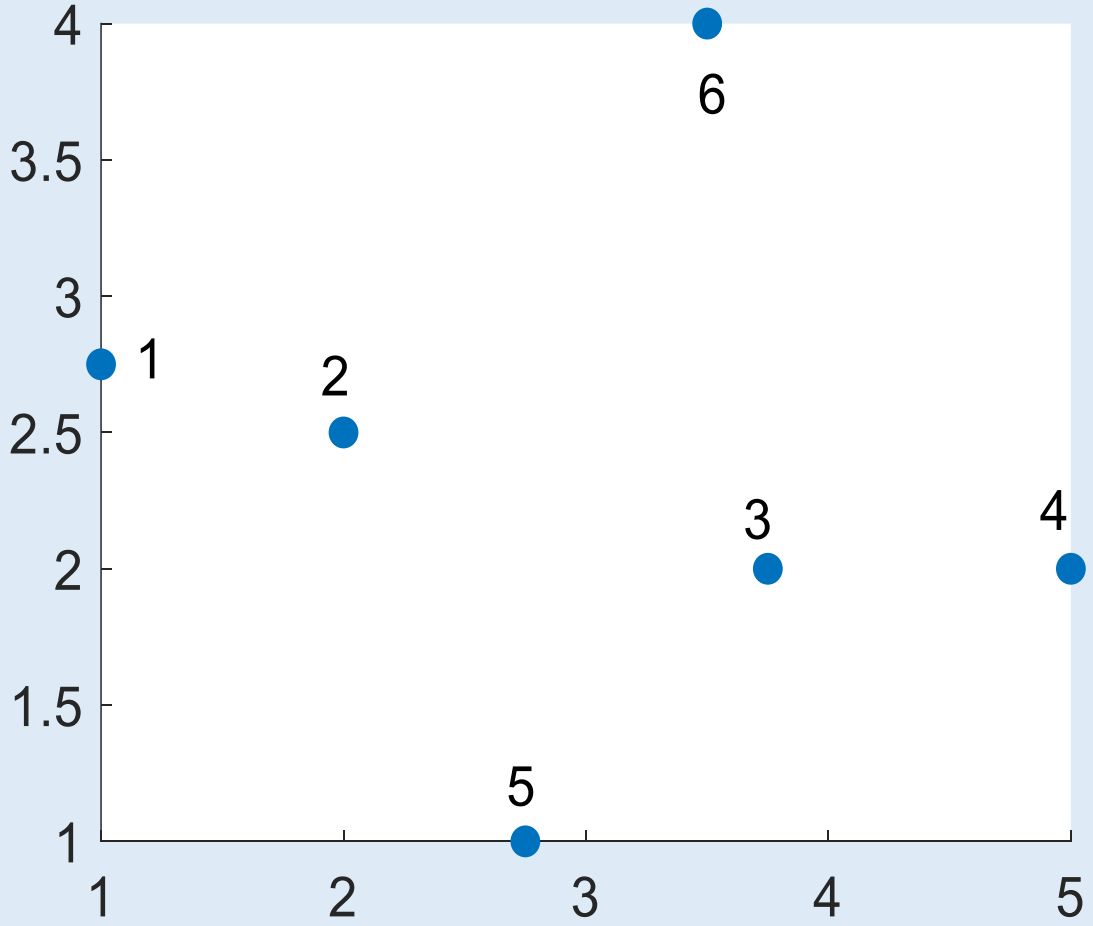
Centroid linkage

המרחק בין שני קלאסטרים שווה למרחק בין הצנטרואידים (הממוצע) של שתי הקבוצות



דוגמא

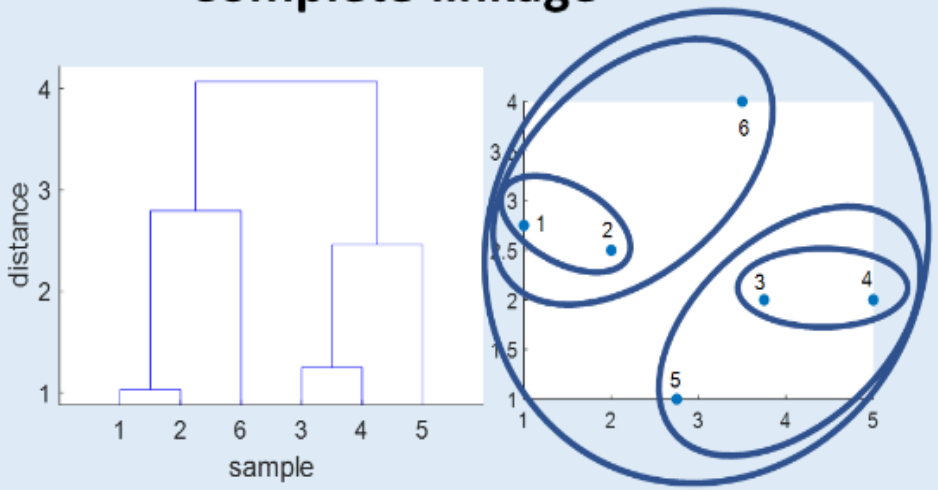
	1	2	3	4	5	6
1	0	1.0308	2.8504	4.0697	2.4749	2.7951
2	1.0308	0	1.8200	3.0414	1.6771	2.1213
3	2.8504	1.8200	0	1.2500	1.4142	2.0156
4	4.0697	3.0414	1.2500	0	2.4622	2.5000
5	2.4749	1.6771	1.4142	2.4622	0	3.0923
6	2.7951	2.1213	2.0156	2.5000	3.0923	0



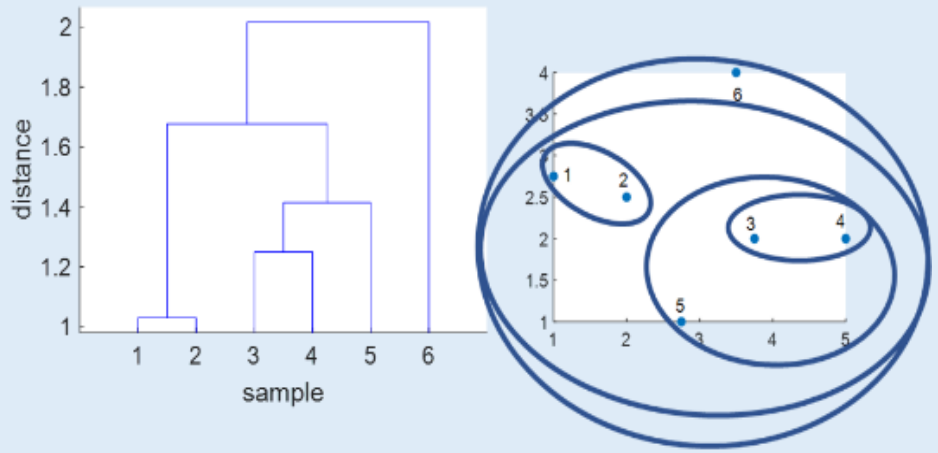
לפי הדנדרוגרמה, אם נרצה לחלק את הדוגמאות ל- 2 קבוצות, איך תתחלקנה הקבוצות?

דוגמא

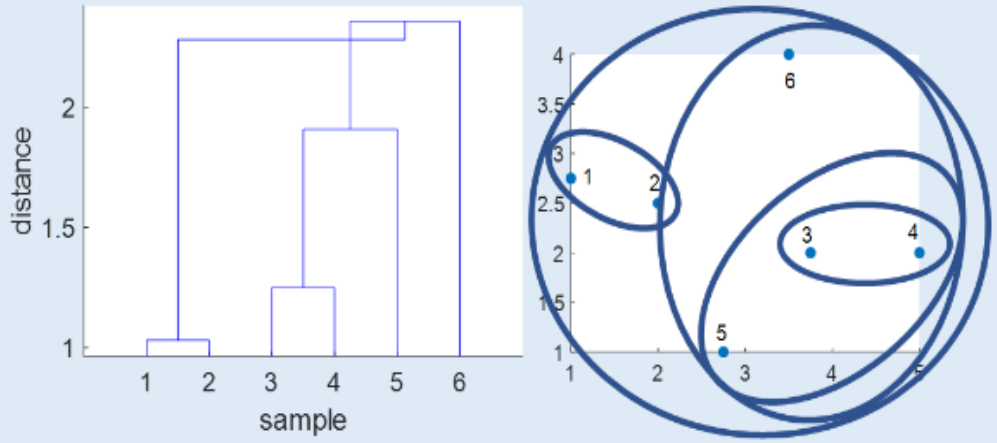
Complete linkage



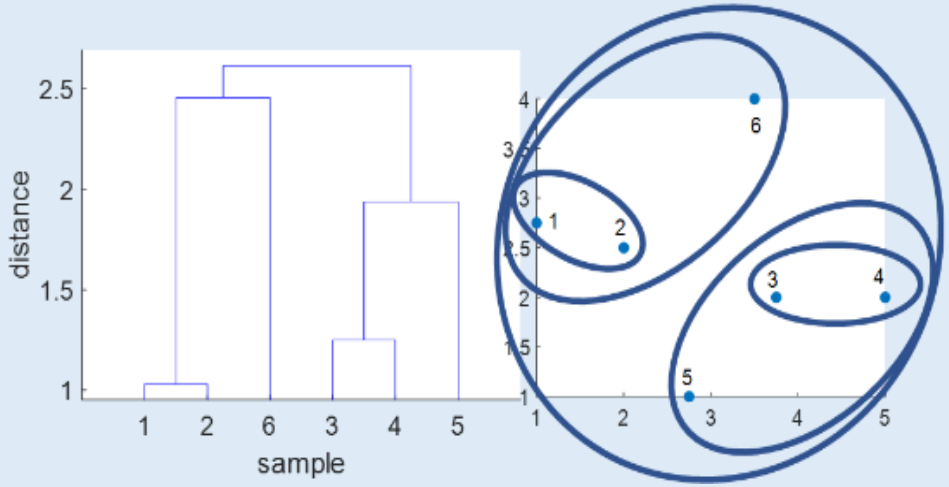
Single linkage



Centroid linkage



Average linkage

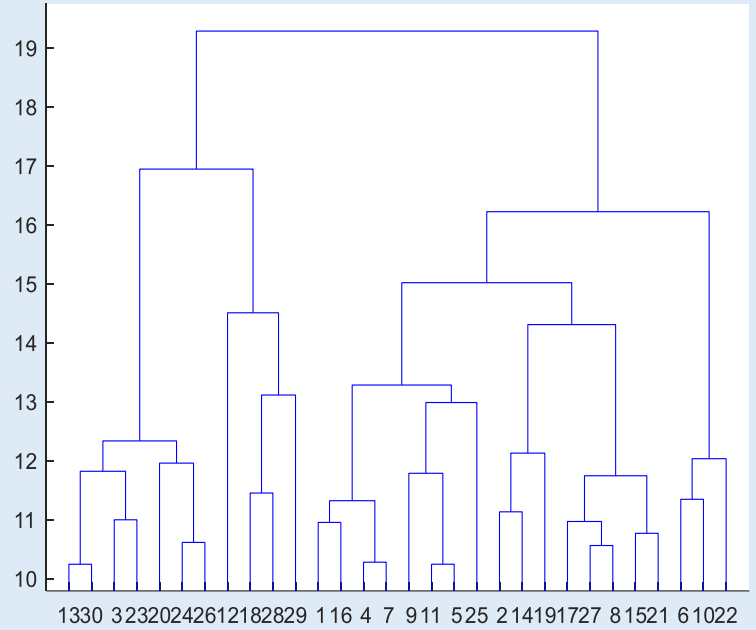


עוד דוגמא

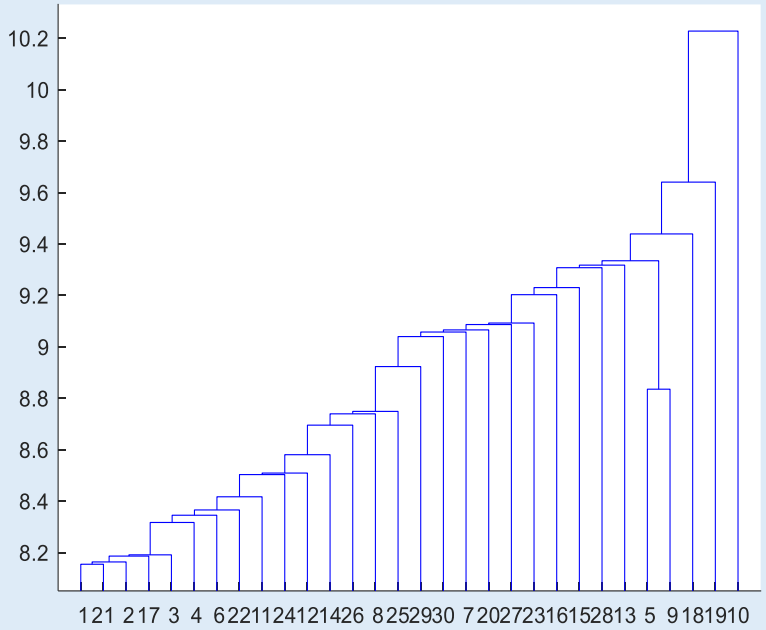
בקובץ נתונים רמות ביטוי של 500 גנים אצל 130 חולי לימפומה. חלקו את **החולים** לשתי קבוצות מובחנות בשיטת אִשכול היררכי.

פה מוצגים רק 30 החולים השונים ביותר זה מזה

Complete linkage

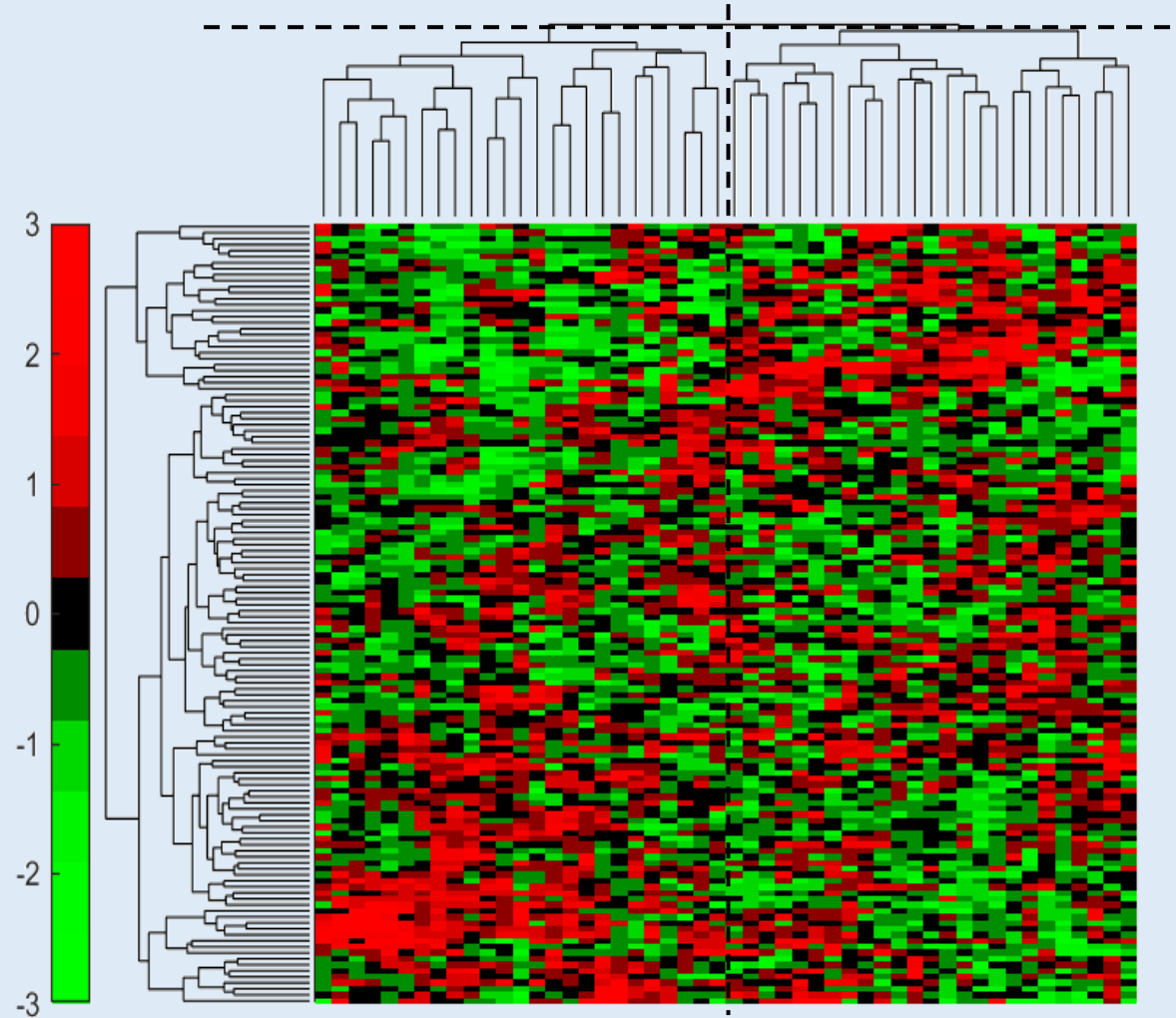


Single linkage



עוד דוגמא

בקובץ נתונים רמות ביטוי של 500 גנים אצל 130 חולי לימפומה. חלקו את החולים לשתי קבוצות מובחנות בשיטת אִשכול היררכי.



עם heatmap ניתן לראות גם אשכול של גנים וגם אשכול של דוגמאות (מוצגים רק 50 גנים)

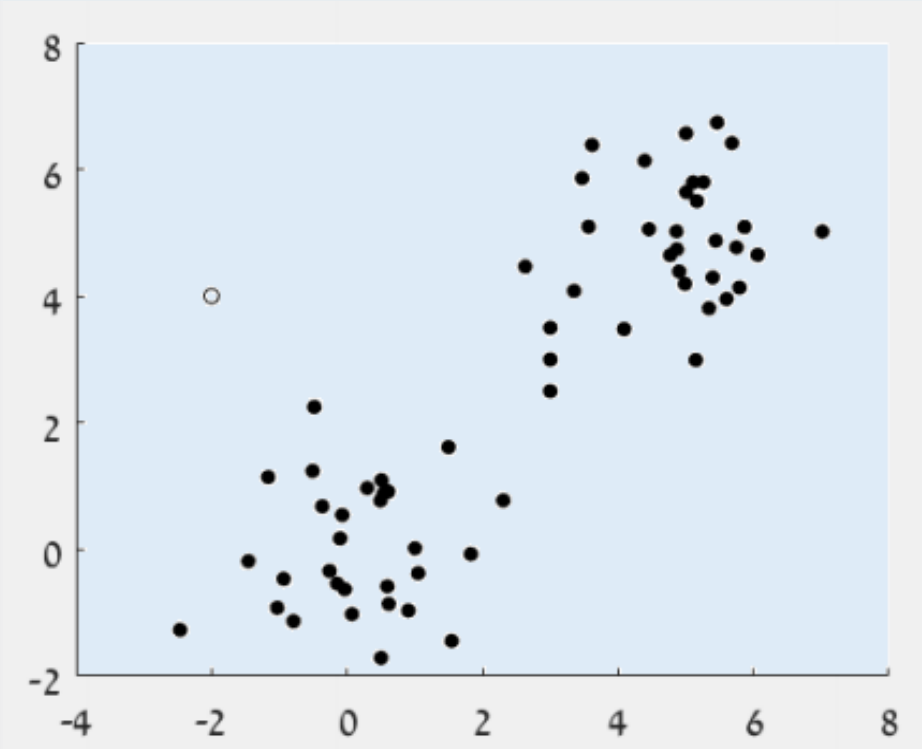
במקרה ויש גנים שסקלת הביטוי שלהם שונה מהשאר ואנחנו לא רוצים שהשוני בסקאלה יכתיב את האשכול:

$$x_{normalized} = \frac{(x_i - \bar{x})}{\sigma}$$

נותן ממוצע 0 וסטיית תקן 1

שאלה ממבחן

עשינו אשכול הררכי לדאטא דו מימדי. איזה שיטת linkage תיתן את החלוקה הבאה לשני אשכולות?



* שימו לב, שני הקלאסטרים בשאלה מסומנים בעיגולים מלאים או ריקים

א. Single linkage
ב. Complete linkage
ג. Centroid linkage
ד. Single או Complete
ה. Complete או Centroid

שאלה ממבחן

16. לפניכם קואורדינטות של 6 נקודות, וכן המרחקים ביניהן. איזה מהנדנדרוגרמות הבאות מייצגות hierarchal clustering לפי complete linkage?

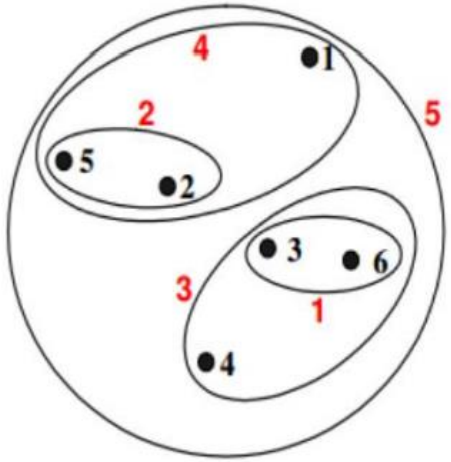
point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

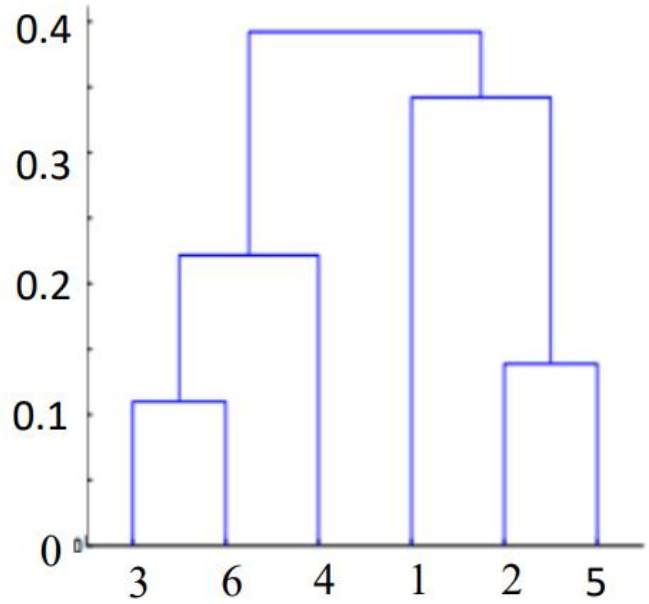
Table : X-Y coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

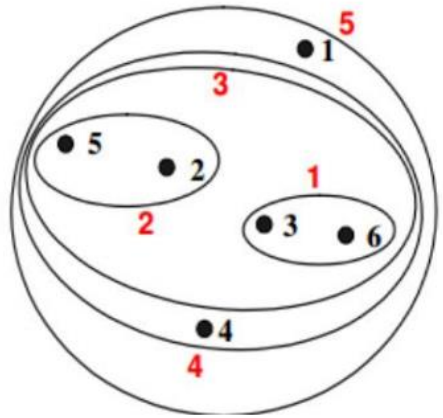
Table : Distance Matrix for Six Points

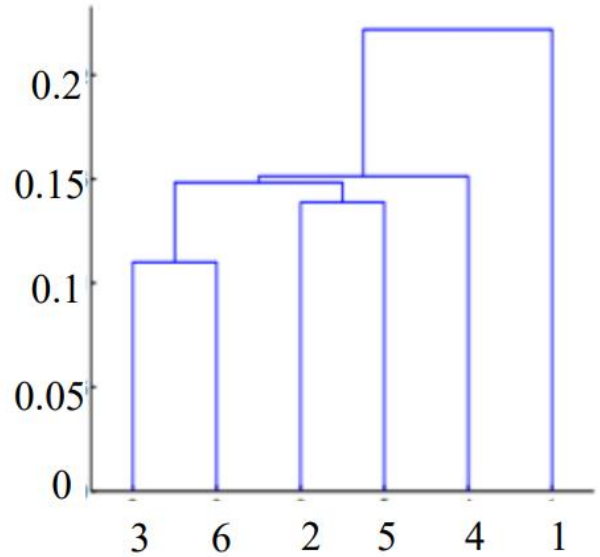
ב


4
1
5
2
5
2
3
6
3
1
4

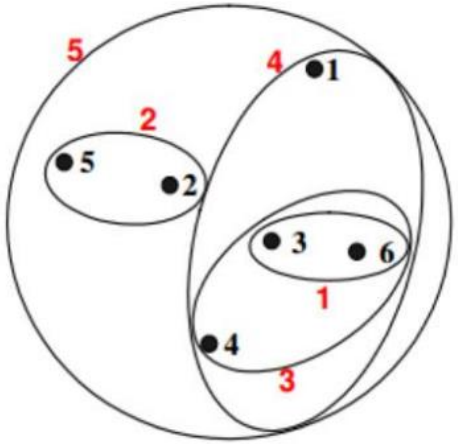

0.4
0.3
0.2
0.1
0
3 6 4 1 2 5

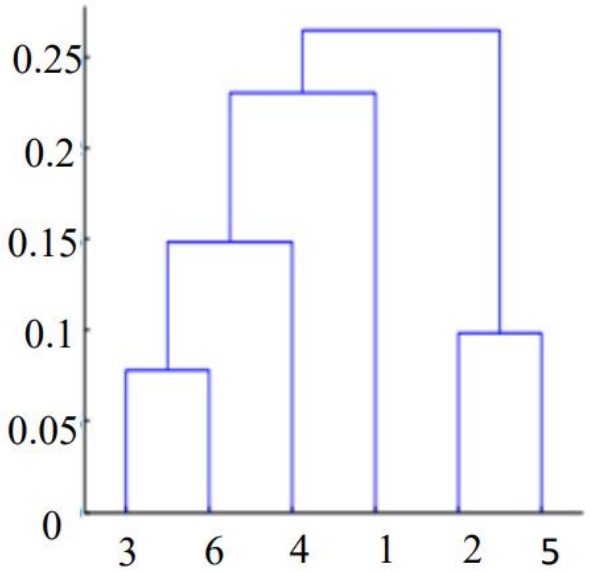
א


5
1
3
5
2
1
2
3
6
4
4

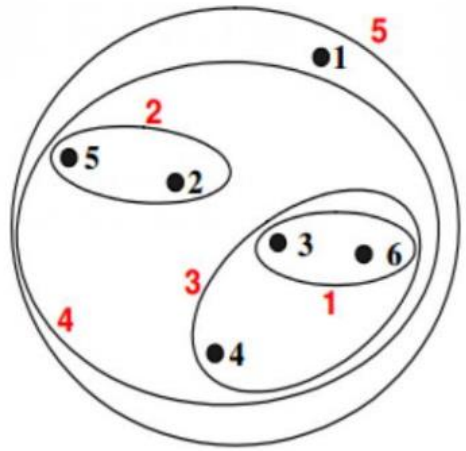

0.2
0.15
0.1
0.05
0
3 6 2 5 4 1

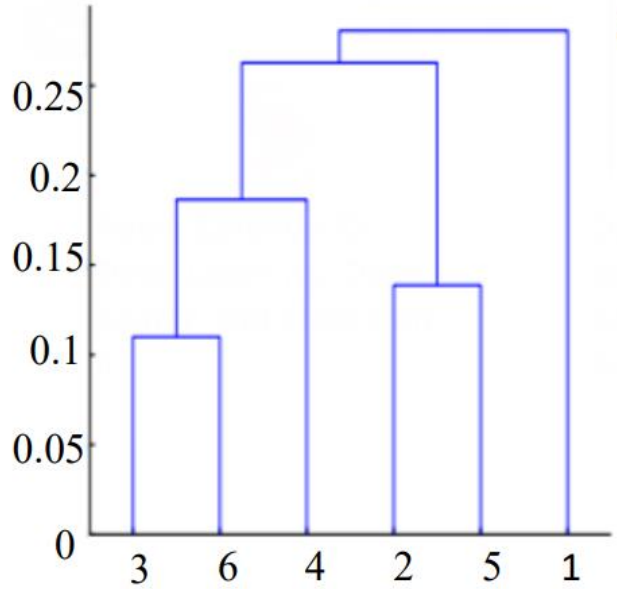
ד


5
4
1
2
5
2
3
6
1
4
3


0.25
0.2
0.15
0.1
0.05
0
3 6 4 1 2 5

ג


5
1
2
5
2
3
6
3
1
4
4


0.25
0.2
0.15
0.1
0.05
0
3 6 4 2 5 1