**תרגיל כיתה 11- אשכול היררכי**

בדיקת ביטוי גנטי מאפשרת מדידת רמת הביטוי של כמעט כל הגנום של הנבדק, כלומר הנתונים המתקבלים הם מממד גבוה. עלינו, אם כן, למצוא דרך לנתח מידע כזה בדרכים יעילות. שיטות ניתוח נתונים במרחב רב-ממדי מתחלקות לשתי קבוצות עיקריות:

**1.Supervised learning (למידה מונחית)-** בלמידה מונחית הדאטא מחולק לקבוצות ידועות מראש (למשל, חולים ובריאים). המטרה שלנו היא למצוא מודל שיכול לשייך דוגמא לאחת הקבוצות על סמך ההבדלים בין הקבוצות. בעיות הסיווג שלמדנו בסמסטר הקודם הם דוגמה לניתוח נתונים מסוג זה.

**2.Unsupervised learning (למידה לא מונחית)-** בלמידה לא מונחית הדאטא אינו מחולק לקבוצות ידועות מראש (למשל, אנחנו לא יודעים מראש מי החולים ומי והבריאים). המטרה שלנו היא לאתר תבניות וקבוצות בדאטא על סמך הדמיון בין הדוגמאות. בסוג למידה זה, אנו לומדים את המבנה הכללי של הנתונים מתוך הנתונים עצמם.

אחת הדרכים המקובלות לניתוח נתונים בגישת למידה שאינה מונחית היא **אִשכּוּל** (Clustering)– חלוקה של הנתונים לאשכולות (קלסטרים), צברים של נקודות דומות. מטרת שיטות האשכול השונות היא לחלק את הדוגמאות לקבוצות במרחב רב-ממדי, 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.

היום נלמד על אשכול לפי אלגוריתם **אִשכּול היררכי**, בשבוע הבא נלמד על אלגוריתם **K-means**.

**אִשכּול היררכי (Hierarchical clustering)**

אשכול היררכי מתבצע בשיטת ה- **bottom-up**: מאחים את הנקודות הדומות מלמטה למעלה. מתחילים בקבוצות קטנות שמתאחדות בהדרגה עד שלבסוף נקבל קבוצה המכילה את כל הנקודות ביחד. בסיום הרצת האלגוריתם, מתקבלת דנדרוגרמה – תצוגה של הנתונים בצורת עץ המתאר את סדר האיחויים.

האלגוריתם:

1. בחר הגדרת "מרחק".
2. אתחול: חשב את המרחק בין כל שתי נקודות.
3. עד שלא נותר מה לאחד:
   1. מצא את שתי הקבוצות הקרובות ביותר. שימו לב – באלגוריתם זה אנו מתייחסים גם לנקודות בודדות בתור קלסטרים (המכילים נקודה אח., קלסטרים אלה נקראים לעיתים singletons).
   2. אחה אותם.
   3. חשב מחדש את המרחקים בין כל הקבוצות.

**הדנדרוגרמה**: הדנדרוגרמה מציגה את סדר האיחוי בין כל שתי קבוצות. ככל שאיחוי נעשה בשלב מאוחר יותר הוא יופיע גבוה יותר בהיררכיה. כל דוגמה היא **עלה** בדנדרוגרמה; כאשר שתי קבוצות מתאחות, הן מתחברות ל**ענף** אחד. העלים מסודרים בתחתית הדנדרוגרמה, וככל שעולים למעלה הענפים מתמזגים עם עלים (דוגמאות) או עם ענפים אחרים. ככל שהמיזוג מתרחש בשלב נמוך יותר בדנדרוגרמה, כך הדמיון בין הקבוצות המתאחדות גבוה יותר. לעומת זאת, איחויים מאוחרים וקרובים לפסגה הם איחויים של קבוצות שונות יחסית זו מזו.

לכן על מנת להסיק כמה שתי דוגמאות דומות זו לזו, נלך מתחתית הדנדרוגרמה עד לפסגתה ונבדוק מתי מתרחש האיחוד הראשון בין שתי הדוגמאות, מתי הן מתאחדות לקבוצה אחד. דוגמאות המתאחדות בתחתית העץ הן דוגמאות דומות יחסית לאלו המתאחדות בפסגת העץ.

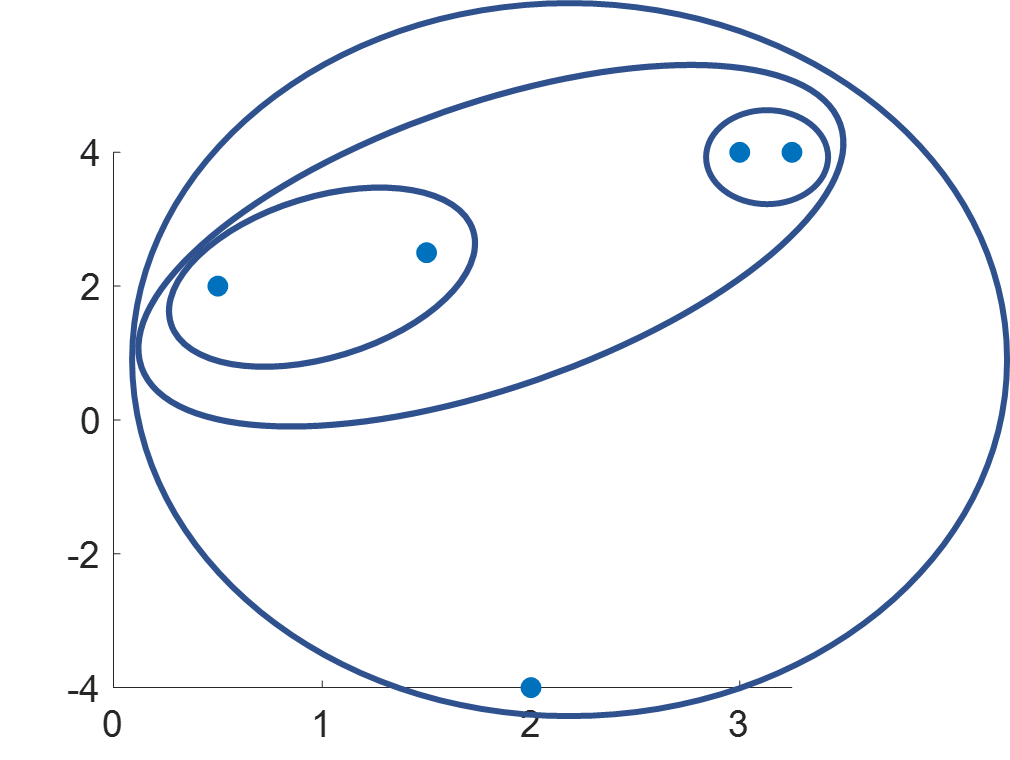
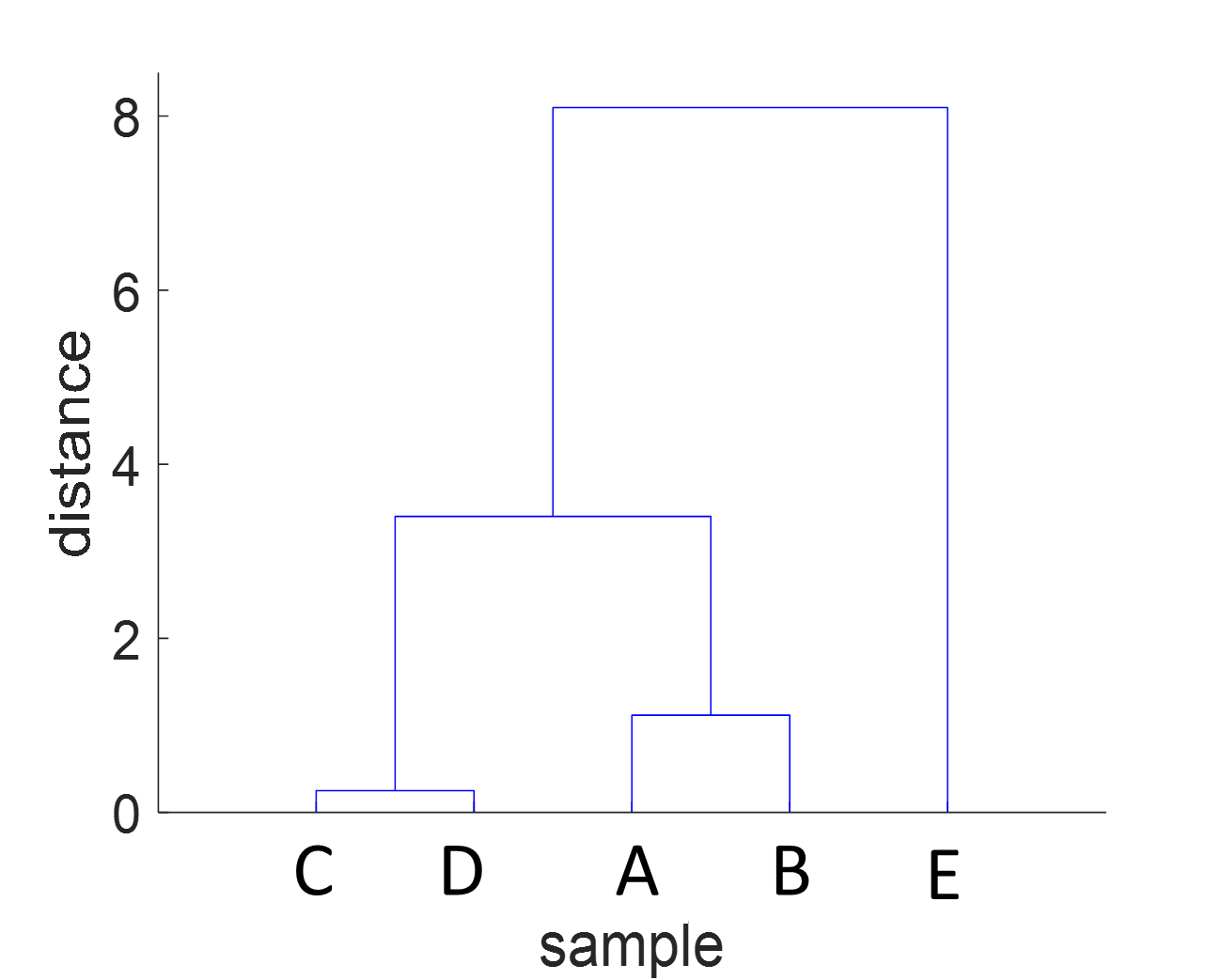
אחרי ששרטטנו את הדנדרוגרמה, נוכל לחלק את הדוגמאות לקבוצות על ידי "חיתוך" הדנדרוגרמה בגובה מסוים, בהתאם למספר הקבוצות הרצוי.

**דוגמה סכמתית**

צייר דנדרוגרמה עבור הנקודות הבאות.

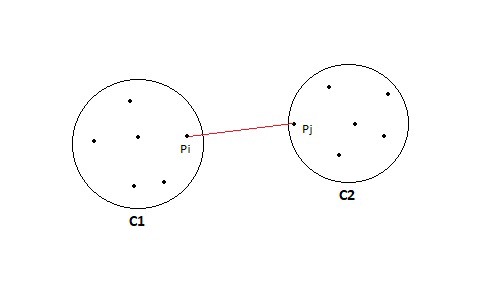


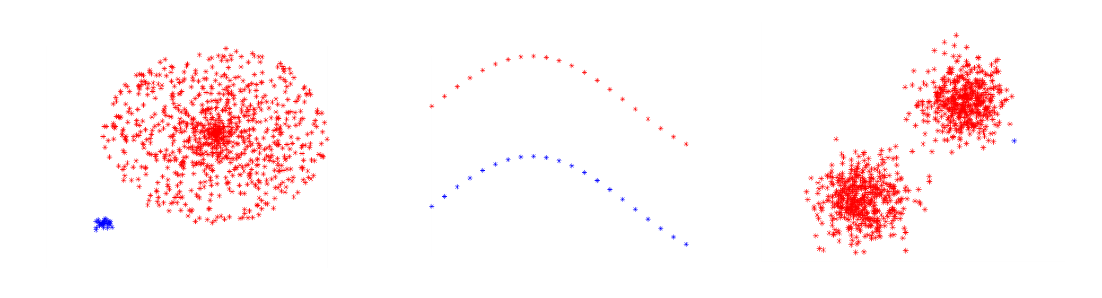
**פתרון:**

** **

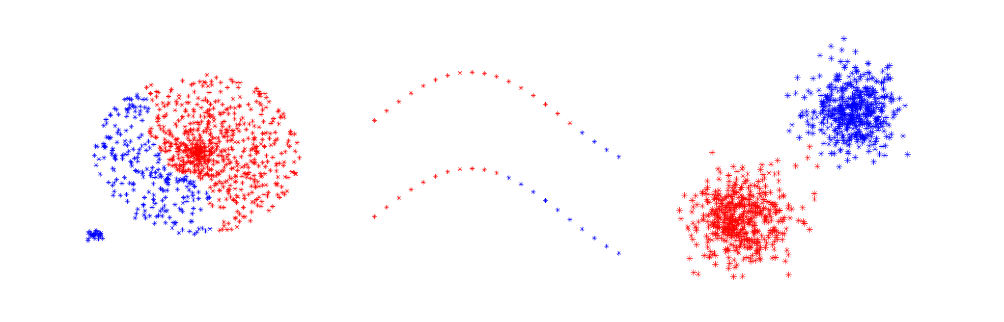
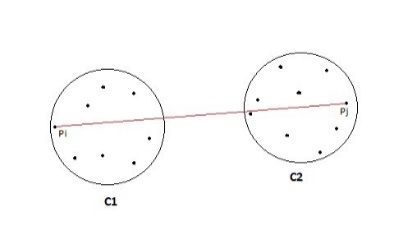
כדי לממש את האלגוריתם צריך להגדיר איך מחשבים **מרחק בין קלסטרים**. קיימות מספר שיטות לכך:

1. Single linkage: המרחק בין שני קלסטרים שווה **למרחק המינימאלי** בין הנקודות בקבוצה האחת לנקודות בקבוצה השנייה. יכול להפריד בין אשכולות לא עגולים אם המרחק ביניהם מספיק גדול. לא עובד טוב כאשר יש "רעש" בין האשכולות:

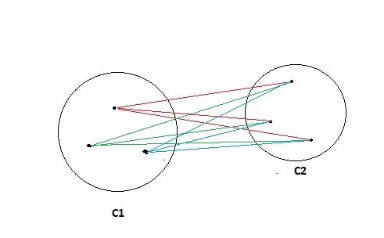




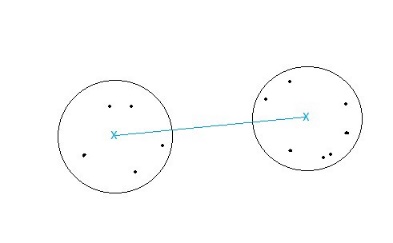
1. Complete linkage: המרחק בין שתי קבוצות שווה **למרחק המקסימלי** בין הנקודות בקבוצה האחת לנקודות בקבוצה השנייה. עובד טוב גם עם "רעש". עלול להפריד אשכול גדול לשניים.



1. Average linkage: המרחק בין שתי קבוצות שווה **למרחק הממוצע** בין הנקודות בקבוצה האחת לנקודות בקבוצה השניה.



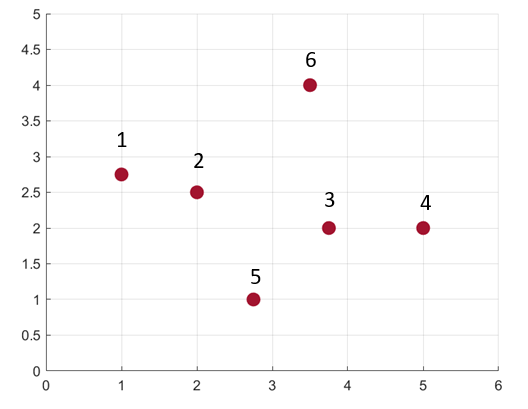
1. Centroid linkage: המרחק בין שתי קבוצות שווה למרחק בין הצנטרואידים (נקודות הממוצע) של שתי הקבוצות.



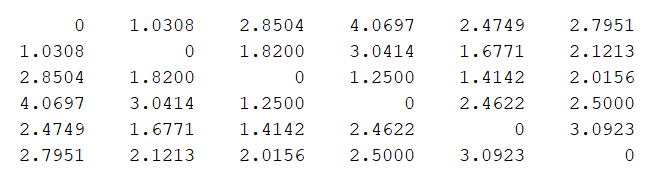
Complete linkage, average linkage and centroid lingkage כולם פחות מושפעים מרעש ו-outliers מאשר single linkage. הם עובדים יותר טוב עם אשכולות עגולים ויש להם נטייה לשבור אשכולות גדולים.

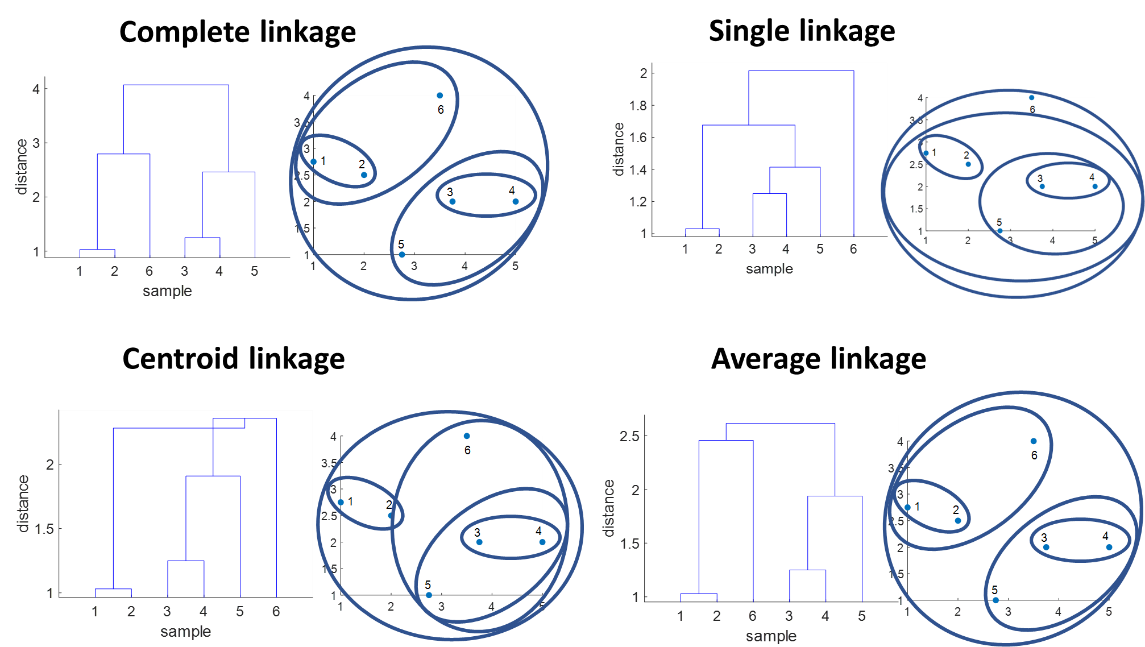
**דוגמא**

נתונות 6 נקודות במרחב דו-ממדי:



מטריצת המרחקים בין הנקודות הנתונות:





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

**תרגיל**

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

**פתרון**

כברירת מחדל מוצגת דנדרוגרמה חלקית, הכוללת רק את 30 הקבוצות השונות ביותר זו מזו:

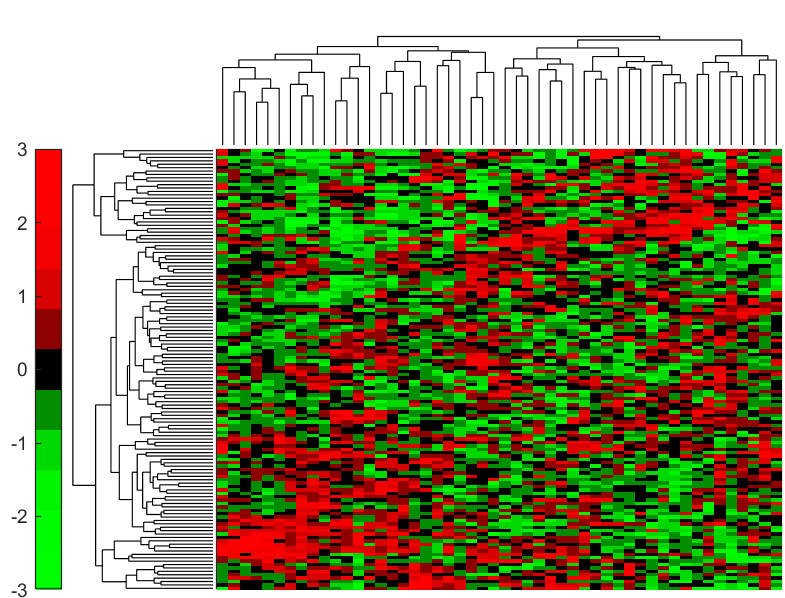
Complete linkage Single linkage

ב single linkage מתקבל מבנה מעוות לנתונים. בעוד שב complete linkage מתקבל מבנה שיותר מתאים לשני אשכולות.

נציג דרך להמחשה ויזואלית של חלוקת הנתונים המקוריים לקבוצות, על ידי קלסטרגרמה אשר דרכה תוצג המטריצה המקורית של הנתונים, ולידה אשכּוּל של השורות ואשכּוּל של העמודות. כדי לא להעמיס יותר מדי על הגרף ניקח רק את 50 הגנים הראשונים.

כדי שהוויזואליזציה תהיה מוצלחת, נרמלנו כל שורה לממוצע 0 וסטיית תקן 1 (כל גן).



נרמול הדאטא: במקרה שיש גנים שסקלת הביטוי שלהם שונה מהשאר ואנחנו לא מעוניינים שהשוני בסקלה יכתיב את השכול, נצטרך לנרמל את הדאטא. אחד הדרכים לנרמול הדאטא היא להחסיר את הממוצע של הגן ולחלק בסטיית התקן:

חלוקת החולים לשתי הקבוצות יכולה לעזור לנו למצוא גנים שמתבטאים בצורה שונה בין שתי הקבוצות (למשל, ע"י מבחן t) וכך אפשר לזהות רשת בקרה גנטית שמאפיינת אחת הקבוצות של החולים ולפתח עבורה תרופה.

**שאלות ממבחן:**

