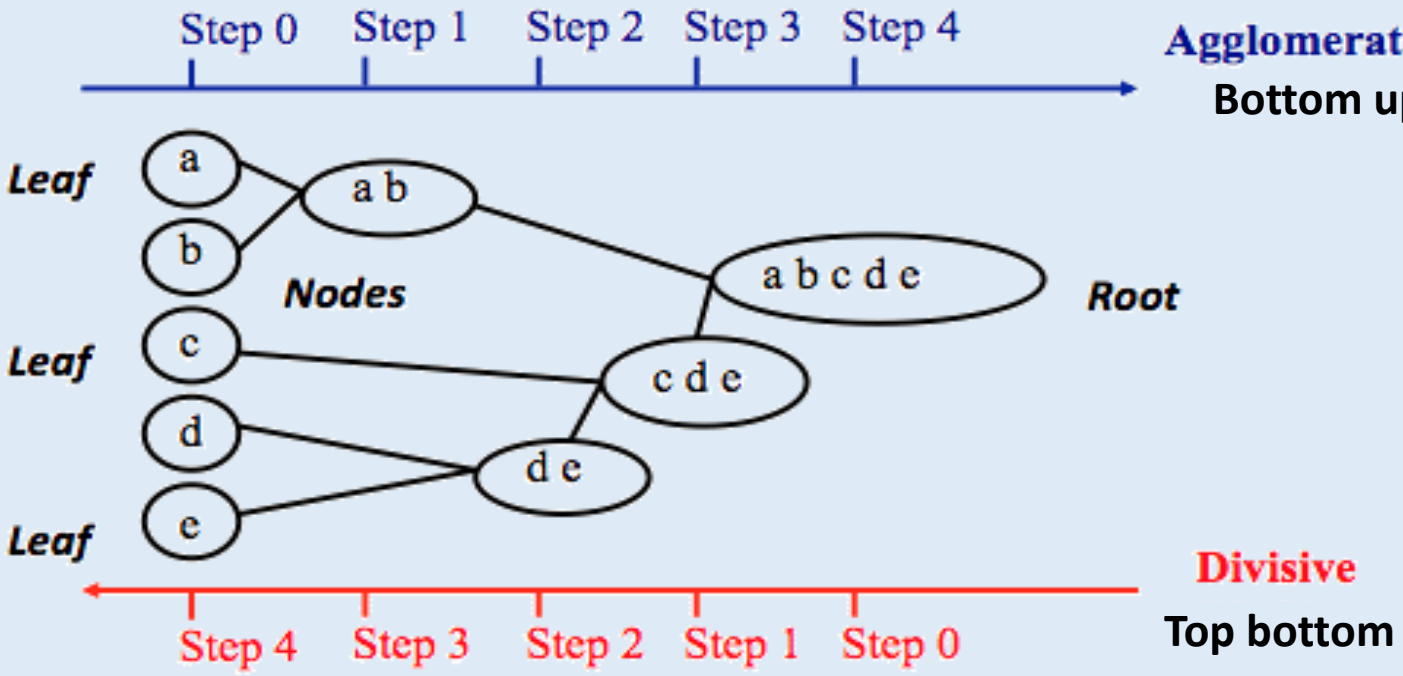
אִשכּוּל (Clustering) רב-מימדי

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

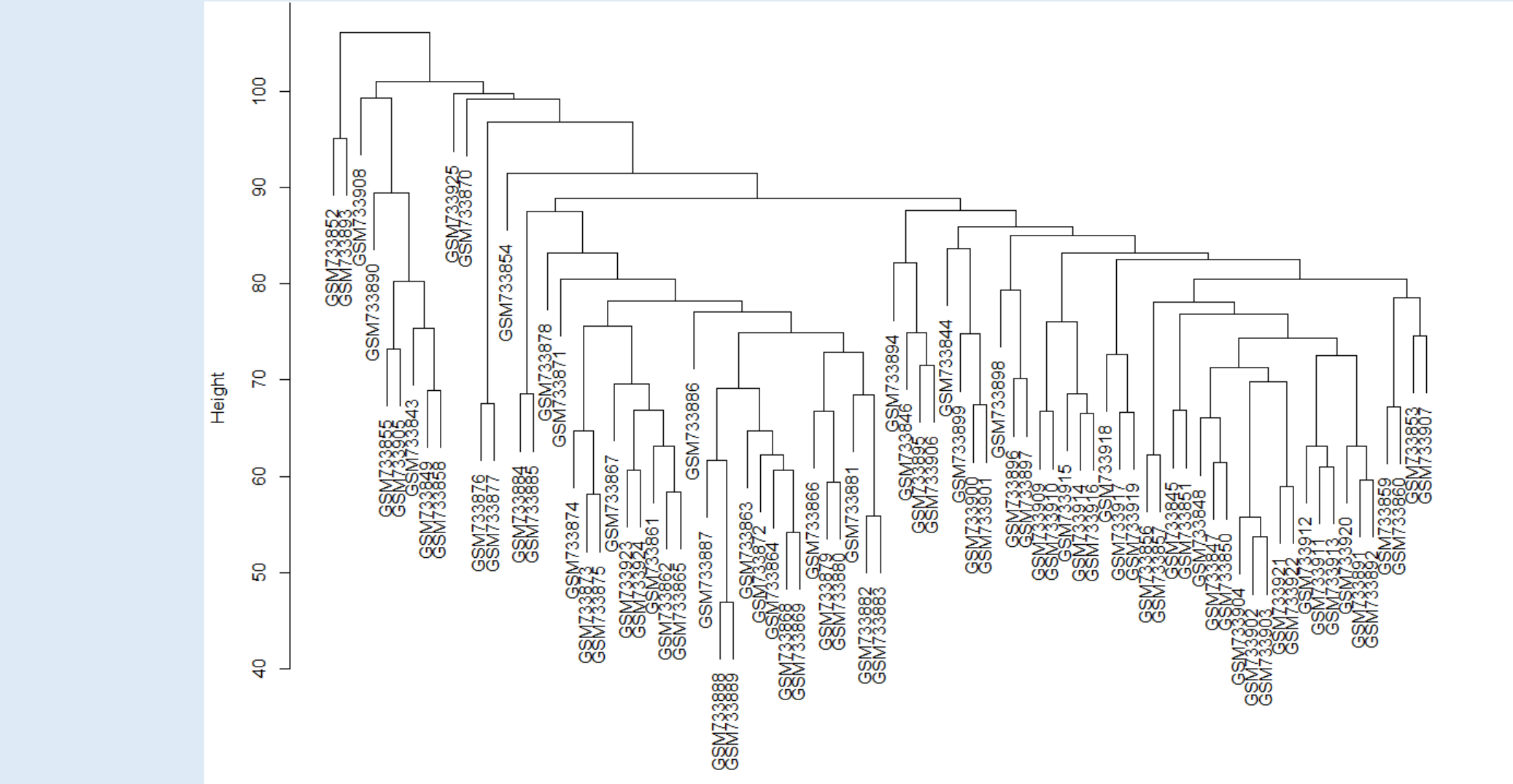
קלאסטרינג היררכי:

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

* 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

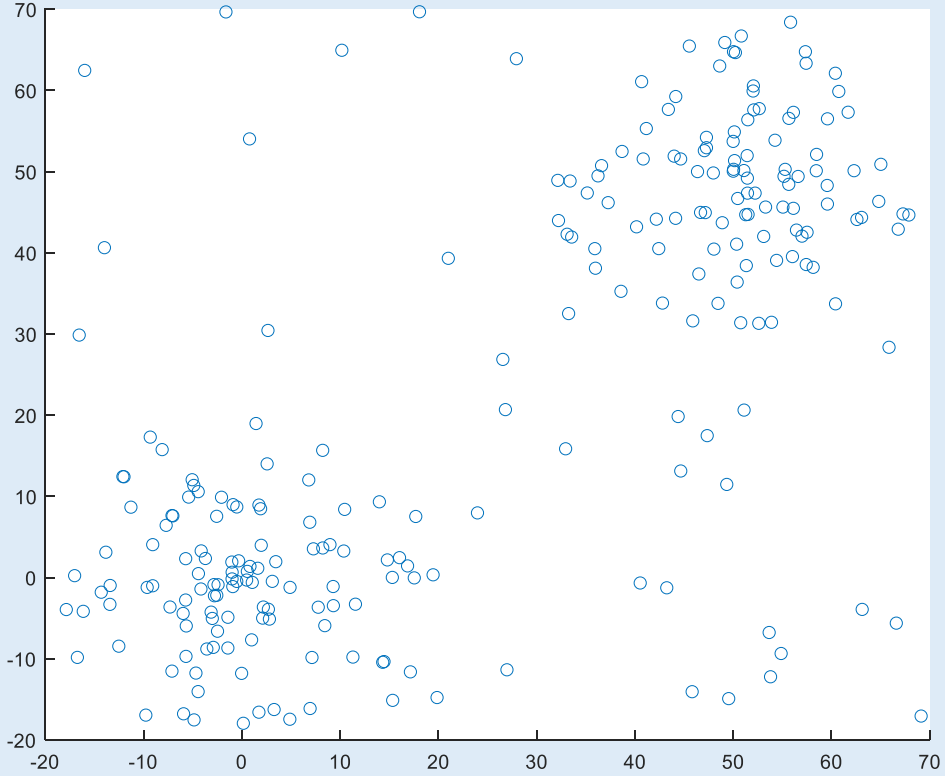
אלגוריתם K-means

עד כה תרגלנו שיטת האִשכּול היררכי, שהיא דוגמא לגישה 'מלמטה למעלה' (bottom-up). עכשיו נתרגל חלוקה לקבוצות בשיטת k-means, שהיא דוגמא לגישה ההפוכה – 'מלמעלה למטה' (top-down).

מטרת האלגוריתם למזער את השונות בתוך כל קבוצה.

האלגוריתם:

1. אתחול: תייג כל דוגמה במספר קלסטר אקראי.
2. חזור עד להתכנסות:
   a. לכל קלסטר חשב צנטרואיד (הנקודה ה'ממוצעת' בקלסטר).
   b. תייג כל נקודה בהתאם לתיוג הצנטרואיד הקרוב אליה ביותר.



אלגוריתם K-means

תנאי עצירה

כדי שהאלגוריתם יעצר מתישהו, צריך להגדיר תנאי עצירה (הגדרה להוראה "חזור עד *להתכנסות*" לעיל).

בהרצאה ראיתם 3 אפשרויות לתנאי עצירה:

1. כשבכל ריצה של הלולאה אין יותר מעבר של דוגמאות מקבוצה לקבוצה.
2. כשבכל ריצה של הלולאה השינוי בממוצע המרחקים של הדוגמאות מהצנטרואידים קטן מערך מסויים.
3. לאחר מספר קבוע מראש של הרצות.

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

ועכשיו....קצת R

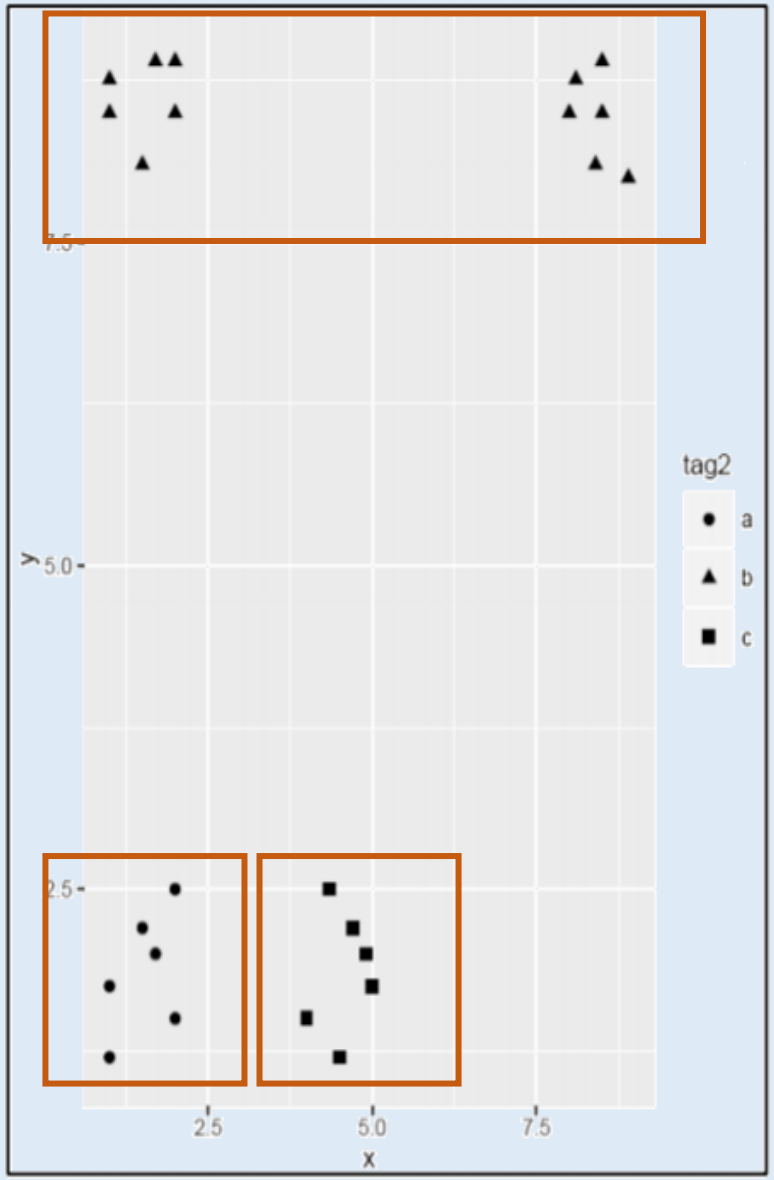
איך עושים את זה ב-R?

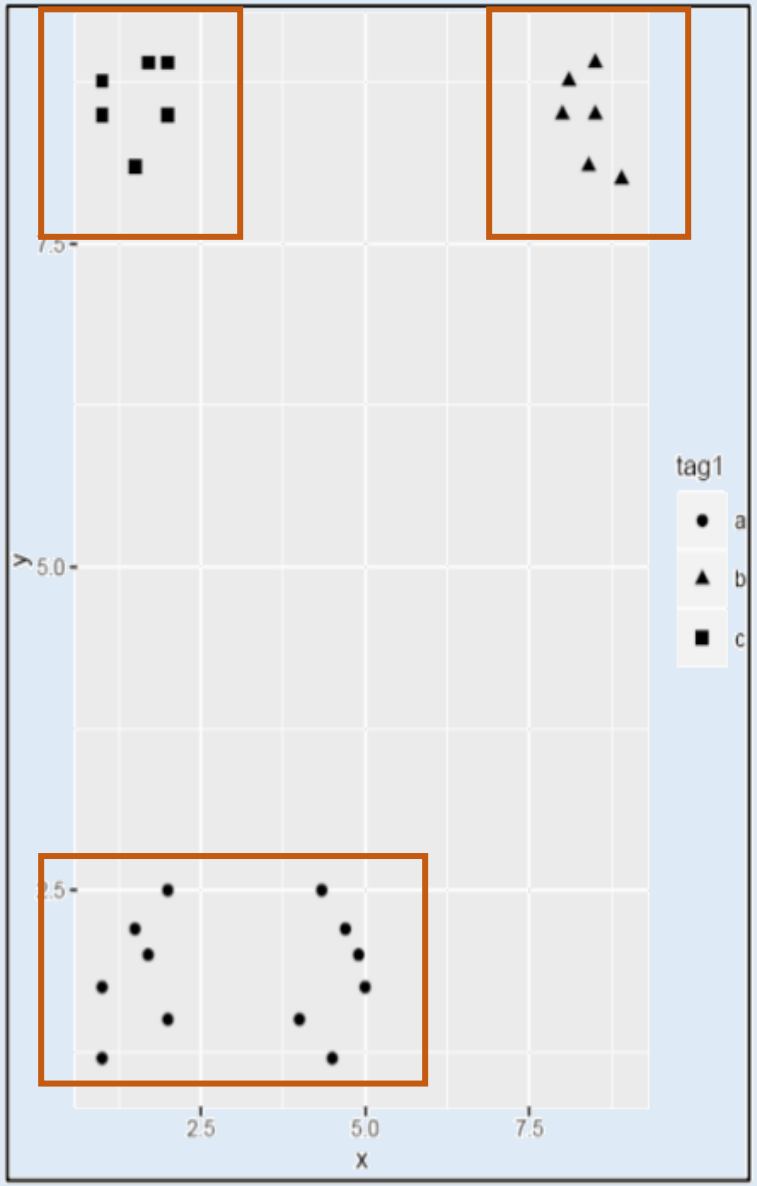
ביצוע אלגוריתם kMeans

```r
# ---------- K-Means ------------
Data <- read.csv("Data.csv", header = F)

plot(Data)

kData = kmeans(Data, centers = 2)

plot(Data, col=kData$cluster)
```

שורה	מה עושים בשורה?
2	קריאת המטריצה Data. כל שורה היא דוגמא, כל עמודה היא מדד.
4	ייצור גרף מנתוני Data. שימו לב שמכיוון שלא נתנו ערכי x וערכי y לפוקציה, הוא צייר את כל האפשרויות
6	ביצוע kMeans על נתוני Data תוך קביעת 2 קלאסטרים. האלגוריתם מתייחס לשורות כדוגמאות אותם צריך לאשכל.
8	ייצור גרף מנתוני Data תוך צביעה לפי מספר הקלאסטר כפי שנקבע ב-kMeans

אלגוריתם K-means





חסרונות האלגוריתם

1. תלות בתנאי התחלה

נתונות 24 נקודות במרחב דו-מימדי. הדוגמאות עברו clustering פעמיים באמצעות שיטת k-means עם k=3 תחת תנאים זהים. נתונים להלן תנאי ההתחלה של האלגוריתם בשני המקרים (צורה שונה מסמלת תיוג התחלתי שונה.

איך תהיה החלוקה לקלאסטרים לאחר התכנסות האלגוריתם?

למה זה קורה?

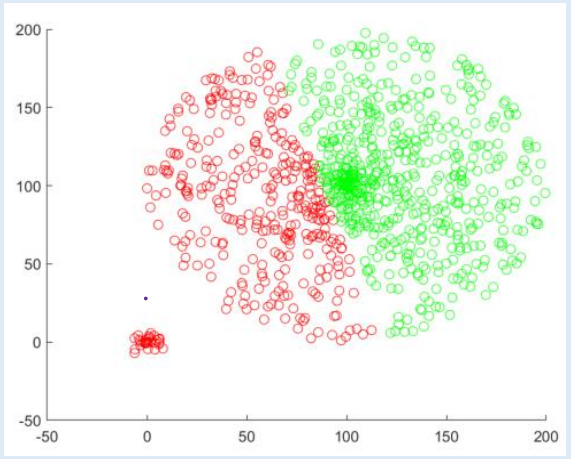
זה קורה מכיוון ש-k-means הוא אלגוריתם אופטימיזציה שמוצא <u>מינימום מקומי</u> כתלות בתנאי ההתחלה.

כדי להגיע ל'מינימום גלובלי' (פתרון האופטימלי), צריך לנסות מספר רב של תנאי התחלה שונים, ולקחת את הפתרון הכי טוב מבין כולם – זה שבו סך המרחקים של הדוגמאות מהצנטרואידים הוא הקטן ביותר.

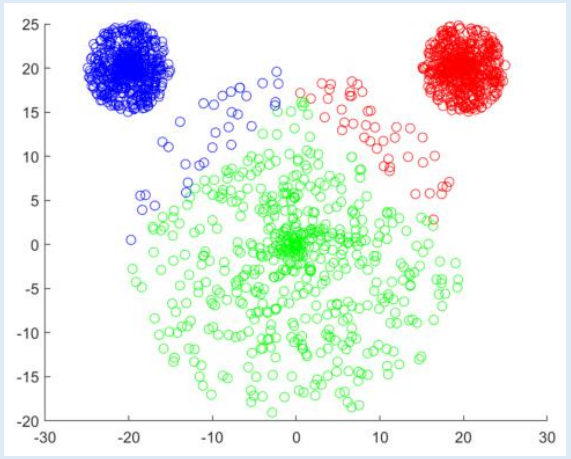
אלגוריתם K-means

חסרונות האלגוריתם

2. גודל הקבוצות לא משנה



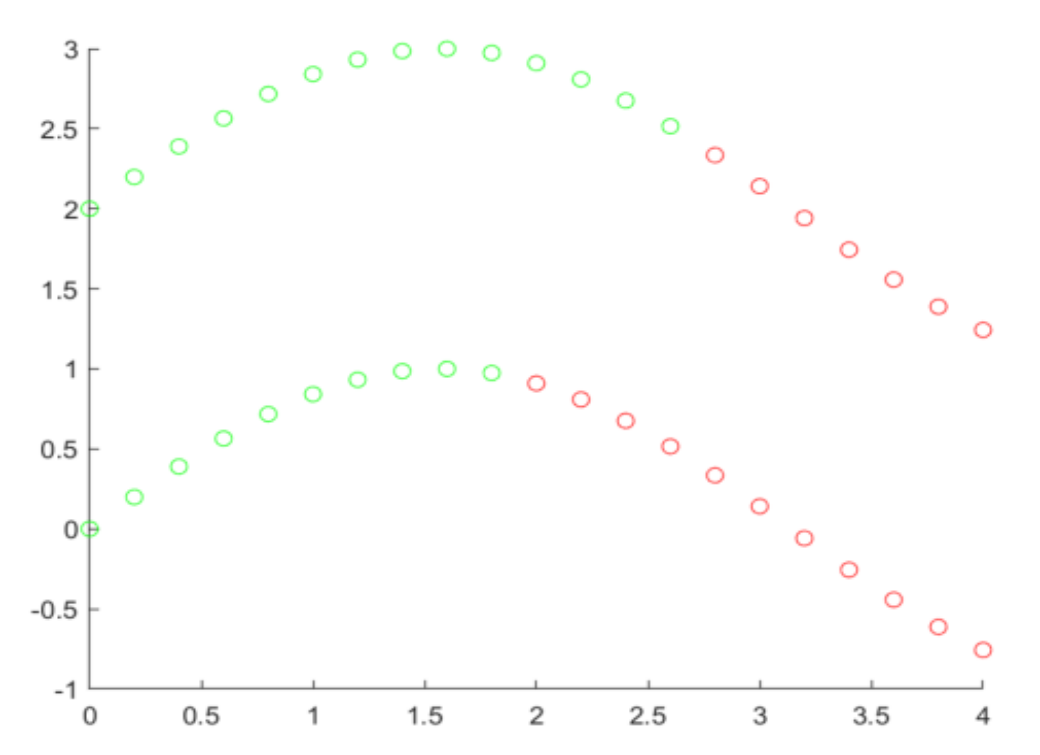
3. צפיפות הקבוצות לא משנה



אלגוריתם K-means

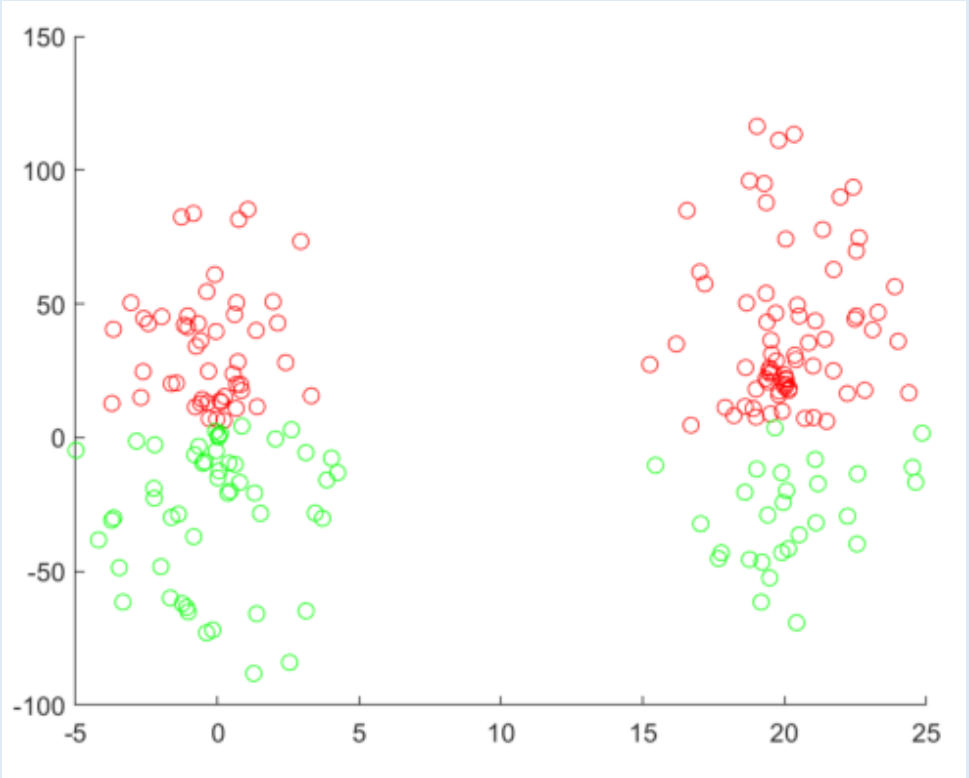
חסרונות האלגוריתם

4. צורת הקבוצה לא משנה (אם היא לא עגולה למשל)

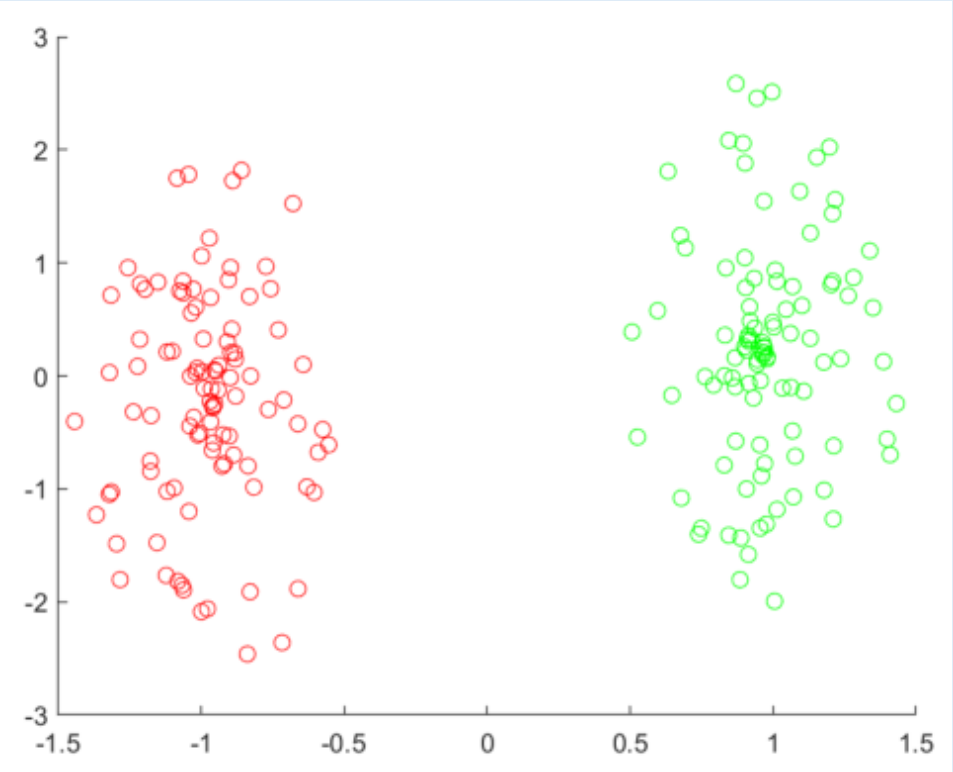


למה חשוב לנרמל?

לפני הנרמול



אחרי הנרמול



שימו לב לסקאלות: ציר X יחסית קטן וציר Y גדול ולכן הצנטרואידים ההתחלתיים יהיו למעלה ולמטה

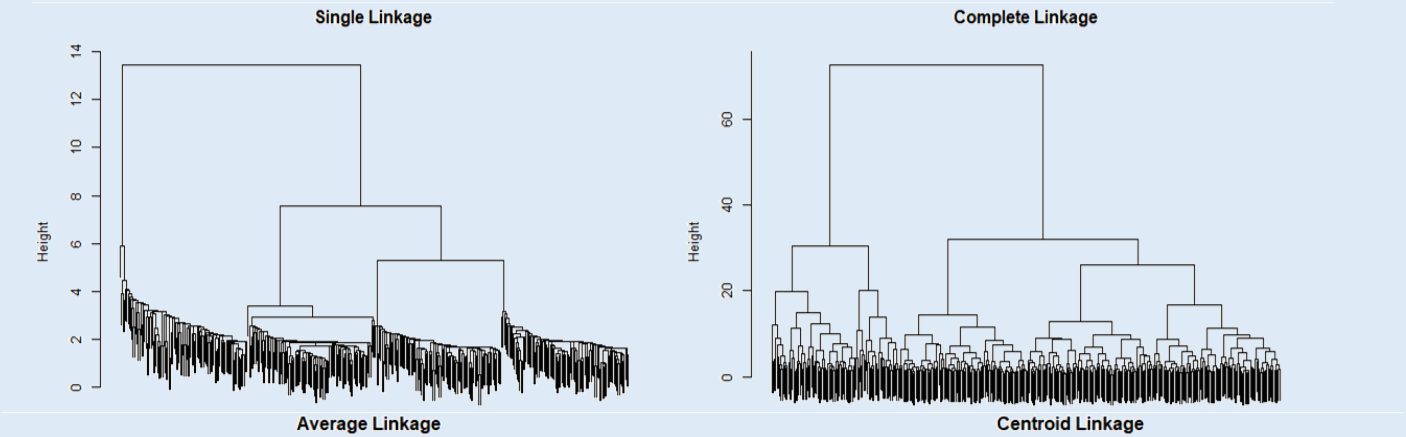
איך נקבע את מספר הקבוצות האופטימלי?

איך נקבע את מספר הקבוצות האופטימלי?

תשובה: נבדוק כמה יש

בניגוד לאשכול היררכי שבו החלוקה לקבוצות 'צומחת' מתוך הנתונים עצמם, ב k-means החלוקה לקבוצות נקבעת על ידי המשתמש. אך כיצד נדע מלכתחילה לכמה קבוצות יש לחלק את נתונים?

1. בעזרת אשכול היררכי

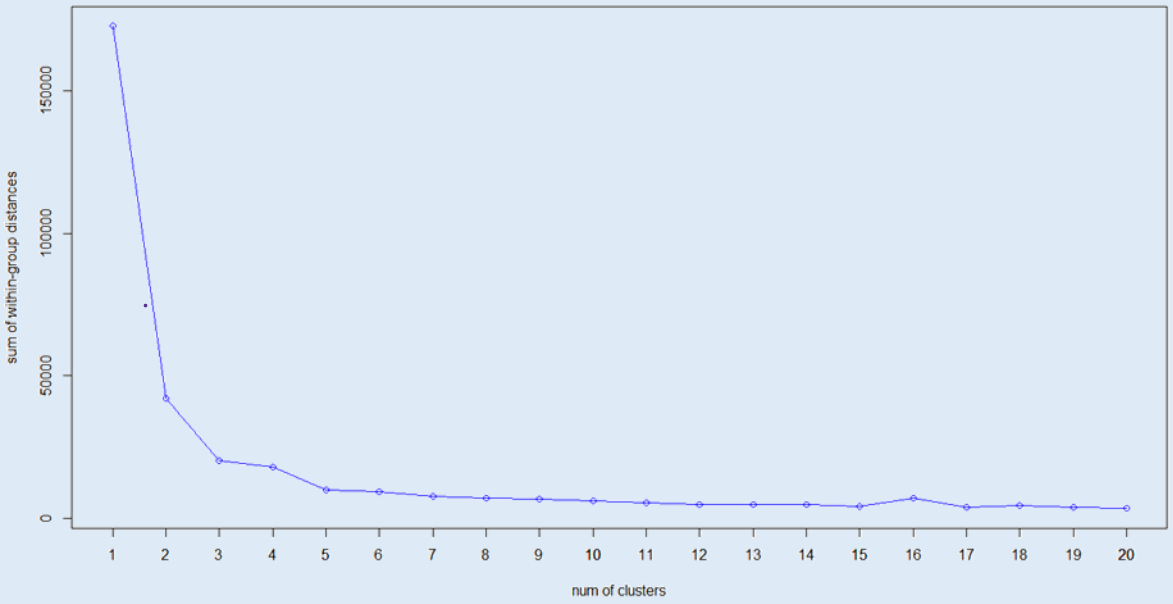


איך נקבע את מספר הקבוצות האופטימלי?

2. בעזרת גרף Scree

גישה כמותית. מציירים גרף של סכום השונויות בתוך כל קבוצה.

מתחילים מקלאסטר (קבוצה) אחד ובודקים את השונות מוסיפים עוד קלאסטר (k=2) ובודקים את השונות כך ממשיכים ויוצרים את הגרף Scree

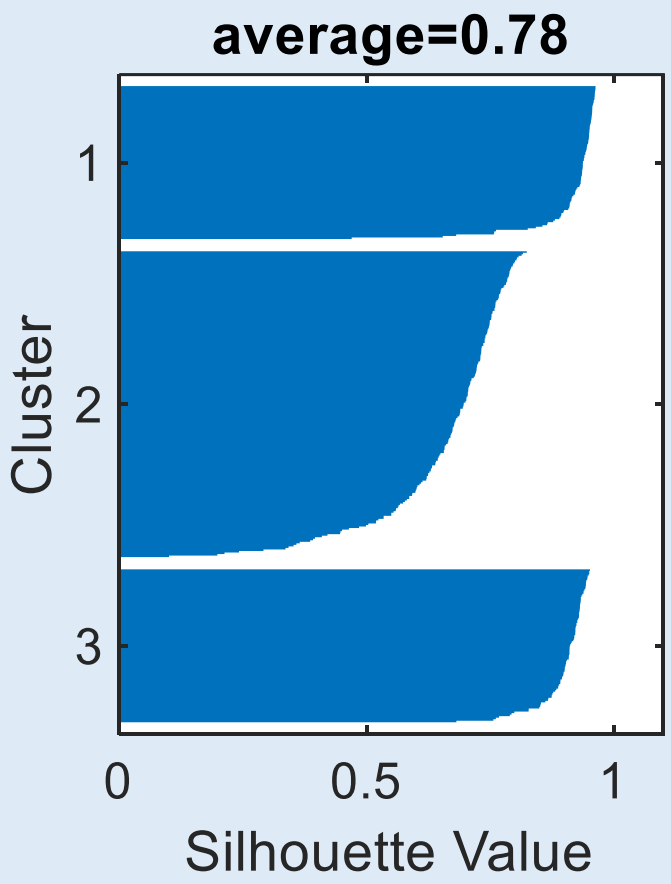


בשלב מסויים תוספת של קבוצה נוספת לא מורידה באופן משמעותי את השונות.

שימו לב: התוצאה לא תמיד חד משמעית, אבל היא נותנת כיוון כללי

איך נקבע את מספר הקבוצות האופטימלי?

3. בעזרת גרף Silhouette



גישה כמותית. בודקים כמה כל דוגמא שייכת לקלאסטר בו היא נמצאת. עבור כל דוגמא:

$$s(i) = \begin{cases} 1 - \frac{a}{b} & a \leq b \\ \frac{b}{a} - 1 & a \geq b \end{cases} \qquad -1 \leq s(i) \leq 1$$

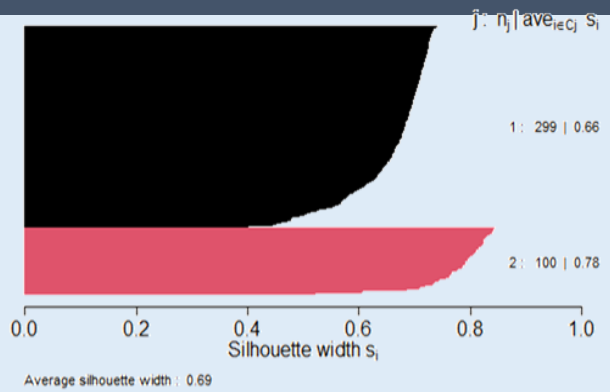
s – ערך ה-silhouette

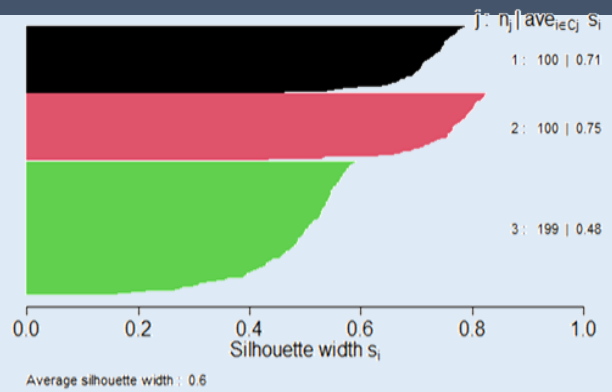
a – **מדד הלכידות:** ממוצע המרחקים בין הדוגמא לדוגמאות האחרות באותו הקלאסטר

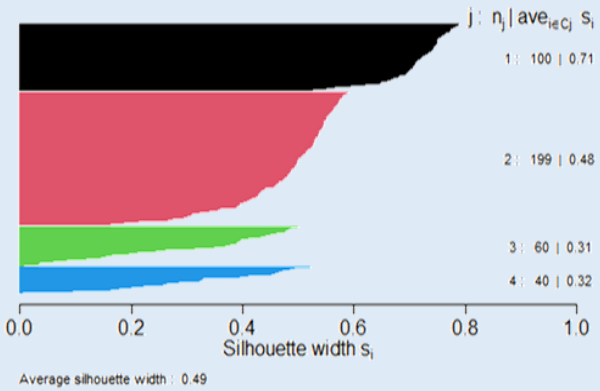
b – **מדד ההפרדה:** ממוצע המרחקים בין הדוגמא לדוגמאות בקלאסטר הכי קרוב אליה

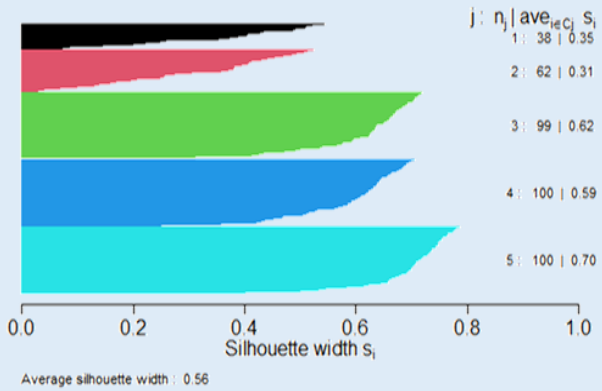
אם מדד ההפרדה קטן ממדד הלכידות, נקבל ערך סילואט שלילי

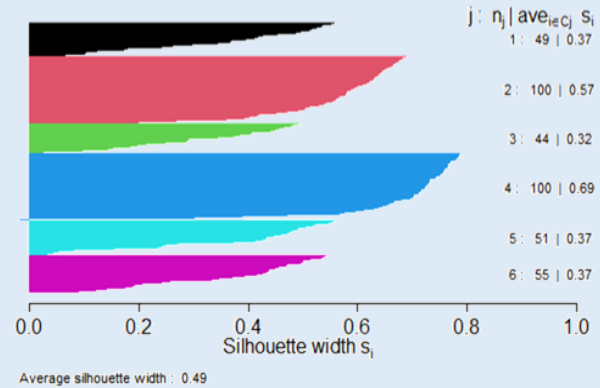
איך נקבע את מספר הקבוצות האופטימלי?











3. בעזרת גרף Silhouette

נתון סט נתונים אותו חילקנו ל-2-6 קבוצות

לכל חלוקה עשינו גרף silhouette

$$s(i) = \begin{cases} 1 - \frac{a}{b} & a \leq b \\ \frac{b}{a} - 1 & b \leq a \end{cases}$$

$$-1 \leq s(i) \leq 1$$

שימו לב: לכל קלאסטר יש ערך s משלו, ויחד יש להם ערך silhouette ממוצע שמצביע על כמה טוב נעשה הקלאסטרינג.

קשה...

לסיכום

עד עכשיו השתמשנו בעיקר בנתונים חד/דו מימדיים. אבל מה קורה ברב מימד?

לא קל לחפש קבוצות ברב-מימד – קשה למצוא איזה מימדים רלוונטיים לנו ואילו סתם מפריעים.

שבוע הבא נלמד על PCA שעוזר לנו למצוא את המימדים הרלוונטיים ביותר לנתונים שלנו.