BigData

**אלגוריתמי אישכול**

**אשכול (cluster)** – אוסף נקודות דומות אחת לשנייה.

**אשכול (clustering)** – בהינתו קובוצת לא מסווגות של נקודות האשכול הוא חלוקת הקבוצה לתתי קבוצות עם נקודות דומות הנמצאות באותו אשכול ונקודות שונות נמצאות באשכולות שונים.

**אלגורתם אשכול** – רצץ פעולות המבצעת את האישכול.

**אלגוריתם Kmeans**

* **Partitional clustering** – חלוקה למספר אשכולות.
* **הלגוריתם מבוסס מרכזים** - עיקר העבודה שלו היא שיוך הנקודות למרכזי האשכולות הקרובים ביותר אליהן.
* שיווך על ידי חישוב דמיון או מרחקים.
* מספר האשכולות נקבע על ידי המשתמש (K).

***תיאור האלגוריתם***

* המשתמש בחור את מספר האשכולות הרצוי K.
* לאחר מכן האלגוריתם בחור באופן אקראי K נקודות מתוך קבוצות הנקודות הנתונה.
* האלגוריתם ישייך את כל הנקודות לפי המרכז הכי קרוב אליהם.
  + תהליך השיוך
  + תשייך את הנקודות לפי המרכז הקרוב ביותר.
  + חשב את המרכזים מחדש של כל אשכול.
* כל עוד המרכזים משתנים האלגוריתם חוזר על תהליך השיוך.

**תהליך השיוך**

* כדי לשייך נקודה למרכז הקרוב ביותר צריך לחשב את המחרקים בין הנקודות לכל המרכזים הקיימים. ואחר מכל למצוא מיהו המרכז הקרוב ביותר ולשייך אליו את הנקודה.
* נשתמש בפונקציית מרחק כדי לחשב את המרחקים המרכזים הקיימים.
* **נוסחת המרחק האוקלידי:**
  + בהינתן שתי נקודות : A=(X1,X2,…,Xn),B=(X1,X2,…,Xn)
  + המרחק האוקלידי בין שתי הנקודות B,A שווה:



<https://www.calculatorsoup.com/calculators/geometry-plane/distance-two-points.php>

**תהליך עדכון המרכזים**

* **מרכז של קבוצות נקודות**: המרכז הוא נקודה אשר סכום המרחקים הריבועי שלה מאשר הנקודות הוא מנימלי.

**דוגמא :** בההנה שנתונים המספרים {1,4,3,7,10} נרצה לחשב את המרכזים שלהם

מחפשים את הערך המינמלי של הפונקציה *f*

כדי למצוא את המינימום יש לחשב את הנגזרת של הפונקציה *f ולהשוות לאפס.*

*מסקנה :*

*המרכז של m נקודות {*P1,P2,…..,Pm} *שווה ל-:*

*דוכמא:*

*בהינתן 3 נקודות ב R2 :*

P1=(1,2), P2=(5,3), P3=(4,1)

[*https://people.revoledu.com/kardi/tutorial/kMean/Online-K-Means-Clustering.html*](https://people.revoledu.com/kardi/tutorial/kMean/Online-K-Means-Clustering.html)

***דומגא להרצת האלגוריתם של k-means***

*בהינתן 5 נקודות :*

|  |  |  |
| --- | --- | --- |
| *Att1* | *Att1* |  |
| *0* | *0* | P1 |
| *1* | *0* | P2 |
| *1* | *1* | P3 |
| *4* | *3* | P4 |
| *4* | *4* | P5 |

*נבחר ב 2=K כלומר חלוקה לשני אשכולות.*

***האלגוריתם:***

1. *בחירת 2 נקודות מתוך כל הנקודות באופן אקראי. נבחר למשל את 2 הנקודות הראשונות.*
2. *שיוך הנקודות למרכזים הקרובים אליהם.*

*בהתחלה יש לחשב את המרחק בין כל הנקודות לשני המרכזים הנתונים.*

*רואים כי הנקודה קרובה למרכז הראשון ולכן הנקודה שייכית לאשכול הראשון.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Dist m2* | *Dist m1* | *Att1* | *Att1* |  |
| *1* | *0* | *0* | *0* | P1 |
| *0* | *1* | *1* | *0* | P2 |
| *1* | *2* | *1* | *1* | P3 |
| *18* | *25* | *4* | *3* | P4 |
| *25* | *32* | *4* | *4* | P5 |

*ניתן לראות כי באשכול הראשון נמצאת רק הנקודה הראשונה ובאשכול השני שאר הנקודות.*

1. *עדכון מכרזים*

*לכל אשכול יש לחשב את המרכז החדש שלו.*

1. *מרכזים השתנו ולכן נחזור ל(2) תהליך השיוך הנקודות למרכזים החדשים.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Dist m2* | *Dist m1* | *Att1* | *Att1* |  |
| *10.25* | *0* | *0* | *0* | P1 |
| *6.25* | *1* | *1* | *0* | P2 |
| *3.25* | *2* | *1* | *1* | P3 |
| *3.25* | *25* | *4* | *3* | P4 |
| *6.25* | *32* | *4* | *4* | P5 |

*נחשב מרכזים מחדש :*

*תהליך השיוך הנקודות למרכזים החדשים:*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Dist m2* | *Dist m1* | *Att2* | *Att1* |  |
| *28.25* | *0.555* | *0* | *0* | P1 |
| *21.25* | *0.222* | *1* | *0* | P2 |
| *15.25* | *0.555* | *1* | *1* | P3 |
| *0.25* | *18.222* | *4* | *3* | P4 |
| *0.25* | *24.555* | *4* | *4* | P5 |

*תוצאת השיוך היא :*

*המרכזים לא ישתנו ולכן האלגוריתם עוצר ותוצאה הסופית היא*

***בידקת איכות האשכול***

*לצרוך בידקת איכות האשכול נשתמש במדד הסוכם את הטעות בריבוע (Sum square error SSE)*

*המדד מגדיר את הטעות כמרחק הנקודה ממרכז האשכול שלה. כלומר ככל שהנקודה קרובה למרכז של האשכול אליו היא שייכת, אז האשכול הוא יותר בטוח בתוצאה שלו.*

*לצרוך הדגמה נחשב את SSE עבור התוצאת האשכול שקיבלנו בדוגמא הקודמת :*

*הערה: ככל שSSE קטן אז השגיאה קטנה ולכן התוצאה היא יותר טובה. מרצים מספר רב של פעמים את האלגוריתם עם מרכזים התחלתים אקראים ובוחרים את המרכזים הטובים ביותר לפי התוצאה הנמוכה ביותר של ה SSE.*

***אשכולות ריקים***

*האלגוריתם K-means יכול לקבל במקרים מסוימים אשכולות ריקים. וזה עבור נקודות התחלה מסוימות.*

*כדי להתגבר על הבעיה הזו יש לעשות את תהליך העיבוד לאחר קבלת התוצאה.*

*נעובר על התוצאות המתקבלות לאחר האשכול ונחשב כמה נקודות נמצאות בכל אשכול.*

*את האשכלות הריקים יש למחוק, ואז קיימות שתי אפשרויות למספר האשכולות הסופי:*

1. *מורידים את כמות האשכולות ונאשרים עם האשכולות המכיל מספר סביר של נקודות.*
2. *לבחור את האשכול הגדול מבין שני האשכולות שנשארו ונפצל אותו לשניים ע"י הרצת האלגוריתם K-means עם 2=K עבור האשכול הגדול.*

***Bisecting k-means clustering algorithm***

*הרעיון המרכזי של הארלגוריתם:*

*האלגוריתם מתבסס על האלגוריתם המקורי של k-means בך שבכל אינטרציה מרצים k-means עם K=2*

*איך האלגוריתם עובד:*

1. *בהתחלה מניחים כי כל הנתונים שייכים לאשכול אחד*
2. *כל עוד שלא קיבלנוK אשכולות:*
   1. *לבחור אחד האשכולות הקיימים בתוך הנתונים שיש לפצל/*
   2. *להריץ את האלגוריתם k-means עם K=2*

*שאלות איך בחורים את האשכול שנרצה לפצל?*

1. *דרך אחת היא לבחור את האשכול בעל הSSE הגדול ביותר.*
2. *ניתן להסתכל כל המספר הנקודות השייכות לכל אשכל ובחור את המספר הגדול ביותר.*
3. *לחשב את השונות/סטיית תקן של מרחקי הנקודות ממרכז האשכול שלהן ונבחר את האשכול בעל סטיית התקן הגדולה ביותר.*

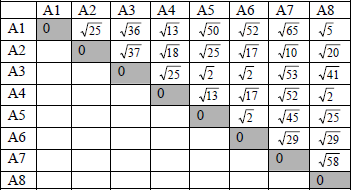
*מה ההבדל העקרי בין השיטות :*

1. *בשיטה זו קיימת פחות השפעה על בחירת המרכזים ההתחלתיים. כי באלגוריתם המקורי בחירת K המרכזים ההתחלתיים מתבצעת בפעם הראשונה וכולם נבחרים פעם אחת. לעומת זאת באלגוריתם החדש בחירת המרכזים נשעשת באופן יותר מושכל.*
2. *האלגוריתם המוצע הוא אלגוריתם היררכי, אשר חלוקת הנותנים לאשכולות מתבצעת באופן היררכי.*

**תרגול - K-means clustering**

תשתמש באלגוריתם האשכול clustering K-means ואת המרחק האוקלידי לאשכול 8 הדגימות הבאות לשלושה אשכולות שונים: 𝐴1(2,10), 𝐴2(2,5), 𝐴3(8,4), 𝐴4(5,8), 𝐴5(7,5), 𝐴6(6,4), 𝐴7(1,2), 𝐴8(4,9)

מטריצת המרחקים המתבססת על מרחק אוקלידי נתונה להלן:



בהנחה שהמרכזים ההתחלתיים הם: 𝐴7 𝐴4, 𝐴1, תפעיל אלגוריתם clustering K-means עבור

איטרציה בודדת בלבד. בסוף האיטרציה הראה:

1. האשכולות החדשים הדגימות השייכות לכל אשכול
2. המרכזים של האשכולות החדשים.
3. שרטט את 8 הנקודות במערכת צירים, ותסמן את האשכולות אחרי האטירציה הראשונה
4. כמה עוד איטרציות צריך בכדי שהאלגוריתם יתכנס.

## הפתרון:

## המרכזים :

מרחקים מהמרכזים:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Dist m3* | *Dist m2* | *Dist m1* | *Att2* | *Att1* |  |
|  |  | *0* | *10* | *2* | A1 |
|  |  |  | *5* | *2* | A2 |
|  |  |  | *4* | *8* | A3 |
|  |  |  | *8* | *5* | A4 |
|  |  |  | *5* | *7* | A5 |
|  |  |  | *4* | *6* | A6 |
|  |  |  | *2* | *1* | A7 |
|  |  |  | *9* | *4* | A8 |

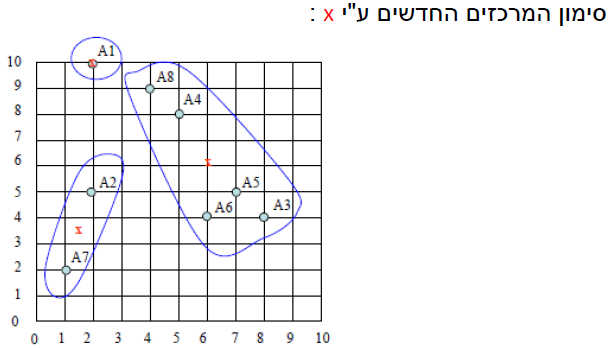
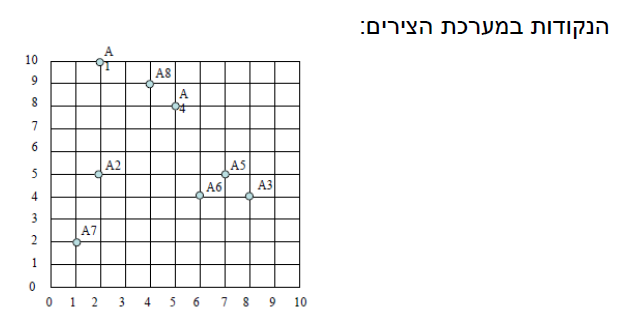
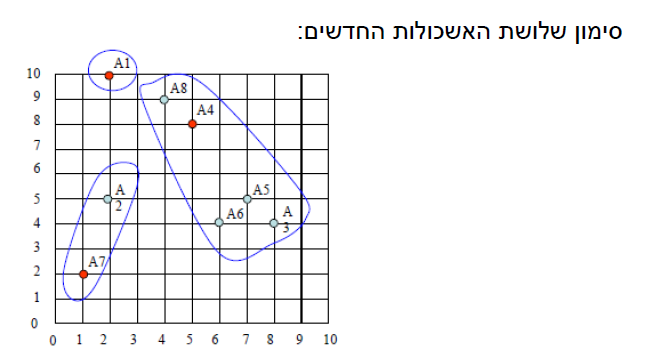
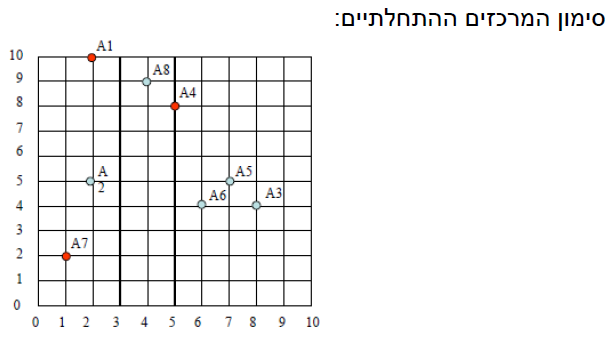
אשכולות:

**סימון המרכזים החדשים ע"י x**

שלושת האשכולות החדשים

סימון המרכזים ההתחלתיי

הנקודות במערכת הצירים



ד.

אנחנו נצטרך לעוד שתי איטרציות נוספות עד שהאלגוריתם שלנו יתכנס. אפשר לראות בלי חישוב שבעוד איטרציה הנקודה 𝐴8 תתווסף לאשכול הראשון. ובעוד איטרציה הנקודה 𝐴4 תתווסף גם לאשכול הראשון. ואז נקבל סופית אשכולות שהמרכזים בהם מתכנסים. תוכלו לעשות את החישוב המתמטי בעצמכם ואז תקבלו:

.1 אחרי איטרציה שנייה:

𝑪𝒍𝒖𝒔𝒕𝒆𝒓𝟏: {𝐴1, 𝐴8}, 𝑪𝒍𝒖𝒔𝒕𝒆𝒓𝟐: {𝐴3, 𝐴4, 𝐴5, 𝐴6}, 𝑪𝒍𝒖𝒔𝒕𝒆𝒓𝟑: {𝐴2, 𝐴7}

עם המרכזים:

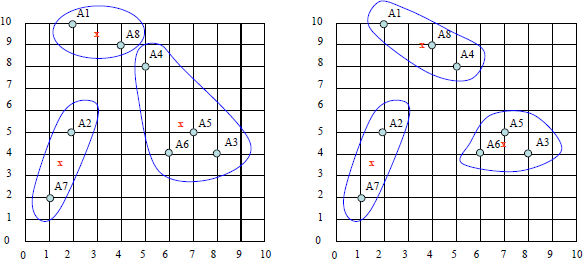
𝐶1 = (3, 9.5), 𝐶2 = (6.5, 5.25), 𝐶3 = (1.5, 3.5).

.2 אחרי איטרציה שלישית :

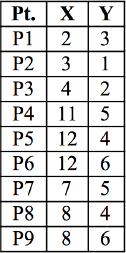
𝑪𝒍𝒖𝒔𝒕𝒆𝒓𝟏: {𝐴1, 𝐴4, 𝐴8}, 𝑪𝒍𝒖𝒔𝒕𝒆𝒓 𝟐: {𝐴3, 𝐴5, 𝐴6}, 𝑪𝒍𝒖𝒔𝒕𝒆𝒓𝟑: {𝐴2, 𝐴7}

עם המרכזים:

𝐶1 = (3.66, 9), 𝐶2 = (7, 4.33), 𝐶3 = (1.5, 3.5).



בהינתן טבלת הנתונים הבאה :



א. ממש את אלגוריתם האשכול k-means עבור המרכזים ההתחלתיים P2 ו' .P8

**ב.** חשב את סכום ריבועי השגיאות SSE של האשכול שקבלת בסעיף א.'

**ג.** בחר שני מרכזים התחלתיים כך שתקבל שני אשכולות שונים ממה שקבלת בסעיף א.'

**ד.** חשב את סכום ריבועי השגיאות SSE של האשכול שקיבלת בסעיף ג.'

**ה.** איזה אשכול טוב יותר ? נמק!

## הפתרון:

## המרכזים :

מרחקים מהמרכזים:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Dist m2* | *Dist m1* | *Att2* | *Att1* |  |
|  |  | *3* | *2* | P1 |
|  |  | *1* | *3* | P2 |
|  |  | *2* | *4* | P3 |
|  |  | *5* | *11* | P4 |
|  |  | *4* | *12* | P5 |
|  |  | *6* | *12* | P6 |
|  |  | *5* | *7* | P7 |
|  |  | *4* | *8* | P8 |
|  |  | *6* | *8* | P9 |

אשכולות:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Dist m2* | *Dist m1* | *Att2* | *Att1* |  |
|  |  | *3* | *2* | P1 |
|  |  | *1* | *3* | P2 |
|  |  | *2* | *4* | P3 |
|  |  | *5* | *11* | P4 |
|  |  | *4* | *12* | P5 |
|  |  | *6* | *12* | P6 |
|  |  | *5* | *7* | P7 |
|  |  | *4* | *8* | P8 |
|  |  | *6* | *8* | P9 |

*ב.*

## ג. המרכזים :

מרחקים מהמרכזים:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Dist m2* | *Dist m1* | *Att2* | *Att1* |  |
| *6.70* | *5.385* | *3* | *2* | P1 |
| *7.07* | *5.656* | *1* | *3* | P2 |
| *5.65* | *4.242* | *2* | *4* | P3 |
| *3.16* | *4* | *5* | *11* | P4 |
| *4.472* | *5.1* | *4* | *12* | P5 |
| *4* | *5.1* | *6* | *12* | P6 |
| *1.414* | *0* | *5* | *7* | P7 |
| ***2*** | *1.414* | *4* | *8* | P8 |
| *0* | *1.414* | *6* | *8* | P9 |

אשכולות:

מרחקים מהמרכזים:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Dist m2* | *Dist m1* | *Att2* | *Att1* |  |
| *9.871* | *3.333* | *3* | *2* | P1 |
| *9.545* | *3.073* | *1* | *3* | P2 |
| *8.232* | *1.666* | *2* | *4* | P3 |
| *0.666* | *6.01* | *5* | *11* | P4 |
| *1.05* | *6.74* | *4* | *12* | P5 |
| *1.05* | *7.31* | *6* | *12* | P6 |
| *4.66* | *2.60* | *5* | *7* | P7 |
| *3.8* | *2.848* | *4* | *8* | P8 |
| *3.8* | *3.655* | *6* | *8* | P9 |

אשכולות:

1. *האישכול הראשון יותר טוב הSSE שלו נמוך יותר כמות השגיאות בריבוע קטן יותר זאת אומר שהאשכולות יותר צפופים והאישכול יותר טוב.*

**אלגוריתם אישכול היררכי**

***אשכול היררכי –*** *כאשר האשכולות מתקבלים עם יחס היררכי בין האשכולות השונים. האשכולות מתקבלים מפיצול/קיבוץ אשכולות קייימים כך שניתן לראות את ההיררכיה בין האשכולות השונים.*

***קיימות שתי שיטות מרכזיות לאשכול היררכי :***

1. ***פיצול:*** *בהיתנן קבוצה של נתונים לא מסווגים, מתחילים מההנחה שכל הנקודות נמצאות באשכול אחד, ולאחר מכן באפון איטרטיבי מתחילים לפצל כל פעם את אחד האשכולות.*
2. ***קיבוץ:*** *הינתן קבוצה של נתונים לא מסווגים, מתחילים מההנחה של נקודה מהווה אשכול בפני עצמו, ולאחר מכן כל פעם בוחרים שני אשכולות ומקבצים אותם כך שבכלל איטרציה מתקבל אשכול פחות ממה שקיים.*

***צורת ההצגה של האשכול ההיררכי:***

*מכיוון שניתן לשמור את ההיסטוריה של פיצול/איחוד האשכולות אז ניתן יהיה להציג את התוצאה בעזרת עץ היררכיה. לעץ הזה קוראים Dendrogram.*

*במסגרת הקורס נתמקד בשיטת הקיבוץ. מתחילים ממספר אשכולות כמספר הנקודות ונמשיך לאחד כל פעם ונמשיך לאחד כל פעם שני אשכולות שמקבלים עד שנקבל את המספר האשכולות הרצוי או שמגיעים לאשכול אחד המכיל את כל הנקודות*

***תיאור האלגוריתם:***

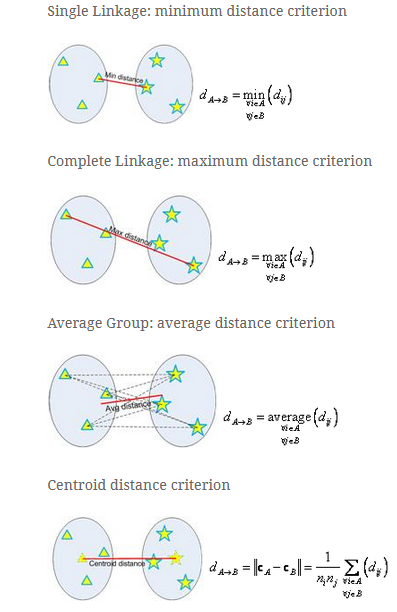
1. *נניח כי כל נקווה מהווה אשכול בפני עצמה.*
2. *מחשבים את המרחקים בין כל שני אשכולות שונים – ומסדרים אותם בטבלת מרחקים.*
3. *כל עוד שלא קבילנו את המספר האשכולות הרצוי או לא הגענו לאשאכול אחד*
4. *תיאור האלגוריתם:*
   1. *נאחד את שני האשכולות הקרובים ביותר אחד לשני.*
   2. *מעדכינים את טבלת המרחקים בין האשכולות שנאשרו.*

***איך מחשבים מרחק בין שני אשכולות ?*** *(שונה מחישוב מחרקים בין שתי נקודות)*

* *ההבל הוא שאשכול יכול להכיל מספר נקודות, ואז המרחק לא מוגדר באופן טריוויאלי כמו המרחק בין שתי נקודות.*

***מרחק בין שתי נקודות*** *- הוא פונקציה d(A,B)פונקציה d מחשבת את המרחק בין שתי הנקודות A וB צריכה לקיים 4 תנאים:*

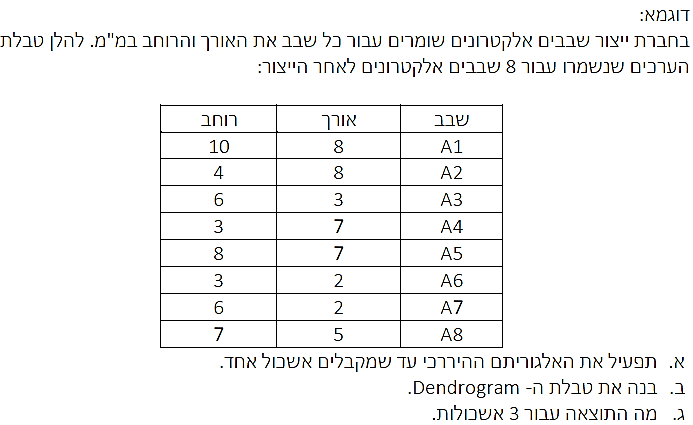
1. *עובר כל שתי נקודות מתקיים כי 0 < d(A,B)*
2. *המרחק מנקודה לעצמה שווה ל-0*
3. *סמטריה, המרחק בין A לB- שווה למרחק בין Bל-A*
4. *עבור כל 3 נקודות A,B,C מתקיים :* 𝑑(𝐴, 𝐵) + 𝑑(𝐵, 𝐶) ≥ 𝑑(𝐴, 𝐶)  *- אי שיוויון המשולש*



***מרחק בין שתי אשכולות*** *- כדי להגדיר את המרחק בין שתי קובצות של נקודות קיימות 4 שטיות מרכזיות :*

1. ***שיטת המינימום*** *- מרחק בין שני אשכולות מוגדר כמרחק בין שתי נקודות קרובות ביותר שנמצאות באשכולות שונים.*
2. ***שיטת המקסימום*** *- מרחק בין שני אשכולות מוגדר כמרחק בין שתי נקודות הרחוקות ביותר שנמצאות באשכולות שונים.*
3. ***שיטת ממוצע מרחקים*** *- המרחק בין שני אשכולות מחשוב ע"י ממוצע המרחקים בין כל שתי נקודות מאשכולות שונים.*
4. ***שיטת המרכזים*** *- מרחק בין שני אשכולות מוגדר כמרחק בין מרכז האשכול הראשון למרכז האשכול השני.*

***דוגמא שיטת המינימום***



1. *בשלב הראשון מנילים כי כל נקודה מהווה אשכול. )כלומר קיימים 8 אשכולות)*

*חישוב כל המרחקים בין כל האשכולות :*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
| A1 | 0 | 6 | 6.4031 | 7.0711 | 2.2361 | 9.2195 | 7.2111 | 4.2426 |
| A2 | 6 | 0 | 5.3852 | 1.4142 | 4.1231 | 6.0828 | 6.3246 | 4.2426 |
| A3 | 6.4031 | 5.3852 | 0 | 5 | 4.4721 | 3.1623 | 1 | 2.2361 |
| A4 | 7.0711 | 1.4142 | 5 | 0 | 5 | 5 | 5.831 | 4.4721 |
| A5 | 2.2361 | 4.1231 | 4.4721 | 5 | 0 | 7.0711 | 5.3852 | 2.2361 |
| A6 | 9.2195 | 6.0828 | 3.1623 | 5 | 7.0711 | 0 | 3 | 5 |
| A7 | 7.2111 | 6.3246 | 1 | 5.831 | 5.3852 | 3 | 0 | 3.1623 |
| A8 | 4.2426 | 4.2426 | 2.2361 | 4.4721 | 2.2361 | 5 | 3.1623 | 0 |

*עכשיו נתחיל באיחוד האשכולות השונים לפי הקרבה ביניהם, נחשב את מרחק בין האשכולות בשיטת המינימום*

*המרחק הקטן ביותר ביותר הוא בין 3A ל7A (1) לכן נאחד ביניהם.*

*עדכון טבלאת המרחקים :*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A2 | A4 | A5 | A6 | A7UA3 | A8 |
| A1 | 0 | 6 | 7.0711 | 2.2361 | 9.2195 | 6.4031 | 4.2426 |
| A2 | 6 | 0 | 1.4142 | 4.1231 | 6.0828 | 5.3852 | 4.2426 |
| A4 | 7.0711 | 1.4142 | 0 | 5 | 5 | 5 | 4.4721 |
| A5 | 2.2361 | 4.1231 | 5 | 0 | 7.0711 | 5.3852 | 2.2361 |
| A6 | 9.2195 | 6.0828 | 5 | 7.0711 | 0 | 3 | 5 |
| A7UA3 | 6.4031 | 5.3852 | 5 | 4.4721 | 3 | 0 | 2.2361 |
| A8 | 4.2426 | 4.2426 | 4.4721 | 2.2361 | 5 | 2.2361 | 0 |

*עדיין אין אשכול אחד לכן נמשיך באיחוד לפי הבטלאה המעודכנת*

*המרחק הקטן ביותר ביותר הוא בין 4A ל2A (*1.4142*) לכן נאחד ביניהם.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A4UA2 | A5 | A6 | A7UA3 | A8 |
| A1 | 0 | 6 | 2.2361 | 9.2195 | 6.4031 | 4.2426 |
| A2UA4 | 6 | 0 | 4.1231 | 6.0828 | 5.3852 | 4.2426 |
| A5 | 2.2361 | 4.1231 | 0 | 7.0711 | 5.3852 | 2.2361 |
| A6 | 9.2195 | 5 | 7.0711 | 0 | 3 | 5 |
| A7UA3 | 6.4031 | 5 | 4.4721 | 3 | 0 | 2.2361 |
| A8 | 4.2426 | 4.4721 | 2.2361 | 5 | 2.2361 | 0 |

*המרחק הקטן ביותר ביותר הוא בין 1A ל5A או 8A ל5A או* A7UA3 *ל8A (*2.2361*) לכן נבחר אחד לאחד ביניהם .*

*נאחד את A1 עם A5*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A1UA5 | A4UA2 | A6 | A7UA3 | A8 |
| A1UA5 | 0 | 4.1231 | 7.0711 | 4.4721 | 2.2361 |
| A2UA4 | 4.1231 | 0 | 6.0828 | 5.3852 | 4.2426 |
| A6 | 7.0711 | 5 | 0 | 3 | 5 |
| A7UA3 | 4.4721 | 5 | 3 | 0 | 2.2361 |
| A8 | 2.2361 | 4.4721 | 5 | 2.2361 | 0 |

*נאחד את* A1UA5 *עם A8*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A1UA5UA8 | A4UA2 | A6 | A7UA3 |
| A1UA5UA8 | 0 | 4.1231 | 5 | 2.2361 |
| A2UA4 | 4.1231 | 0 | 6.0828 | 5.3852 |
| A6 | 5 | 5 | 0 | 3 |
| A7UA3 | 2.2361 | 5 | 3 | 0 |

*נאחד את* A7UA3 *עם* A1UA5UA8

|  |  |  |  |
| --- | --- | --- | --- |
|  | A1UA5UA8UA7UA3 | A4UA2 | A6 |
| A1UA5UA8UA7UA3 | 0 | 4.1231 | 3 |
| A2UA4 | 4.1231 | 0 | 6.0828 |
| A6 | 3 | 5 | 0 |

*נאחד את* A1UA5UA8UA7UA3 *עם* A6

|  |  |  |
| --- | --- | --- |
|  | A1UA5UA8UA7UA3UA6 | A4UA2 |
| A1UA5UA8UA7UA3UA6 | 0 | 4.1231 |
| A2UA4 | 4.1231 | 0 |

*נאחד את* A1UA5UA8UA7UA3UA6 *עם*  A4UA2

*נקבל אשכול אחד :* 𝐴1 ∪ 𝐴5 ∪ 𝐴8 ∪ 𝐴7 ∪ 𝐴3 ∪ 𝐴6 ∪ 𝐴4 ∪ 𝐴2

1. *ציור הdendrogram בהתחלה יש לראות את הסדר האיחוד של האשכולות השונים*

1. 𝐴7 ∪ 𝐴3

2. 𝐴4 ∪ 𝐴2

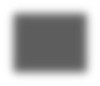
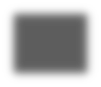
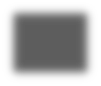
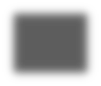
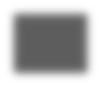
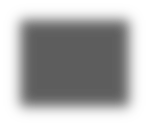
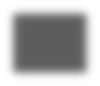
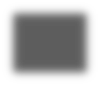
3. 𝐴1 ∪ 𝐴5

4. 𝐴1 ∪ 𝐴5 ∪ 𝐴8

5. 𝐴1 ∪ 𝐴5 ∪ 𝐴8 ∪ 𝐴7 ∪ 𝐴3

6. 𝐴1 ∪ 𝐴5 ∪ 𝐴8 ∪ 𝐴7 ∪ 𝐴3 ∪ 𝐴6

*7.* 𝐴1 ∪ 𝐴5 ∪ 𝐴8 ∪ 𝐴7 ∪ 𝐴3 ∪ 𝐴6 ∪ 𝐴4 ∪ 𝐴2



A2

A4

A6

A3

A7

A8

A5

A1

3=K

1. התוצאה עבור 3 אשכולות:

𝐶1 = {𝐴1, 𝐴5, 𝐴8, 𝐴7, 𝐴3}, 𝐶2 = {𝐴6}, 𝐶3 = {𝐴4, 𝐴2}

אותה דוגמה עם שיטת **המקיסימום:**

מחלקים לפי המרחק הקטן ביותר ובחורים את המרחק הגדול

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
| A1 | 0 | 6 | 6.4031 | 7.0711 | 2.2361 | 9.2195 | 7.2111 | 4.2426 |
| A2 | 6 | 0 | 5.3852 | 1.4142 | 4.1231 | 6.0828 | 6.3246 | 4.2426 |
| A3 | 6.4031 | 5.3852 | 0 | 5 | 4.4721 | 3.1623 | 1 | 2.2361 |
| A4 | 7.0711 | 1.4142 | 5 | 0 | 5 | 5 | 5.831 | 4.4721 |
| A5 | 2.2361 | 4.1231 | 4.4721 | 5 | 0 | 7.0711 | 5.3852 | 2.2361 |
| A6 | 9.2195 | 6.0828 | 3.1623 | 5 | 7.0711 | 0 | 3 | 5 |
| A7 | 7.2111 | 6.3246 | 1 | 5.831 | 5.3852 | 3 | 0 | 3.1623 |
| A8 | 4.2426 | 4.2426 | 2.2361 | 4.4721 | 2.2361 | 5 | 3.1623 | 0 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A2 | A4 | A5 | A6 | A7UA3 | A8 |
| A1 | 0 | 6 | 7.0711 | 2.2361 | 9.2195 | 7.2111 | 4.2426 |
| A2 | 6 | 0 | 1.4142 | 4.1231 | 6.0828 | 6.3246 | 4.2426 |
| A4 | 7.0711 | 1.4142 | 0 | 5 | 5 | 5.831 | 4.4721 |
| A5 | 2.2361 | 4.1231 | 5 | 0 | 7.0711 | 5.3852 | 2.2361 |
| A6 | 9.2195 | 6.0828 | 5 | 7.0711 | 0 | 3.1623 | 5 |
| A7UA3 | 7.2111 | 6.3246 | 5.831 | 5.3852 | 3.1623 | 0 | 3.1623 |
| A8 | 4.2426 | 4.2426 | 4.4721 | 2.2361 | 5 | 3.1623 | 0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A2UA4 | A5 | A6 | A7UA3 | A8 |
| A1 | 0 | 7.0711 | 2.2361 | 9.2195 | 7.2111 | 4.2426 |
| A2U A4 | 7.0711 | 0 | 5 | 6.0828 | 6.3246 | 4.4721 |
| A5 | 2.2361 | 5 | 0 | 7.0711 | 5.3852 | 2.2361 |
| A6 | 9.2195 | 6.0828 | 7.0711 | 0 | 3.1623 | 5 |
| A7UA3 | 7.2111 | 6.3246 | 5.3852 | 3.1623 | 0 | 3.1623 |
| A8 | 4.2426 | 4.4721 | 2.2361 | 5 | 3.1623 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A1 | A2UA4 | A5UA8 | A6 | A7UA3 |
| A1 | 0 | 7.0711 | 4.2426 | 9.2195 | 7.2111 |
| A2U A4 | 7.0711 | 0 | 5 | 6.0828 | 6.3246 |
| A5UA8 | 4.2426 | 5 | 0 | 7.0711 | 5.3852 |
| A6 | 9.2195 | 6.0828 | 7.0711 | 0 | 3.1623 |
| A7UA3 | 7.2111 | 6.3246 | 5.3852 | 3.1623 | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A1 | A2UA4 | A5UA8 | A6UA7UA3 |
| A1 | 0 | 7.0711 | 4.2426 | 9.2195 |
| A2U A4 | 7.0711 | 0 | 5 | 6.3246 |
| A5UA8 | 4.2426 | 5 | 0 | 7.0711 |
| A6UA7UA3 | 9.2195 | 6.3246 | 7.0711 | 0 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | A2UA4 | A5UA8UA1 | A6UA7UA3 |
| A2UA4 | 0 | 7.0711 | 6.3246 |
| A5UA8UA1 | 7.0711 | 0 | 9.2195 |
| A6UA7UA3 | 6.3246 | 9.2195 | 0 |

|  |  |  |
| --- | --- | --- |
|  | A2UA4A6UA7UA3 | A5UA8UA1 |
| A2UA4A6UA7UA3 | 0 | 7.0711 |
| A5UA8UA1 | 7.0711 | 0 |

אישכול סופי : 𝐴2 ∪ 𝐴4 ∪ 𝐴6 ∪ 𝐴7 ∪ 𝐴3 ∪ 𝐴5 ∪ 𝐴8 ∪ 𝐴1

𝐴7 ∪ 𝐴3

𝐴2 ∪ 𝐴4

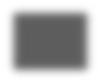
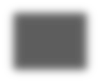
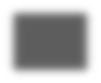
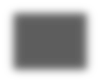
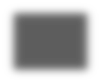
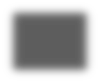
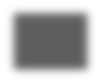
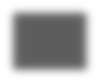
𝐴5 ∪ 𝐴8

𝐴6 ∪ 𝐴7 ∪ 𝐴3

𝐴5 ∪ 𝐴8 ∪ 𝐴1

𝐴2 ∪ 𝐴4 ∪ 𝐴6 ∪ 𝐴7 ∪ 𝐴3

𝐴2 ∪ 𝐴4 ∪ 𝐴6 ∪ 𝐴7 ∪ 𝐴3 ∪ 𝐴5 ∪ 𝐴8 ∪ 𝐴1



A1

A8

A5

A3

A7

A4

A23

A6

3=K

𝐶1 = {𝐴3, 𝐴7, 𝐴6}, 𝐶2 = {𝐴2, 𝐴4}, 𝐶3 = {𝐴1, 𝐴5, 𝐴8}

אותה דוגמה עם שיטת ***ממוצע המרחקים***:

מחלקים לפי המרחק הקטן ביותר ועושים ממוצע בין המרחקים

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
| A1 | 0 | 6 | 6.4031 | 7.0711 | 2.2361 | 9.2195 | 7.2111 | 4.2426 |
| A2 | 6 | 0 | 5.3852 | 1.4142 | 4.1231 | 6.0828 | 6.3246 | 4.2426 |
| A3 | 6.4031 | 5.3852 | 0 | 5 | 4.4721 | 3.1623 | 1 | 2.2361 |
| A4 | 7.0711 | 1.4142 | 5 | 0 | 5 | 5 | 5.831 | 4.4721 |
| A5 | 2.2361 | 4.1231 | 4.4721 | 5 | 0 | 7.0711 | 5.3852 | 2.2361 |
| A6 | 9.2195 | 6.0828 | 3.1623 | 5 | 7.0711 | 0 | 3 | 5 |
| A7 | 7.2111 | 6.3246 | 1 | 5.831 | 5.3852 | 3 | 0 | 3.1623 |
| A8 | 4.2426 | 4.2426 | 2.2361 | 4.4721 | 2.2361 | 5 | 3.1623 | 0 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A2 | A3UA7 | A4 | A5 | A6 | A8 |
| A1 | 0 | 6 | 6.8071 | 7.0711 | 2.2361 | 9.2195 | 4.2426 |
| A2 | 6 | 0 | 5.8549 | 1.4142 | 4.1231 | 6.0828 | 4.2426 |
| A3UA7 | 6.8071 | 5.8549 | 0 | 5.4155 | 4.92865 | 3.08115 | 2.6992 |
| A4 | 7.0711 | 1.4142 | 5.4155 | 0 | 5 | 5 | 4.4721 |
| A5 | 2.2361 | 4.1231 | 4.92865 | 5 | 0 | 7.0711 | 2.2361 |
| A6 | 9.2195 | 6.0828 | 3.08115 | 5 | 7.0711 | 0 | 5 |
| A8 | 4.2426 | 4.2426 | 2.6992 | 4.4721 | 2.2361 | 5 | 0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A2UA4 | A3UA7 | A5 | A6 | A8 |
| A1 | 0 | 6.53555 | 6.8071 | 2.2361 | 9.2195 | 4.2426 |
| A2UA4 | 6.53555 | 0 | 5.6352 | 4.56155 | 5.5414 | 4.35735 |
| A3UA7 | 6.4031 | 5.6352 | 0 | 4.4721 | 3.1623 | 2.2361 |
| A5 | 2.2361 | 4.56155 | 4.92865 | 0 | 7.0711 | 2.2361 |
| A6 | 9.2195 | 5.5414 | 3.08115 | 7.0711 | 0 | 5 |
| A8 | 4.2426 | 4.35735 | 2.6992 | 2.2361 | 5 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A1 | A2UA4 | A3UA7 | A5UA8 | A6 |
| A1 | 0 | 6.53555 | 6.8071 | 3.23935 | 9.2195 |
| A2UA4 | 6.53555 | 0 | 5.6352 | 4.45945 | 5.5414 |
| A3UA7 | 6.8071 | 5.6352 | 0 | 3.81393 | 3.08115 |
| A5UA8 | 3.23935 | 4.45945 | 3.81393 | 0 | 6.03555 |
| A6 | 9.2195 | 5.5414 | 3.08115 | 6.03555 | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A1 | A2UA4 | A3UA7UA6 | A5UA8 |
| A1 | 0 | 6.53555 | 7.61123 | 3.23935 |
| A2UA4 | 6.53555 | 0 | 5.60393 | 4.45945 |
| A3UA7UA6 | 7.61123 | 5.60393 | 0 | 4.55447 |
| A5UA8 | 3.23935 | 4.45945 | 4.55447 | 0 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | A5UA8U A1 | A2UA4 | A3UA7UA6 |
| A5UA8UA1 | 0 | 5.15148 | 5.57339 |
| A2UA4 | 5.15148 | 0 | 5.60393 |
| A3UA7UA6 | 5.57339 | 5.60393 | 0 |

|  |  |  |
| --- | --- | --- |
|  | A5UA8UA1UA2UA4 | A3UA7UA6 |
| A5UA8UA1UA2UA4 | 0 | 5.57339 |
| A3UA7UA6 | 5.57339 | 0 |

אישכול סופי : 𝐴5 ∪ 𝐴8 ∪ 𝐴1 ∪ 𝐴2 ∪ 𝐴4 ∪ 𝐴3 ∪ 𝐴7 ∪ 𝐴6

𝐴3 ∪ 𝐴7

𝐴2 ∪ 𝐴4

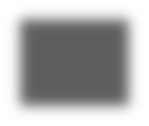
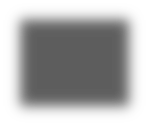
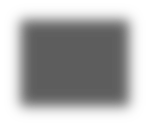
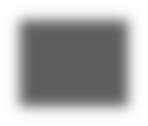
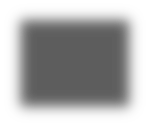
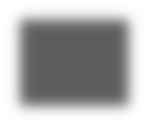
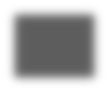
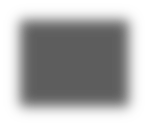
𝐴5 ∪ 𝐴8

𝐴3 ∪ 𝐴7 ∪ 𝐴6

𝐴5 ∪ 𝐴8 ∪ 𝐴1

𝐴5 ∪ 𝐴8 ∪ 𝐴1 ∪ 𝐴2 ∪ 𝐴4

𝐴5 ∪ 𝐴8 ∪ 𝐴1 ∪ 𝐴2 ∪ 𝐴4 ∪ 𝐴3 ∪ 𝐴7 ∪ 𝐴6



A6

A7

A3

A4

A2

A1

A8

A5

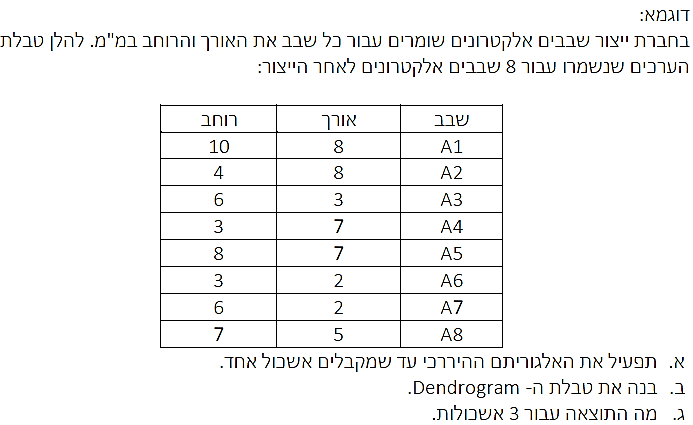
3=K

ג.

𝐶1 = {𝐴5, 𝐴8, 𝐴1}, 𝐶2 = {𝐴2, 𝐴4}, 𝐶3 = {𝐴3, 𝐴7, 𝐴6}

אותה דוגמה עם שיטת ***המרכזים***:

מחלקים לפי המרחק הקטן ביותר מחשבים לפי המרכזים



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
| A1 | 0 | 6 | 6.4031 | 7.0711 | 2.2361 | 9.2195 | 7.2111 | 4.2426 |
| A2 | 6 | 0 | 5.3852 | 1.4142 | 4.1231 | 6.0828 | 6.3246 | 4.2426 |
| A3 | 6.4031 | 5.3852 | 0 | 5 | 4.4721 | 3.1623 | 1 | 2.2361 |
| A4 | 7.0711 | 1.4142 | 5 | 0 | 5 | 5 | 5.831 | 4.4721 |
| A5 | 2.2361 | 4.1231 | 4.4721 | 5 | 0 | 7.0711 | 5.3852 | 2.2361 |
| A6 | 9.2195 | 6.0828 | 3.1623 | 5 | 7.0711 | 0 | 3 | 5 |
| A7 | 7.2111 | 6.3246 | 1 | 5.831 | 5.3852 | 3 | 0 | 3.1623 |
| A8 | 4.2426 | 4.2426 | 2.2361 | 4.4721 | 2.2361 | 5 | 3.1623 | 0 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A2 | A3UA7 | A4 | A5 | A6 | A8 |
| A1 | 0 | 6 | 5.852 | 7.0711 | 2.2361 | 9.2195 | 4.2426 |
| A2 | 6 | 0 | 2.061 | 1.4142 | 4.1231 | 6.0828 | 4.2426 |
| A3UA7 | 5.852 | 2.061 | 0 | 5.408 | 4.924 | 3.041 | 2.692 |
| A4 | 7.0711 | 1.4142 | 5.408 | 0 | 5 | 5 | 4.4721 |
| A5 | 2.2361 | 4.1231 | 4.924 | 5 | 0 | 7.0711 | 2.2361 |
| A6 | 9.2195 | 6.0828 | 3.041 | 5 | 7.0711 | 0 | 5 |
| A8 | 4.2426 | 4.2426 | 2.692 | 4.4721 | 2.2361 | 5 | 0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A2UA4 | A3UA7 | A5 | A6 | A8 |
| A1 | 0 | 6.519 | 5.852 | 2.2361 | 4.301 | 4.2426 |
| A2UA4 | 6.519 | 0 | 5.590 | 4.527 | 5.522 | 4.2426 |
| A3UA7 | 5.852 | 5.590 | 0 | 4.924 | 3.041 | 2.692 |
| A5 | 2.2361 | 4.527 | 4.924 | 0 | 7.0711 | 2.2361 |
| A6 | 9.2195 | 5.522 | 3.041 | 7.0711 | 0 | 5 |
| A8 | 4.2426 | 4.301 | 2.692 | 4.4721 | 5 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A1UA5 | A2UA4 | A3UA7 | A6 | A8 |
| A1UA5 | 0 | 5.5 | 5.830 | 6.264 | 3.201 |
| A2UA4 | 5.5 | 0 | 5.590 | 5.522 | 4.2426 |
| A3UA7 | 5.830 | 5.590 | 0 | 3.041 | 2.692 |
| A6 | 6.264 | 5.522 | 3.041 | 0 | 5 |
| A8 | 3.201 | 4.301 | 2.692 | 5 | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A1UA5 | A2UA4 | A3UA7UA8 | A6 |
| A1UA5 | 0 | 5.5 | 4.946 | 6.264 |
| A2UA4 | 5.5 | 0 | 5.038 | 5.522 |
| A3UA7UA8 | 4.946 | 5.038 | 0 | 3.590 |
| A6 | 6.264 | 5.522 | 3.590 | 0 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | A1UA5 | A2UA4 | A3UA7UA8UA6 |
| A1UA5 | 0 | 5.5 | 5.700 |
| A2UA4 | 5.5 | 0 | 4.924 |
| A3UA7UA8UA6 | 5.700 | 4.924 | 0 |

|  |  |  |
| --- | --- | --- |
|  | A1UA5 | A2UA4UA3UA7UA8UA6 |
| A1UA5 | 0 | 5.5 |
| A2UA4UA3UA7UA8UA6 | 5.5 | 0 |

אישכול סופי : 𝐴2 ∪ 𝐴4 ∪ 𝐴3 ∪ 𝐴7 ∪ 𝐴8 ∪ 𝐴6 ∪ 𝐴1 ∪ 𝐴5

𝐴3 ∪ 𝐴7

𝐴2 ∪ 𝐴4

𝐴1 ∪ 𝐴5

𝐴3 ∪ 𝐴7 ∪ 𝐴8

𝐴3 ∪ 𝐴7 ∪ 𝐴8 ∪ 𝐴6

3=K

𝐴3 ∪ 𝐴7 ∪ 𝐴8 ∪ 𝐴6 ∪ 𝐴2 ∪ 𝐴4



A5

A1

A4

A2

A6

A8

A7

A3

A3 ∪ 𝐴7 ∪ 𝐴8 ∪ 𝐴6 ∪ 𝐴2 ∪ 𝐴4 ∪ 𝐴1 ∪ 𝐴5

𝐶1 = {𝐴3, 𝐴7, 𝐴8, 𝐴6}, 𝐶2 = {𝐴2, 𝐴4}, 𝐶3 = {𝐴1, 𝐴5,}

סיבוכיות מקום וזמן

* סיבוכיות המקום היא 2Nכאשר N מהווה את מספר הנתונים.
* סיבוכיות זמן
  + N אינטרציות כל פעם יורדים בכמות האשכולות
  + ובכל אינטרציה מחשבים את מטריצת המרחק 2N
  + מכאן נובע כי סיבוכיות הזמן שווה ל- 3N

היתרונות וחסרונות

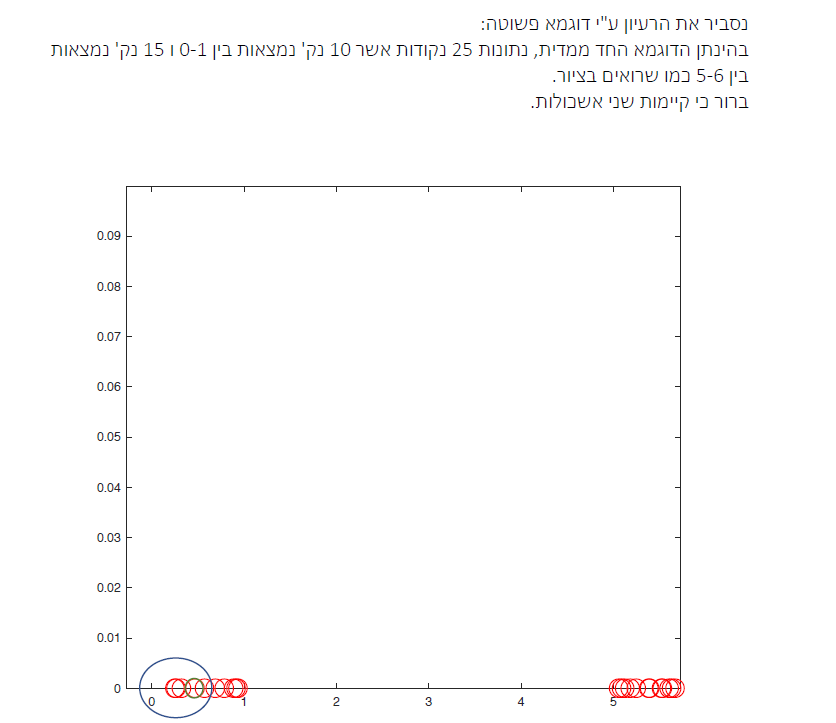
היתרונות :

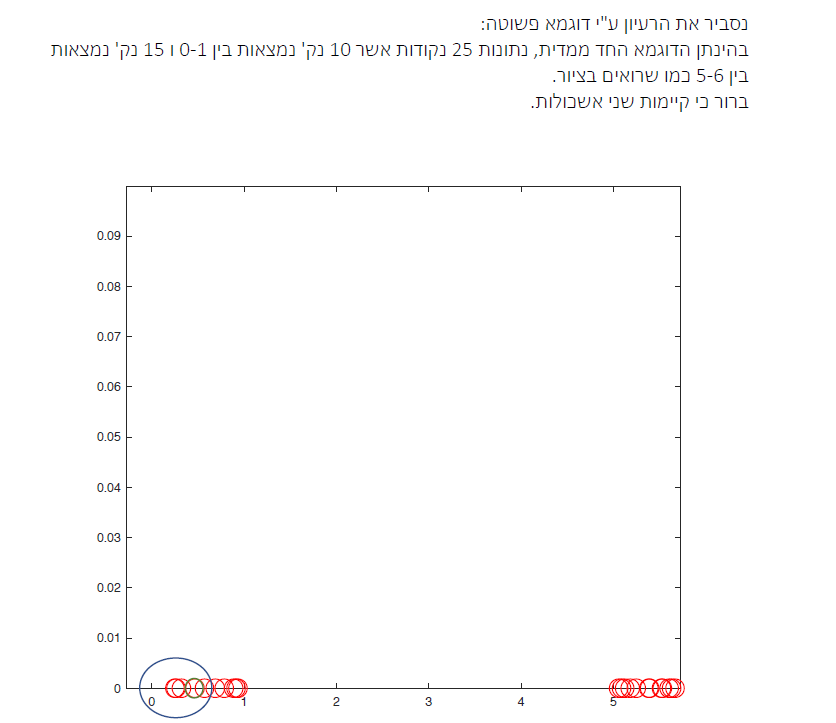
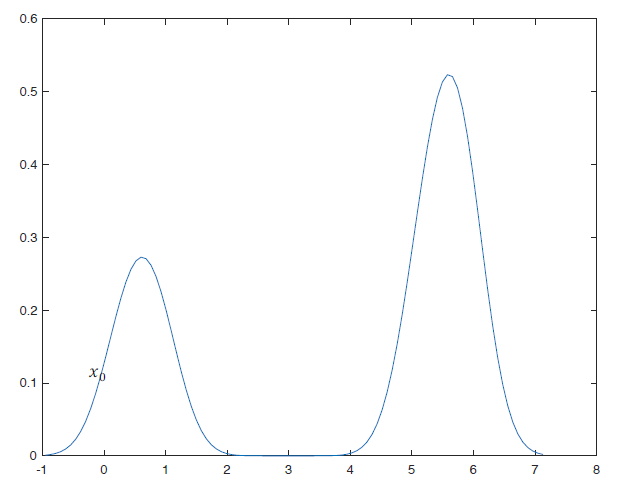
* לא נדרש מספר האשכולות מראש
* האלגוריתם לא צריך את הנקודות האמתיות אפשר להתספק במרחקים בין כל הנקודות.
* קיימת משמעות לתוצאה המתקבלת, כלומר ניתן להבין דרך הפיצול וחלוקת האשכולות ומזה לגזור את הקשר בין האשכולות..

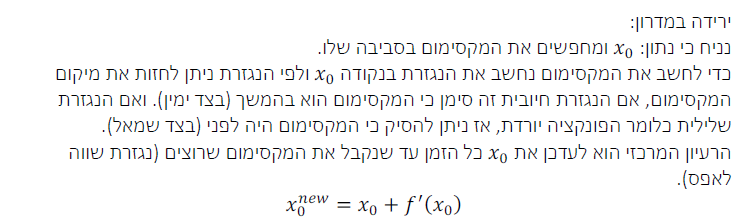
חיסרונות :

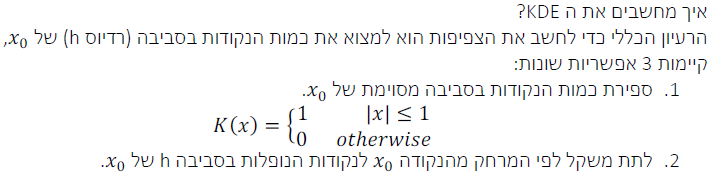
* השפעת השלבים הראשונים על כל האלגוריתם היא קריטית ומבובן שאם שני אשכולות התחברו בהתחלה אז לא ניתן לשנות את זה במהלך האלגוריתם.
* ביצוע האלגוריתם תלויים בפונקציית המרחק שאנו בחורים לכל פונקציה קיימים חסרונות שלה
  + **פונקציית המינימום :**טובה עבור אשכולות רציפים, וגם אם הם לא היו מבוססי מרכזים או מעגליים. לעמות זאת היא חלשה כאשר יש רעש בתוך הנתונים
  + **פונקציית המקסימום:** טובה כאשר הנתונים מכילים רעש פחות רגישה לרעש. ומצד שני היא נוטה לפצל אשכולות רחבים.

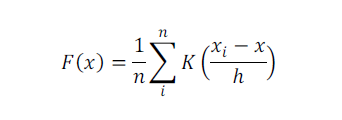
**אלגוריתם Mean Shift**



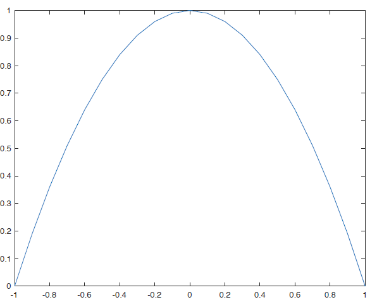


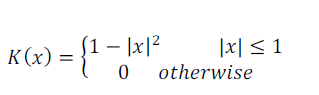


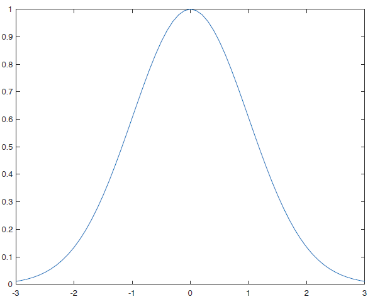


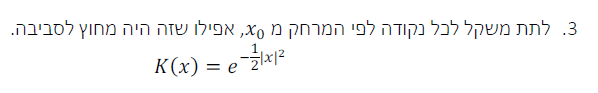


בהינתן X נרצה לספר את כמות הנקודות בתוך הרדיוס H.עבור כל נקודה X נחשב את המרק שלה מ X ונחליט האם להכיליל אותה בחישוב או לא לפי גודל המרחק ובמידה והמרחק גדול מH לא סופרים אותה בחישוב הצפיפות.

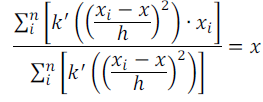


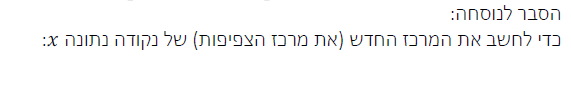


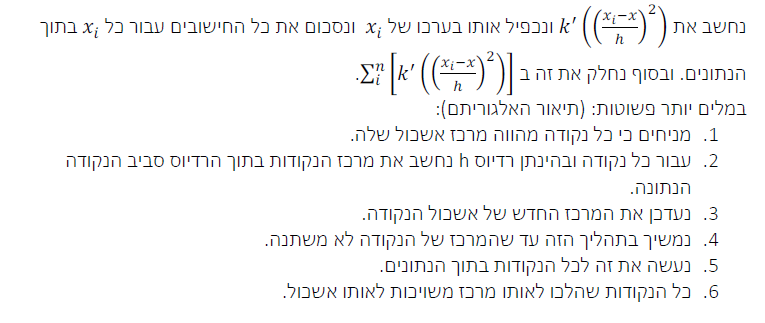


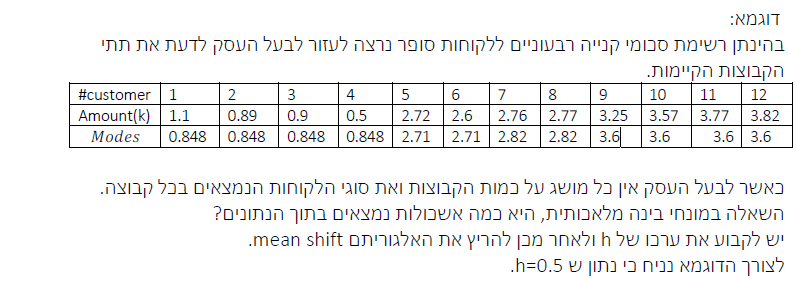
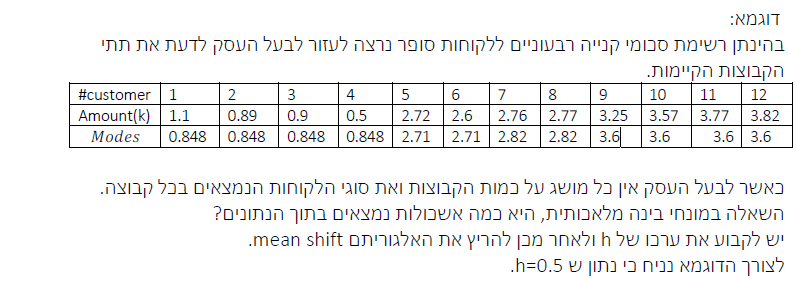


נוסחת עדכון מרכזים :









פתרון :

|  |  |  |
| --- | --- | --- |
| *Dist c1* | *Amount* |  |
| *0* | *1.1* | C1 |
| *0.21* | *0.89* | C2 |
| *0.2* | *0.9* | C3 |
| *0.6* | *0.5* | C4 |
| *1.62* | *2.72* | C5 |
| *1.5* | *2.6* | C6 |
| *1.66* | *2.76* | C7 |
| *1.67* | *2.77* | C8 |
| *2.15* | *3.25* | C9 |
| *2.47* | *3.57* | C10 |
| *2.67* | *3.77* | C11 |
| *2.72* | *3.82* | C12 |

עבור כל נקודה יש לחשב את המרכז הסופי אליו הוא יתכנס.

בוסף כל הלקוחות אשר יש להם אותו מרכז (או מכרז מאוד דומה) יהיו באותו אשכול

עבור רדיוס נתון h=0.5 נרצה לחשב את המרכז החדש

נתחיל עם הלקוח הראשון C1 = 1.1

רואים כי הלקוחות הנמצאים בתוך הרדיוס של C1 הם C1, C2 , C3ולכן נחשב את המרכז החדש לפי שלושת הלקוחות האלה ונקבל:

כרגע מסתכלים על המרכז החדש  *בתור הנקודה החדשה אשר רוצים לה את החשב לה את המרכז הבאה.*

*נחשב שוב את המרקחים בין המרכז החדש לבין כל הנקודות. ומוצאים את כל הנקודות אשר המרחק שלהם מהמרכז החדש פחות או שווה לh-*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *2.857* | *2.807* | *2.607* | *2.287* | *1.807* | *1.797* | *1.637* | *1.757* | *0.463* | *0.063* | *0.073* | *0.137* | *Dist* |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *2.972* | *2.922* | *2.722* | *2.402* | *1.922* | *1.912* | *1.752* | *1.872* | *0.348* | *0.052* | *0.042* | *0.252* | *Dist* |

*מצאנו כי הנקודות שנמצאות בתוך הרדיוס הן* C1, C2 , C3*,* C4*והמרכז החדש אינו מתשנה ולכן האלגוריתם מתכנס עבור הלקוח הראשון וקיבלנו כי מרכז הצפיפות המתאים לו זה :*

*המרכז הצפיפות של* C1 *מתכנס ל*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *2.93* | *2.88* | *2.68* | *2.36* | *1.88* | *1.78* | *1.71* | *1.83* | *0.39* | *0.01* | *0* | *0.21* | *Dist* C1 |

*המרכז הצפיפות של* C2 *מתכנס ל*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *2.92* | *2.87* | *2.67* | *2.35* | *1.87* | *1.86* | *1.7* | *1.82* | *0.4* | *0* | *0.01* | *0.2* | *Dist* C3 |

*המרכז הצפיפות של* C3 *מתכנס ל*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *3.32* | *3.27* | *3.07* | *2.75* | *2.27* | *2.26* | *2.1* | *2.22* | *0* | *0.4* | *0.39* | *0.6* | *Dist* C4 |

*המרכז הצפיפות של* C4 *מתכנס ל*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *1.1* | *1.05* | *0.85* | *0.53* | *0.05* | *0.04* | *0.12* | *0* | *2.22* | *1.82* | *1.83* | *1.62* | *Dist* C5 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *1.11* | *1.06* | *0.86* | *0.54* | *0.06* | *0.05* | *0.11* | *0.01* | *2.21* | *1.81* | *1.82* | *1.61* | *Dist* |

*מרכז הצפיפות של* C5 *מתכנס ל*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *1.22* | *1.17* | *0.97* | *0.65* | *0.17* | *0.16* | *0* | *0.12* | *2.1* | *1.7* | *1.71* | *1.5* | *Dist* C6 |

*מרכז הצפיפות של* C6 *מתכנס ל*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *1.06* | *1.01* | *0.81* | *0.49* | *0.01* | *0* | *0.16* | *0.04* | *2.26* | *1.86* | *1.87* | *1.66* | *Dist* C7 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *1* | *0.95* | *0.75* | *0.43* | *0.05* | *0.06* | *0.22* | *0.1* | *2.23* | *1.93* | *1.93* | *1.72* | *Dist* |

*מרכז הצפיפות של* C7 *מתכנס ל*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *1.05* | *1* | *0.8* | *0.48* | *0* | *0.01* | *0.16* | *0.05* | *2.27* | *1.87* | *1.88* | *1.67* | *Dist* C8 |

*מרכז הצפיפות של* C8 *מתכנס ל*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *0.57* | *0.52* | *0.32* | *0* | *0.48* | *0.49* | *0.65* | *0.53* | *2.75* | *2.35* | *2.36* | *2.15* | *Dist* C9 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *0.732* | *0.682* | *0.48* | *0.26* | *0.31* | *0.32* | *0.48* | *0.36* | *2.58* | *2.18* | *2.17* | *1.98* | *Dist* |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *0.88* | *0.83* | *0.63* | *0.31* | *0.17* | *0.18* | *0.34* | *0.22* | *2.44* | *2.04* | *2.05* | *1.84* | *Dist* |

*מרכז הצפיפות של* C9 *מתכנס ל*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *0.25* | *0.2* | *0* | *0.32* | *0.8* | *0.81* | *0.97* | *0.85* | *3.07* | *2.67* | *2.68* | *2.47* | *Dist* C10 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *0.22* | *0.17* | *0.03* | *0.35* | *0.83* | *0.84* | *1* | *0.88* | *3.1* | *2.93* | *2.71* | *1.5* | *Dist* |

*מרכז הצפיפות של* C10 *מתכנס ל*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *0.05* | *0* | *0.2* | *0.52* | *1* | *1.01* | *1.17* | *1.05* | *3.27* | *2.87* | *2.88* | *2.76* | *Dist* C11 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *0.1* | *0.05* | *0.15* | *0.47* | *0.95* | *0.95* | *0.9* | *1.12* | *3.22* | *2.82* | *2.83* | *2.62* | *Dist* |

*מרכז הצפיפות של* C11 *מתכנס ל*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C12 | C11 | C10 | C9 | C8 | C7 | C6 | C5 | C4 | C3 | C2 | C1 |  |
| *3.82* | *3.77* | *3.57* | *3.25* | *2.77* | *2.76* | *2.6* | *2.72* | *0.5* | *0.9* | *0.89* | *1.1* | *Amount* |
| *0* | *0.05* | *0.25* | *0.57* | *1.05* | *1.06* | *1.22* | *1.1* | *3.32* | *2.92* | *2.93* | *2.72* | *Dist* C12 |

*מרכז הצפיפות של* C12 *מתכנס ל*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *#customer* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* |
| *Amount(k)* | *1.1* | *0.89* | *0.9* | *0.5* | *2.72* | *2.6* | *2.76* | *2.77* | *3.25* | *3.57* | *3.77* | *3.82* |
| *𝑀𝑜𝑑𝑒𝑠* | *0.848* | *0.848* | *0.848* | *0.848* | *2.71* | *2.71* | *2.82* | *2.82* | *2.82* | *3.6* | *3.6* | *3.6* |

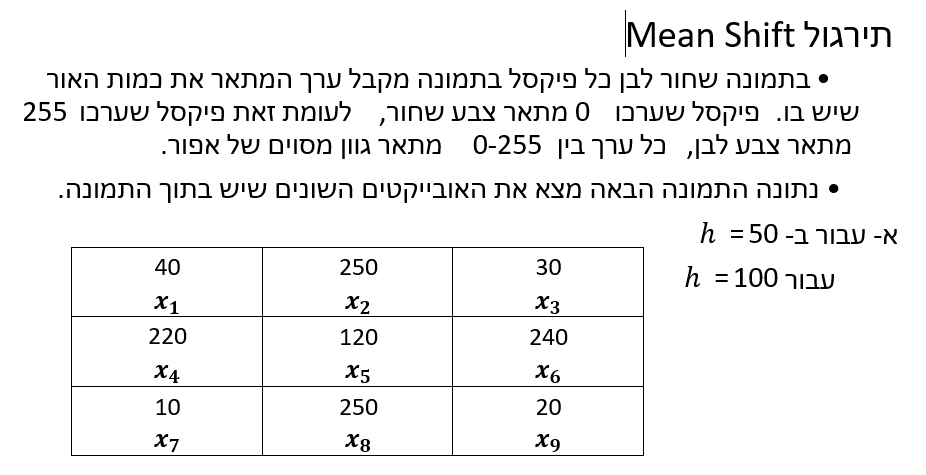
*Cluster A = {c1,c2,c3,c4} Cluster B = {c5,c6}*

*Cluster C = {c7,c8,c9} Cluster D = {c10,c11,c12}*

*שאלה :*

*מה היה קורה למספר האשכולות אם היינו מגדלים את h*

*בדרך כלל כאשר מגדילים את h נקבל פחות אשכולות, וכאשר מקטינים את הערך של h נקבל נקבל יותר אשכולות כאשר המיספר המקסימלי של האשכולות הוא כמספר הנקודות.*



*עבור h = 50*

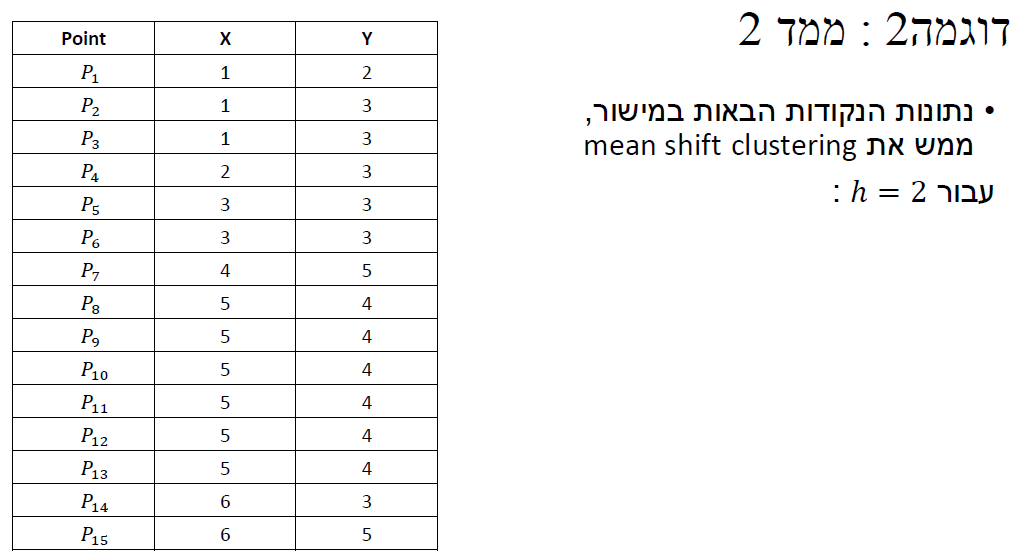
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X9 | X8 | X7 | X6 | X5 | X4 | X3 | X2 | X1 |  |
| *20* | *250* | *10* | *240* | *120* | *220* | *30* | *250* | *40* | *כמות אור* |
| *20* | *200* | *30* | *200* | *80* | *180* | *10* | *210* | *0* | *D(*X1*) = 40* |
| *5* | *225* | *15* | *215* | *215* | *195* | *5* | *225* | *15* | *D( = 25* |
| *230* | *0* | *240* | *10* | *130* | *30* | *220* | *0* | *210* | *D(*X2*) = 250* |
| *220* | *10* | *230* | *0* | *120* | *20* | *210* | *10* | *200* | *D(= 240* |
| *10* | *230* | *20* | *210* | *90* | *190* | *0* | *230* | *10* | *D(*X3*) = 30 == 25* |
| *200* | *30* | *210* | *20* | *100* | *0* | *190* | *30* | *180* | *D(*X4*) = 220== 240* |
| *100* | *130* | *110* | *120* | *0* | *100* | *90* | *130* | *80* | *D(*X5*) = 120* |
|  |  |  |  |  |  |  |  |  | *D(*X6*) == 240* |
| *10* | *240* | *10* | *230* | *110* | *210* | *20* | *240* | *30* | *D(*X7*) = 10 == 25* |
| *230* | *0* | *240* | *10* | *230* | *30* | *220* | *20* | *210* | *D(*X8*) = 250 == 240* |
|  |  |  |  |  |  |  |  |  | *D(*X9*) = 20 == 25* |



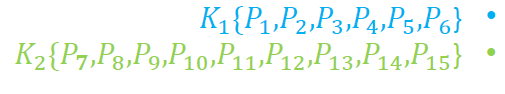
*עבור h = 100*

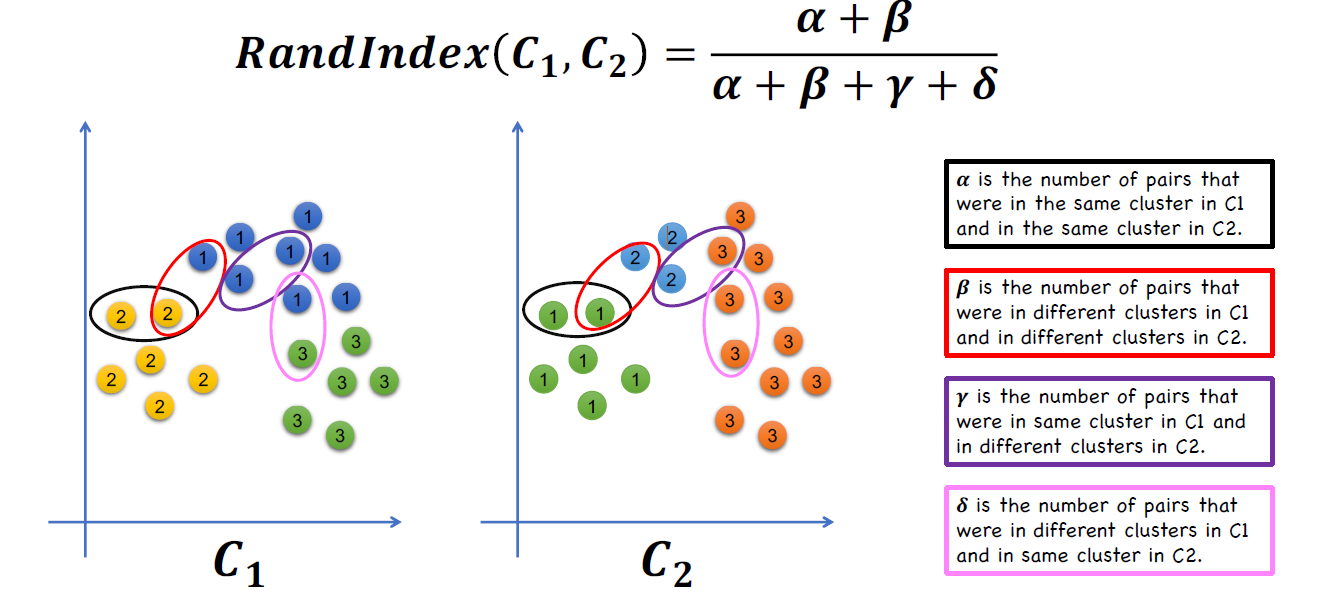
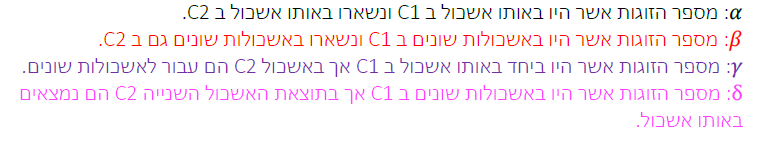
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X9 | X8 | X7 | X6 | X5 | X4 | X3 | X2 | X1 |  |
| *20* | *250* | *10* | *240* | *120* | *220* | *30* | *250* | *40* | *כמות אור* |
| *20* | *200* | *30* | *200* | *80* | *180* | *10* | *210* | *0* | *D(*X1*) = 40* |
| *24* | *206* | *34* | *196* | *76* | *176* | *14* | *206* | *4* | *D( = 44* |
| *230* | *0* | *240* | *10* | *130* | *30* | *220* | *0* | *210* | *D(*X2*) = 250* |
| *220* | *10* | *230* | *0* | *120* | *20* | *210* | *10* | *200* | *D(= 240* |
| *10* | *230* | *20* | *210* | *90* | *190* | *0* | *230* | *10* | *D(*X3*) = 30 == 44* |
| *200* | *30* | *210* | *20* | *100* | *0* | *190* | *30* | *180* | *D(*X4*) = 220* |
| *196* | *34* | *206* | *24* | *96* | *4* | *186* | *34* | *196* | *D( = 216* |
| *100* | *130* | *110* | *120* | *0* | *100* | *90* | *130* | *80* | *D(*X5*) = 120* |
| *66* | *164* | *7* | *154* | *34* | *134* | *56* | *164* | *46* | *D(=86 = D(= 44* |
| *220* | *10* | *230* | *0* | *120* | *20* | *210* | *10* | *200* | *D(*X6*) == 240* |
| *10* | *240* | *10* | *230* | *110* | *210* | *20* | *240* | *30* | *D(*X7*) = 10* |
| *5* | *225* | *15* | *215* | *95* | *195* | *5* | *225* | *15* | *D(=25 = D(= 44* |
| *230* | *0* | *240* | *10* | *230* | *30* | *220* | *20* | *210* | *D(*X8*) = 250 == 240* |
| *0* | *230* | *10* | *220* | *100* | *200* | *10* | *230* | *20* | *D(*X9*) = 20 == 44* |



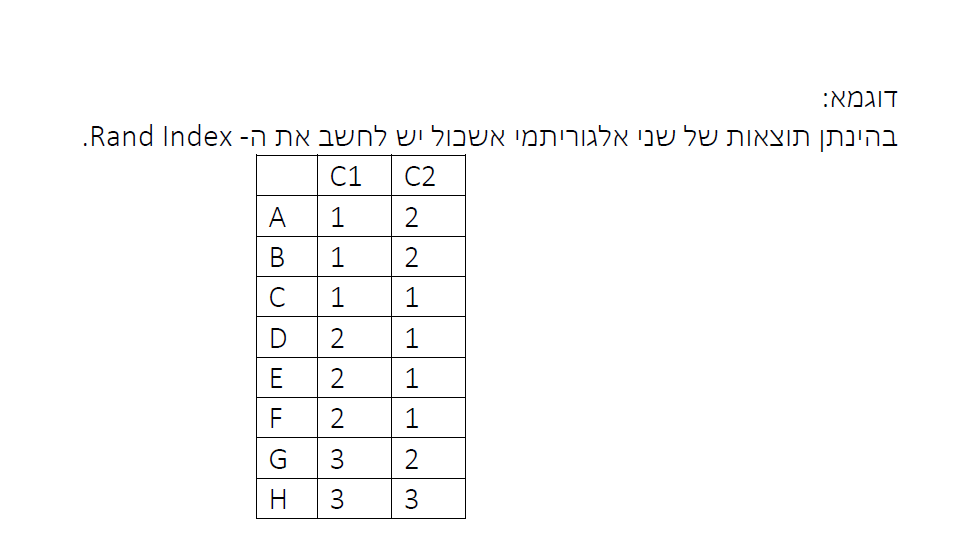


|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| P15 | P14 | P8,9...13 | P7 | P5,6 | P4 | P2,3 | P1 |  |
| *(6,5)* | *(6,3)* | *(5,4)* | *(4,5)* | *(3,3)* | *(2,3)* | *(1,3)* | *(1,2)* | *point* |
|  |  |  |  | *2.23* | *1.41* | *1* | *0* | *D(C*1*) = (1,2)* |
|  |  |  | *3.55* | *1.767* | *0.79* | *0.354* | *0.79* | *D() =(1.25,2.75)* |
|  |  |  | *3.07* | *1.18* | *0.24* | *0.847* | *1.17* | *D() =(1.83,2.83)* |
|  |  |  | *3.60* | *2* | *1* | *0* | *1* | *D(C*2,3*) = (1,3)* |
|  |  |  | *2.82* | *1* | *0* | *1* | *1.41* | *D(C*4*) = (2,3)* |
|  |  |  | *2.23* | *0* | *1* | *2* | *2.23* | *D(C*5,6*) = (3,3)* |
|  |  |  |  |  |  |  |  | *D() = (2,3)* |
| *2* | *2.82* | *1.41* | *0* | *2.23* | *2.82* | *3.60* | *4.24* | *D(C*7*) = (4,5)* |
| *1.25* | *1.6* | *0.25* | *1.25* | *2.35* | *3.25* | *4.19* | *4.58* | *D() = (5,4.25)* |
| *1.25* | *1.42* | *0.15* | *1.42* | *2.38* | *3.30* | *4.25* | *4.62* | *D() = (5.111,4.111)* |
| *1.41* | *1.41* | *0* | *1.41* | *2.24* | *3.16* | *4.12* | *4.47* | *D(C*8...13*) = (5,4)* |
| *2* | *0* | *1.41* | *2.83* | *3* |  |  |  | *D(C*14*) = (6,3)* |
| *1.25* | *1.25* | *0.25* | *1.6* | *2.46* |  |  |  | *D() = (5.25,4)* |
| *0* | *2* | *1.41* | *2* | *3.6* |  |  |  | *D(C*15*) = (6,5)* |

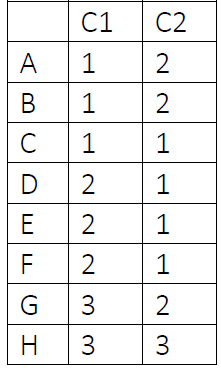


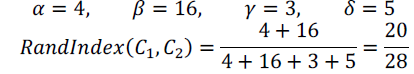
**Rand Index**

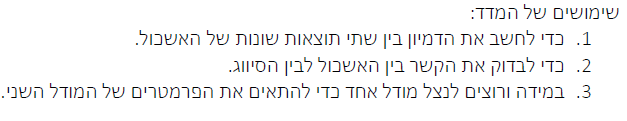
*המדד Rand Index מודד את הדימיון בין שתי תוצאות אשכול. בהיתן שתי תצאות C*2 *, C*1 *נרצה לחשב את הדמיון בניהם .*



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | E | F | G | H |
| A | **0** | **a** | **g** | **b** | **b** | **b** | **d** | **b** |
| B |  | **0** | **g** | **b** | **b** | **b** | **d** | **b** |
| C |  |  | **0** | **d** | **d** | **d** | **b** | **b** |
| D |  |  |  | **0** | **a** | **a** | **b** | **b** |
| E |  |  |  |  | **0** | **a** | **b** | **b** |
| F |  |  |  |  |  | **0** | **b** | **b** |
| G |  |  |  |  |  |  | **0** | **g** |
| H |  |  |  |  |  |  |  | **0** |

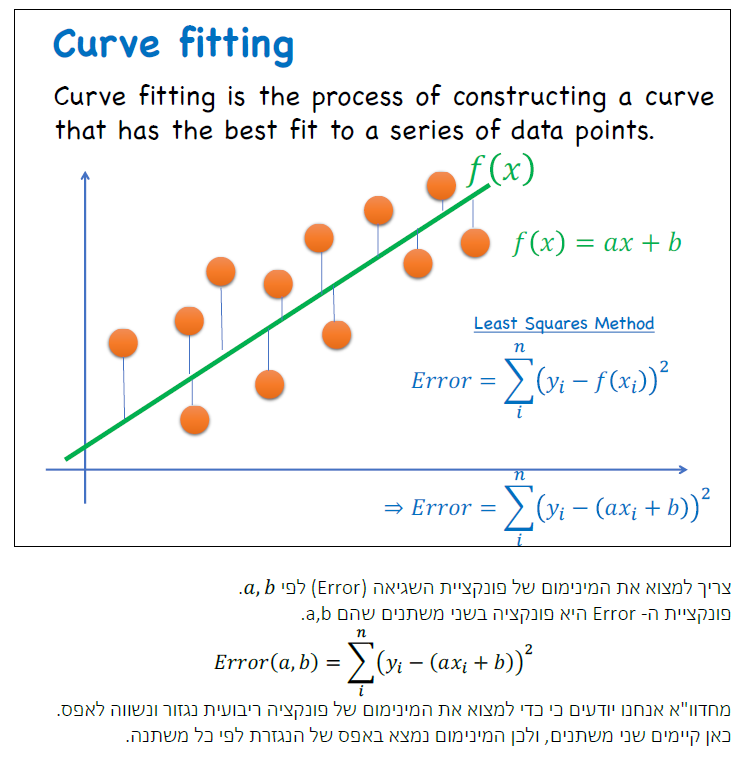


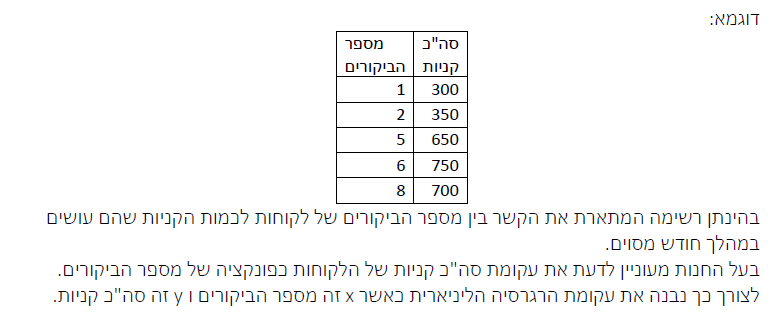


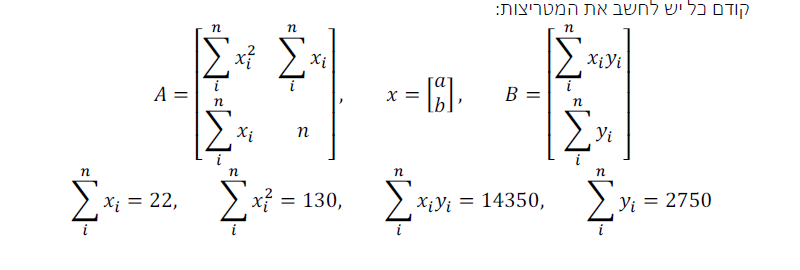


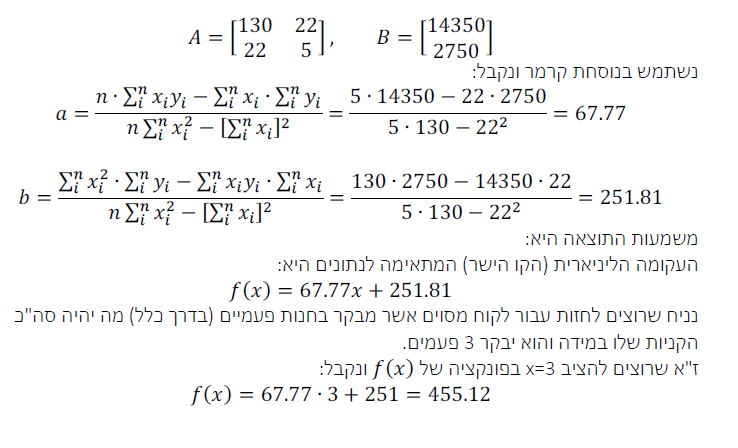
*כלל שה- Rand Index קרוב ל-1 ככה האשכולות יותר דומים.*

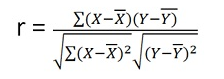
**רגריסיה ליניארית**

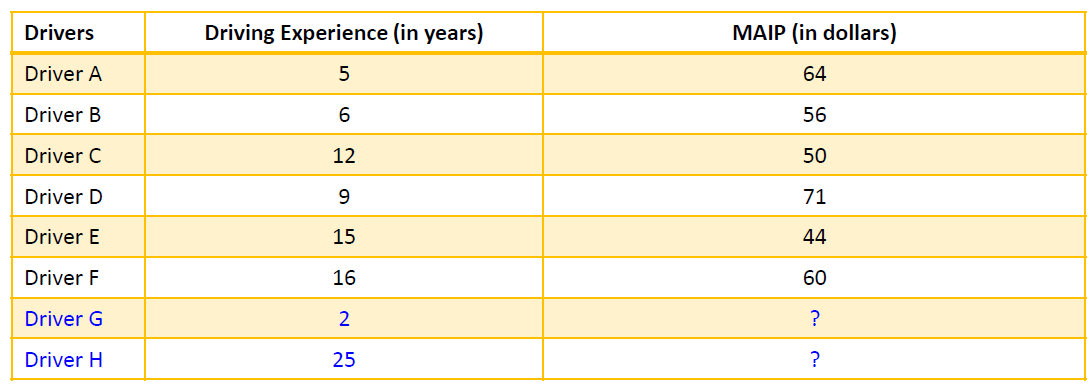










*תירגול :*

*Y=-1.02X+68.2*

*G=-1.02\*2+68.2=66*

*H=-1.02\*25+68.2=43*

*קשר לינארי שלילי הוא קשר הפוך*

**דגימת תכונות**

תהליך דיגמת *התוכנות הוא תהליך אשר בור בחוחרים את התוכנות החשובות ביותר לייצג קובץ נתונים (מסווגים).*

*ידוע שככל שיש יותר נתונים* ***איכותים*** *התוצאות היו טובות יותר, ולכן צריך להתמקד בבחירת התכונות המשמעותיות ביותר לצורך פתרון הבעיה נתונה.*

*קיימות מספר סיבות לתהליך בחירת התכונות:*

1. ***שיפור******ביצועים***
   1. *על ידי דגימת העמודות החשובות ביותר ניתן לשפר את ביצוע את האלגוריתם ואת הדיוק שלו.*

*מאחר ובתהליך דיגמת העמודות נסנן את חלק מהעמודות המוסיפות רעש לנתונים.*

* 1. *חלק מהאלגוריתמים רגישים לכמות העמודות כמו knn.*

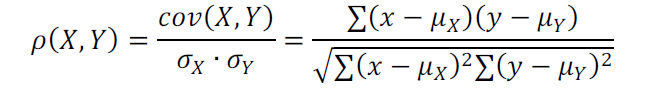
1. ***הצגת הנתונים*** *: אנו יכולים להציג נתונים מספרים ב ולכן ככל שיש יותר עמודות דבר זה יגביל את היכולות להציג את הנתונים בצורה מובנת.*
2. ***הפחת כמות הנתונים :*** *חיסכון באיחסון וגם בזמן ריצה של האלגוריתמים.*

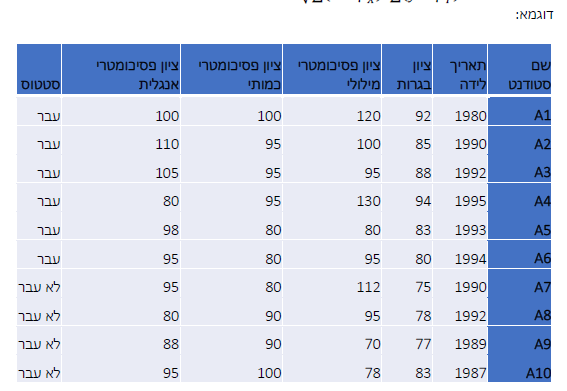
*קיימות 3 קטגוריות מרכזיות של דיגמת עמודות:*

1. ***סינון:*** *נעבור על כל העומודות ונסן את העמודות הלא רלוונטיות על ידי שימוש במדדים סטטיסטיים למדידת הקשר בין ערכי העמודה לבין הסוג של כל רשומה.**והאלגורתמים מסננים את העמודות בעלי קשר נמוך.*
2. ***שיטת העטיפה:*** *המטרה שלה היא לבצע חיפוש על העמודות הטובות ביותר והרעיון הכללי הוא בהינתן קבוצה של N עמודות נצרה לבחור את הקבוצה של העמודות אשר תייצג את הנתונים בצורה הטובה ביותר.*
3. ***שיטות משלובות אלגוריתמי בחירת עמודות:*** *חלק מהאלגורתמים מכילים בתוכם את תהליך דגימת הנקודות. לדוגמא את אלגוריתם עצי ההחלטה מכיל בתוכו תהליך של בחירת עמודות. כלומר לאחר שבנינו את העץ על סמך קובצת האימון ניתן להשתמש רק בעמודות הכלולות בתוך על ההחלטה שנבנה.*

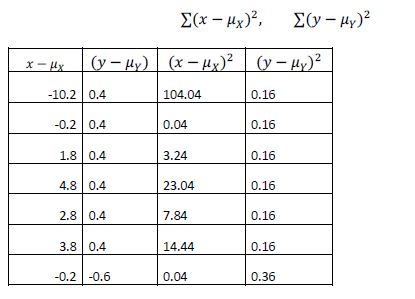
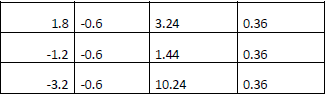
***אלגוריתם בחירת עמודות:***

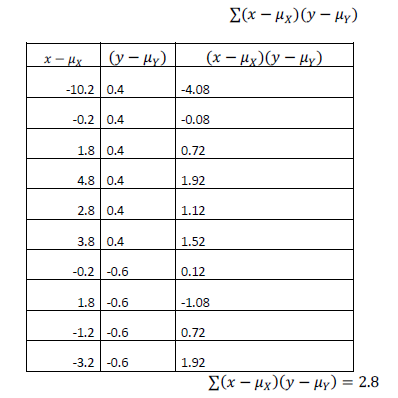
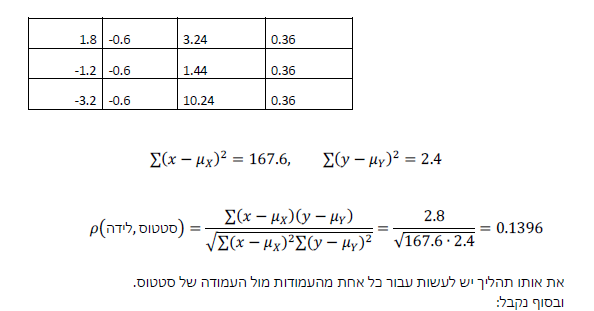
1. *האלגוריתם הראשון* ***שיטת הקורילציה*** *לפי שיטה זאת יש לחשב את הקורילציה בין כל עמודה מספרית בנתונים לבין עמודת הסוגים. ככל שהמספר בערך מוחלט גובה יותר זה אומר שהקשר הוא חזק יותר. ובסוף האלגוריתם מעדיף את העמודות בעלות הקשר החזק עם עמודת הסוגים.*

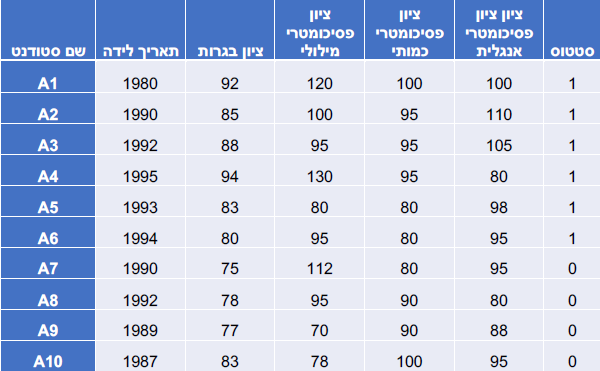


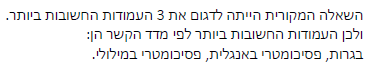


*<https://www.socscistatistics.com/tests/pearson/default2.aspx>*

[](https://www.socscistatistics.com/tests/pearson/default2.aspx)

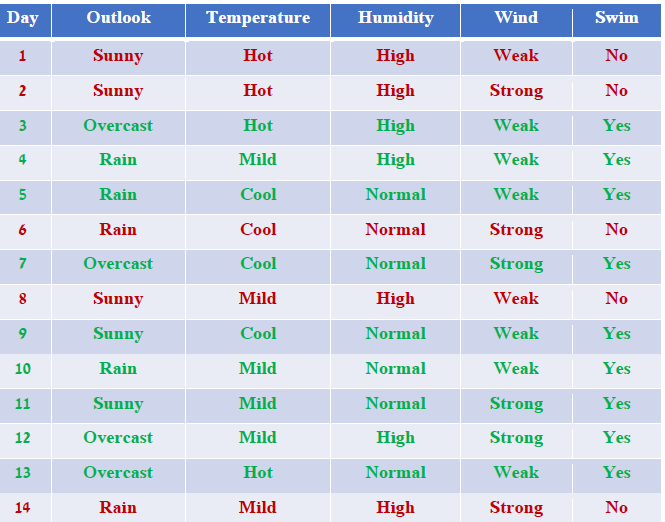




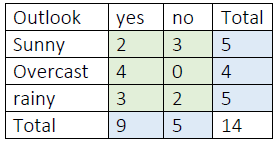


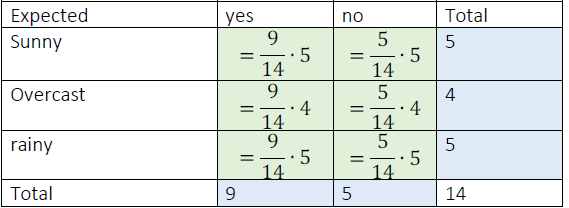
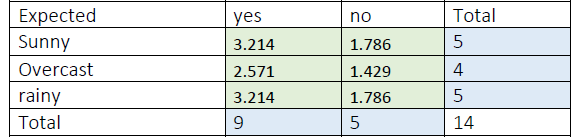
1. *האלגוריתם השני* ***מדד 2X*** *בדומה להתהליך עם המקדם המתאם נשתמש במדד זה ושוב נבחר את העמודות*

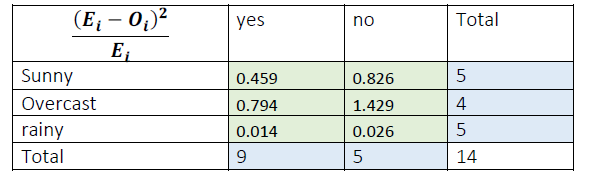
*בעלות הקשר החזק ביותר עם העמודות הסוגים.*

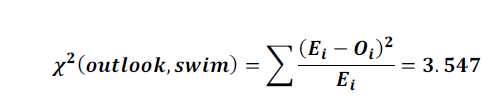


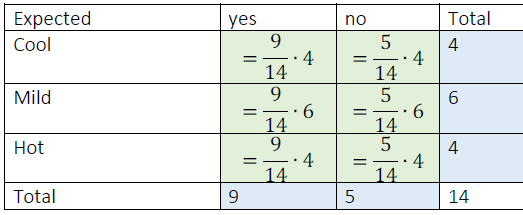
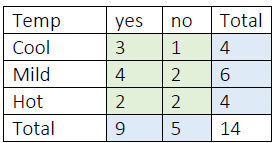
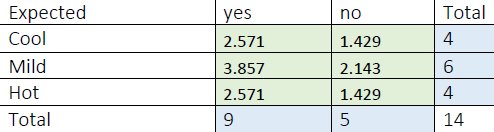
[*https://www.socscistatistics.com/tests/chisquare2/default2.aspx*](https://www.socscistatistics.com/tests/chisquare2/default2.aspx)

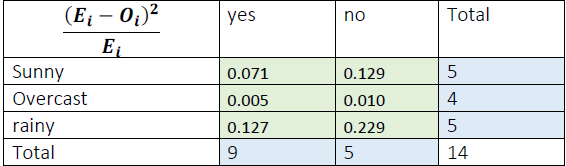
[*http://www.quantpsy.org/chisq/chisq.htm*](http://www.quantpsy.org/chisq/chisq.htm)

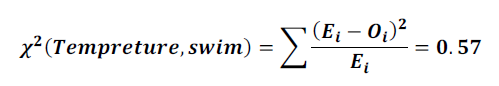


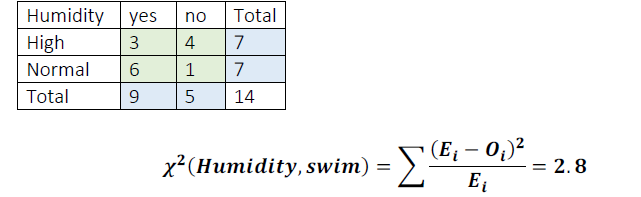
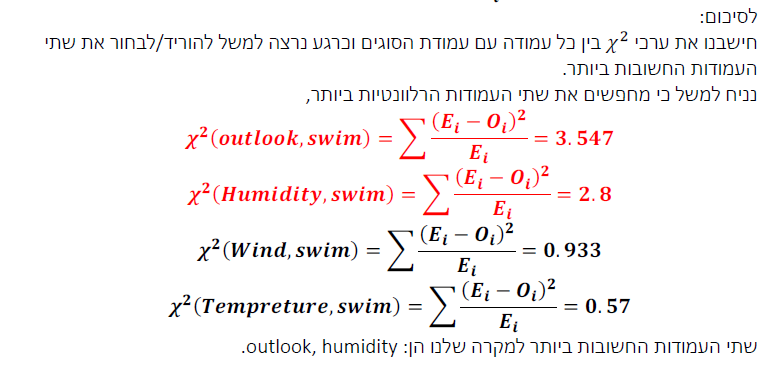












***שיטת העטיפה :***

*המטרה שלנו באלגוריתם של הבחירה עמודות היא דגימת תת קבוצת העמודות האופטימלית.*

*ולכן הבעיה שאנו מנסים לפתור הינה בעיית חיפוש של תת הקבוצה.*

*בשיטת העטיפה אנחנו נשתמש במסווג הנתון כמדד / מצביע על איכות תת הקבוצה הנבחרת.*

*כלומר נעיזר באיכות התוצאה דיוק/ f-score על הקבוצת הבקרה ואז נדגום את קבוצת העמודות הגומרת לתוצאה הטובה ביותר.*

*נשאלת השאלה על כמות הבידקות שיש לעשות כדי לבחור את התת הקבוצה הטובה ביותר?*

