

ב' י"ג ז' אלול

מחבר : יוסף משה

קורט : ב' י"ג ז' אלול

נושא : חזרה אלמחן

K-means:

בהינתן קבוצת נתונים:

Point	x	y
A ₁	8	8
A ₂	8	6
A ₃	3	5
A ₄	1	3
A ₅	2	1

ובהינתן טבלת המרחקים בהיבול בין כל שני נקודות:

#	A ₁	A ₂	A ₃	A ₄	A ₅
A ₁	0	4	34	74	85
A ₂	4	0	26	58	61
A ₃	34	26	0	8	17
A ₄	74	58	8	0	5
A ₅	85	61	17	5	0

א. הראה חישוב מלא של K-means עבור $K=2$ כאשר המרכזים ההתחלתיים הם A₁, A₅

ב. חשב את ה- SSE עבור תוצאת הסופית

ג. הרצת אלגוריתם של K-means עבור $K=3$. איך לך נראה על SSE?

ד. האם K-means צריך נחל? ?

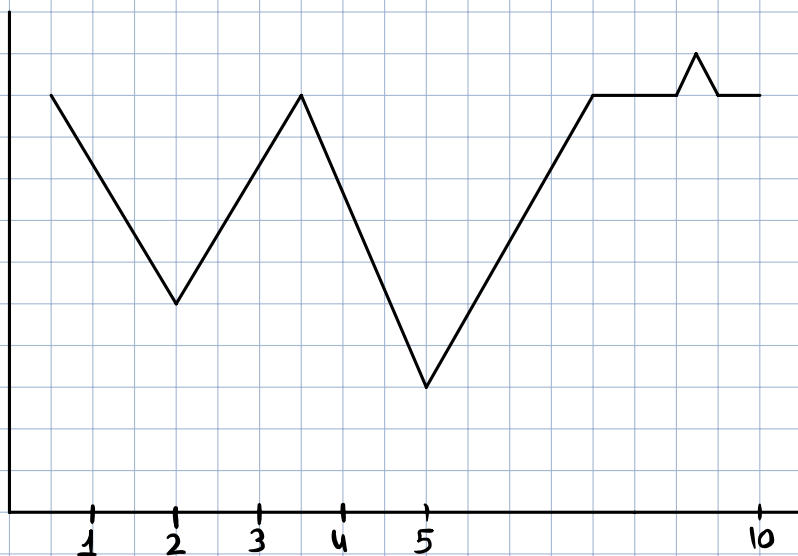
ה. אם מגדלים את K, איך לך מראה על SSE?

ו. עבור $K=5$, מה תהיה תוצאת ה- SSE?

ז. עבור $K=1$, מה תהיה תוצאת ה- SSE?

ה) הרצנו K-means עם $K=5$, 10 בלטים. ציר X מתאר את מ"ס הניסוי, וציר Y מתאר את ה-SSE.

כעת איתנו 5 קבוצות. מהי התוצאה הכי טובה? באיזה איזה הרצה הכי טובה?



פתרון:

© נתון $B=A$ ואם ציין 2 מרכזים התחלתיים. נתון לנו ש-2 מרכזים הנתונים הם A_1, A_5 .

המטרה היא לחשב עבור כל נקודה את המרחק משני המרכזים, כשיוק אותה אל המרכז הקרוב אליה, ובסוף מוצגים מרכזים.

נתון לנו שבירת המרחקים וביתיים המרכזים: $center_2 = A_5 = (2, 1)$ $center_1 = A_1 = (8, 8)$

A_1 : שייכת למרכז של A_1 (המרכז הוא הנקודה עצמה).

A_2 : קרוב יותר ל C_1

A_3 : קרוב יותר ל C_2

A_4 : קרוב יותר ל C_2

A_5 : שייכת למרכז של C_2 (הנקודה עצמה).

עזרון מרכזים:

$$C_1 = \{A_1, A_2\}$$

$$C_2 = \{A_3, A_4, A_5\}$$

$$C_1(x) = \frac{8+8}{2} = 8$$

$$C_2(x) = \frac{3+1+2}{3} = 2$$

$$C_1(y) = \frac{8+6}{2} = 7$$

$$C_2(y) = \frac{5+3+1}{3} = 3$$

$$C_1(x, y) = (8, 7)$$

$$C_2(x, y) = (2, 3)$$

A1:

$$d^2(A1, C1) = (8-8)^2 + (8-7)^2 = 1$$

$$d^2(A1, C2) = (8-2)^2 + (8-3)^2 = 61$$

$$A1 \in C1$$

A2:

$$d^2(A2, C1) = (8-8)^2 + (6-7)^2 = 1$$

$$d^2(A2, C2) = (8-2)^2 + (6-3)^2 = 45$$

$$A2 \in C1$$

A3:

$$d^2(A3, C1) = (3-8)^2 + (5-7)^2 = 29$$

$$d^2(A3, C2) = (3-2)^2 + (5-3)^2 = 5$$

$$A3 \in C2$$

A4:

$$d^2(A4, C1) = (1-8)^2 + (3-7)^2 = 65$$

$$d^2(A4, C2) = (1-2)^2 + (3-3)^2 = 1$$

$$A4 \in C2$$

A5:

$$d^2(A5, C1) = (2-8)^2 + (1-7)^2 = 72$$

$$d^2(A5, C2) = (2-2)^2 + (1-3)^2 = 4$$

$$A5 \in C2$$

תוצאה סופית:

הנהוגות בתוך האסטמיות לא השתנו, כאשר אם נחשב מרכזים, נקבל:

$$C_1 = \{A_1, A_2\}$$

$$C_2 = \{A_3, A_4, A_5\}$$

$$C_1(x) = \frac{8+8}{2} = 8$$

$$C_2(x) = \frac{3+1+2}{3} = 2$$

$$C_1(y) = \frac{8+6}{2} = 7$$

$$C_2(y) = \frac{5+3+1}{3} = 3$$

$$C_1(x, y) = (8, 7)$$

$$C_2(x, y) = (2, 3)$$

הבולן אותם מרכזים ולכן האלמנטים נשברו. כמו כן, היינו יכולים להגיד שהבולן אותם

(הנקודות האסטמיות ולכן האלמנטים נשברו) (כלי הישוב מרכזים).

Ⓢ

$$SSE = \sum dist^2(m_1, P) + \sum dist^2(m_2, P) + \dots + \sum dist^2(m_n, P) = \sum_{i=1}^K \sum_{P \in C_i} dist^2(m_i, P)$$

↑
סכמת כל המרחקים מרכז של כל תחום באסטמיות

לכן שמסתכלים על התוצאה הסופית, אנו רואים כאשכול הראשון ובדקוים כמה הנחיות החבות / רחוקות ממרכז האסטמיות... אנו רואים כאשכול השני ובדקוים כמה הנחיות החבות / רחוקות ממרכז האסטמיות וכן הלאה... עבור כל האסטמיות, במקרה שלנו היינו 2 אסטמיות, ונקבל:

$$SSE = d^2(A_1, C_1) + d^2(A_2, C_1) + d^2(A_3, C_2) + d^2(A_4, C_2) + d^2(A_5, C_2) = 1+1+5+1+4 = 12$$

Ⓢ - ה- SSE יהיה 0 מכיון שיש לנו סך 5 נקודות, ובמקרה של 5-4 אנו כל נקודת תחום באסטמיות.

Ⓢ אם אנו רוצים להיות ברורים, אנו רוצים להיות ברורים, כי זה מתבסס על חישובי

מרחקים. אם כמובן שני המרחקים מאותו סט, אז לא צריך כי סטיית התבונן שלהם צריכה

הוא לא ישאל שאם שנייה אחת בו סטיית תהיה, אם ישאל אז אחר יהיה באפסים והשני באפסים.

ה) כד פער שמעדרים את א, ה- SSE ערביות הוא קטן/שווה. וש מקרים מיוחדים שאם מעדרים את א אז מרבית SSE יגדל.

① אין צורך לחשב, מכיוון שכבר $k=5$, כל נקודה תהיה אשכול, ובמיוחד כלל המרחק של כל נקודה מהאשכול שלה = 0

כי כל אשכול מכיל רק נקודה אחת, ולכן מרחק של כל אשכול לאת הנקודה עצמה, ומרחק הריבוע של נקודה אל עצמה = 0.

SSE זה סכום מרחק הריבוע של כל נקודה ממרכז האשכול אליו היא שייכת, ואז נקבל $SSE = 0+0+0+0+0 = 0$.

ב) במקרה של $k=1$, נמצא והיו באותו אשכול (או משנה מי המרכזים הכתובותיים). כי את בחינו עצבנו עם אשכול אחד, אז ברור

שנמצא והיו באותו אשכול.

מבקשים להשיג SSE של מרחק הריבוע של כל נקודה ממרכז האשכול שלה (במקרה שלנו נמצא שייכים באותו אשכול).

לכן, צריך לחשב את המרחק של האשכול (כל הנקודות) ואז לסכום את מרחקי הריבוע של כל נקודה ממרכז.

נחשב מרחק

$$\left(\frac{A_1(x)+A_2(x)+A_3(x)+A_4(x)+A_5(x)}{5}, \frac{A_1(y)+A_2(y)+A_3(y)+A_4(y)+A_5(y)}{5} \right) = (4.4, 4.6)$$

נסמן מרחק הריבוע של כל נקודה, ממרכז האשכול $(4.4, 4.6)$:

$$SSE = d^2(A_1, C_1) + d^2(A_2, C_1) + d^2(A_3, C_1) + d^2(A_4, C_1) + d^2(A_5, C_1) + d^2(A_6, C_1)$$

$$= (8-4.4)^2 + (8-4.6)^2 + \dots + (2-4.4)^2 + (1-4.6)^2 = 74.4$$

② SSE נמוך מדי, על תוצאה טובה יותר. לכן נרצה לבחור את ה- SSE הנמוך ביותר, ולכן התוצאה

הטובה ביותר זה עבור הרצה 5.

אופן הידור:

בהינתן קבוצת נתונים:

Point	x	y
A ₁	8	8
A ₂	8	6
A ₃	3	5
A ₄	1	3
A ₅	2	1

ובהינתן טבלת המרחקים בהיבול בין כל שני נקודות:

	A ₁	A ₂	A ₃	A ₄	A ₅
A ₁	0	4	34	74	85
A ₂	4	0	26	58	61
A ₃	34	26	0	8	17
A ₄	74	58	8	0	5
A ₅	85	61	17	5	0

Ⓐ הראה חישוב מלא עבור שיטת המינימום

Ⓑ מהי תוצאת האלגוריתם עבור 2 אשפוזים.

פתרון:

א) מידור על שיטת מנימוס, ואם גם עדכן הטבלה וגם החיבור עם המנימוס.

אז נילכד בדרכי הפעולה:

1) כל נהדר מהווה אשכול.

2) מחשבים מרחקים בין כל שני אשכולות ושומרים את כל המרחקים במטריצת מרחקים.

3) כל עוד שזא קיבלנו אשכול אחד:

תמיד עם המרחק הקטן ביותר.

3.1. מאחדים את שני האשכולות הקטנים ביותר.

3.2. מעדכנים את מטריצת המרחקים. → ואת העדכון עושים לפי השיטה הניתנה בתנאי.

בשלב זה עדיין לא חיברנו אף אשכול עם אשכול אחר.

כל נהדר הוא אשכול. היותו בשלב זה 5 אשכולות.

אז באופן חד בעלי אנחנו בונים את מטריצת המרחקים שתלווה אותנו לאורך כל הפתרון.

ואז כל פעם נחבר אשכולות תמיד עם המרחק הקטן ביותר ונעדכן מרחקים עם השיטה.

המוצא בתנאי, במקרה של שיטת המנימוס. נעשה כך עד שנקבל אשכול אחד.

מטריצת המרחקים (טבלת מרחקים בין אשכולות) חד בזמני:

	A1	A2	A3	A4	A5
A1	0	4	34	74	85
A2	4	0	26	58	61
A3	34	26	0	8	17
A4	74	58	8	0	5
A5	85	61	17	5	0

חיבור גשמי:

המרחק המינימלי הוא בין A_2 ל- A_1 ואם נחבר אותם. החיבור נעשה תמיד לפי המינימום.

נקבל $A_1 \cup A_2$

לדבן טבלה:

השיטה המוצעת היא מינימום ואם אדגון הצדדים נעשה לפי המינימום.

	$A_1 \cup A_2$	A_3	A_4	A_5
$A_1 \cup A_2$	0	26	58	61
A_3	26	0	8	17
A_4	58	8	0	5
A_5	61	17	5	0

חיבור גשמי:

המרחק המינימלי הוא בין A_4 ל- A_5 ואם נחבר אותם. החיבור נעשה תמיד לפי המינימום.

נקבל $A_4 \cup A_5$

לדבן טבלה:

	$A_1 \cup A_2$	A_3	$A_4 \cup A_5$
$A_1 \cup A_2$	0	26	58
A_3	26	0	8
$A_4 \cup A_5$	58	8	0

חיבור אגרגציה:

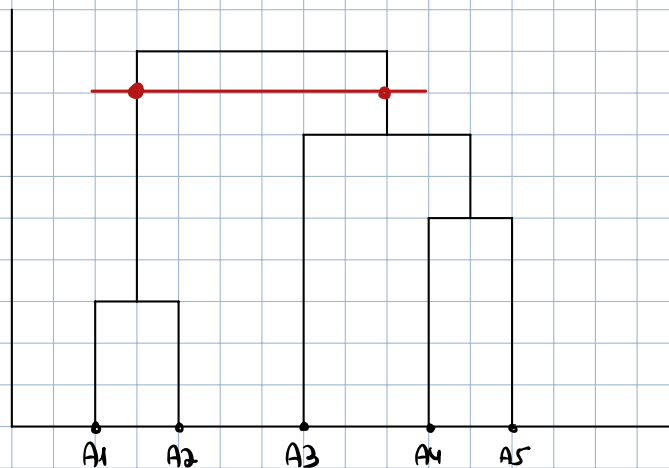
המרחק המינימלי הוא בין A_4 ל- A_5 ואם נחבר אותם. החיבור נעשה תמיד לפי המינימום.

נקבל $A_4 \cup A_5$

דגון טבלה:

$A_3, A_4 \cup A_5$		
	$A_1 \cup A_2$	$A_4 \cup A_5 \cup A_3$
$A_1 \cup A_2$	0	26
$A_4 \cup A_5 \cup A_3$	26	0

🕒 נצטרך קונקטור ונראה קו איבר שיש חיתוך עם A_3 הוויס, כי חוצים A_3 אגרגציות.



$$cluster_1 = \{A_1, A_2\}$$

$$cluster_2 = \{A_3, A_4, A_5\}$$

mean shift:

בהינתן קבוצת נתיבים:

Point	x	y
A ₁	8	8
A ₂	8	6
A ₃	3	5
A ₄	1	3
A ₅	2	1

ובהינתן טבלת המרחקים בהיבול בין כל שני נקודות:

#	A ₁	A ₂	A ₃	A ₄	A ₅
A ₁	0	4	34	74	85
A ₂	4	0	26	58	61
A ₃	34	26	0	8	17
A ₄	74	58	8	0	5
A ₅	85	61	17	5	0

א) הראה חישוב מלאן של ה- mean shift עם $H=5$

ב) מצא את מ"מ האשכולות ואת הנקודות בתוך כל אשכול.

ג) הראה מבחניב וזנואית את התוצאה (כלי חישובים).

ד) מה יקרה עבור $H=6$?

ה) מה יקרה עבור $H=0.5$?

ו) מה יקרה עבור $H=4.5$?

ז) נניח והיה נתון $H=5$ וידוע כי התפוצסה (הקוצר) אחת חזשה: $AC(200, 400)$. מה תהיה תוצאת האשכול ?

במקבץ:

המטרה היא לסבור עבור כל נקודה (נקודה) האילו מרכז צפיפות הוא שייכת. כל שם הודות אילו נקודות נמצאות בסביבת

H, מחשבים את המרכז של הנקודות בסביבה ונקודות מרכז חדש. למרכז החדש שם הודות אילו נקודות נמצאות בסביבתה,

מחשבים כך עד שהמרכז לא משתנה. הנושם של נמצאים בסביבתה: זה אילו נקודות מרחקים הם ממרכז מאשר

הרדיוס. אם המרחק ממרכז המעגל קטן מ-H אז סמן שהם "בתוך המעגל" ונמצאים בסביבה.

A1:

נתחיל עם A1 ונקודות שנמצאות בסביבתה במרחק H=5 ואשר להיחלר בשלם זה במטריצת המרחקים.

הנקודות שנמצאות סביב A1 במרחק H זה: A1, A2

$$\left(\frac{A1(x) + A2(x)}{2}, \frac{A1(y) + A2(y)}{2} \right) = (8, 7)$$

נלקח את המרכז החדש:

כעת, נבדוק אילו נקודות נמצאות בסביבת המרכז החדש (8,4). אפשר לראות בהקשר שהנקודות A1, A2 נמצאות

בסביבה, ואילו הנקודות A3(3,5), A4(1,3), A5(2,1) ממחית.

ללא שהמרכז לא השתנה ואם A1 שייכת למרכז הצפיפות (8,7).

A2:

הנקודות שנמצאות סביב A2 במרחק H זה: A1, A2 ואילו בקצו כלית מטריצת המרחקים.

$$\left(\frac{A1(x) + A2(x)}{2}, \frac{A1(y) + A2(y)}{2} \right) = (8, 7)$$

נלקח את המרכז החדש:

כעת, נבדוק אילו נקודות נמצאות בסביבת המרכז החדש (8,4). אפשר לראות בהקשר שהנקודות A1, A2 נמצאות

בסביבה, ואילו הנקודות A3(3,5), A4(1,3), A5(2,1) ממחית (ללא מרחקים < H היתכן).

ללא שהמרכז לא השתנה ואם A2 שייכת למרכז הצפיפות (8,7).

הצרה:

אפשר לראות שם נשווה בין איטרציה ראשונה של A1 ואיטרציה ראשונה של A2, הובלע אותם נקודות

שנמצאות בדיוק. ללא שגם שאר האיטרציות והיו דומות, כי באיטרציה הראשונה היו זהם אותם קבוצת שברים

ואם הן שייכות לאותו אשט.

A3:

הנהודות שנמצאות סביב A_3 במרחק H זה A_3 ואותן בקטנו האלמן מטרצת המחוקים.

אבל, זה לא נכון להגיד ככה, כי הסבאר התורה זה המחוקים המבואר, ואם יש לשים שורש עבור כל האנשים

הסבאר, כי השותפים $H=5$ אל מתכונים למרחק $=5$ ואם למרחק הרבוא.

לכן, הנהודות שנמצאות בסביבת A_3 במרחק של $H=5$ הן: A_3, A_4, A_5

$$\left(\frac{A_3(x) + A_4(x) + A_5(x)}{3}, \frac{A_3(y) + A_4(y) + A_5(y)}{3} \right) = (2, 3)$$

כעת, נבדוק איזה נהודות נמצאות בסביבת המרכז החדש $(2,3)$ אפשר לראות בקלות שהנהודות A_3, A_4, A_5 נמצאות

בסביבה, ואילו הנהודות A_1, A_2 מחוץ.

כל שרמרכז לא השתרע ואם A_3 שויפת למרכז הצביות $(2,3)$

A4:

הנהודות שנמצאות סביב A_4 במרחק H זה A_3, A_4, A_5 ואותן בקטנו האלמן מטרצת המחוקים.

$$\left(\frac{A_3(x) + A_4(x) + A_5(x)}{3}, \frac{A_3(y) + A_4(y) + A_5(y)}{3} \right) = (2, 3)$$

כעת, נבדוק איזה נהודות נמצאות בסביבת המרכז החדש $(2,3)$ אפשר לראות בקלות שהנהודות A_3, A_4, A_5 נמצאות

בסביבה, ואילו הנהודות A_1, A_2 מחוץ.

כל שרמרכז לא השתרע ואם A_4 שויפת למרכז הצביות $(2,3)$

A5:

הנהודות שנמצאות סביב A_5 במרחק H זה A_3, A_4, A_5 ואותן בקטנו האלמן מטרצת המחוקים.

$$\left(\frac{A_3(x) + A_4(x) + A_5(x)}{3}, \frac{A_3(y) + A_4(y) + A_5(y)}{3} \right) = (2, 3)$$

כעת, נבדוק איזה נהודות נמצאות בסביבת המרכז החדש $(2,3)$ אפשר לראות בקלות שהנהודות A_3, A_4, A_5 נמצאות

בסביבה, ואילו הנהודות A_1, A_2 מחוץ.

כל שרמרכז לא השתרע ואם A_5 שויפת למרכז הצביות $(2,3)$

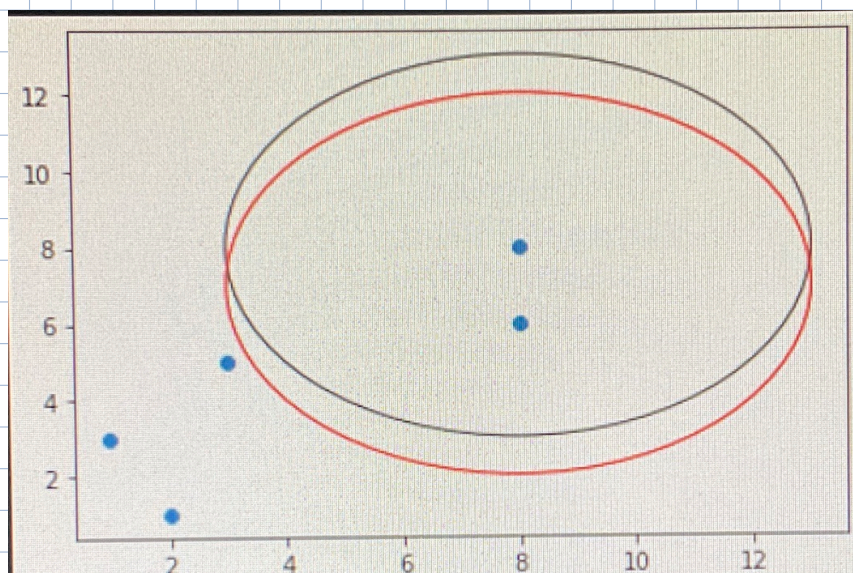
ב) היבטו של מרכז צפיפות שונים. הנקודות ששוויות לאותו מרכז צפיפות נמצאות באותו אשכול, ואם

נהיה שהאשכולות הם: $Cluster = \{A_3, A_4, A_5\}$ $Cluster = \{A_1, A_2\}$

ג) נתמיה עם הנקודה A_1 ונראה שמי שנמצא הרדיוס שלו זה רק A_1, A_2 .

אנחנו מחשבים מרכז ופיתחים סביבו עוד רדיוס נמצא (בצבא אדום) ואפשר לראות שהוא מכיל אותם נקודות

כלומר מרכז הצפיפות לא השתנה.



ניתן לראות שההתארה A_1 הוא המרכז של המאסף ואי שנמצאים ברדיוס הם רק A_1 ! A_2

כי מרחקם מ- A_1 (המרכז) קטן מ- H .

היחסים עבור הנקודות ברדיוס (A_1, A_2) את המרכז וסביבו פתחנו מאסף (בצבא אדום) כאשר המרכז

הוא הממוצע בין A_1 ל- A_2 . אפשר לראות שהרדיוס נשאר A_1 ! A_2 ואם זה מרכז הצפיפות של A_1 .

אזכיר כך עבור כל הנקודות.

ד) עבור $H=6$, אם נסתכל בדוגמא של A_1 , אז סתם A_3 נמצא בסביבתו. עבור $H=6$ ואז אנחנו מקבלים שהשכנים

של A_1 : A_1, A_2, A_3 וזה כן אלוף לשניה את התוצאה.

אזכור דבר גלוי. A_3 אם $H=6$ אז עבור A_3 , כל הנהודות נמצאות בדיוק.

ה) במקרה כזה, כל הנהודות תהיה אשכול. כאשר 5 אשכולות. מכיון שאזכור כל הנהודות, מרכז הצפיפות שלה הייתה

הנהודות עצמה.

① במקרה כזה, כולם נמצאים בסביבה של כלם ואם עבור כל הנהודות, הם ישתייכו למרכז הצפיפות:

$$\frac{A_1(x) + A_2(x) + A_3(x) + A_4(x) + A_5(x)}{5}, \quad \frac{A_1(y) + A_2(y) + A_3(y) + A_4(y) + A_5(y)}{5}$$

כאמור, היינו מקבלים אשכול אחד גדול, המכיל את כל הנהודות.

ג) הנהודות שהתקיימו היא הנהודות חריגה. המטרה היא לעבור הנהודות נהודות ולחשב את מרכז הצפיפות של כל אחד.

אז חזקתם הנהודות, והיבולו לפני שאזכור $H=5$ שני אשכולות. כעת משימים A_6 וקבוצה שהיא הנהודות חריגה

אז אין אז הנהודות סביבה ואם נקבע בסוף 3 אשכולות, כאשר A_6 היא אשכול בפני עצמה. ניקח לראות שהנהודות

החריגה לא הוסרה את האשכולות הנהודות.

$$\text{נקבע שהאשכולות הם: } \text{cluster}_1 = \{A_1, A_2\} \quad \text{cluster}_2 = \{A_3, A_4, A_5\} \quad \text{cluster}_3 = \{A_6\}$$

: Rand index

בהינתן תוצאות של שני אלגוריתמי סיווג, כיצד נמדדת רמת הסכמה ביניהם? - Rand index

#	c1	c2
A	1	2
B	1	2
C	1	1
D	2	1
E	2	3
F	2	3
G	2	4
H	2	4

מבחן:

נבנה טבלה בו הוצגו האותיות: $\alpha, \beta, \gamma, \delta$

	A	B	C	D	E	F	G	H
A	-	α	γ	β	β	β	β	β
B	-	-	γ	β	β	β	β	β
C	-	-	-	δ	β	β	β	β
D	-	-	-	-	γ	γ	γ	γ
E	-	-	-	-	-	α	γ	γ
F	-	-	-	-	-	-	γ	γ
G	-	-	-	-	-	-	-	α
H	-	-	-	-	-	-	-	-

* לדוגמה, (A,B) היו ביחד ב c_1 ושאריו ביחד ב c_2 ולכן לה α

* לדוגמה, (A,C) היו ביחד ב c_1 ולא שאריו ביחד ב c_2 ולכן לה γ

$$\alpha = 3 \quad \beta = 14 \quad \gamma = 10 \quad \delta = 1$$

$$\text{Rand index}(C_1, C_2) = \frac{17}{28}$$

תרגיל נוסף:

אם $a = b$! $\gamma = \delta$ אז הוכחנו:

(א) $\text{Rand index} = 1$

(ב) $\text{Rand index} < 1$

(ג) $\text{Rand index} = 0.5$

(ד) לא ניתן לדעת

פתרון:

סיוף ד' נכונה, כי $a = b$! $\gamma = \delta$ לא אומר כי שים דבר.

לדוגמא, $a = b = 5$! $\gamma = \delta = 0$ ובמקרה כזה אקראי $\text{Rand index} = 1$ אז פסלתי את ב' ו' $\frac{1}{2}$

אם $a = b = 5$! $\gamma = \delta = 1$ אז $\text{Rand index} = \frac{5}{6}$ אז פסלתי את א'.

כן, סיוף ד' נכון. כי a ! b הם מקדמים שחזרים ומשברים את התוצאה וזה שיהיה שווה, זה לא

אומר כי משהו נכדי התוצאה הסופית, אותו דבר נכדי $\gamma = \delta$.

תרגיל נוסף:

אם $a = 5$! $\beta = \gamma$ אז הוכיחו:

(א) $\text{Rand index} = 1$

(ב) $\text{Rand index} < 1$

(ג) $\text{Rand index} = 0.5$

(ד) אם ניתן, סדור

פתרון:

הוכיחו, סדרו ל נכון, כי אם $a = 5$! $\beta = \gamma$ אז המונח תמיד יהיה חצי ממוצע.

סדר למשל, $a = 5 = 4$! $\beta = \gamma = 2$ אז $\text{Rand index} = \frac{4+2}{4+2+4+2} = 0.5$

דוגמה נוספת, $a = 5 = 1$! $\beta = \gamma = 0$ אז $\text{Rand index} = \frac{1+0}{1+0+1+1} = 0.5$

רגרסיה

שאלות שיטתיות לבחינה אל רגרסיה:

לא באות מצבים מאותו מצב כלל מתיצית ולעצא את ההבדל וכן הלאה...

השאלות ברגרסיה יכולות להיות שאלות הבנה:

1) בהינתן הכול נתונים (x, y) האם רגרסיה יכולה מתאימה או קווקא אינאחית?

2) המקדמ של x^2 חיובי או שלילי? אם הנתונים זולים ואחר כך וורדים אז זה מניח, ואם הנתונים וורדים

ואז זולים, אז המקדמ חיובי.

3) שאלות אל רגרסיה לבחינה כן יכולות להיות שאלות בשאלות כמו למצוא את הו הרגרסיה האינאחית כי שם בתימנו

את העססה (אפשר למצוא בהליות את a ! b האופטימליים)

שאלת קוואד רגרסיה:

מספר חולה	כמות ייצור	איכות %
1	1	100
2	6	95
3	11	80
4	16	77
5	21	85
6	26	95
7	41	99

מצא את נוסחת הרגרסיה המתאר את איכות הייצור כפונקציה של הכמות:

א. $\hat{y} = -1.2645x + 99.324$

ב. $y = 2.1x^2 + 7.3x + 5$

ג. $y = -2.1x^2 + 7.3x + 5$

ד. $\hat{y} = 1.2645x - 99.324$

אנחנו מזהים שהרגרסיה לא אינאחית כי כל כסף ש-x (כמות) גדלה אז האיכות (y) לא גדלה בהכרח.

אם בסעיף את א' ! ד'. בעל, אנחנו מזהים שההתחלה יש ורידה, מתיצית למינאחית, ואז יש עלייה ואם

מקור אל רגרסיה יכולה להתקדם x^2 שמו חיובית ואם התשובה הוא ב'.

תרגיל:

בהינתן הנתונים הבאות:

	לג (הכנסה) (x)	לג (הוצאה) (y)
p ₁	25	23
p ₂	28	27
p ₃	35	31
p ₄	40	35
p ₅	29	25
p ₆	30	30
p ₇	37	32
p ₈	29	32
p ₉	28	25
p ₁₀	32	32

א) האם הרגרסיה שמשמרת היא רגרסיה אינארית או הומוגית? בנה את פונקציית הרגרסיה

ב) מהו הגודל של error?

ג) ע"י הרגרסיה שבנית, עבור $x=35$ (הכנסה), איזה $\hat{y}$ מתאים? נשים לב כי $x=35$ יש אותו בתעורים

האמיתיים וה- y שלו 31

ד) ע"י הרגרסיה שבנית, עבור לג (הכנסה) $x=31$, איזה x מתאים? נשים לב כי $y=32$ (הוצאה) בתעורים

האמיתיים וה- x שלו 31.

ה) האם הרגרסיה שמשמרת בו מבטא את ה- error הנמוך ביותר?

פתרון:

Ⓐ אפשר לזהות שזהו כל x שצד, y בצד גס. לכן, רגרסיה ליניארית היא מתאימה.

גם מנת למצוא את פונקציית הרגרסיה הליניארית הכי טובה (את a ! b האופטימליים), נשתמש:

$$a = \frac{\sum_{i=1}^n y_i x_i \cdot n - \sum_{i=1}^n y_i \cdot \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 \cdot n - \left[\sum_{i=1}^n x_i \right]^2} \quad b = \frac{\sum_{i=1}^n x_i^2 \cdot \sum_{i=1}^n y_i - \sum_{i=1}^n y_i x_i \cdot \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 \cdot n - \left[\sum_{i=1}^n x_i \right]^2}$$

$$\sum_{i=1}^n x_i = 25 + 28 + \dots + 32 = 313$$

$$\sum_{i=1}^n y_i = 23 + 27 + \dots + 32 = 292$$

$$\sum_{i=1}^n y_i x_i = 25 \cdot 23 + 28 \cdot 27 + \dots + 32 \cdot 32 = 9277$$

$$\sum_{i=1}^n x_i^2 = 25^2 + 28^2 + \dots + 32^2 = 9993$$

$$\left[\sum_{i=1}^n x_i \right]^2 = [25 + 28 + \dots + 32]^2 = 97969$$

$$a = \frac{9277 \cdot 10 - 292 \cdot 313}{9993 \cdot 10 - 97969} = 0.70$$

$$b = \frac{9993 \cdot 292 - 9277 \cdot 313}{9993 \cdot 10 - 97969} = 7.269$$

לכן, פונקציית הרגרסיה הליניארית היא: $\hat{y} = 0.70x + 7.269$

Ⓑ error נחשו זכור את ההמשל בין y של כל x אמיתי אל מול $\hat{y}$ הנחשו.

כל שזכור לזכור כל x נתון, זכור מה מתקבל עבור כל x ($\hat{y}$) ולעומת ההמשל.

$$\sum (y - \hat{y})^2 : \text{error (ובמקרה)}$$

② נציב $x=35$ (גיל היתן) ונראה איזה $\hat{y}$ (גיל הבנה) מקבלים: $S(35) = 0.70(35) + 7.269 = 31.77$

נראה כי בתנאים האמיתיים, $x=35$ נים, $y=32$ ואז הרגסיה שהבאנו $\hat{y}=31.77$ שזה קרוב מאוד.

③ המטרה היא זרזים $\hat{y}=31$ ונמצא את x (גיל היתן) המשותף:

$$31 = 0.70x + 7.269 \longrightarrow -0.70x = -23.73 \longrightarrow x = 33.9$$

נראה כי בתנאים האמיתיים, $y=31$ נים, $x=35$ ואז הרגסיה שהבאנו $x=31.77$ שזה קרוב מאוד.

④ כן, כי בייחוד את המסחר ועדי זה היכלו את a ! b ראובטמאיים שיתעם את הרגסיה אם ה- Error

העוק הייתה.

תרגיל נוסף:

בהינתן הנתונים הבאים, איזה מן התשובות נכונות?

x	y
3	21
1	20
5	19
9	7
0	19

Ⓐ $y = 3.2x + 8.3$

Ⓑ $y = -3.2x + 8.3$

Ⓒ $y = -0.30x^2 + 1.39x + 19.04$

Ⓓ $y = 0.30x^2 + 1.39x + 19.04$

פתרון:

המטרה היא לא באמת להשיג. אנחנו צריכים להסתכל על הנתונים ולראות שמדובר בהרסיה הכוללת את מקדמי

שלישי. מכיון שההתנהגות של x גורמת לעיירה ב- y ואז נעזרים למקסימום ואז לעיירה ב- x גורמת ירידה ב- y.

ללא שקל להיות שזה מתאים לנו להרסיה אנחנו אלא הכוללת את א' ב'.

ומכיון שהקדם יש לעיירה ואז ירידה ללא שיש נקודת מקסימום ואם המקדם של x^2 חיובית להיות שלילית.

תשובה ג' נכונה.

שאלות הבנה:

1) מה ההבדל בין רזל זהים ?

רזל הוא זרז שלא אמיתי אשר יטל להיות קוטר זרזות האמיתיות. מנגד, זרז חזק הוא זרז אמיתי אבל היצני.

2) האם חייבים לנתח K -means ! $mean - shift$

צריך לנתח K -means מכיוון שזה מתבסס על חישובי מרחקים. אלא אם כן סייט תהיו של האלמנטים. זהה, אין צורך לנתח K -means.

3) בהנחה של "מרכז", איזה א מועיל ?



$K=6$ או $K=3$ וכן האם יכולים להיות טובים.

4) האם באפשרותי הודע את וסל להספק חק בטכניקה המרחקים, כל עצם את הנקודות האמיתיות ?

כן, כי אנחנו מאחדים לפי מרחק ואז מאמצים את המרכז. אין בעיה שימוש בנקודות עצמם.

5) אם היינו $SS=10$ עבור נערים לסיומים ! $SS=20$ עבור נערים אחרים.

האם אפשר לראות בודאות ש- $SS=10$ יותר טוב ?

לא, SS אינו כלל מדד זהירות בין שני נערים שונים. וסל להיות שהיבט $SS=20$ בעצם שכיחות א נמוך מדי.

ואם היינו בוחרים א אבטיחו, היינו מקבלים SS נמוך מ- 10. כמו כן, יכול להיות כי היבט $SS=20$ מכיון

שיש זי כמות נערים גדולה וסלם החיבים למרכז שלהם אבל כמות הנקודות גדולה ואם מיקולם SS והסוף עבור

שאלות הבנה למבחן

שאלה 1:

מה מכין האופציות הבאות, וכל להיות טבעי לאחר הדעת אלגוריתם?

(א) נוצא ערכים חריגים

(ב) נשלים ערכים חסרים עם ערך אחד מהענפים

(ג) נחבר נהודות הנחות

(ד) נחבר אטכולות צמודים

פתרון:

אין צורך להוציא ערכים חריגים לאחר שהדעת נשלים ערכים חסרים זה רק לפני ההדעת.

התשובה - ד'. מכיון שאם נהבא אשכולות לאחר הדעת, אנחנו כן יכולים לחבר אותם אשכולות.

שאלה 2:

מה מכין האנשים הבאות נכון?

(1) צריך לקבוע ערך k מראש עבור האלגוריתם k -means.

(2) האלגוריתם k -means מחפש k ומחזיר לך אשכולות באופן צורך לספק לו k מראש.

(3) k -means טובה יותר בשימוש של סלמנטציות תמונות מאשר k -means.

(4) התשובות 1,2 נכונות

(5) התשובות 1,2,3 נכונות

שאלה 3:

הרצנו אלגוריתם של k -means ומסלף ההרצה היבט 3 אשטיות. בנוסף, הרצנו גם אותם נתונים $k=3$ גם

האלגוריתם k -means. היבטנו כי ה- SS של k -means נמוכה יותר מל k -means.

מה מבין הטענות הבאות נכונה ?

א) הייח סיבוי שלבור טע האלגוריתם, היבטנו 3 אשטיות לזהב.

ב) SS אינו מדד לשני אלגוריתם טונים.

ג) בוודאות לא מדובר גם אותם אשטיות בין שני התבאות המובות.

ד) התבאה המבית של k -means טובה יותר מל k -means.

ה) תשובות א' ! ד נכונות.

פתרון:

טענה א' לא נכונה - אין סיבוי. אם היינו מקבלים את אותם אשטיות בדיוק, אז ארץ ה- SS היה לזה בשני האלגוריתם.

טענה ב' לא נכונה - SS הוא כן מדד כל אחד מדובר גם אותם נתונים.

טענה ג' נכונה - נכון, כי ה- SS לא לזה.

טענה ד' נכונה - נכון, כי היבטנו SS נמוך יותר.

שאלה 4:

בהינתן 5 נהדרות, נענה על משיבת העיניים בהיבוא:

	A_1	A_2	A_3	A_4	A_5
A_1	0	15	14	19	28

מצא מי נמצא בסביבת $H=4$ עבור A_1 , זה האלמנטים $mean-smist$

Ⓐ לא ניתן לדעת כי מבקשים מרחק של 4 אבל הטבלה הנענה היא בהיבוא.

Ⓑ זה מנת לדעת את השנים של A_1 , אפשר לזכות שורש עבור כל דרך בטבלה ולבדוק מי קטן מ-4.

Ⓒ זה מנת לדעת את השנים של A_1 , אפשר לזכות חזרה ל-4 ולבדוק מי קטן מ-16

Ⓓ השנים של A_1 : $\{A_1, A_2, A_3\}$

Ⓔ תשובות ב', ג', ל' נענות

פתרון:

ידוע כי האלמנטים $mean-smist$ מתבססת על מציאת נהדרות "ברדיוס". הנדרות שמרחקם קטן מ-4

נמצאות בסביבה. מכיון שאנחנו רוצים לדעת מי נמצא בסביבת A_1 והטבלה הנענה היא בהיבוא, אפשר

לבדוק האם A_1 קטן מ-4 (כמו הטבלה) או לשים שורש על ידי הטבלה ולבדוק את החלקה.

בשני המקרים נקבל שהשנים של A_1 : A_1, A_2, A_3 .

שאלה 5:

מה מבין האנטי הבאות נכונה לגבי Rand index ?

- Ⓐ - כמות הזוגות שהיו ביחד ונשארו ביחד
- Ⓑ - כמות הזוגות שלא היו ביחד ונשארו לא ביחד
- Ⓒ - כמות הזוגות שהיו ביחד ונשארו לא ביחד.
- Ⓓ - כמות הזוגות שלא היו ביחד ונשארו ביחד.
- Ⓔ - כמות התשובות נכונות.

שאלה 6:

אם $0.5 < \alpha + \beta$, בחר את התשובה הכי מדויקת:

Ⓐ $\text{Rand index} = 0.5$

Ⓑ $\text{Rand index} = 1$

Ⓒ $0 < \text{Rand index} < 1$

Ⓓ $0.5 < \text{Rand index} < 1$

פתרון:

טענה א לא נכונה - אם $\alpha + \beta = 1$ אז $\text{Rand index} = 0.5$

טענה ב לא נכונה - אם $\alpha + \beta = 0$! $\alpha + \beta \geq 1$ אז $\text{Rand index} = 1$

טענה ג נכונה - תמיד מתקיים אבל לא התשובה הכי מדויקת.

טענה ד נכונה - אם $\alpha + \beta < 1$ אז התוצאה בהכרח בין 0.5 ל-1.

שאלה 7:

מה מבין הטענות הבאות נכונה?

א) האלגוריתם K -means לא טוב עבור נתונים שאין מבוסס, מרכזים

ב) האלגוריתם K -means מושג מבחינת מרכזים התפלגותיים

ג) האלגוריתם K -means Bisection אינו מושג ממרכזים התפלגותיים.

ד) כל התשובות נכונות

שאלה 8:

מה מבין הטענות הבאות נכונה?

א) כאשר מחשבים $f(x)$, התוצאה חייבת להיות הטענה או שווה ל-1.

ב) אטמא מבוסס מרכזים ביחסו של נהדר, המרכז הכי קרוב אליה זה המרכז אלווה הוא שייכת.

ג) האלגוריתם K -means מבוסס על צפיפות

ד) בהינתן נהדרות שמראות נתונים בצורה אינאליי, אפשר לבנות רגרסיה ליניארית לרצף ויתר מרגרסיה ריבועית

למרות שהנתונים הם אינאליי

ה) אין תשובה לא נכונה

ו) תשובות א, ב, ג נכונות

במבחן:

טזנה א' נכונה - מבין שמחזירים ב' ו' אז אנחנו דואלים שזה יהיה בטוח של 1-5.

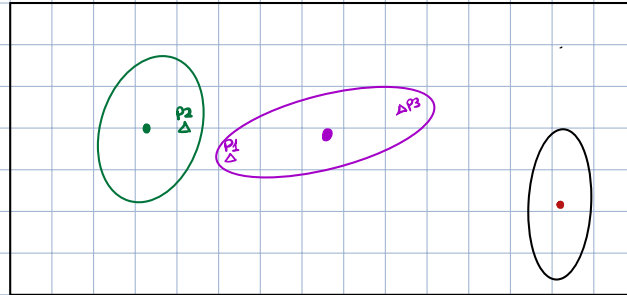
טזנה ב' נכונה - זה בירוש של אטמא מבוסס מרכזים

טזנה ג' לא נכונה - K -means הוא זה שמבוסס צפיפות

טזנה ד' נכונה - למ אב הנענים מראים הו אינאליי, אנחנו וכלים לבנות רגרסיה ליניארית לא אובטימיזציה בכלל ורמיקה מאוד

מהנהדרות האמיתיות

התקבלו האשמות הבאות:



יש לחשוב שכל אשמה מניח הורה נהוגות בתוכו. (הנהגות האדומות הם מרכזים)

מה מבין האשמות הבאות נכונה?

☐ א. מופרד היטב ומבוסס מרכזים

☐ ב. מופרד היטב אבל לא מבוסס צפיפות

☐ ג. לא מופרד היטב אבל מבוסס צפיפות

☐ ד. לא מופרד היטב ולא מבוסס צפיפות

בתשובה:

זה לא מופרד היטב כי צפיפות P_1 החברה ויתר P_2 (שהיא מאשמת אחר) מאשר P_3 .

זה לא מבוסס מרכזים כי P_1 החברה ויתר למרכז של האשמה היחיד, מאשר המרכז שלה.

שאלה 10:

מה מבין הטענות הבאות נכונה ?

- א) אם נתון כי נתונים מכו מומד, אז פונקציית הצפיפות תהיה מכו מומד.
- ב) אם נתון כי נתונים מכו מומד, אז פונקציית הצפיפות תהיה מחד מומד.
- ג) אם נתון כי נתונים מכו מומד, אז פונקציית הצפיפות תהיה מתפת מומד.
- ד) אם נתון כי נתונים חד מומד, אז פונקציית הצפיפות תהיה מחד מומד.

פתרון:

תמיד, פונקציית הצפיפות דוגה אחת מוהנתיב. אם הנתונים שלי הם חד מומד (אמור אחת), אז פונקציית הצפיפות תהיה מכו מומד, וכן הלאה...