כלל אצבע - כמה מחלקות יש בסוף

$$\text{Information gain} = \text{Info} - \text{Info}_{\text{new}} \quad \text{אבל} \quad \text{Info} = -\sum p \log p = \sum p \log \frac{1}{p}$$

- קל יותר לחשב את ה-Information gain כי אין צורך לחשב את ה-Info של כל הנתונים.
- כמות פעמים שצריך לחשב את ה-Info תהיה כמו מספר הערכים השונים של התכונה.
- כמה יחידות מידע מכילה התכונה ביחס לכל הנתונים. זה ה-Info.

לפני חלוקה לפי התכונה

מה ההבדל בין dendrogram לבין histogram? ב-dendrogram יש ציר y שמייצג מרחק בין אשכולות, ובהיסטוגרמה יש ציר y שמייצג כמות.

מה ההבדל בין אשכול לבין אשכולות?
אשכול: קבוצה של נקודות. אשכולות: קבוצה של אשכולות.

מה ההבדל בין אשכול h לבין אשכולות h?
אשכול h: קבוצה של נקודות שהמרחק ביניהן קטן מ-h.
אשכולות h: חלוקה של כל הנתונים לאשכולות כך שהמרחק בין כל שתי נקודות באשכול שונה הוא לפחות h.

מה ההבדל בין אשכול מרכזי לבין אשכול צפיפות? אשכול מרכזי: כל הנקודות קרובות למרכז. אשכול צפיפות: כל נקודה קרובה לנקודה אחרת באשכול.

- כ"כ מאד - מקרה מיוחד

- הוא שהמטרה של האלגוריתם היא למצוא כמה שיותר קלאסטרים.
  יש לזכור כי, לכל מחלקה יש קלאסטר אחד או יותר שלה בלבד.
- ניתן לחשב את האלגוריתם עבור מספר k שונים ולבחור את הטוב ביותר לפי מדד כלשהו. (אשר k הכי טוב עבור הנתונים).
- לא ניתן לדעת מראש מהו k הטוב ביותר, אלא לפי מה שמתאים לנתונים.
  כמה נתונים מסוימים מראה מקרה שבו אלגוריתם לא טוב כ"כ.

* כאשר k הוא גדול מדי כמות הנקודות בכל קלאסטר תהיה קטנה מדי, ולכן יהיה קשה לזהות מבנים כלליים יותר בנתונים.

- יתרון של k-means - מהירות ויעילות גבוהה.
  חסרון - יכולת ביצועים לא טובה כאשר הנתונים מורכבים.
- WSS - מדד לבדיקת איכות האשכולות (מרחקים)
  כלומר, סכום כל המרחקים בין כל נקודה לבין מרכז הקלאסטר שלה.
  ערכים נמוכים מעידים על קלאסטרים יותר הומוגניים.
- WSS מדד המשמש לחישוב הפרש בין ערכים k כך ש-
  שיפור משמעותי בערך המדד יתקבל.
- silhouette - מדד המשמש לבדיקת איכות האשכולות ומודד את מידת הדמיון של נקודה לקלאסטר
  שלה ביחס לקלאסטרים אחרים.
  ערכים גבוהים מעידים על כך שהנקודה מתאימה לקלאסטר שלה.

* כאשר הנתונים מכילים רעש קשה לזהות את מרכז הקלאסטר ולכן יהיה קשה לחשב את ערך המדד.

- יש להשתמש בשיטות מיוחדות לטיפול בערכים חריגים.
  יש לזכור כי, לכל מחלקה יש לפחות נקודה אחת.
- mcs DB מהו המרחק הגבוה ביותר שאליו יש לחפש שכנים לנקודה מסוימת.
  (אשר k הכי טוב עבור הנתונים).
  לעומת זאת, קשה מדי.
- לעיתים קשה לדעת מראש מהו k הטוב ביותר עבור הנתונים.
  כאשר הנתונים מכילים הרבה רעש קשה לזהות את מרכז הקלאסטר ולכן יהיה קשה לחשב את ערך המדד.
- כאשר k הוא גדול מדי כמות הנקודות בכל קלאסטר תהיה קטנה מדי, ולכן יהיה קשה לזהות מבנים כלליים יותר בנתונים.

כריית מידע - תרגול

FT-Tree

כאשר נתונים עבור ה dataset הבא בנה FP-tree עבור (minsup = 2):

{A,B} {A,C,D} {B,D} {A,B}

A = 3 סופרים שמופיעים

B = 3

C = 1 לא מופיע ב minsup

D = 2

אם נשאל: האם יתכן כאשר מעמיקים בעץ נמצא ש minsup(אחיך שלך) < minsup(אבא שלך)?
התשובה היא: לא יתכן. אחיך מופיע לפחות פעמים כמו אבא שלך.
כלומר אבא שלך מופיע לפחות פעמים כמו כל אחד מהילדים שלו.

עץ שלם:

הערה: ייתכן שקבוצת פריטים מסוימת תופיע בעץ פעמים.

{A} {A,B} {A,C} {A,B,C} :FP-tree כאשר נתונים כאן minsup=2

A : 4 כולם מופיעים מעל ה minsup

B : 2

C : 2

{A} {A,B} {A,B,C} {A,B,C,D}

cout = $\frac{1}{4}$

conf = $\frac{3}{4}$

conf = $\frac{1}{1}$

conf = $\frac{2}{3}$

confidence = $\frac{\text{מספר הפעמים שהכלל מתקיים}}{\text{מספר הפעמים שהתנאי מתקיים}}$

$X \to y$ $\quad conf = \frac{p(x \cup y)}{p(x)}$ $\quad Confidence(A \Rightarrow B) = p(A|B) = \frac{p(A \cup B)}{P(A)}$

1. $A \to D$
2. $A \to B$
3. $D \to A$ ✓
4. $B \to C$

3. נמצא את NN הקרוב ביותר שיש לו תווית שונה מהנתון.
   כל הנקודות שביניהם.

* 3. נמצא את NN הקרוב ביותר שיש לו תווית שונה מהנתון.
   (ניתן לבחור גם את ה-k הקרובים ביותר שיש להם תווית שונה).
   מה נעשה בכל אחד? (כיצד נסווג נתון חדש שנמצא ליד הנתון שבחרנו).
   נסווג את הנתון החדש בתווית של NN שלו.
   נסווג את הנתון החדש בתווית של רוב ה-k השכנים שלו.

* א. מהו ההבדל בין 1-NN ל-3-NN?
   1. Simple-X (כמו בתרגיל הבית) - מחשב את ה clustering (ה-centroid של כל אשכול).
   2. RIPPER - לומד כללים מתוך כלל ה clustering.
   3. EM - מבצע clustering
   4. Cobweb - לומד ה clustering

* מהן הנקודות החשובות בהרצת האלגוריתם DBSCAN:
   א. כל נקודה שלא נמצאת בטווח של אף נקודת core היא outlier.
   ב. כל נקודה שנמצאת בטווח של נקודת core, אך היא עצמה לא נקודת core, היא נקודת border.
   ג. כל נקודה שנמצאת בטווח של יותר מנקודת core אחת תשויך לאשכול של אחת מהן.

* הסבר את הרעיון מאחורי ה-kernel trick ב-SVM.
   הרעיון הוא למפות את הנתונים למרחב בעל ממד גבוה יותר, כך שבמרחב החדש הנתונים יהיו ניתנים להפרדה לינארית.
   ה-kernel trick מאפשר לנו לחשב את המכפלה הפנימית במרחב החדש מבלי לדעת את הפונקציה המדויקת שממפה את הנתונים למרחב החדש.

* ב WAS חלוקה ל KAWA ONS

- כלל יד הלוט - שיטה שם אחד לכל המחלקה ופעל ומחזיר ערכים שונים (מחלקה 1>2).
לכן כאן יש רמזים שם הפעלת הפונקציה. זה ב AWA אבל בפונקציה אחרת שם הפונקציה, לא ממש ניתן לדעת, רמזים שם x מאפיין לא כאן אבל בפונקציה.

* נוסחאות AWA - אלגוריתם לחישוב הסתברויות, רמזים שם x מאפיין לא כאן אבל בפונקציה.

סמכות: חוזק כלל $\Rightarrow$ חוזק $sup <$ (כלל) dns

כלל 1: support של חוזק כלל יהיה מספר האובייקטים שם support שם כלל בפונקציה - זהו.

כלל 2: support של חוזק כלל מספר האובייקטים יהיו יחסית שם support שם B (כלל $\leq$ B).

כלל 3: support של חוזק כלל מספר האובייקטים יהיו יחסית שם support שם A (חוזק > A)

כלל 4: support של חוזק כלל יהיה מספר האובייקטים יהיו יחסית שם support שם A ו B יחד. (חוזק > כלל $\Rightarrow$ סמכות: sup (כלל) < sup (חוזק))

3. A ו B יחד. itemset אחד, dataset, אחד מופיעים יחד.

support (חוזק כלל) $\geq$ support (כלל) (כלל חוזק יהיה יחסית לכלל כמה support יש לו).

* support - אחוז מהפעמים שבהם ה dataset שלנו מכיל אותו.

תלמידה לכל אחד שיטת חוזק שונה. - לאחר שהיה אחד מהם - כאשר הוא משתמש כאחד מהשיטות לחישוב חוזק כלל, הוא מכפיל כלפי כל אחד מהשיטות את הערך שלו ומחלק בסכום כל הערכים שלו, כדי שהסכום יהיה 1.

לאחר מכן הוא מחשב את הממוצע של כל אחד מהשיטות ומחזיר את הכלל עם הממוצע הגבוה ביותר.

Laplace Smoothing - הוסף מספר של מופעים כל ערך ו/או כל data set לחישוב חוזק כלל. כאשר יש ערכים שלא מופיעים כלל בדוגמאות, הוסף 1 לכל ערך כדי שלא יהיה 0 בחישוב.

M-estimate Smoothing - הוסף מספר של מופעים כל ערך ו/או כל data set לחישוב חוזק כלל. כאשר יש ערכים שלא מופיעים כלל בדוגמאות, הוסף $\frac{1}{k}$ כדי שלא יהיה 0 בחישוב.

* WAS - אלגוריתם לחישוב - אבל שהוא יודע כמה מהשיטות לחישוב חוזק כלל, הוא מחשב את הממוצע של כל אחד מהשיטות ומחזיר את הכלל עם הממוצע הגבוה ביותר.

אלגוריתם - כללי אלגוריתם - רק עם נתונים

כלי עזר - מילות מפתח.

* מילות עצירה Stop Words - מילים שלא נרצה שיופיעו בענן המילים, כי הן לא מעניינות אותנו.

- מילות קישור - מילים שמחברות בין מילים אחרות, כמו: אבל, אם, כי, אז, ועוד.
- מילות Stop Words - מילים שכיחות מאוד בשפה, כמו: את, של, על, עם, ועוד.

ג. אפשר להשתמש ב dataset מוכן מראש בתוך WEKA בשביל לבדוק את accuracy של המודל.
ד. מהי מילות Stop Words? מילים שכיחות מאוד בשפה, שלא נרצה שיופיעו בענן המילים.

נשאלת השאלה כי $\frac{5}{100} < 5\%$ (כלומר פחות מ 5% מהמילים הן מילות עצירה).

* DBSCAN - אלגוריתם 3 פרמטרים (מספר נקודות מינימלי, רדיוס ε) 1 epsilon (מספר נקודות מינימלי, רדיוס ε).
core - קודקוד שיש בסביבתו 3 או יותר epsilon. (קודקוד שיש בסביבתו 3 או יותר epsilon).
border - קודקוד שיש בסביבתו פחות מ 3 נקודות אך הוא שכן של קודקוד core (קודקוד שיש בסביבתו 3 או יותר epsilon).
noise - קודקוד שלא שייך לאף אחת מהקבוצות.

- מציין שקודקוד הוא קודקוד core ומספר השכנים שלו בתוך epsilon.

* מספר הקודקודים בגרף הוא 4 ולכן המטריצה היא בגודל 4×4.
לא נתון לנו מספר הקשתות בגרף, אך ניתן לחשב אותו לפי סכום כל הערכים במטריצה חלקי 2.
כל קשת מופיעה פעמיים במטריצה, פעם אחת בשורה של הקודקוד הראשון ופעם אחת בשורה של הקודקוד השני.
לכן, כדי לקבל את מספר הקשתות בגרף, יש לחלק את סכום כל הערכים במטריצה ב 2.

- מטריצת הסמיכויות של גרף לא מכוון היא תמיד סימטרית ביחס לאלכסון הראשי.
- מטריצת הסמיכויות של גרף לא מכוון היא תמיד סימטרית ביחס לאלכסון הראשי, כי אם יש קשת בין קודקוד i לקודקוד j, אז יש קשת בין קודקוד j לקודקוד i.
- מטריצת הסמיכויות של גרף מכוון לא בהכרח סימטרית ביחס לאלכסון הראשי, כי אם יש קשת מקודקוד i לקודקוד j, לא בהכרח יש קשת מקודקוד j לקודקוד i.

* PCA - אלגוריתם שמטרתו להפחית את מספר המאפיינים במודל, תוך שמירה על המידע החשוב ביותר.
כדי לבצע PCA יש לבצע נרמול לנתונים, כי אלגוריתם PCA רגיש לסקאלה של המאפיינים.
לאחר מכן, מחשבים את מטריצת השונות המשותפת (Covariance Matrix) בין המאפיינים.
לאחר מכן, מחשבים את הערכים העצמיים והווקטורים העצמיים של מטריצת השונות המשותפת.
הווקטורים העצמיים מייצגים את הכיוונים שבהם השונות בנתונים היא הגדולה ביותר.
הערכים העצמיים מייצגים את כמות השונות שמוסברת על ידי כל וקטור עצמי.
לבסוף, בוחרים את k הווקטורים העצמיים עם הערכים העצמיים הגדולים ביותר, ומשתמשים בהם כדי להקרין את הנתונים למרחב חדש עם k מאפיינים.

* מטריצת השכנויות של גרף מכוון היא תמיד סימטרית ביחס לאלכסון הראשי.
מטריצת השכנויות של גרף לא מכוון היא תמיד סימטרית ביחס לאלכסון הראשי, כי אם יש קשת בין קודקוד i לקודקוד j, אז יש קשת בין קודקוד j לקודקוד i.
מטריצת השכנויות של גרף מכוון לא בהכרח סימטרית ביחס לאלכסון הראשי, כי אם יש קשת מקודקוד i לקודקוד j, לא בהכרח יש קשת מקודקוד j לקודקוד i.

כ"ר מ"ף - סדנ"ל תשפ"ה

3. כה: נמצא את ה-MVS עבור הנקודות הבאות:

$x_1=(0,0)$	$y_1=-1$	שלילי
$x_2=(0,1)$	$y_2=1$	חיובי
$x_3=(0,2)$	$y_3=1$	
$x_4=(0,3)$	$y_4=1$	
$x_5=(0,4)$	$y_5=1$	

מה הוקטורים התומכים?

1. $\alpha_1=2 \quad \alpha_2=-2 \quad \alpha_3=0 \quad \alpha_4=0 \quad \alpha_5=0$
2. $\alpha_1=1 \quad \alpha_2=1 \quad \alpha_3=0 \quad \alpha_4=0 \quad \alpha_5=0$
3. $\alpha_1=2 \quad \alpha_2=2 \quad \alpha_3=0 \quad \alpha_4=0 \quad \alpha_5=0$
4. $\alpha_1=1 \quad \alpha_2=2 \quad \alpha_3=3 \quad \alpha_4=4 \quad \alpha_5=5$

- $\alpha_i>0$ אם ורק אם x_i הוא וקטור תומך. אחרת $\alpha_i=0$.
- $0<\alpha<C$ הוא ולכן שולחה הוא ו לוגו $y_i(w\cdot x_i+b)=1$ שהוא מישור תומך.
- כי מינימליות אף וקטור לא נכנס.
- (כדי לוודא את הנתונים שהוא למעלה מהם מ"ב!)
- סכום ה-α שווה ל $\alpha_1=\alpha_2$ ולכן שווה גם ל-α_3 שהוא 0.

$\alpha=2$ כל הנתונים נכונים כי!

$-\alpha^2+2\alpha=1$ $\quad$ $\frac{1}{2}\alpha^2-2\alpha=-1$ $\quad$ $\alpha-2=0$

$L=2\alpha-\frac{1}{2}\cdot\alpha^2\cdot 1\cdot 1\cdot(0,1)(0,1)=2\alpha-\frac{1}{2}\alpha^2=1$

$L=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\cdot\alpha_j\cdot y_i\cdot y_j\cdot(x_i\cdot x_j)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\|w\|^2$

- נוסחה לחישוב: $(x_1,y_1),(x_2,y_2),(x_3,y_3)$ כאשר x_3 הוא וקטור תומך ו α_3 שווה ל-0.
- נוסחה לחישוב: $(x_1,y_1),(x_2,y_2),(x_3,y_3)$ כאשר x_3 הוא וקטור תומך ו α_3 שווה ל-0.

שאלות אמריקאיות - כריית מידע

8 נתונים אלו נבנתה bootstrap על dataset בגודל
{A,A,B,B,B,D} {A,B,A,C,C,B} {D,D,C,A,A,A}

א. אחת מהאפשרויות הבאות אינה יכולה להיות ה dataset
1. {A,B,C,D,E,F,G,H}
2. {A,A,B,B,C,C,D,D}
3. {A,B,C,C,B,A,F,D}
4. {A,B,C,E,F,G,H}

ב. הסבר: בכלל שכזאת אין שום סיכוי שהערך D יהיה כפול יותר משתי פעמים בדאטה המקורי.

* * *

1 {M,O,N,K,E,Y}
2 {D,O,N,K,E,Y}
3 {M,A,K,E}
4 {M,U,C,K,Y}
5 {C,O,O,K,I,E}

דוגמה: בקובץ 5 הפריטים הבאים
min_sup = 60% , min_conf = 75%

- עלינו למצוא את כללי האסוציאציה שעבורם מתקיים
קבוצת הפריטים.
- שימו לב שהסדר אינו חשוב.
- בכל שלב נשמיט את כל הפריטים שאינם עומדים ב min_sup

M = 3/5 = 60%
O = 4/5 = 80%
~~N = 2/5 = 40%~~
K = 5/5 = 100%
E = 4/5 = 80%
Y = 3/5 = 60%
~~D = 1/5 = 20%~~
~~A = 1/5 = 20%~~
~~U = 1/5 = 20%~~
~~C = 1/5 = 20%~~
~~I = 1/5 = 20%~~

{M,O} = 1/5 = 20%
{M,K} = 3/5 = 60%
~~{M,E} = 2/5 = 40%~~
~~{M,Y} = 2/5 = 40%~~
{O,K} = 3/5 = 60%
{O,E} = 3/5 = 60%
~~{O,Y} = 2/5 = 40%~~
{K,E} = 4/5 = 80%
~~{E,Y} = 2/5 = 40%~~
{K,Y} = 3/5 = 60%

~~{M,K,O} = 1/5 = 20%~~
~~{M,K,Y} = 2/5 = 40%~~
{O,K,E} = 3/5 = 60%
~~{Y,K,O} = 2/5 = 40%~~

$(X \to y) : \frac{support(X \cup y)}{support(X)}$

$M \to K = \frac{3}{3} = 1 > 75\%$ ✓
$O \to K = \frac{3}{4} = 60\% < 75\%$ ✗
$O \to E = \frac{3}{4} = 100\%$ ✓
$K \to E = \frac{4}{5} = 80\%$ ✓
$K \to Y = \frac{3}{5} = 60\%$ ✗ לכן

כל הצמדות שנותרו (מלבד K→M ו M→K) להלן שאינם מתקיימים.
ג) רשום את כל ה frequent itemset שהתקבלו בסעיף הקודם (הצמד המקסימלי) וחשב את confidence עבור כל אחד.

- sklearn - ספרייה מלאה בדגמים (וגם כלי עזר שמשמשים גם לשלב EM)
- EM - בכל שלב חישוב, בדומה לכך שמבצעים (הערכה של פרמטרים מוכללים).
  ולאחר מכן מרווח מלאה בחלוקה.
- DBSCAN - יוצר צורות מלא מאוד מהירות ומהירות כי מסתכל על הפיזור, מתעלמים מנקודות

- average distance - הוא ממוצע המרחק בין הנקודה לכל שאר הנקודות בקלאסטר ו בין כל הנקודות בקלאסטר השני. מחשבים את כל המרחקים ומחלקים במספר הנקודות, כלומר מחשבים את ממוצע המרחקים.
- min/max - בדיוק ההפוך מה שקורה ב- average distance.

1. RandomForest - מסווג טוב.
2. SMOTE - מסווג clustering
3. SVC - מסווג טוב
4. OneR - לוקח מסווג אחד.

* מי יבין בשם מסווג המחלקה הטובה?

ניתן לחלק מסווגים ל 2 סוגים שונים: מחלקה ספציפית שניתן לה את כל המידע. לבין מסווגים שבהם יש לנו מחלקה מסוימת שבה כל הנקודות, ולא נשמרים שום מידע על הנקודות.

- כשיש לנו Naive Bayes ונרצה לחשב מסווג מסוים, נחשב את ההסתברות לכל אחד מהם.
  מה נשמרים בכל? נשמרים רק הפרמטרים.
  לכל instance שנרצה לסווג, נחשב את ההסתברות לכל אחד מהם.
  ואז, ה instance יסווג לאותה מחלקה שבה ההסתברות הכי גבוהה.

כלי מיון - שיטות מתקדמות

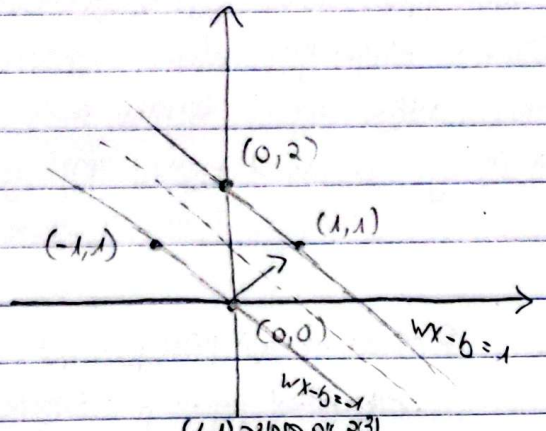
تمرين ٣: (٤ نقاط)

نعتبر المعطيات التالية:

$x_1=(0,0)$ $x_2=(1,1)$ $x_3=(0,2)$ $x_4=(1,-1)$ ، $w=(1,1)$ و $b=1$

1) ما رأيك في q؟

1	0
2	1
3	-1
4	2



$wx-b=1 \Rightarrow (1,1)(1,1)-b=1 \Rightarrow 2-b=1 \Leftrightarrow \underline{\underline{b=1}}$

$wx-b=-1 \Rightarrow (1,1)(0,0)-b=-1 \Rightarrow \underline{\underline{b=1}}$

$(0,2)(1,1)-b=1 \Rightarrow 2-b=1 \Rightarrow \underline{\underline{b=1}}$: أخذ من مجموع الجهة الأخرى

$(-1,1)(1,1)-b=-1 \Rightarrow -1+1-b=-1 \Leftrightarrow \underline{\underline{b=1}}$

- لماذا يمكن أن نأخذ أحدهما

* * *

أعطى الأستاذ الجدول

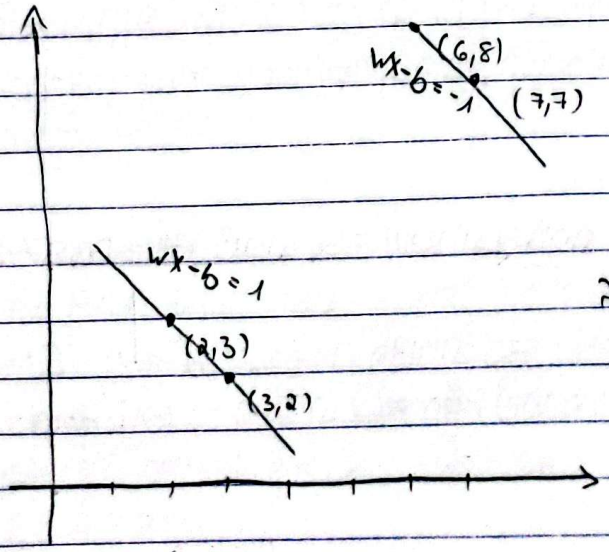
$x_1=(2,3)$ $y=1$

$x_2=(3,2)$ $y=1$

$x_3=(6,8)$ $y=-1$

$x_4=(7,7)$ $y=-1$

هذا كل ما يتذكره. هل يمكن حل التمرين؟



$wx-b=1$

$wx-b=-1$

x	y
0	1
1	3
0	0
3	1
3	3
3	0
5	2.5

MinPts = 2
ε = 1.5