Clustering

- Clustering = קיבוץ אובייקטים דומים.
- נרצה לא רק לקבץ אובייקטים דומים אלא גם להרחיק זה מזה אובייקטים שונים.

Hierarchical Clustering

- Start with every point is its own cluster.
- Find the most similar pair of clusters.
- Merge the most similar pair of clusters into a parent cluster.
- Repeat until you've reached the number of clusters you wish.

שאלה

נתונות נקודות הנמצאות על קו (מרחב חד-מימדי) במספרים הראשוניים: 2,3,5,7,11,13,17,19,23,29.

לצורך חישוב מרחק נשתמש במרחק אוקלידי. נרצה לבצע hierarchical clustering לנקודות הנתונות, תוך שימוש ב-centroid לצורך ייצוג מיקום של cluster. מצא את ה-clusters כאשר מספר ה-clusters הוא 3.

Similarity between clusters

$$x(i)=(x_1(i), x_2(i), \ldots, x_p(i))$$

- Euclidean distance:

$$d_E(i,j)=\sqrt{\sum_{k=1}^{p}(x_k(i) \quad x_k(j))^2}$$

- Jaccard distance:

$$d_J(i,j)=\frac{a}{a+b+c}$$

$$a=\left|\{t|x_t(i)=1 \wedge x_t(j)=1\}\right|$$
$$b=\left|\{t|x_t(i)=0 \wedge x_t(j)=1\}\right|$$
$$c=\left|\{t|x_t(i)=1 \wedge x_t(j)=0\}\right|$$

- Cosine distance:

$$d_C(x(i),x(j))=\frac{\sum_{k=1}^{p} x_k(i)x_k(j)}{\sqrt{\sum_{k=1}^{p} x_k^2(i)\sum_{k=1}^{p} x_k^2(j)}}$$

שאלה

נתונים 5 ווקטורים המייצגים רכישות ספרים ע"י 10 לקוחות:

0101010101=A, 0011001100=B, 1110010001=C, 1001001001=D, 0001111100=E.

א. חשב Cosine distance בין הוקטורים.

ב. חשב Jaccard distance בין הוקטורים.

K-means

- Choose an integer k (number of clusters).
- Randomly guess k cluster center locations.
- For each data point – find the center it's closest to.
- For each center – find the centroid of the points it owns.
- For each data point – relocate the point to the centroid it's closest to.

ID3

- ID3 זוהי תכנית ללמידת מסווגים.
- התכנית מקבלת כקלט קבוצת דוגמאות ומוציאה כפלט עץ החלטה המסווג נכון את כל הדוגמאות שנצפו.
- התכנית מתחילה משורש העץ ומפתחת תתי-עצים.
- צומת יפוצל אם הוא מכיל דוגמאות חיוביות ושליליות.

ID3- cont'

- כל צומת מתפצל על תכונה מסוימת. התכנית יוצרת תת עץ לכל ערך אפשרי של התכונה. קבוצת הדוגמאות מתחלקת לפי ערכי התכונה ותת-הקבוצות מועברות לתת העצים.
- ההחלטה על איזו תכונה לפצל מתבססת על יוריסטיקה המשתמשת בתורת האינפורמציה: נבחרת התכונה המביאה לתוספת מקסימלית של אינפורמציה.

ID3- cont'

- בכל צומת ID3 בודקת על איזו תכונה כדאי לפצל.
- לכל תכונה נמדדת תוספת האינפורמציה

האינפורמציה הגדולה ביותר נבחרת לפיצול.

Let: n_t *be the total number of examples*

n_b *be the total number of examples in branch b*

n_{bc} *be the number of examples in branch b that belong to class c*

$$Average\ disorder = \sum_b \frac{n_b}{n_t}\left(\sum_c \frac{n_{bc}}{n_b} \log_2 \frac{n_{bc}}{n_b}\right)$$

שאלה

ידרג אותו בתור 1, אחרת – 0. הסטודנטית מיכל יודעת אילו דיסקים היא אוהבת ואילו לא. ברצונה להשתמש בדירוג של ארבעת המומחים כדי להחליט האם יתכן כי היא אוהבת דיסק מסוים.

1101, 0100, 0111 (הביט ה-i מציין האם המומחה ה-i אהב את הדיסק)

א. ברצוננו לבנות עץ החלטה עבור מיכל, איזו תכונה (דעה של איזה מומחה) נשים בתור שורש?

ב. באיזו תכונה נשתמש לבניית הרמה השנייה בעץ עבור כל אחד מהבנים של השורש?

שאלה מסכמת

הצגנו מספר שיטות שונות לכריית נתונים: מציאת פריטים תדירים, מציאת זוגות דומים, clustering ובניית עצי החלטה.

עבור הסיטואציות הבאות, הצע את השיטה המתאימה ביותר לכריית הנתונים:

- רשות שדות התעופה שמרה מידע על כל בעלי דרכון ישראלי שיצאו לחו"ל ב-10 השנים האחרונות. נתונות 1000רשומות של בעלי דרכון ישראלי החשודים כטרוריסטים ו-1000 רשומות של בעלי דרכון אשר אינם חשודים כטרוריסטים. ברצוננו לדעת האם אדם מסויים הוא טרוריסט עפ"י קבוצת המדינות בהן ביקר.
- משרד הנסיעות Orbitz.com שמר מידע על המקומות בהם ביקרו כל לקוחותיהם. ברצונם להחליט, בהתבסס על המקומות בהם בילה הלקוח בעבר איזו חבילה להציע לו כך שירצה לקנות אותה.

שאלה מסכמת - המשך

1. Amazon.com מציעים ללקוחותיהם עסקאות tie-in: כאשר לקוח בוחן ספר A, הם מציעים לו לקנות את הספרים A ו-B במחיר של X$ בלבד. כיצד יצליחו לבחור בשילוב ספרים מוצלח לעיסקה?
2. eBay שואפת לשפר את סיווג המוצרים למכירה ע"י בחינת המוצרים אותם קנה לקוחאשר נבדקו אך לא ניקנו ע"י אותו לקוח.
3. eTrade מעוניינת למצוא זוגות של מניות כך שכשאחת מהן עולה בזמן מסויים, האחרת אינה עולה בתקופת זמן קצרה.