Market-Basket Model

Market-Basket Model

- נייצג את המידע כמטריצה בוליאנית.
- שורה = סל (basket)
- עמודה = מוצר (item)
- 1 בכניסה (i,j) מציין כי סל (לקוח) i קנה את מוצר j, 0 בכניסה זו מציין כי מוצר j לא נקנה בסל i.

Market-Basket Model-cont'

- **Support** עבור קבוצת מוצרים I = מספר הסלים המכילים את I.
- מוצרים שכיחים (frequent items) = בהינתן s, סף ה-support, מוצרים שכיחים הם קבוצות מוצרים המופיעים בלפחות s סלים.

Market-Basket Model-cont'

- נתעניין במציאת חוקים: **association rules** – המקשרים בין קניית מוצרים.

$$\textit{Association Rule: } \{i_1, i_2, \ldots, i_k\} \Rightarrow j$$

- מידת הביטחון של association rule היא ההסתברות של j בהינתן $i_1, \ldots, i_k$.

A-Priory Algorithm

- Pass 1: Read baskets and count in main memory the occurrences of each item.

- Pass 2: Read baskets again and count in main memory only those pairs both of which where found in pass 1 to have occurred at least s times.

שאלה 1

- s=10,000
- קיימים מיליון עצמים המיוצגים ע"י השלמים 0,1,...,999999
- קיימים 25,000 פריטים תדירים
- קיימים מיליון זוגות אשר מופיעים 10,000 פעמים או יותר
- קיימים ביליון זוגות המופיעים בדיוק פעם אחת. חצי מזוגות אלה מכילים 2 פריטים תדירים, בחצי השני – בכל זוג קיים לפחות פריט אחד שאינו תדיר
- לא קיימים זוגות נוספים (מלבד המתואר לעיל)
- מספרים שלמים מיוצגים תמיד ע"י 4 בתים

מהו גודל הזיכרון הראשי המינימאלי הדרוש לצורך הרצת אלגוריתם a-priory על הנתונים הנ"ל ?

Hash based A-Priory (PCY)

- Pass 1:
  - Count items
  - Hash each pair to a bucket and increment its count by 1
- Pass 2:
  - Summarize buckets by a bitmap:

    1 = frequent //count ≥ s

    0 = not frequent
  - Count only those pairs that are both:

    (a) frequent

    (b) hash to frequent bucket

שאלה 2

- כל אחד מהזוגות התדירים ממופה לדלי שונה
- זוגות שאינם תדירים מתפרשים באופן שווה בין הדליים (כולל אלה אשר מכילים זוגות תדירים)
- כל אחד מהזוגות שאינם תדירים (בפרט זוג אשר מכיל 2 פריטים תדירים) יכול להתמפות באופן אחיד לדלי כלשהו

Low-Support High-Correlation Mining

- אלגוריתם A-priory יעיל רק כאשר מתעניינים בקשרים בין מוצרים המופיעים בתדירות מאוד גבוהה.
- קיימים מספר יישומים (למשל: clustering) בהם ניתנת חשיבות לחוקים שאינם שכיחים (low support).

Set similarity

- תזכורת: המידע נשמר במטריצה בוליאנית כאשר $a_{ij}=1$ מציין כי מוצר i נמצא בקבוצה (סל) j.
- מידת הדמיון בין עמודות נמדדת ע"י:

Jaccard measure:

$$sim_J(c_i, c_j) = \frac{|c_i \cap c_j|}{|c_i \cup c_j|}$$

Set similarity - cont'

- נרצה למצוא את כל זוגות העמודות בהן מידת הדמיון היא מעל סף מסוים s ≤ $sim(c_i, c_j^*)$: s^*.

- נבצע 3 שלבים:
  - חישוב חתימות
  - יצירת מועמדים
  - גיזום המועמדים

Signatures

- כל עמודה c_i נמפה למחרוזת מידע קטנה יותר באמצעות hashing – חתימה של c_i $sig(c_i)$.
- נאמוד את $sim_J(c_i, c_j)$ ע"י $sim_H(sig(c_i), sig(c_j))$.

Min -Hashing Schemes

- Randomly permute rows
- Hash $h(c_i)$ = index of first row with 1 in column c_i

טענה: לכל זוג עמודות (c_i, c_j) מתקיים:

$$p[h(c_i)=h(c_j)]=sim_J(c_i, c_j)$$

Min-Hash Signatures

- Pick P random row permutations
- Min-hash signature:
  $sig(c_i)$ = list of P indexes of first rows with 1 in column c_i

אבחנה: $E(sim_H(sig(c_i), sig(c_j))) = sim_J(c_i, c_j)$.

חסרונות Min-Hashing

- נדרש לייצר p ערכים של min-hash לכל עמודה — מגדיל את זמן ריצת האלגוריתם (באופן ליניארי למספר ערכי ה-min-hash)

- במקרה הגרוע יצירת ה"מועמדים" היא ריבועית למספר העמודות.

Locality-Sensitive Hashing (LSH)

- Think of the signatures as columns of integers.
- Partition the rows of the signatures into bands, say l bands of r rows each.
- Hash the columns in each band into buckets. A pair of columns is candidate-pair if they hash to the same bucket in any band.
- Verify each candidate-pair (c_i, c_j) by sim (sig(c_i), sig(c_j)).

שאלה

נתונה מטריצה עם 1,000,000 עמודות. בכל עמודה בדיוק שלושה 1'ים. ישנם 1000 זוגות של עמודות בעלי 50%similarity , 1,000,000 זוגות בעלי 20%similarity ולשאר זוגות העמודות אין שורות דומות.

מחשבים min-hash signatures כאשר p=8 ולאחר מכן משתמשים ב-LSH כדי למצוא זוגות של עמודות "מועמדים" ל-similarity.

א. כמה זוגות "מועמדים" נצפה למצוא כאשר בוחרים LSH סכמה המורכבת מ- 4 bands בהינתן שזוג הוא מועמד כאשר מתקיים 50%similarity, 20%similarity ?

ב. פתור את סעיף א' עבור LSH סכמה המורכבת מ-2 bands.