מסחר אלקטרוני 096211 - פתרון תרגיל בית 2

המרצה: פרופ' משה טננהולץ. המתרגל: מר אלון אלטמן.

שאלה 1

שאלה זו עוסקת ב- market-basket.

נתונים הסלים הבאים ומספר הלקוחות שקנו כל סל:

סל	מספר לקוחות	סל	מספר לקוחות
$\{C\}$	1,000	$\{A,B,C\}$	1,100
$\{D\}$	2,000	$\{A,D\}$	500
$\{E\}$	500	$\{B,E\}$	2,000
$\{A,B\}$	3,000	$\{B,D,E\}$	1,800
$\{A,C\}$	1,200	$\{A,C,D,E\}$	100
$\{B,C\}$	500		

שימו לב: ספירת הלקוחות של תת-סל לא כוללת את הלקוחות שקנו סלים שמכילים אותו. סלים שאינם מצויינים לעיל לא נקנו כלל.

מעוניינים להשתמש באלגוריתם PCY למציאת frequent pairs, כאשר זוג נפוץ הוא כזה שקנו אותו לפחות 4,000 לקוחות.

א. אילו מוצרים עומדים בתנאי A priori? יש לפרט את החישוב.

ב. נתונה פונקציית ערבול h המערבלת את הזוגות לדליים לפי הטבלה הבאה:

דלי 1	דלי 2	דלי 3	דלי 4	דלי 5
$\{A,B\}$	$\{A,C\}$	$\{A,D\}$	$\{A,E\}$	$\{C,D\}$
$\{C,E\}$	$\{B,D\}$	$\{B,E\}$	$\{B,C\}$	$\{D,E\}$

בשלב השני של האלגוריתם משתמשים בפונקציית הערבול h על הזוגות שנותרו מהפעלת A priori. אילו דליים ייחשבו נפוצים? יש לפרט את החישוב.

ג. בשלב השלישי סופרים את הזוגות מהסלים הנפוצים. מהם הזוגות הנפוצים בסיום האלגוריתם?

פתרון

א. כל המוצרים עומדים בתנאי A priori:

מוצר	כמות
A	$6,000$
B	$8,400$
C	$3,900$
D	$4,500$
E	$4,500$

ב. ספירת דליים (נתעלם מזוגות המכילים את C; דליים 1,3 נפוצים):

דלי	כמות
1	$4,100$
2	$1,800$
3	$4,400$
4	100
5	$1,900$

ג. זוג $\{A, B\}$ הוא הזוג הנפוץ היחיד.

שאלה 2

שאלה זו עוסקת באלגוריתם multistage PCY למציאת frequent pairs.

א. הוצע, לשם יעילות, להשתמש באותה פונקציית hash בכל שלבי האלגוריתם. כיצד תשפיע פעולה זו על ביצועי האלגוריתם? נמקו את תשובתכם!

ב. לשם חסכון בזיכרון, הוצע שבכל שלב יישמר רק ה- bitmap מהשלב הקודם ולא יישמרו bitmaps אחרים ו/או רשימת ה- frequent items. כיצד תשפיע פעולה זו על ביצועי האלגוריתם? נמקו את תשובתכם!

ג. הוצע לשפר את האלגוריתם ולשמור בכל שלב את תוצאות הספירה עצמה של ה- hash table ולא רק bitmap. כיצד ישפיע שינוי זה על ביצועי האלגוריתם, אם אנו מניחים שכל הזיכרון הפנוי מנוצל לאחסון ה- hash table?

פתרון

א. שימוש באותה פונקציית hash בכל שלבי האלגוריתם יגרור שכל ה- bins בכל הסיבובים יהיו זהים וכך נספור בדיוק את אותם זוגות, ולא נצמצם

את מספר ה- candidate pairs ובך האלגוריתם יעבוד רק כמו האלגוריתם הדו-שלבי.

ב. במקרה זה יהיה לנו בכל שלב רק קריטריון אחד לבחינת הזוגות אותם אנו סופרים ולא נוכל להביא לצמצום משמעותי בכמות ה- candidate pairs.

ג. אם נשמור את תוצאות הספירה עצמה, לא יישאר זיכרון לאחסון תוצאות הספירה של השלב הבא ולא נוכל להמשיך להריץ את האלגוריתם.

שאלה 3

יש לבצע clustering לקבוצת הנקודות הבאה לפי אלגוריתם k-means תוך שימוש במרחק אוקלידי (L_2) עבור $k = 3$. יש לפרט את שלבי ההרצה ואת ה- clusters המתקבלים.

הנקודות: $\{(2,20),(2,5),(8,4),(5,8),(7,5),(6,4),(1,2),(4,9)\}$.

יש להתחיל עם המרכזים ההתחלתיים הבאים $x_1 = (2,20); x_2 = (2,5); x_3 = (8,4)$.

פתרון

בשלב הראשון נוסיף כל נקודה ל- cluster הקרוב אליה:

נקודה	x_1	x_2	x_3	סיווג
$(2,20)$	0	15	17.09	x_1
$(2,5)$	15	0	6.08	x_2
$(8,4)$	17.09	6.08	0	x_3
$(5,8)$	12.37	4.24	5	x_2
$(7,5)$	15.81	3.81	1.41	x_3
$(6,4)$	16.49	3.53	1.58	x_3
$(1,2)$	18.03	5.15	6.44	x_2
$(4,9)$	11.18	4.22	5.55	x_2

ה- clusters המתקבלים לאחר שלב זה הנם

$\{(2,20)\},\{(2,5),(5,8),(1,2),(4,9)\},\{(8,4),(7,5),(6,4)\}$.

הצנטרואידים המתקבלים הנם $(2,20)$, $(3,6)$, ו- $(7,4.33)$. נחשב מרחקים שנית:

נקודה	$(2,20)$	$(3,6)$	$(7,4.33)$
$(2,20)$	0	14.04	16.45
$(2,5)$	15	1.41	5.04
$(8,4)$	17.09	5.39	1.05
$(5,8)$	12.37	2.83	4.18
$(7,5)$	15.81	4.12	0.67
$(6,4)$	16.49	3.61	1.05
$(1,2)$	18.03	4.47	6.44
$(4,9)$	11.18	3.16	5.55

ה- clusters הסופיים המתקבלים מהרצת האלגוריתם הנם

$\{(2,20)\}, \{(2,5),(5,8),(1,2),(4,9)\}, \{(8,4),(7,5),(6,4)\}$.

שאלה 4

הטבלה הבאה מכילה נתונים לגבי קונפיגורציות של מחשבים בהם פגע ה- worm לאחרונה וכאלה שלא נפגעו:

מערכת הפעלה	דפדפן	תוכנת דואר	נפגע?
Linux	Firefox	Thunderbird	לא
Windows XP	Internet Explorer	Outlook	כן
Windows XP	Firefox	Outlook	כן
Windows XP	Internet Explorer	Thunderbird	לא
Windows 2000	Internet Explorer	Outlook	כן
Windows 2000	Firefox	Outlook	לא
Windows 2000	Internet Explorer	Thunderbird	כן

א. יש לבנות עץ החלטה ע"י שימוש באלגוריתם ID3 על מנת לקבוע האם מחשב בעל קונפיגורציה מסוימת יפגע מה-worm או לא.

יש לבחור את הרמה הבאה לבניית העץ לפי האנטרופיה המשוקללת של הקבוצות המתקבלות (ולא האנטרופיה המקסימלית)

ב. בהסתמך על עץ ההחלטה שבנית, יש לקבוע האם מחשב בעל מערכת הפעלה Linux שמריץ Firefox ו- Outlook ייפגע מה-worm. ומה לגבי מחשב בעל Windows XP שמריץ Firefox ו- Thunderbird?

פתרון

א. נחשב אנטרופיה עבור כל אחד מהסיווגים האפשריים:

- עבור מערכת הפעלה:

$$\begin{aligned}
E(Linux) &= 0 \\
E(WinXP) &= -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918 \\
E(Win2k) &= -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918 \\
E &= \frac{1}{7}E(Linux) + \frac{3}{7}E(WinXP) + \frac{3}{7}E(Win2k) = 0.787
\end{aligned}$$

- עבור דפדפן:

$$\begin{aligned}
E(Firefox) &= \frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3} = 0.918 \\
E(IE) &= \frac{3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4} = 0.811 \\
E &= \frac{3}{7}E(Firefox) + \frac{4}{7}E(IE) = 0.857
\end{aligned}$$

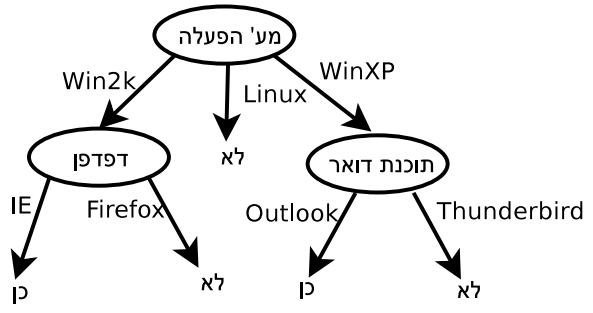
- עבור תוכנת דואר

$$\begin{aligned}
E(Thunderbird) &= \frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3} = 0.918 \\
E(Outlook) &= \frac{3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4} = 0.811 \\
E &= \frac{3}{7}E(Thunderbird) + \frac{4}{7}E(Outlook) = 0.857
\end{aligned}$$

בשלב הראשון נבחר להפריד לפי מערכת הפעלה.

- עבור Linux, סיימנו – התוצאה היא לא.
- עבור WinXP, עדיף להבחין לפי תוכנת דואר – במקרה של Outlook התוצאה היא כן, ובמקרה של Thunderbird לא.
- עבור Win2k, עדיף להבחין לפי דפדפן – במקרה של Internet Explorer התוצאה היא כן, ובמקרה של Firefox לא.

העץ המתקבל הנו:



ב. מחשב בעל מערכת הפעלה Linux שמריץ Firefox ו- Outlook לא ייפגע.
מחשב בעל Windows XP שמריץ Firefox ו- Thunderbird גם כן לא ייפגע.