Lecture 8

Thursday, September 26, 2024 10:05 AM

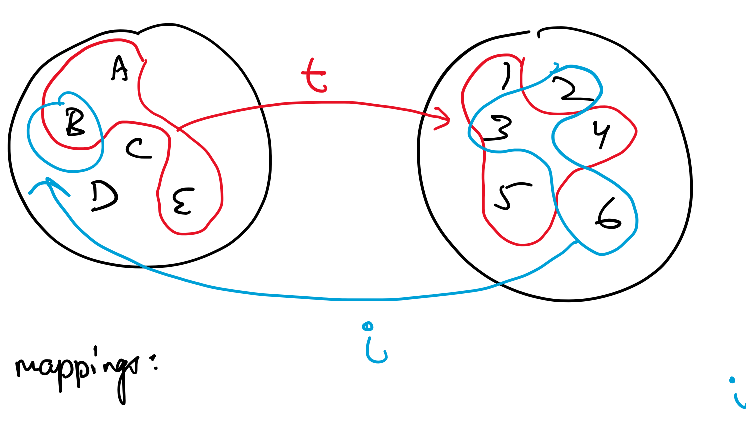
frequent itemsets

$X \subseteq \underbrace{I}_{\text{all the items}}$

$Y \subseteq \underbrace{T}_{\text{set of samples / instance / transactions ID}}$

dataset

1. ABDE
2. BCE
3. ABDE
4. ABCE
5. ABCDE
6. BCD




two mappings:

$X \subseteq I$

define $t(X) = \{$ set of all transactions that contain $X \}$

$Y \subseteq T$

$i(Y) = \{$ items that appear in all the tids in $Y \}$

$i(236) = B$

$t(\{A,B,E\}) = t(ABE) = \{1,3,4,5\} = 1345$

$t: 2^I \rightarrow 2^T$

itemset tidset

$t(X) \equiv$ tidset for $X$

$sup(X) = |t(X)|$

$\equiv$ # of transactions containing $X$

support of $X$

frequency of $X$

$\hat{P}(X) = \frac{sup(X)}{|T|}$ — Counting

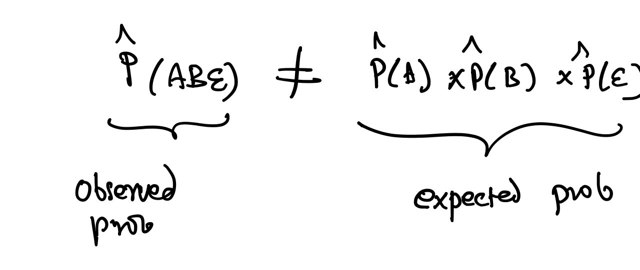
estimate for $\underbrace{\text{prob of } X}_{\text{joint probability of all the items in } X}$

$\hat{P}(ABE) = \frac{4}{6} = \frac{2}{3} = 0.67$

$\underbrace{\hat{P}(ABE)}_{\text{observed prob}} \neq \underbrace{\hat{P}(A) \times \hat{P}(B) \times \hat{P}(E)}_{\text{expected prob under independence}}$

---

1. ABDE
2. BCE
3. ABDE
4. ABCE
5. ABCDE
6. BCD




$t(ABE) = 1345$

$i(t(ABE)) = i(1345)$

$= ABE$

$i(t(ABC)) = i(45)$

$= ABCE$

$i \circ t = c$ ← closure operation

If $c(X) = X$ then we say that $X$ is a closed set

$\downarrow$

$i \circ t(X)$ ← Composition

e.g. ABE is a closed set

$X = ABC$ not closed since $c(ABC) = ABCE$

$c(X)$ is always closed, (even if $X$ is not closed)

Closure operator c:

1) idempotent $c(X) = c(c(X))$
2) extensive $c(X) \supseteq X$
3) Monotone

$X_1 \subseteq X_2$

$\Rightarrow c(X_1) \subseteq c(X_2)$




'X is closed if there is no superset with the same frequency'

---

Eclat: intersections of tidsets

1. ABDE
2. BCE
3. ABDE
4. ABCE
5. ABCDE
6. BCD

$\Rightarrow$

$x \rightarrow t(x)$

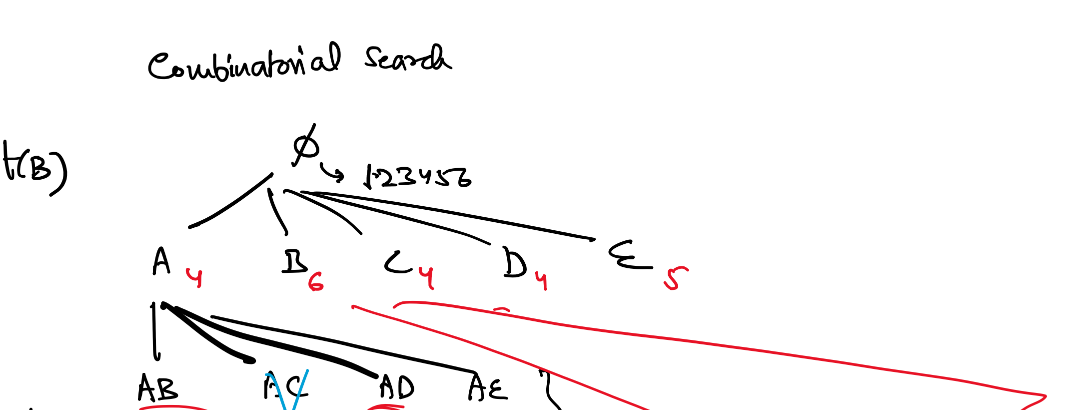
| $x$ | $t(x)$ |
|---|---|
| A | 1345 |
| B | 123456 |
| C | 2456 |
| D | 1356 |
| E | 12345 |

} inverted dataset

Combinatorial search

$t(AB) = t(A) \cap t(B)$

Minsup = 3




$t(ABD) = t(AB) \cap t(AD)$

---

diffset $\equiv$ difference of two tidsets (prefix)

$= t(\text{prefix}) - t(\text{prefix} + \text{last item})$

$d(ABD) = t(AB) - t(ABD)$

initially: $d(A) = t(\phi) - t(A)$

$A = \{\phi A\}$

$= 123456 - 1345$

$= 26$

those transactions that do not contain A

$sup(A) = sup(\phi) - len(d(A))$

$sup(X) = sup(\text{prefix}) - |d(X)|$

all tids: $\underbrace{123456 - 26}_{\substack{1345 \\ \text{all tids that contain A}}}$ — 26: tids that do not contain

$6 - 2$

$= 4$ sup

$d(X) = t(\text{prefix}) - t(X)$

diffsets from diffset

$$\boxed{d(\text{prefix} + \text{item1} + \text{item2}) = d(\text{prefix} + \text{item 2}) - d(\text{prefix} + \text{item 1})}$$

$d(ABD) = d(AD) - d(AB)$

$\parallel$

$t(AB) - t(ABD)$

$\phi - t(\phi) = 123456$

dEclat

| $X$ | $t(X)$ | $d(X)$ |
|---|---|---|
| A | 1345 | 26 |
| B | 123456 | $\phi$ |
| C | 2456 | 13 |
| D | 1356 | 24 |
| E | 12345 | 6 |

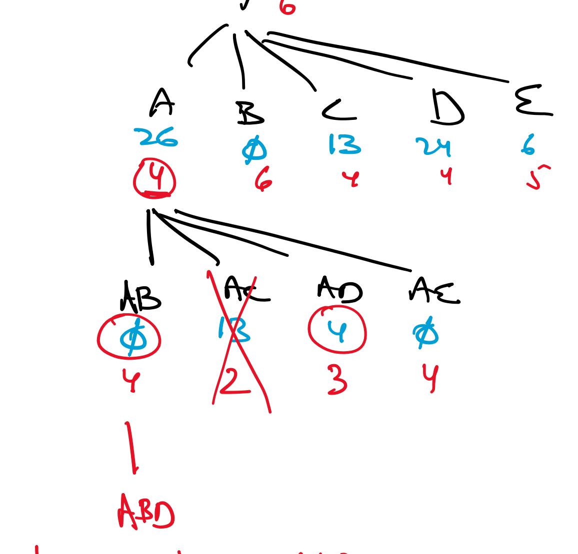
$d(AB) = d(B) - d(A)$

$= \phi - \{2,6\}$

$= \phi$

$\{1,3\} - \{2,6\}$

$= \{1,3\}$




---

closed freq sets are loss-less compressed representation of all freq sets

F = set of all frequent itemsets

Minsup = 3

no superset with same freq $\Rightarrow$ closed

| sup/freq | itemsets |
|---|---|
| 6 | B |
| 5 | E, BE |
| 4 | A, C, D, AB, AE, BC, BD, ABE |
| 3 | AD, CE, DE, ABD, ADE, BCE, BDE, ABDE |

1. ABDE
2. BCE
3. ABDE
4. ABCE
5. ABCDE
6. BCD

F

19 frequent itemsets

$c(B) = B$

$B \rightarrow t(B) = 123456$

$i(t(B)) = B$

$c(E) = i(t(E)) = i(12345)$

$= BE$

$\mathcal{C}$ = set of all closed and frequent itemsets

$= \{ B, BE, BC, BD, ABE, BCE, ABDE \}$

(supports: 6, 5, 4, 4, 4, 3, 3)

we can compute the exact freq/sup of all freq sets

Q. $sup(ABD) = 3$

$sup(AB) = 4$

---

charm: algo for finding/mining closed freq itemsets

↳ tidsets (replace with diffsets)

3 observations

tidsets

1) $t(X_1) = t(X_2)$

$\Rightarrow \underbrace{c(X_1) = c(X_2) = c(X_1 \cup X_2)}_{\text{itemsets}}$

$t(AB) = t(AC)$

$c(AB) = c(AC)$

$= c(ABC)$

$X_1 = AB$

$c(X_1) = i(t(AB))$

$= i(1345)$

$= ABE$

| | | |
|---|---|---|
| A | 1345 | 4 |
| B | 123456 | 6 |
| C | 2456 | 4 |
| D | 1356 | 4 |
| E | 12345 | 5 |

1. ABDE
2. BCE
3. ABDE
4. ABCE
5. ABCDE
6. BCD

2) $t(X_1) \subset t(X_2)$

then $c(X_1) \neq c(X_2)$

but $c(X_1) = c(X_1 \cup X_2)$

3) $t(X_1) \neq t(X_2)$

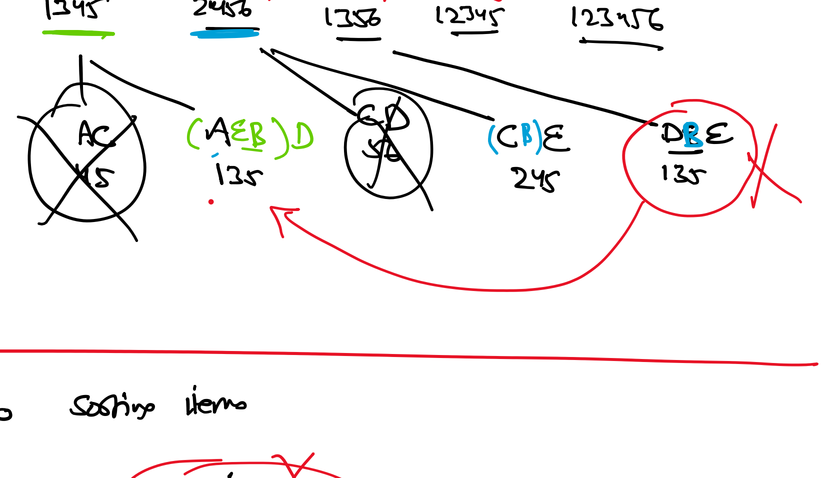
$c(X_1) \neq c(X_2) \neq c(X_1 \cup X_2)$

---

1) Sort items by increasing support

A C D E B

Minsup = 3




---

repeat: w/o sorting items