COSC-254 Data Mining

Lec 04: Pattern Mining

Outline

- Motivation and applications
- Itemsets
- The itemset lattice and the downward closure
- Association rules

You run an on-line store, and want to increase sales.

Associative advertising: show ads of relevant products **before** your users search for these

Need the sets of products that are frequently bought together

Key definitions

- Items
- Dataset
- Transaction
- Itemset
- Support / frequency,
- Minimum support frequency threshold,
- Frequent itemsets

Market basket data

- Items for sale: $\mathcal{I} = \{\text{apple, beer, cola, diapers, eggs}\}$

Transactions: 1: {apple, cola}, 2: {apple, beer, diapers, eggs},
3: {cola, beer, diapers}, 4: {apple, beer, cola, diapers},
5: {apple, cola, diapers}

Transaction IDs

TID	Apple	Beer	Cola	Diapers	Eggs
1	✔	✔	✔	✔	
2	✔	✔	✔		✔
3	✔	✔		✔	
4	✔	✔		✔	✔
5			✔		
6	✔				

Transaction Data as a Binary Matrix

TID	Apple	Beer	Cola	Diapers	Eggs
1	1	1	1	1	0
2	1	1	1	0	1
3	1	1	0	1	0
4	1	1	0	1	1
5	0	0	1	0	0
6	1	0	0	0	0

Any data that can be represented as a binary matrix can be used

Applications –

- **Baskets** = sentences; **Items** = documents containing those sentences
  - Items that appear together too often could represent plagiarism
  - items do not have to be "in" baskets

- **Baskets** = patients; **Items** = drugs & side-effects
  - Has been used to detect combinations of drugs that result in particular side-effects
  - **Requires extension:** Absence of an item needs to be observed as well as presence

Example: Frequent Itemsets

- **Items** = {milk, coke, pepsi, beer, juice}
- **Support threshold** = 3 baskets

B_1 = {m, c, b} B_2 = {m, p, j}

B_3 = {m, b} B_4 = {c, j}

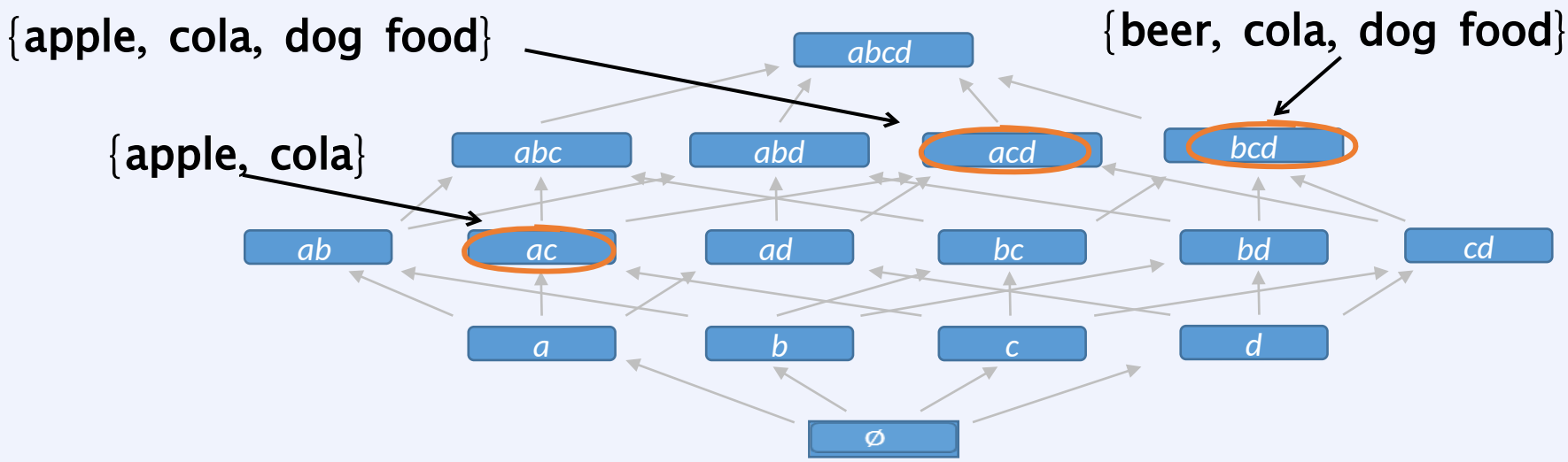
B_5 = {m, p, b} B_6 = {m, c, b, j}

B_7 = {c, b, j} B_8 = {b, c}

- **Frequent itemsets:** {m}, {c}, {b}, {j}, {m,b}, {b,c}, {c,j}

Search space and search strategy

The Itemset Lattice



a: apple
b: beer
c: cola
d: dog food

Naïve search strategy for FIs

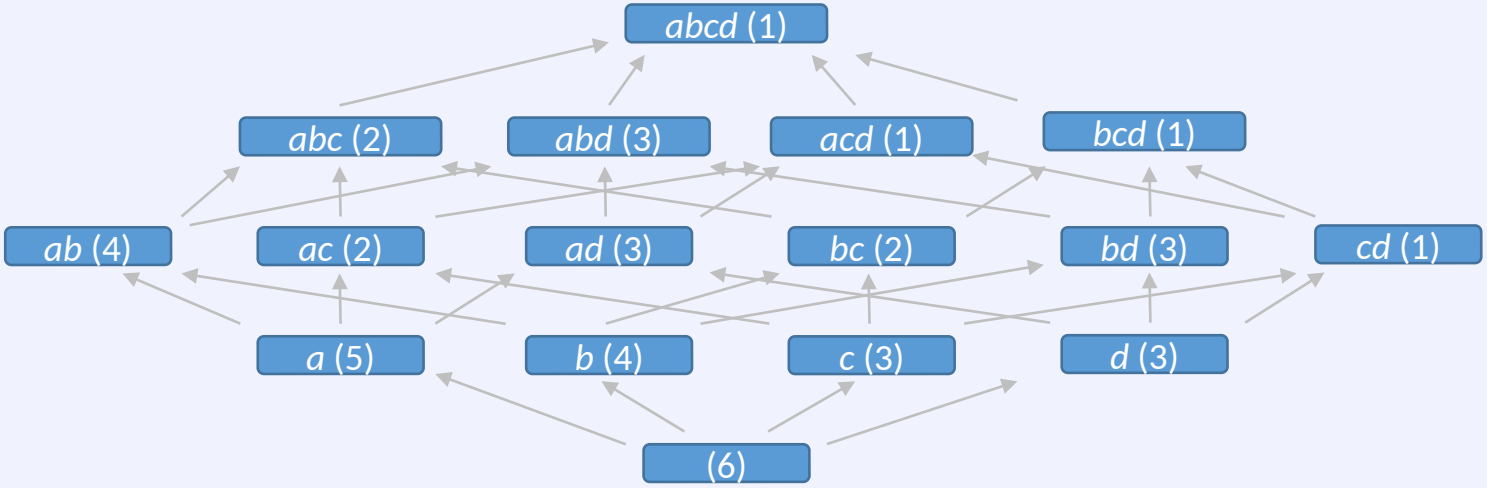
Try every possible itemset and check if it is frequent

Qs:

- How to try every itemset?
- How to check if it is frequent? How long does it take?
- How long does it take to mine all fis?

The Itemset Lattice

a	b	c	d
1	1	1	1
1	1	1	0
1	1	0	1
1	1	0	1
0	0	1	0
1	0	0	0

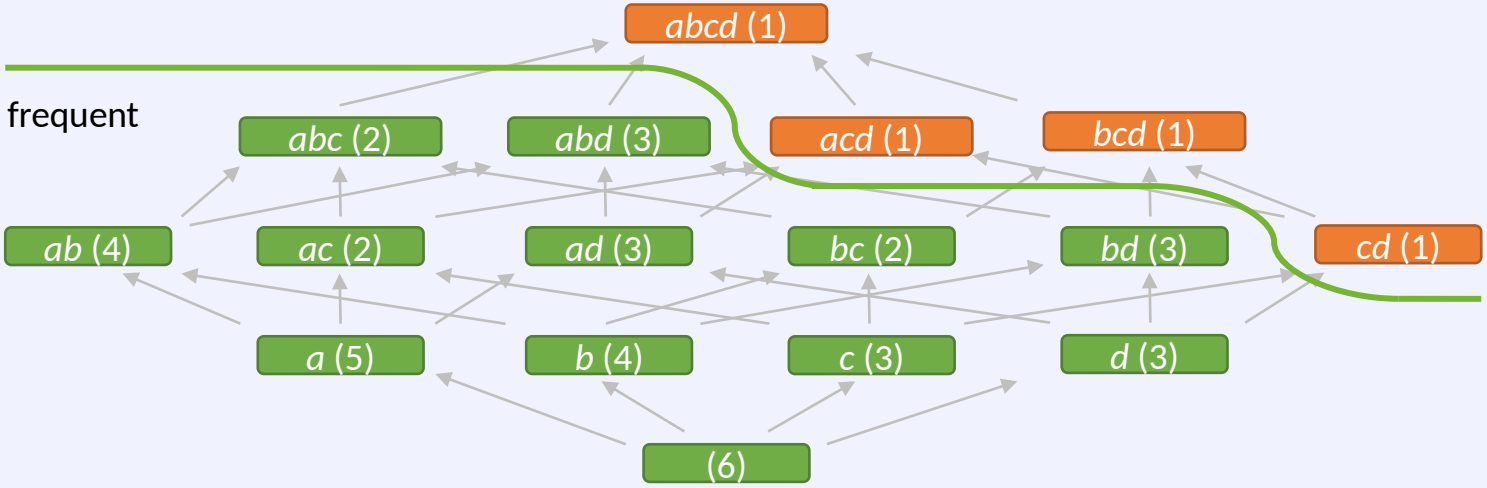


data

itemset lattice

The Itemset Lattice

a	b	c	d
1	1	1	1
1	1	1	0
1	1	0	1
1	1	0	1
0	0	1	0
1	0	0	0



data

itemset lattice

Downward closure property

a.k.a. Antimonotonicity of support