Attribute types, concept hierarchies and negative associations

Lecture 17

Types of attributes

- We were working with asymmetric binary attributes:
  - Binary: Item: 0 – not present, 1 – present
  - Asymmetric: more interested in presence than in absence

- What do we do if attributes are
  - Symmetric binary
  - Categorical
  - Numeric

Attribute type examples

- **Symmetric binary attributes**
  - Gender
  - Computer at Home
  - Chat Online
  - Shop Online
  - Privacy Concerns

Internet survey data with categorical attributes.

Gender	Level of Education	State	Computer at Home	Chat Online	Shop Online	Privacy Concerns
Female	Graduate	Illinois	Yes	Yes	Yes	Yes
Male	College	California	No	No	No	No
Male	Graduate	Michigan	Yes	Yes	Yes	Yes
Female	College	Virginia	No	No	Yes	Yes
Female	Graduate	California	Yes	No	No	Yes
Male	College	Minnesota	Yes	Yes	Yes	Yes
Male	College	Alaska	Yes	Yes	Yes	No
Male	High School	Oregon	Yes	No	No	No
Female	Graduate	Texas	No	Yes	No	No
...	...	...	...	...	...	...

Attribute type examples

- **Nominal (categorical) attributes**
  - Level of Education
  - State

Internet survey data with categorical attributes.

Gender	Level of Education	State	Computer at Home	Chat Online	Shop Online	Privacy Concerns
Female	Graduate	Illinois	Yes	Yes	Yes	Yes
Male	College	California	No	No	No	No
Male	Graduate	Michigan	Yes	Yes	Yes	Yes
Female	College	Virginia	No	No	Yes	Yes
Female	Graduate	California	Yes	No	No	Yes
Male	College	Minnesota	Yes	Yes	Yes	Yes
Male	College	Alaska	Yes	Yes	Yes	No
Male	High School	Oregon	Yes	No	No	No
Female	Graduate	Texas	No	Yes	No	No
...	...	...	...	...	...	...

Transforming attributes into asymmetric binary

- Create a new item for each distinct attribute-value pair.

- E.g., the nominal attribute Level of Education can be replaced by three binary items:
  - Education = College
  - Education = Graduate
  - Education = High School

- Binary attributes such as Gender are converted into a pair of binary items
  - Male
  - Female

Data after binarizing attributes into "items"

Male	Female	Education = Graduate	Education = College	...	Privacy = Yes	Privacy = No
0	1	1	0	...	1	0
1	0	0	1	...	0	1
1	0	1	0	...	1	0
0	1	0	1	...	1	0
0	1	1	0	...	1	0
1	0	0	1	...	1	0
1	0	0	1	...	0	1
1	0	0	0	...	0	1
0	1	1	0	...	0	1
...	...	...	...	...	...	...

Note, that here we are interested in both yes and no values of binary attributes, so we generate a separate item for each: privacy=Yes and privacy=No

Numeric (continuous) attributes

- **Solution**: Discretize

- Example of rules:
  - Age$\in$[21,35) $\wedge$ Salary$\in$[70k,120k) $\rightarrow$ Buy
  - Salary$\in$[70k,120k) $\wedge$ Buy $\rightarrow$ Age: μ=28, σ=4

- Of course discretization isn't always easy.
  - If intervals too large may not have enough confidence
    Age $\in$ [12,36) → Chat Online = Yes (s = 30%, c = 57.7%)
    (minconf=60%)
  - If intervals too small may not have enough support
    Age $\in$ [16,20) → Chat Online = Yes (s = 4.4%, c = 84.6%)
    (minsup=15%)

Statistics-based quantitative association rules

Salary $\in$ [70k,120k) $\wedge$ Buy $\rightarrow$ Age: μ=28, σ=4

Generated as follows:

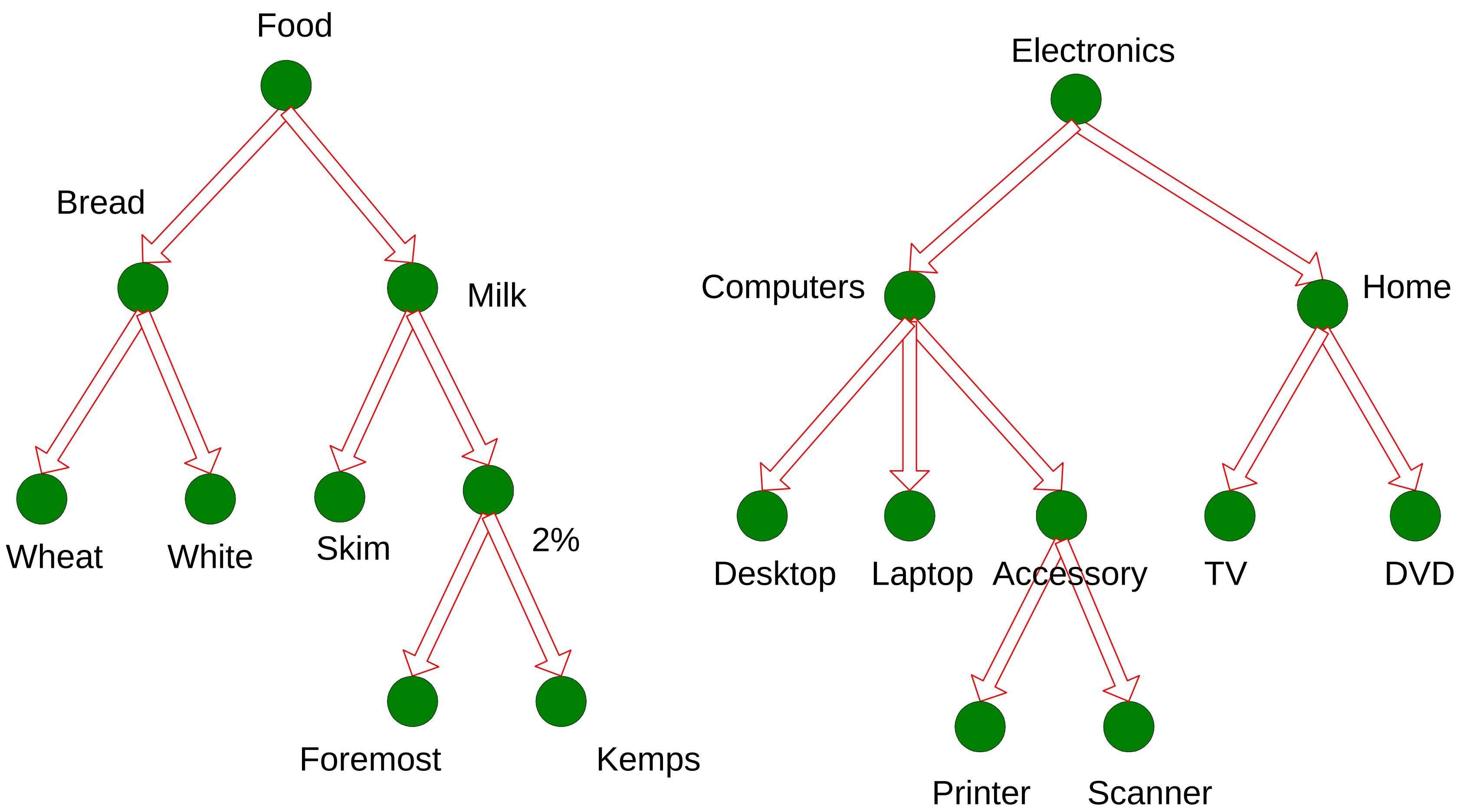
- Specify the target attribute (e.g. Age).
- Withhold target attribute, and "itemize" the remaining attributes.
- Extract frequent itemsets from the itemized data.
  - Each frequent itemset identifies an interesting segment of the population.
- Derive a rule for each frequent itemset.
  - E.g., the preceding rule is obtained by averaging the age of Internet users who support the frequent itemset

    {Annual Income> $100K, Shop Online = Yes}

- Remark: Notion of confidence is not applicable to such rules.

Associations across concept hierarchies

Items: levels of abstraction



Multi-level Association Rules

- Rules about items at lower levels of abstraction can represent a more general rule:

    skim milk $\rightarrow$ white bread,

    2% milk $\rightarrow$ wheat bread,

    skim milk $\rightarrow$ wheat bread, etc.

  are all indicative of association between their generalizations milk and bread

How much to generalize?

- Should we consider correlation between milk and bread, between cream and bagels, or between specific labels of cream and bagels?

- The correlation between specific items can be hard to find because of the low support
- The correlation between more general itemsets can be very low, despite that the support is high

Mining multi-level Association Rules

Approach 1

- Augmenting each transaction with higher level items

  Original Transaction: {skim milk, wheat bread}

  Augmented Transaction:
  {skim milk, wheat bread, milk, bread, food}

- Issue:
  - Items that reside at higher levels have much higher support counts

    if support threshold is low, we get too many frequent patterns involving items from the higher levels

Multi-level Association Rules

Approach 2

- Generate frequent patterns at highest level first.
- Then, generate frequent patterns at the next highest level, and so on, decreasing minsupport threshold
- Issues:
  - May miss some potentially interesting **cross-level** association patterns. E.g.

    skim milk $\rightarrow$ white bread,

    2% milk $\rightarrow$ white bread,

    skim milk $\rightarrow$ white bread

    might not survive because of low support, but

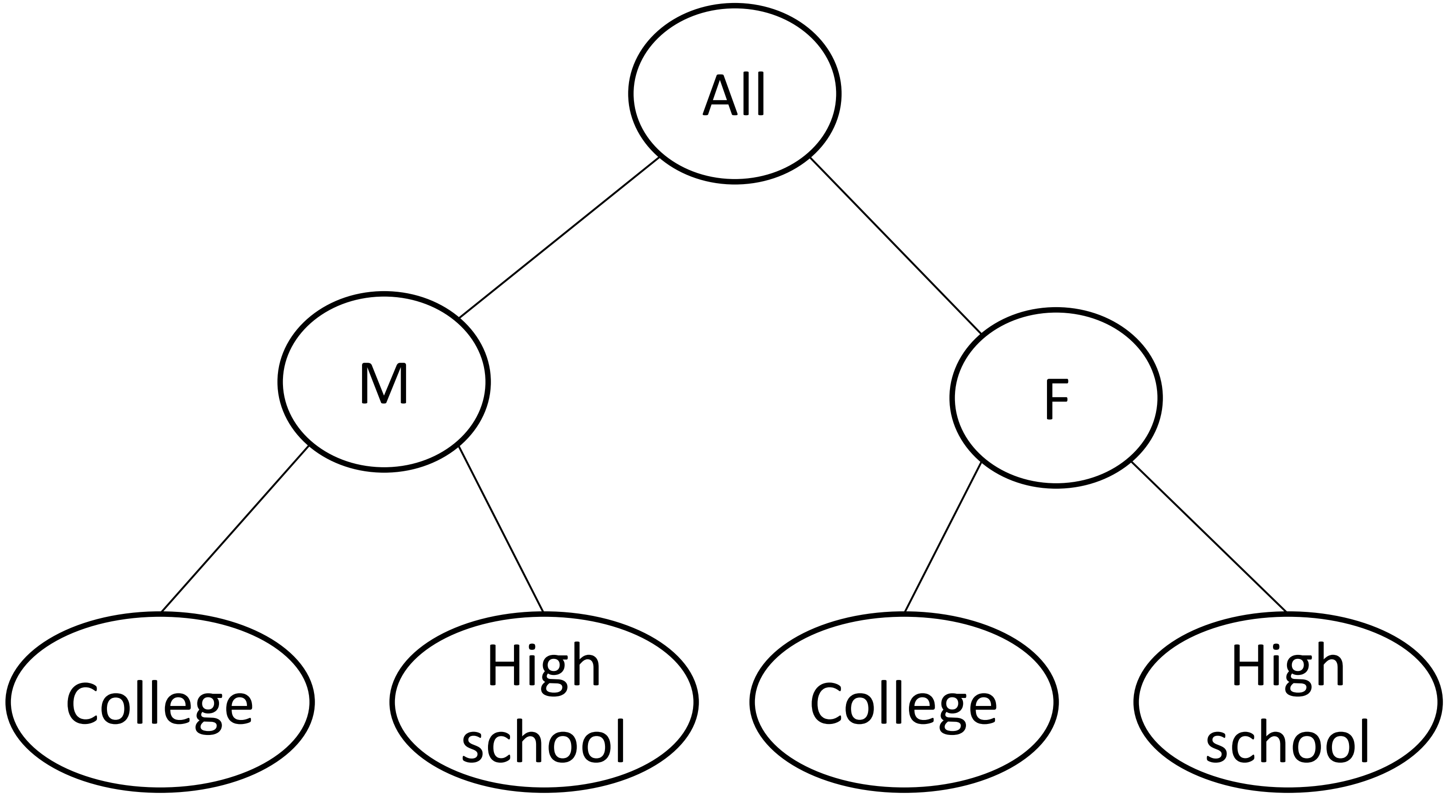
    milk $\rightarrow$ white bread

    could.

    However, we don't generate a cross-level itemset such as

    {milk, white bread}

Transactions also may have hierarchies



Hierarchy of groups: strata

Example (symmetric binary variables)

Buy HDTV	Buy Exercise Machine		
	Yes	No	
Yes	99	81	180
No	54	66	120
	153	147	300

- What's the confidence of the following rules:
  (rule 1) {HDTV=Yes} → {Exercise machine = Yes}
  (rule 2) {HDTV=No} → {Exercise machine = Yes} ?

Confidence of rule 1 = 99/180 = 55%

Confidence of rule 2 = 54/120 = 45%

Conclusion: there is a positive correlation between buying HDTV and buying exercise machines

What if we look into more specific groups

Customer Group	Buy HDTV	Buy Exercise Machine		Total
		Yes	No	
College Students	Yes	1	9	10
	No	4	30	34
Working Adult	Yes	98	72	170
	No	50	36	86

- What's the confidence of the rules for each strata:
  (rule 1) {HDTV=Yes} → {Exercise machine = Yes}
  (rule 2) {HDTV=No} → {Exercise machine = Yes} ?

College students:

Confidence of rule 1 = 1/10 = 10%

Confidence of rule 2 = 4/34 = 11.8%

Working Adults:

Confidence of rule 1 = 98/170 = 57.7%

Confidence of rule 2 = 50/86 = 58.1%

The rules suggest that, for each group, customers who don't buy HDTV are more likely to buy exercise machines, which contradict the previous conclusion when data from the two customer groups are pooled together.

Correlation is reversed at different levels of generalization

At a more general level of abstraction:

{HDTV=Yes} → {Exercise machine = Yes}

College students:

{HDTV=No} → {Exercise machine = Yes}

Working Adults:

{HDTV=No} → {Exercise machine = Yes}

This is called Simpson's Paradox

Importance of Stratification

- The lesson here is that proper stratification is needed to avoid generating spurious patterns resulting from Simpson's paradox.

For example

- Market basket data from a major supermarket chain should be stratified according to store locations, while
- Medical records from various patients should be stratified according to confounding factors such as age and gender.

Explanation of Simpson's paradox

- Lisa and Bart are programmers, and they fix bugs for two weeks

	Week 1	Week 2	Both weeks
Lisa	60/100	1/10	**61/110**
Bart	**9/10**	**30/100**	39/110

Who is more productive: Lisa or Bart?

Explanation of Simpson's paradox

	Week 1	Week 2	Both weeks
Lisa	60/100	1/10	**61/110**
Bart	**9/10**	**30/100**	39/110

If we consider productivity for each week, we notice that the samples are of a very different size

The work should be judged from an equal sample size, which is achieved when the numbers of bugs each fixed are added together

Explanation of Simpson's paradox

	Week 1	Week 2	Both weeks
Lisa	60/100	1/10	**61/110**
Bart	**9/10**	**30/100**	39/110

Simple algebra of fractions shows that even though

a1/A > b1/B
c1/C > d1/D

(a1+c1)/(A+C) can be smaller than (b1+d1)/(B+D) !

This may happen when the sample sizes A, B, C,D are skewed (Note, that we are not adding two inequalities, but adding the absolute numbers)

Simpson's paradox in real life

- Two examples:
  - Gender bias
  - Medical treatment

Example 1: Berkeley gender bias case

Admitted to graduate school at University of California, Berkeley (1973)

	Admitted	Not admitted	Total
Men	3,714	4,727	8,441
Women	1,512	2,808	4,320

- What's the confidence of the following rules:
  (rule 1) {Man=Yes} → {Admitted= Yes}
  (rule 2) {Man=No} → {Admitted= Yes} ?

Confidence of rule 1 = 3714/8441= 44%
Confidence of rule 2 = 1512/4320 = 35%

Conclusion: bias against women applicants

Example 1: Berkeley gender bias case

Stratified by the departments

	Men		Women	
Dept.	Total	Admitted	Total	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	272	6%	341	**7%**

In most departments,
the bias is towards women!

Example 2: Kidney stone treatment

Success rates of 2 treatments for kidney stones

Treatments	Success	Not success	Total
A*	273	77	350
B**	289	61	350

- What’s the confidence of the following rules:
  (rule 1) {treatment=A} → {Success= Yes}
  (rule 2) {treatment=B} → {Success = Yes} ?

(A) Confidence of rule 1 = 273/350= 78%

(B) Confidence of rule 2 = 289/350 = **83**%

Conclusion: treatment B is better

*Open procedures (surgery)

** Percutaneous nefrolithotomy (removal through a small opening)

Example 2: Kidney stone treatment

Success rates of 2 treatments for kidney stones

	Treatment A	Treatment B
Small stones	**93% (81/87)**	87%(234/270)
Large stones	**73%(192/263)**	69%(55/80)
Both	78%(273/350)	**83% (289/350)**

Treatment A is better for both small and large stones,
But treatment B is more effective if we add both groups together

Implications in decision making

- Which data should we consult when choosing an action: the aggregated or stratified?
- Kidney stones: if you know the size of the stone, choose treatment A, if you don’t – treatment B?

Implications in decision making

- Which data should we consult when choosing an action: the aggregated or stratified?

- The common sense: the treatment which is preferred under both conditions should be preferred when the condition is unknown

Implications in decision making

- Which data should we consult when choosing an action: the aggregated or stratified?

- If we always choose to use the stratified data, we can partition strata further, into groups by eye color, age, gender, race … These arbitrary hierarchies can produce opposite correlations, and lead to wrong choices

Implications in decision making

- Which data should we consult when choosing an action: the aggregated or stratified?

- Conclusion: data should be consulted with care and the understanding of the underlying story about the data is required for making correct decisions

Negative correlations and flipping patterns

Negative association rules

- The methods for association mining were based on the assumption that the presence of an item is more important than its absence (asymmetric binary attributes)

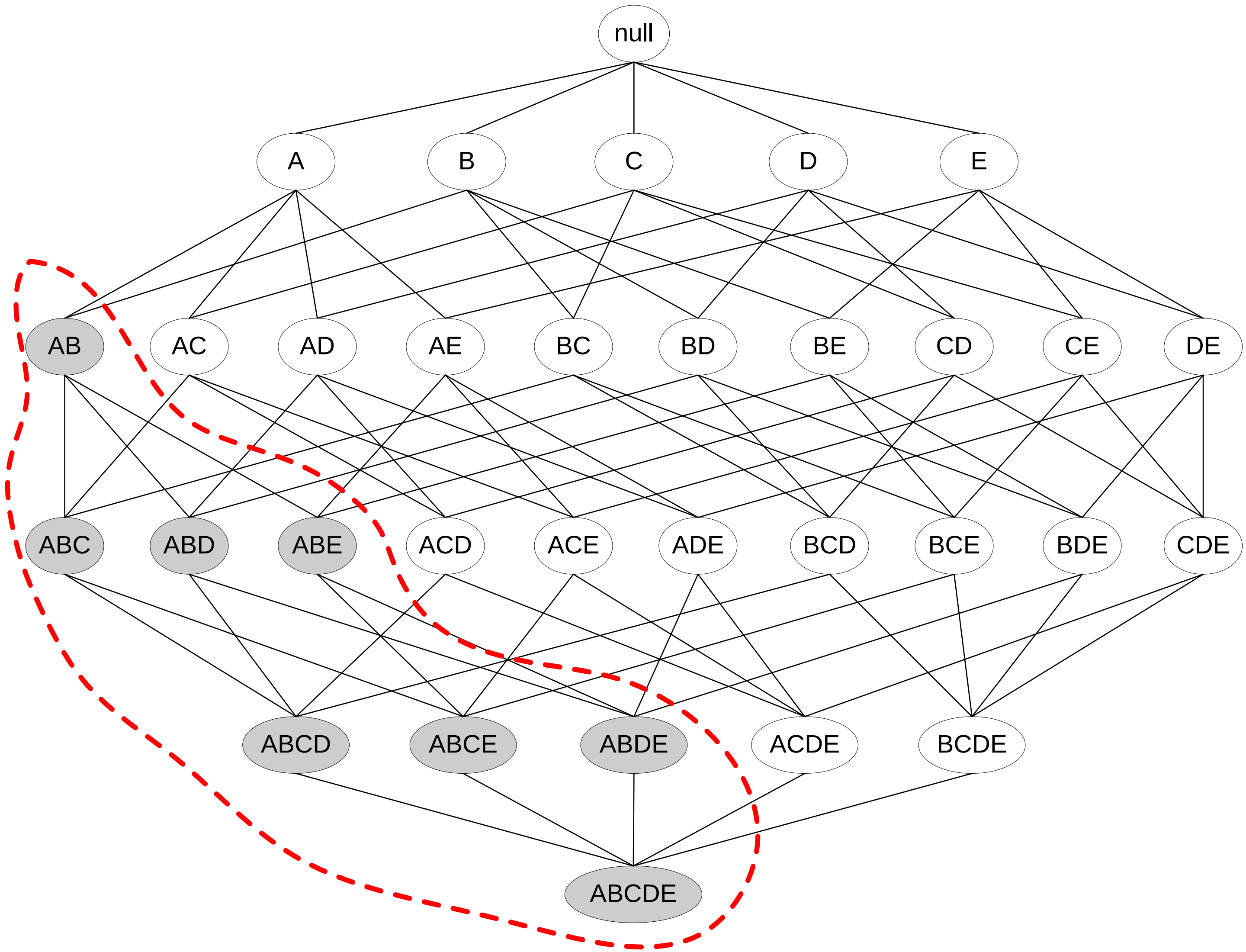
- The negative correlations can be useful:
  - To identify competing items: absence of Blu ray and DVD player in the same transaction
  - To find rare important events: rare occurrence {Fire=yes, Alarm=On}

Mining negative patterns

- Negative itemset: a frequent itemset where at least one item is negated
- Negative association rule: is an association rule between items in a negative itemset with confidence ≥ *minConf*

- If a regular itemset is infrequent due to the low count of some item, it is frequent if we consider the negation (absence) of a corresponding item

Negative patterns = non-positive



Challenging task

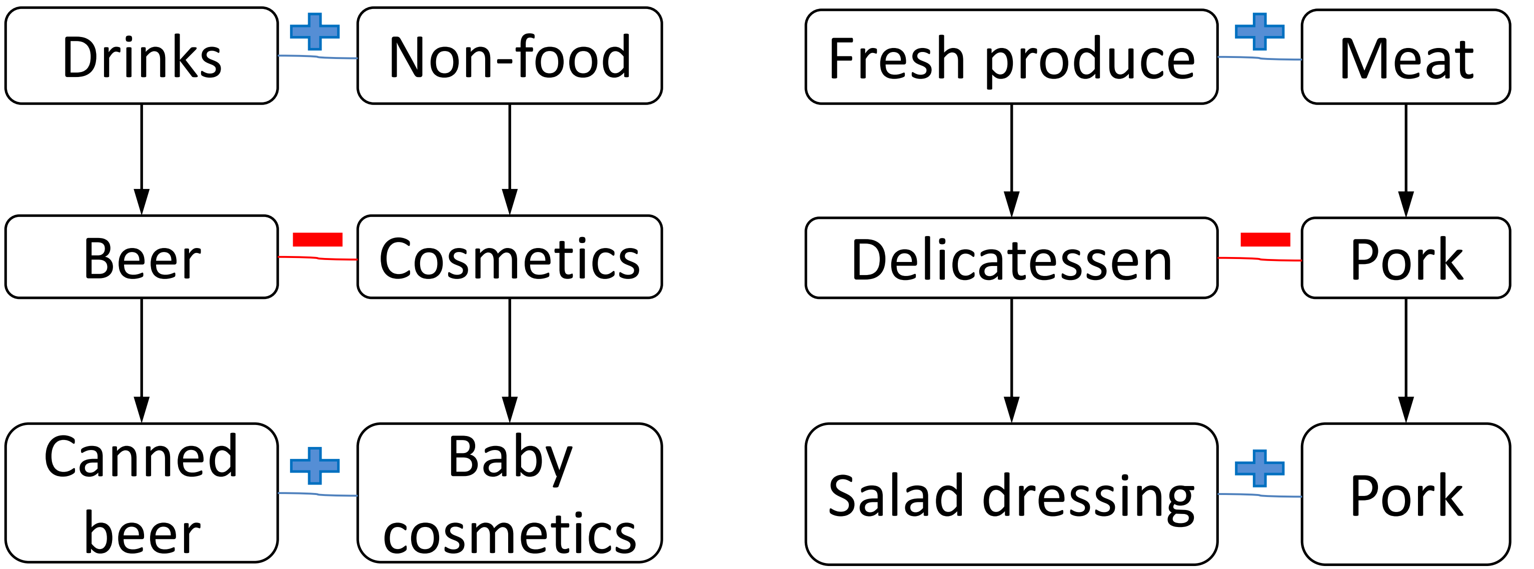
- Positive associations can be extracted only for high-levels of support. Then the set of all frequent itemsets is manageable
- In this case, the complement to all frequent itemsets is exponentially large, and cannot be efficiently enumerated

- But do we need all negative associations?

Flipping patterns

- Flipping patterns are extracted from the datasets with concept hierarchies
- The pattern is interesting if it has positive correlation between items which is accompanied by the negative association of their minimal generalizations, and vice versa
- We call such patterns *flipping patterns*

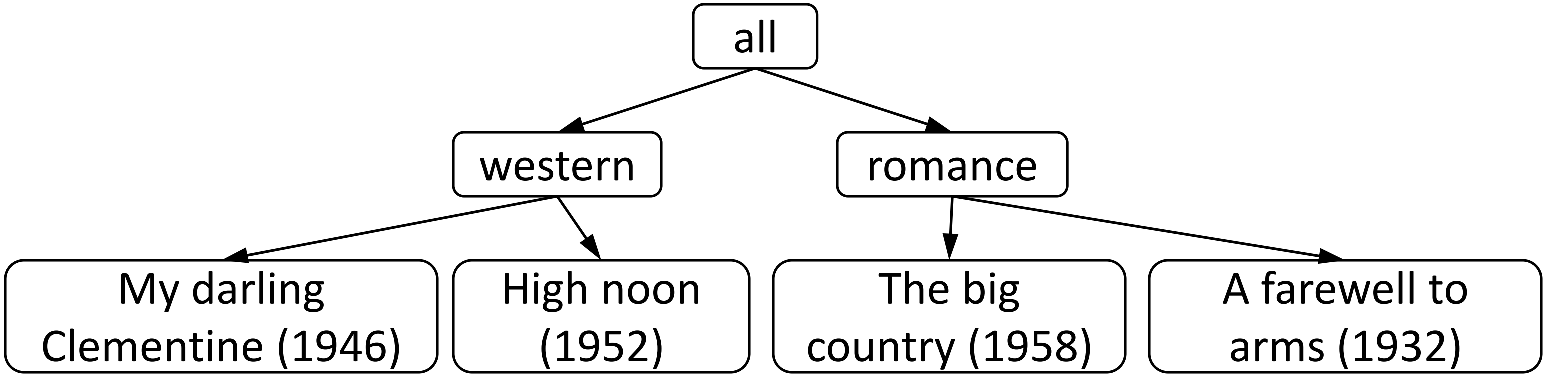
Example from Groceries dataset

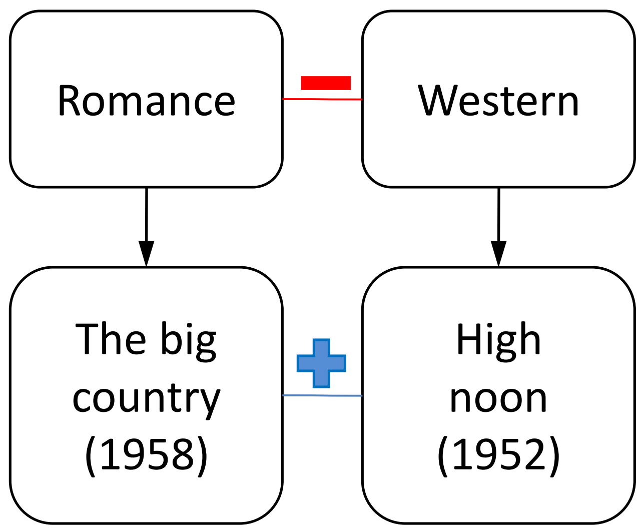


A

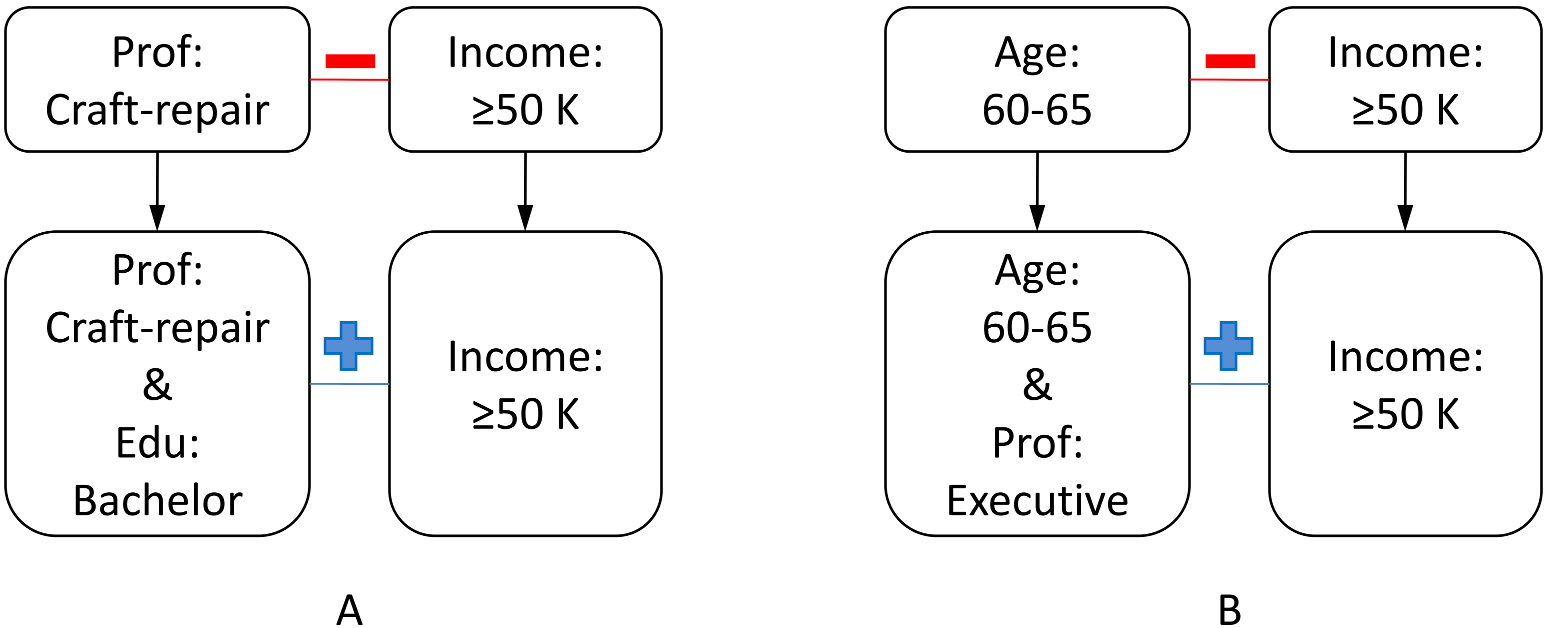
B

Examples from Movie rating dataset





Examples from US census dataset



Examples from medical papers dataset