# 无监督学习

黄书剑

- 无监督学习
- 聚类分析
  - k均值聚类
- 关联分析
  - Apriori
- 异常检测

# 无监督学习

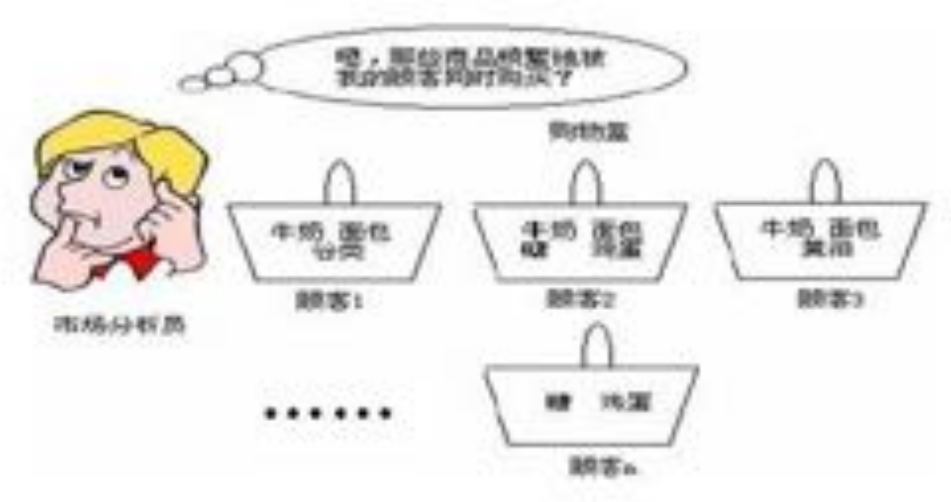
- Unsupervised machine learning algorithms infer patterns from a dataset without reference to known, or labeled, outcomes.

- "Mining" / infer patterns from examples $x_i$

- 维度约简 Dimension Reduction
- 聚类 Clustering
- 关联分析 Association Analysis
- 异常检测 Anomaly Detection

# 关联分析

- 发掘元素集合中潜在的关联性
  - 商品布局、购物习惯分析
  - 网页访问日志
  - 基因关联性



| TID | Items |
|---|---|
| t1 | {牛奶,面包} |
| t2 | {面包,尿布,啤酒,鸡蛋} |
| t3 | {牛奶,尿布,啤酒,可乐} |
| t4 | {面包,牛奶,尿布,啤酒} |
| t5 | {面包,牛奶,尿布,可乐} |
| … | … |

{牛奶,面包,尿布}！

{牛奶,面包} → {尿布}！

# 基本概念

| TID | Items |
| --- | --- |
| t1 | {牛奶,面包} |
| t2 | {面包,尿布,啤酒,鸡蛋} |
| t3 | {牛奶,尿布,啤酒,可乐} |
| t4 | {面包,牛奶,尿布,啤酒} |
| t5 | {面包,牛奶,尿布,可乐} |
| … | … |

- 项/元素（item）
  - 如：面包、牛奶
- 项集（itemset）
  - 如：{面包、牛奶}
- k-项集（k-itemset）
  - 有k个项的项集
- 事务（transaction）
  - 如：$t_2$: {面包,尿布,啤酒,鸡蛋}
  - 事务中项的个数，也称为事务的宽度
  - 给定一系列事务的集合记为T

# 基本概念

- 项/元素、项集、k-项集、事务
- 关联分析
  - 从给定事务集合T中发掘：频繁项集（Frequent Itemset）和关联规则（Association Rule）
- 关联规则
  - X → Y ： X和Y是两个不相交的项集
  - 如：{牛奶,面包} → {尿布}

# 基本概念

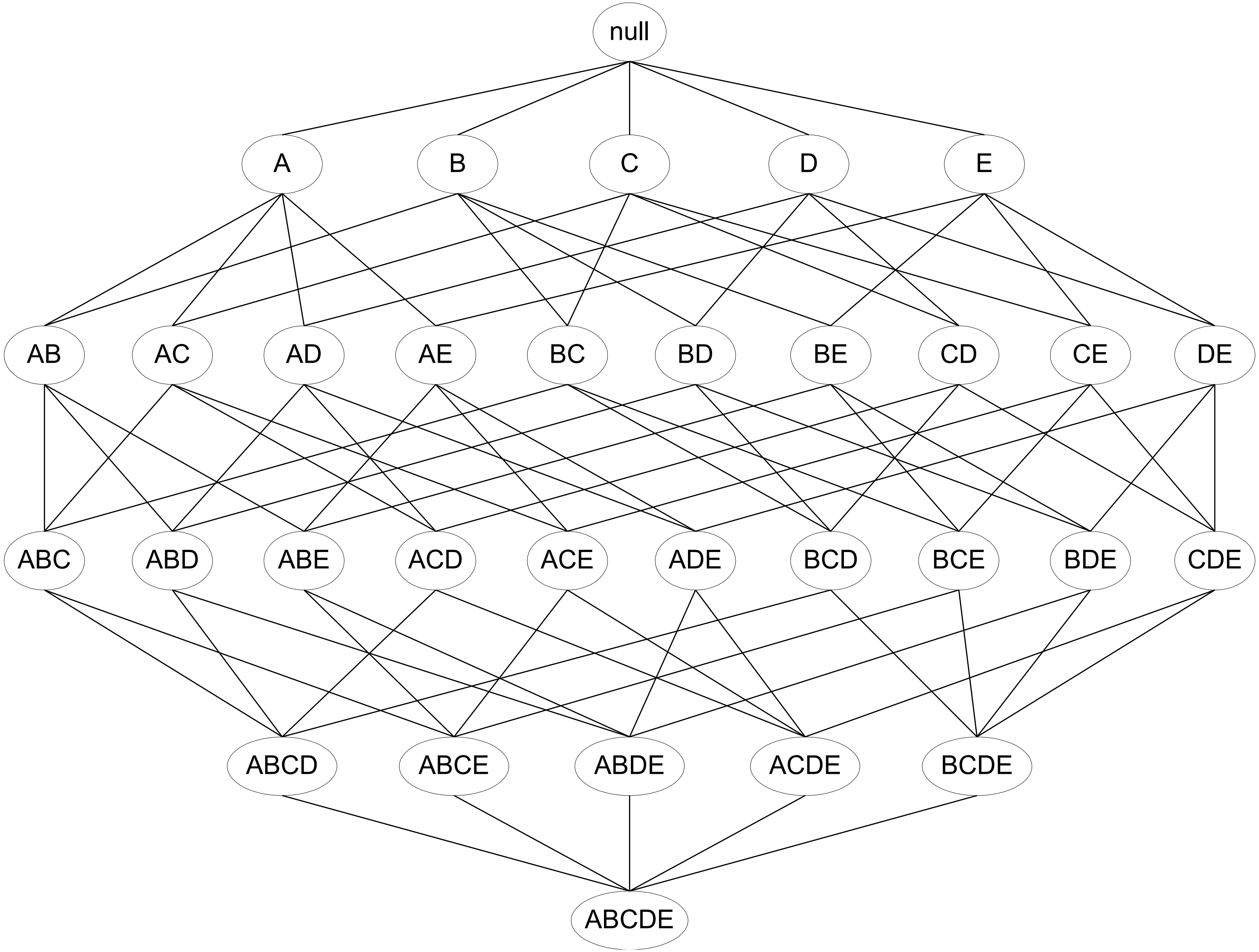
- 项/元素、项集、k-项集、事务、关联规则
- 关联分析: 频繁项集和关联规则
- 重要程度:
  - 在T中出现次数计为$\sigma$:
    - $\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}|$
  - 支持度support:
    - 给定事务集合T中出现的频繁程度（概率p(X)）
    - $s(X) = \frac{\sigma(X)}{N}$ ， $s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$
  - 置信度confidence:
    - 关联规则的可靠程度（条件概率p(Y|X)）
    - $c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$

# 实例:

| TID | Items |
|---|---|
| t1 | {牛奶,面包} |
| t2 | {面包,尿布,啤酒,鸡蛋} |
| t3 | {牛奶,尿布,啤酒,可乐} |
| t4 | {面包,牛奶,尿布,啤酒} |
| t5 | {面包,牛奶,尿布,可乐} |

- 给定右图的事务集合
- 要求 s>0.5, c>0.5
- 频繁项集:
  - {牛奶} 0.8、{面包} 0.8、{尿布} 0.8 、{啤酒} 0.6
  - {牛奶,面包} 0.6 、 {牛奶,尿布} 0.6 、{面包,尿布} 0.6 、 {啤酒,尿布} 0.6
- 关联规则:
  - {牛奶} -> {面包} 0.6，0.75
  - {啤酒} -> {尿布} 0.6，1
  - {尿布} -> {啤酒} 0.6，0.75
  - ……

# 首先考虑频繁项集


null
A
B
C
D
E
AB
AC
AD
AE
BC
BD
BE
CD
CE
DE
ABC
ABD
ABE
ACD
ACE
ADE
BCD
BCE
BDE
CDE
ABCD
ABCE
ABDE
ACDE
BCDE
ABCDE

# 蛮力方法（Brute-force）

- **穷举所有可能的项集，并依次为其计数**
  - 对每个事务，考察其包含的每一个项集
  - O(NMw)
    - M为项集候选数（2^n-1）
    - N为事务数、w为事务的宽度

| TID | Items |
|---|---|
| t1 | {牛奶,面包} |
| t2 | {面包,尿布,啤酒,鸡蛋} |
| t3 | {牛奶,尿布,啤酒,可乐} |
| t4 | {面包,牛奶,尿布,啤酒} |
| t5 | {面包,牛奶,尿布,可乐} |

| 候选项集 | 计数 |
|---|---|
| {xxx} | |
| {xxx} | |
| {xxx} | |
| {xxx} | |
| {xxx} | |
| … | … |

# Apriori原理

- 频繁项集的子集一定是频繁的（Any subset of a frequent itemset must be frequent)
  - 如果{牛奶,尿布,啤酒}是频繁的，{尿布,啤酒}一定是频繁的
  - 任何包含某项集的事务，一定包含其子项集

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

  - 不频繁项集的超集一定是不频繁的

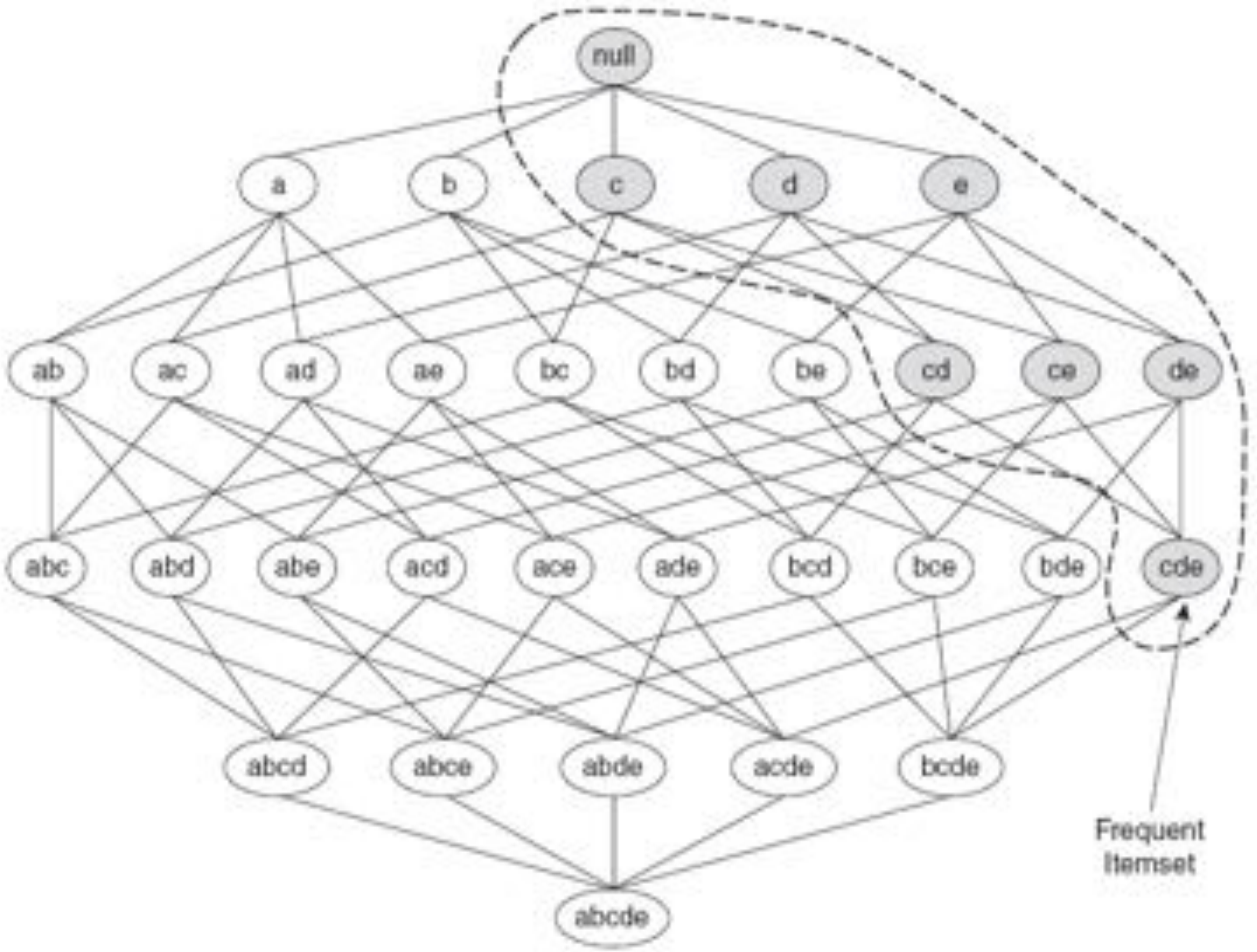
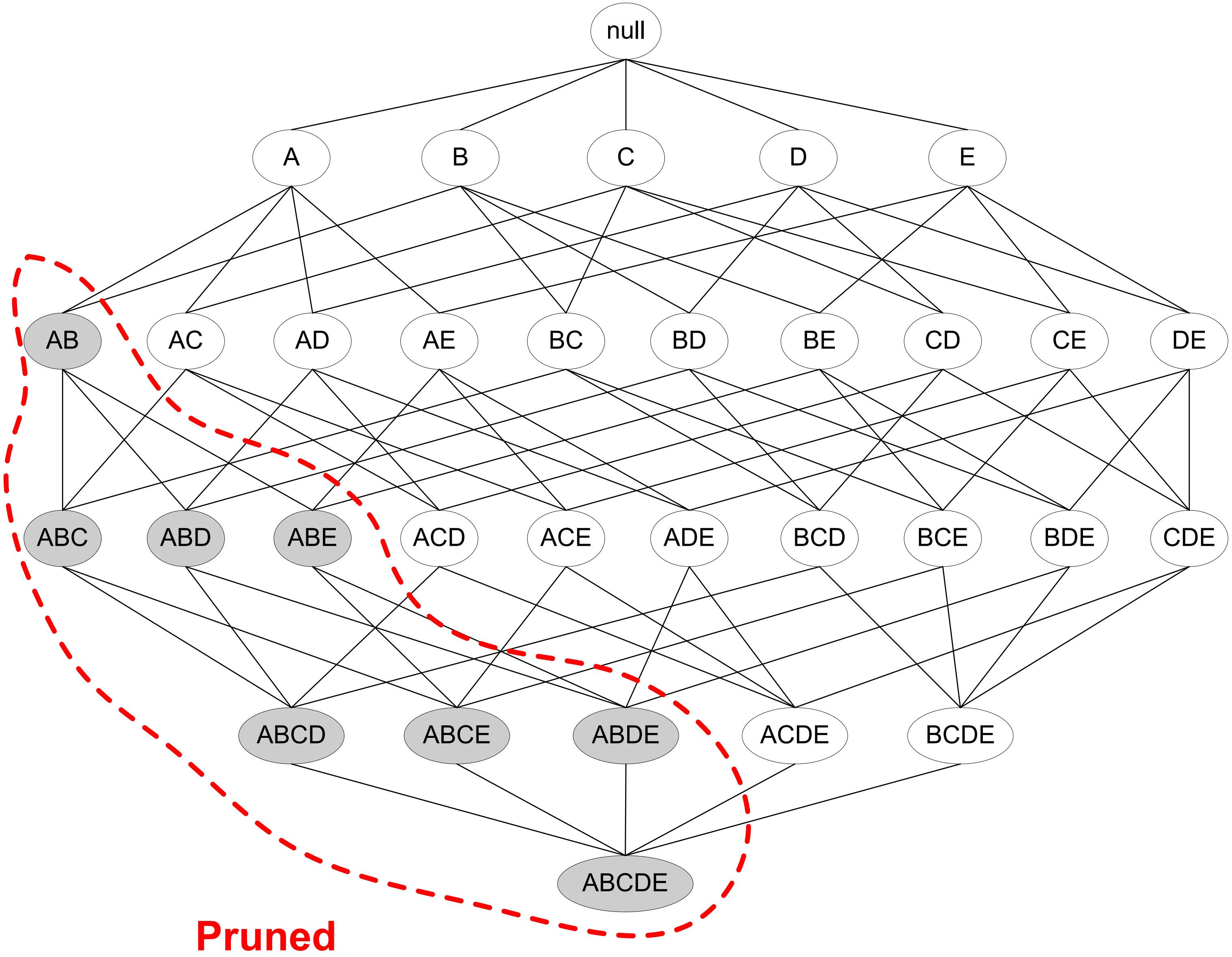



**Figure 6.3.** An illustration of the *Apriori* principle. If $\{c, d, e\}$ is frequent, then all subsets of this itemset are frequent.


null
A
B
C
D
E
AB
AC
AD
AE
BC
BD
BE
CD
CE
DE
ABC
ABD
ABE
ACD
ACE
ADE
BCD
BCE
BDE
CDE
ABCD
ABCE
ABDE
ACDE
BCDE
ABCDE
Pruned

# Apriori: A Candidate Generation & Test Approach

- Apriori pruning principle: If there is any itemset which is infrequent, its superset should not be generated/tested! (Agrawal & Srikant @VLDB'94, Mannila, et al. @ KDD' 94)
- Method:
  - Initially, scan DB once to get frequent 1-itemset
  - Generate length (k+1) candidate itemsets from length k frequent itemsets
  - Test the candidates against DB
  - Terminate when no frequent or candidate set can be generated

From Professor Han https://hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm 

# The Apriori Algorithm—An Example

$Sup_{min} = 2$

Database TDB

| Tid | Items |
|---|---|
| 10 | A, C, D |
| 20 | B, C, E |
| 30 | A, B, C, E |
| 40 | B, E |

$1^{st}$ scan →

$C_1$

| Itemset | sup |
|---|---|
| {A} | 2 |
| {B} | 3 |
| {C} | 3 |
| {D} | 1 |
| {E} | 3 |

→ $L_1$

| Itemset | sup |
|---|---|
| {A} | 2 |
| {B} | 3 |
| {C} | 3 |
| {E} | 3 |

$C_2$

| Itemset |
|---|
| {A, B} |
| {A, C} |
| {A, E} |
| {B, C} |
| {B, E} |
| {C, E} |

$2^{nd}$ scan ←

$C_2$

| Itemset | sup |
|---|---|
| {A, B} | 1 |
| {A, C} | 2 |
| {A, E} | 1 |
| {B, C} | 2 |
| {B, E} | 3 |
| {C, E} | 2 |

← $L_2$

| Itemset | sup |
|---|---|
| {A, C} | 2 |
| {B, C} | 2 |
| {B, E} | 3 |
| {C, E} | 2 |

$C_3$

| Itemset |
|---|
| {B, C, E} |

$3^{rd}$ scan →

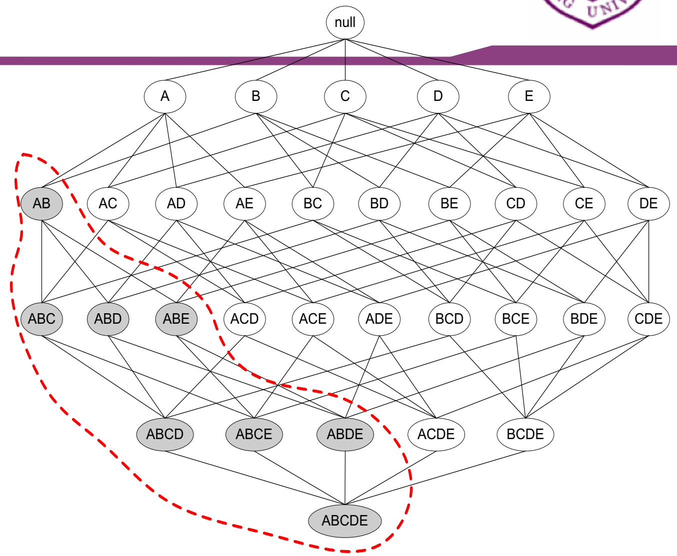
$L_3$

| Itemset | sup |
|---|---|
| {B, C, E} | 2 |

From Professor Han https://hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm

# 如何生成候选集合?

- 蛮力方法
  - 穷举所有可能，并按照前述剪枝
- $F_{k-1} * F_1$
  - 从已有的k-1频繁项集扩展
- $F_{k-1} * F_{k-1}$
  - 所有的k-1子项都应该是频繁的




- 如何更高效的生成候选?
  - 避免重复候选保持字典顺序

# 基本概念（回顾）

- 项/元素、项集、k-项集、事务、关联规则
- 关联分析: 频繁项集和关联规则
- 重要程度:
  - 在T中出现次数计为$\sigma$:
    - $\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}|$
  - 支持度support:
    - 给定事务集合T中出现的频繁程度（概率p(X)）
    - $s(X) = \frac{\sigma(X)}{N}$ ，  $s(X \to Y) = \frac{\sigma(X \cup Y)}{N}$
  - 置信度confidence:
    - 关联规则的可靠程度（条件概率p(Y|X)）
    - $c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$

# 关联规则生成

- **关联规则($X \rightarrow Y$)的项集$X \cup Y$是频繁的**
  - 首先得到符合支持度要求的k-频繁项集（记为Y）
  - 将该项集划分为两个非空子集X和Y-X，则得到关联规则（X→Y-X）
    - 逐层生成，每层规则后件的项数增大
    - 检查置信度要求
- **如果规则X->Y-X不满足置信度要求，则X’->Y-X’也一定不满足，其中X’为X的子集**

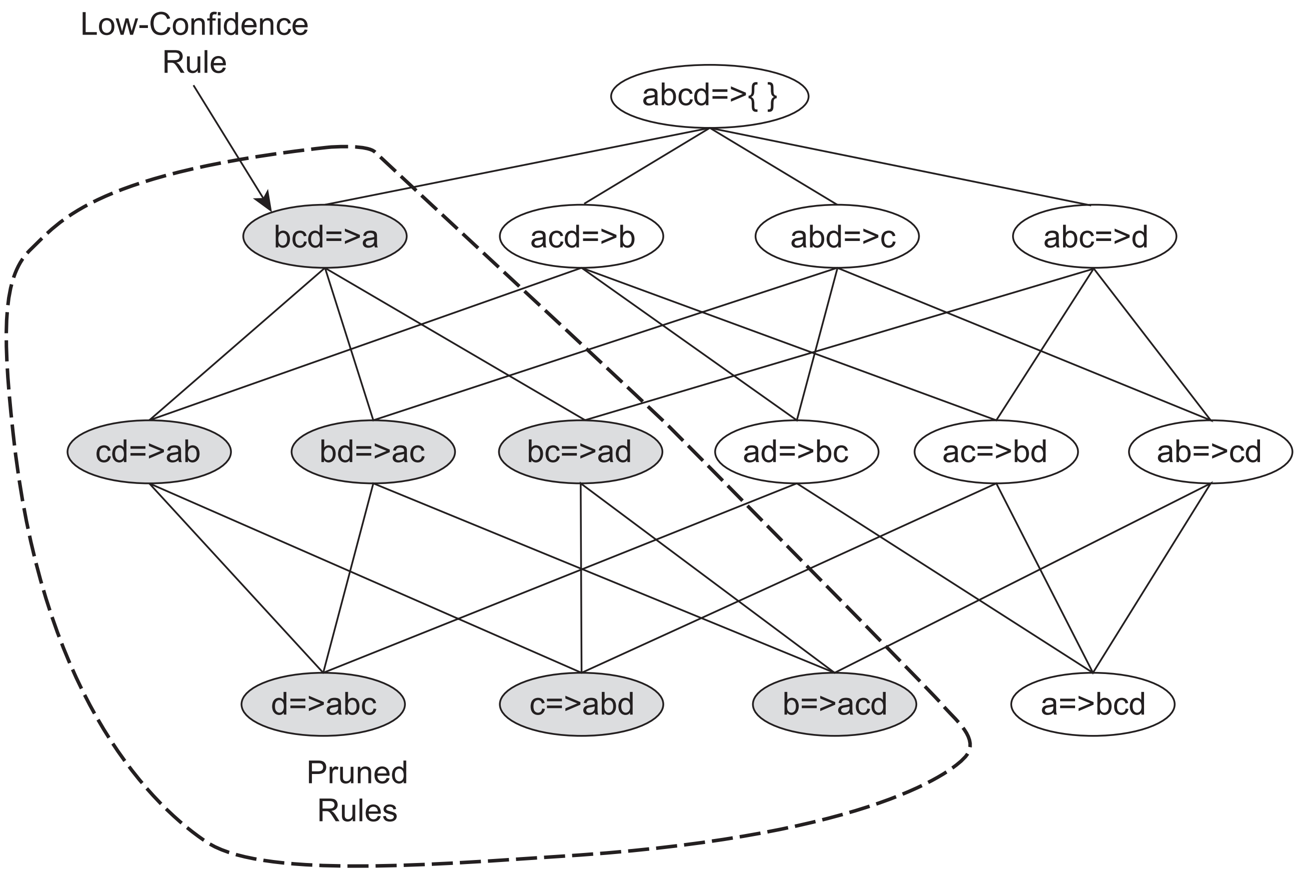

Low-Confidence Rule
abcd=>{ }
bcd=>a
acd=>b
abd=>c
abc=>d
cd=>ab
bd=>ac
bc=>ad
ad=>bc
ac=>bd
ab=>cd
d=>abc
c=>abd
b=>acd
a=>bcd
Pruned Rules

**Table 5.3.** List of binary attributes from the 1984 United States Congressional Voting Records. Source: The UCI machine learning repository.

| | |
|---|---|
| 1. Republican | 18. aid to Nicaragua = no |
| 2. Democrat | 19. MX-missile = yes |
| 3. handicapped-infants = yes | 20. MX-missile = no |
| 4. handicapped-infants = no | 21. immigration = yes |
| 5. water project cost sharing = yes | 22. immigration = no |
| 6. water project cost sharing = no | 23. synfuel corporation cutback = yes |
| 7. budget-resolution = yes | 24. synfuel corporation cutback = no |
| 8. budget-resolution = no | 25. education spending = yes |
| 9. physician fee freeze = yes | 26. education spending = no |
| 10. physician fee freeze = no | 27. right-to-sue = yes |
| 11. aid to El Salvador = yes | 28. right-to-sue = no |
| 12. aid to El Salvador = no | 29. crime = yes |
| 13. religious groups in schools = yes | 30. crime = no |
| 14. religious groups in schools = no | 31. duty-free-exports = yes |
| 15. anti-satellite test ban = yes | 32. duty-free-exports = no |
| 16. anti-satellite test ban = no | 33. export administration act = yes |
| 17. aid to Nicaragua = yes | 34. export administration act = no |

https://archive.ics.uci.edu/ml/datasets/congressional+voting+records

| Association Rule | Confidence |
|---|---|
| $\{$budget resolution $=$ no, MX-missile$=$no, aid to El Salvador $=$ yes $\}$ $\longrightarrow \{$Republican$\}$ | 91.0% |
| $\{$budget resolution $=$ yes, MX-missile$=$yes, aid to El Salvador $=$ no $\}$ $\longrightarrow \{$Democrat$\}$ | 97.5% |
| $\{$crime $=$ yes, right-to-sue $=$ yes, physician fee freeze $=$ yes$\}$ $\longrightarrow \{$Republican$\}$ | 93.5% |
| $\{$crime $=$ no, right-to-sue $=$ no, physician fee freeze $=$ no$\}$ $\longrightarrow \{$Democrat$\}$ | 100% |

https://archive.ics.uci.edu/ml/datasets/congressional+voting+records

# 练习五

- 尝试实现一个简单的Apriori算法，比较不同实现的性能差距
- 尝试观察原有数据中的异常分布

# 参考资料

- 本章大部分内容来源于以下两个课程的相关部分：
  - Introduction to Data Mining (Second Edition) https://www-users.cs.umn.edu/~kumar001/dmbook/index.php
  - Data Mining: Concepts and Techniques, 3rd ed. https://hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm