关联分析-习题（200 分）

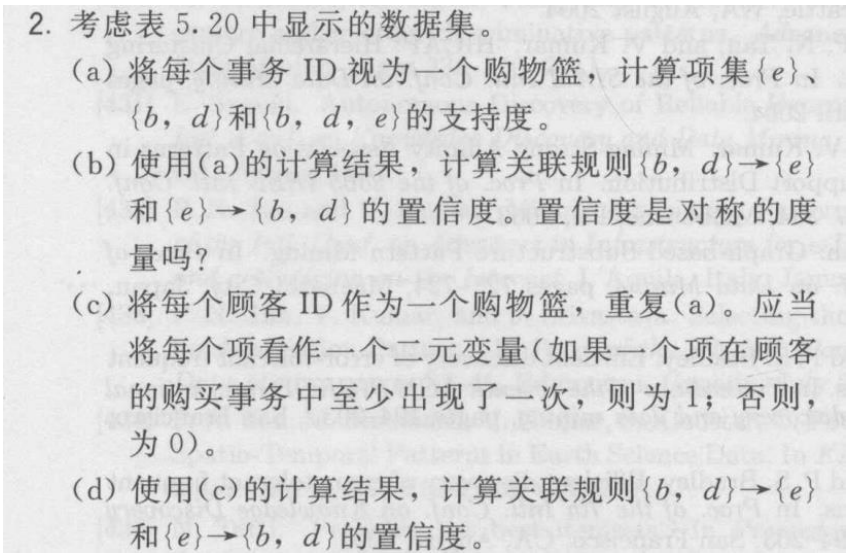
专业班级：18 级软件工程（NIIT）大数据三班

学号：20181620310145

姓名：李辰星

[计算题] （30 分）1. 课本 256 页，第 2 题 (a)-(d)

批注 [s1]: 30



2. 考虑表 5.20 中显示的数据集。

(a) 将每个事务 ID 视为一个购物篮，计算项集{e}、{b，d}和{b，d，e}的支持度。

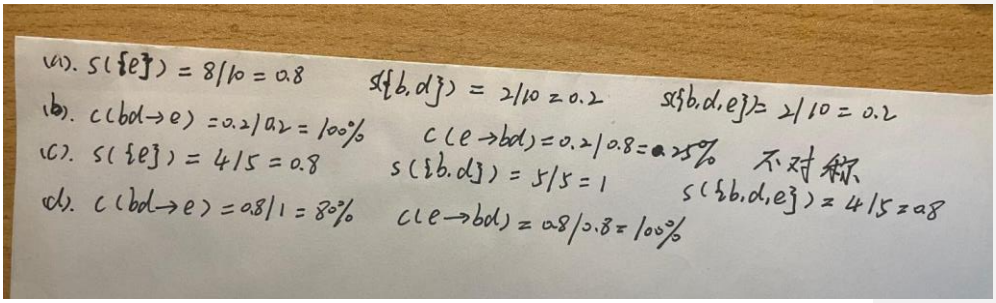
(b) 使用(a)的计算结果，计算关联规则{b，d}→{e}和{e}→{b，d}的置信度。置信度是对称的度量吗?

(c) 将每个顾客 ID 作为一个购物篮，重复(a)。应当将每个项看作一个二元变量(如果一个项在顾客的购买事务中至少出现了一次，则为 1；否则，为 0)。

(d) 使用(c)的计算结果，计算关联规则{b，d}→{e}和{e}→{b，d}的置信度。

表 5.20　购物篮事务的例子

顾客 ID	事务 ID	购买项
1	0001	{a, d, e}
1	0024	{a, b, c, e}
2	0012	{a, b, d, e}
2	0031	{a, c, d, e}
3	0015	{b, c, e}
3	0022	{b, d, e}
4	0029	{c, d}
4	0040	{a, b, c}
5	0033	{a, d, e}
5	0038	{a, b, e}



[选择题] （10 分）2. 根据 Apriori 的先验原理，项集{ABC}的支持度__A__项集{AB}的支持度。

A. 小于或等于　B. 大于或等于

批注 [s2]: 10

[选择题] （10 分） 3. 根据 Apriori 的先验原理，假设{ABCD}是一个频繁 4-项集，规则 ABC → D 的置信度__B__规则 AB → CD 的置信度。

B. 小于或等于　B. 大于或等于

批注 [s3]: 10

[计算题]（**30 分**）4. 一个数据集有 4 次交易，令 min_sup=60% ，min_conf=80%.

批注 [s4]: 30

TID	Date	Items_bought
100	10/15/99	{K,A,D,B}
200	10/15/99	{D,A,C,E,B}
300	10/19/99	{C,A,B,E}
400	10/22/99	{B,A,D}

a) 使用 Aprior 算法找到所有频繁 1-项集，频繁 2-项集和频繁 3-项集。

b) 列出所有强规则，即置信度大于等于 80%的规则。

a) 4x60%=2.4

频繁 1-项集：{A}、{B}、{D}；

频繁 2-项集：{AB}、{AD}、{BD}；

频繁 3-项集：{ABD}

b)AD->B cf=3/3=100%

BD->A cf=3/3=100%

D->AB cf=3/3=100%

[简答题] （30 分）5. 根据表 5.22 的数据集和图 5.33 给出的项集格，用 N 或 F 或 I 在项集格上为每个节点进行标记。N, F, I 的定义如下：

批注 [s5]: 30

Apriori 算法使用产生-计数的策略找出频繁项集。通过合并一堆大小为 k 的频繁项集得到一个大小为 $k+1$ 的候选项集(称作候选产生步骤)。在候选项集剪枝步骤中，如果一个候选项集的任何一个子集是不频繁的，则该候选项集将被丢弃。假定将 Apriori 算法用于表 5.22 所示的数据集，最小支持度为 30%，即任何一个项集在少于 3 个事务中出现就被认为是非频繁的。

表 5.22　购物篮事务的例子

事务 ID	购买项
1	{a, b, d, e}
2	{b, c, d}
3	{a, b, d, e}
4	{a, c, d, e}
5	{b, c, d, e}
6	{b, d, e}
7	{c, d}
8	{a, b, c}
9	{a, d, e}
10	{b, d}

画出表示表 5.22 所示数据集的项集格。用下面的字母标记格中的每个结点。

- N：如果该项集被 Apriori 算法认为不是候选项集。一个项集不是候选项集有两种可能的原因：它不是在候选项集产生步骤中产生，或它在候选项集产生步骤中产生，但是由于它的一个子集是非频繁的而在候选项集剪枝步骤被丢掉。
- F：如果该项集被 Apriori 算法认为是频繁的。
- I：如果经过支持度计数后，该项集被发现是非频繁的。

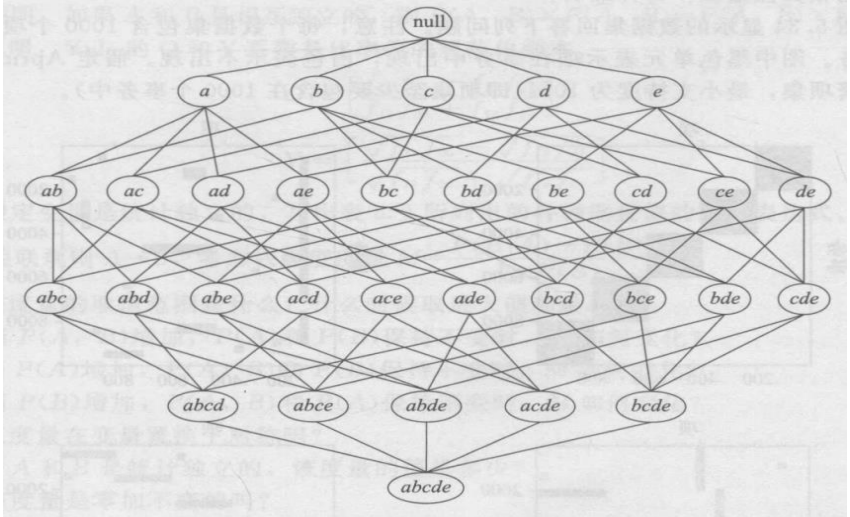
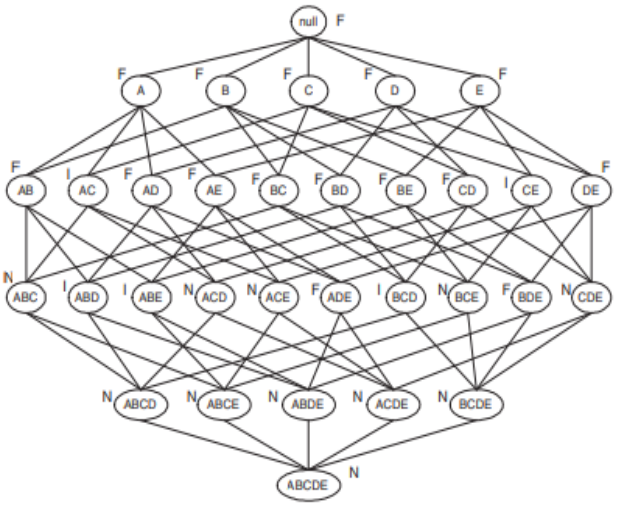


图 5.33　项集的格



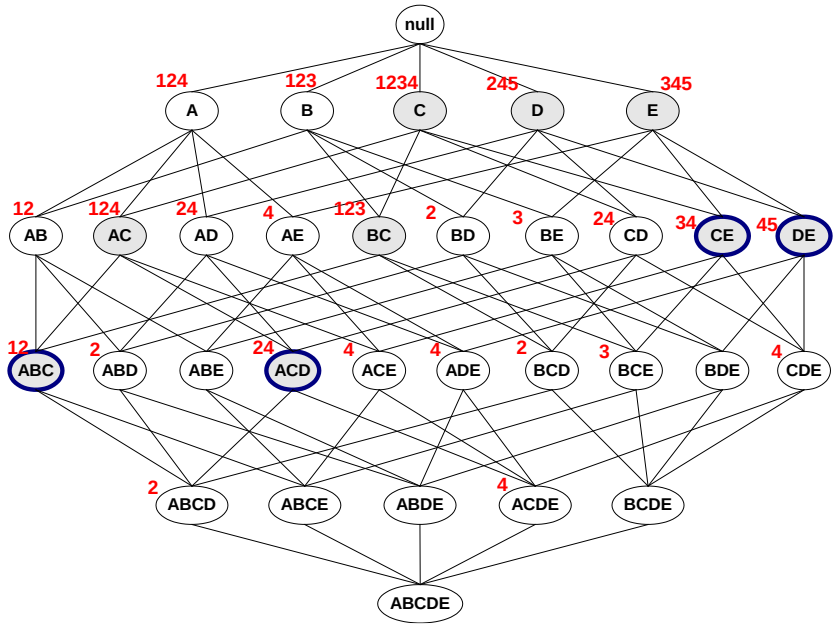
[简答题] （30 分）6. 给定下图的项集格，令最小支持度计数的阈值为 2，

6.1 说明为什么项集{C}是闭项集？

6.2 说明为什么项集{C}不是极大频繁项集？

6.3 说明为什么项集{CE}是极大频繁项集？

批注 [s6]: 30



注：每个项集上方的数字代表该项集出现的交易的 ID. 如项集{A}上方的 124 代表项集{A}出现在交易 1、2、4 之中，即项集{A}的支持度计数为 3.

答：

项集{C}它的直接超集的支持度计数都不等于它本身的支持度计数， 则项集{C}是闭的

项集{C}的所有直接超集都是频繁的

项集{CE}是频繁的，而它的所有直接超集都不是频繁的，那么它就是极大频繁项集

[计算题] （30 分）7.

批注 [s7]: 30

20. 考虑表 5.25 中显示的列联表。

表 5.25 习题 20 的列联表

	B	$\overline{B}$
A	9	1
$\overline{A}$	1	89

a）表Ⅰ

	B	$\overline{B}$
A	89	1
$\overline{A}$	1	9

b）表Ⅱ

（a）对于表Ⅰ，计算关联模式$\{A, B\}$的支持度、兴趣度和 ϕ 相关系数，并计算规则 $A \rightarrow B$和 $B \rightarrow A$ 的置信度。

（b）对于表Ⅱ，计算关联模式$\{A, B\}$的支持度、兴趣度和 ϕ 相关系数，并计算规则 $A \rightarrow B$和 $B \rightarrow A$ 的置信度。

（c）由(a)和(b)的结果可以得出什么结论？

(a) s(A)=0.1, s(B)=0.9,

s(A, B)=0.09. I(A, B) = 9, φ(A, B)=0.89.

c(A −→ B)=0.9, c(B −→ A)=0.9.

(b) s(A)=0.置信度 s(B)=0.9,

s(A, B)=0.89. I(A, B)=1.09, φ(A, B)=0.89.

c(A −→ B)=0.98, c(B −→ A)=0.98.

(c) 兴趣、支持度置信度是变化的，而 ϕ 系数在求逆运算下是不变的。

[计算题] （30 分）8.

批注 [s8]: 30

21. 考虑 5.17 和 5.18 中显示的购买高清晰度电视和购买健身器的顾客之间的联系。

（a）计算两个表的比率值。

（b）计算两个表的 ϕ 系数。

（c）计算两个表的兴趣因子。

对于上述每一个度量，描述当汇总数据取代分层数据后，关联方向的变化。

表 5.17 HDTV 和健身器销售之间的 2 路列联表

买 HDTV	买健身器		
	是	否	
是	99	81	180
否	54	66	120
	153	147	300

表 5. 18　HDTV 和健身器销售之间的 3 路列联表

顾客组	买 HDTV	买健身器		总数
		是	否	
大学生	是	1	9	10
	否	4	30	34
在职人员	是	98	72	170
	否	50	36	86

（a） 优势比：99x66/54x81=1.4938

1x30/4x9=0.833

98x36/72x50=0.98

（b）

（c）兴趣因子 I=99/300/ 180/300 x 153/300=1.0784

兴趣因子 I=1/44/ 5/44 x 10/44=0.88

兴趣因子 I=98/256/ 148/256 x 170/256=0.9971