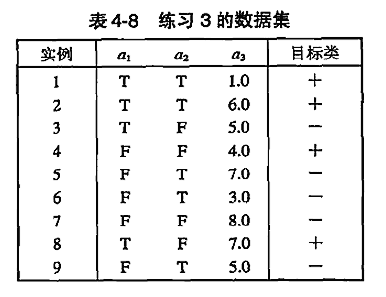
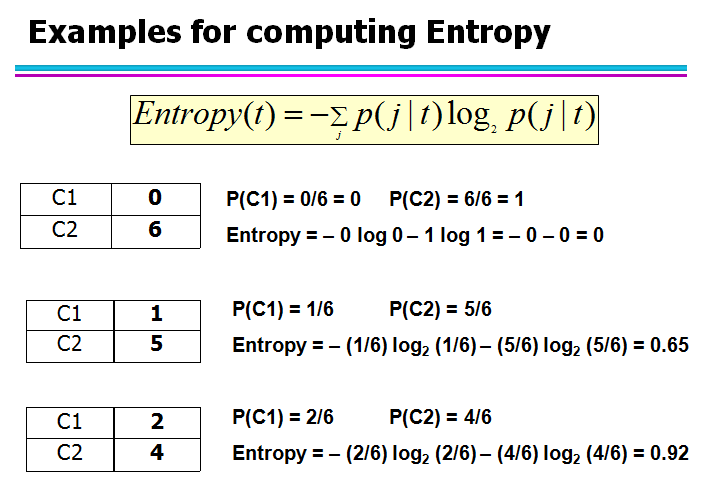
1. 考虑表中二元分类问题的训练样本集



1. 整个训练样本集关于类属性的熵是多少？
2. 关于这些训练集中a1,a2的信息增益是多少？
3. 对于连续属性a3,计算所有可能的划分的信息增益。
4. 根据信息增益，a1,a2,a3哪个是最佳划分?
5. 根据分类错误率，a1,a2哪具最佳？
6. 根据gini指标，a1,a2哪个最佳？

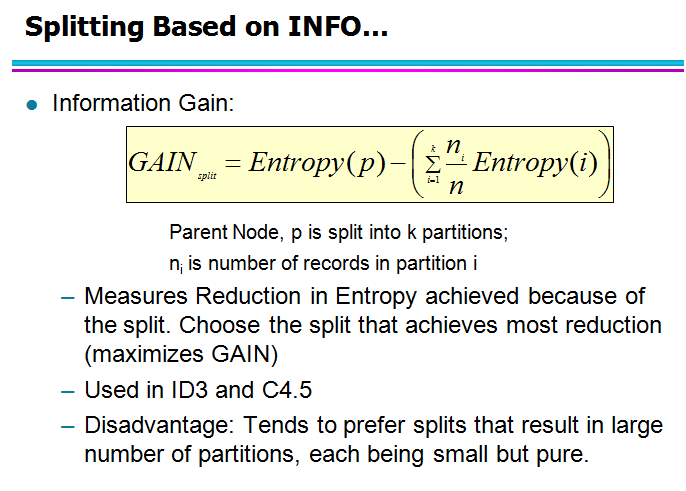
答1.

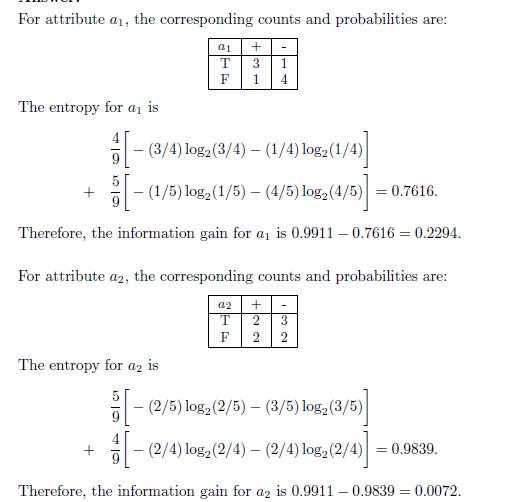


*P*(+) = 4*/*9 and *P*(*−*) = 5*/*9

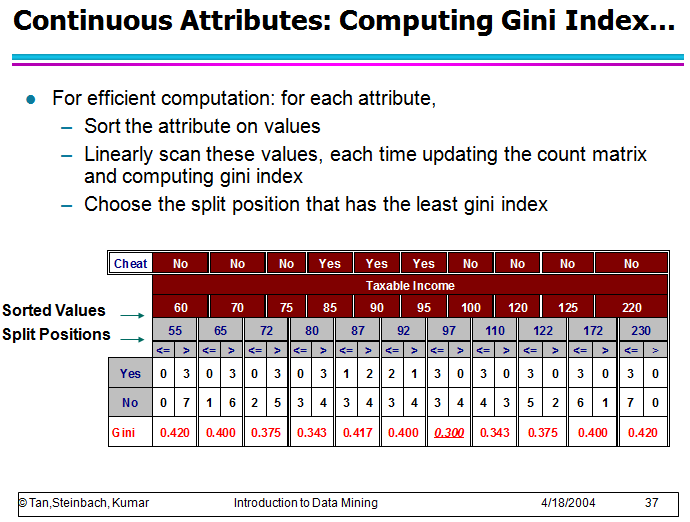
*−*4*/*9 log2(4*/*9) *−* 5*/*9 log2(5*/*9) = 0*.*9911*.*

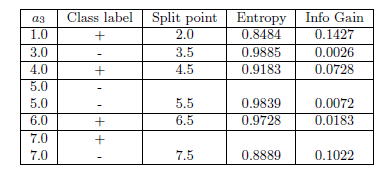
答2：

（估计不考）



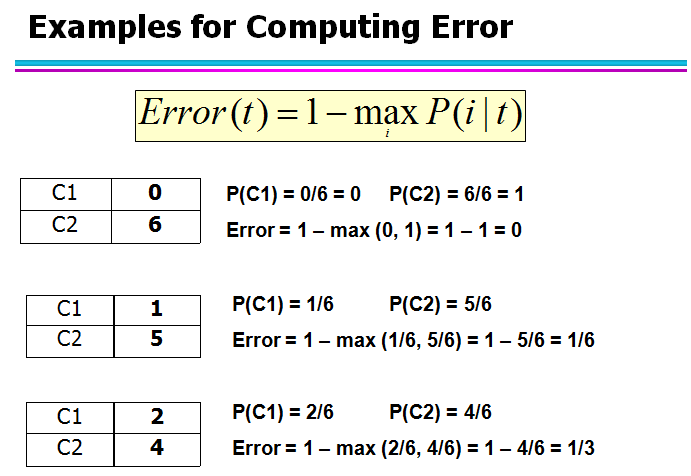
答3：





答4: According to information gain, *a*1 produces the best split.

答5：

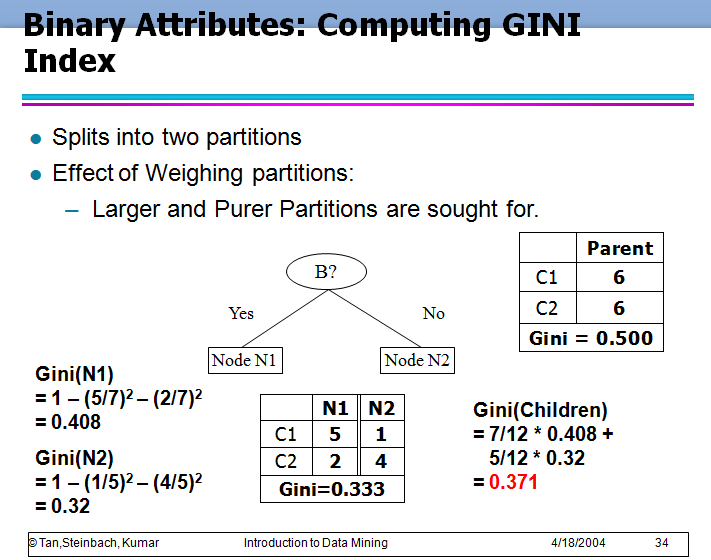


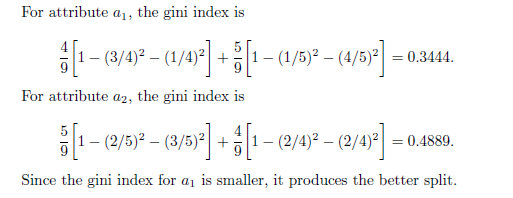
For attribute *a*1: error rate = 2*/*9.

For attribute *a*2: error rate = 4*/*9.

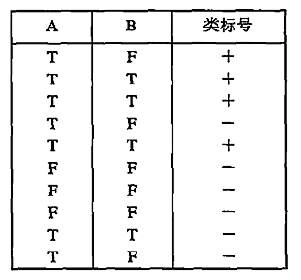
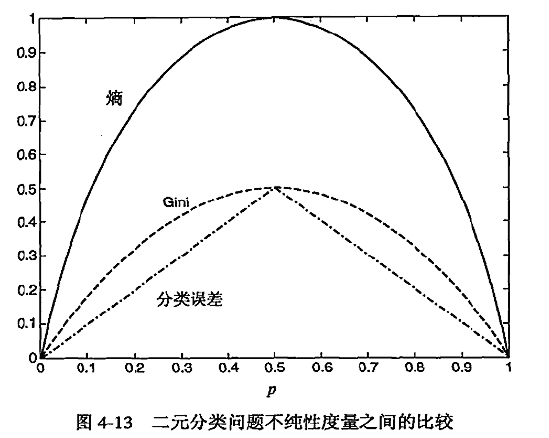
Therefore, according to error rate, *a*1 produces the best split.

答6：

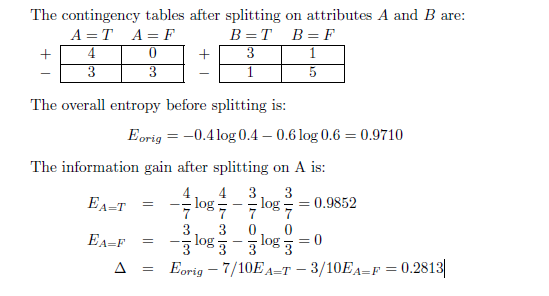


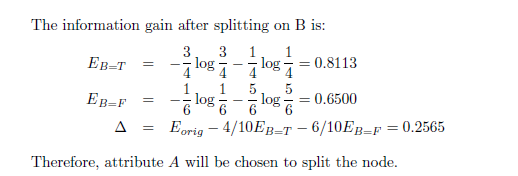


1. 考虑如下二元分类问题的数据集

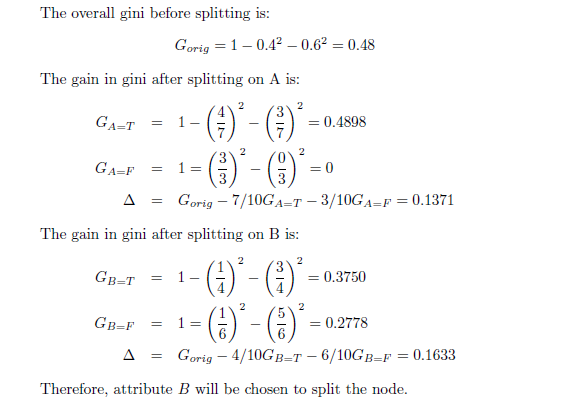
 

1. 计算a.b信息增益，决策树归纳算法会选用哪个属性





1. 计算a.b gini指标，决策树归纳会用哪个属性？



这个答案没问题

1. 从图4-13可以看出熵和gini指标在[0,0.5]都是单调递增，而[0.5,1]之间单调递减。有没有可能信息增益和gini指标增益支持不同的属性？解释你的理由

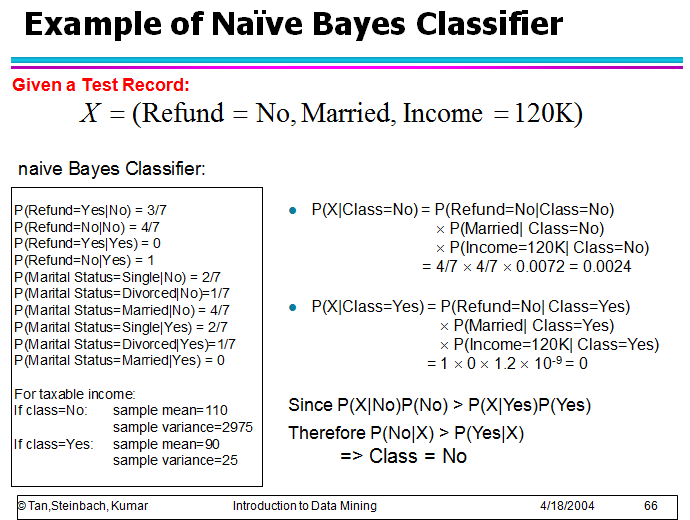
Yes, even though these measures have similar range and monotonous

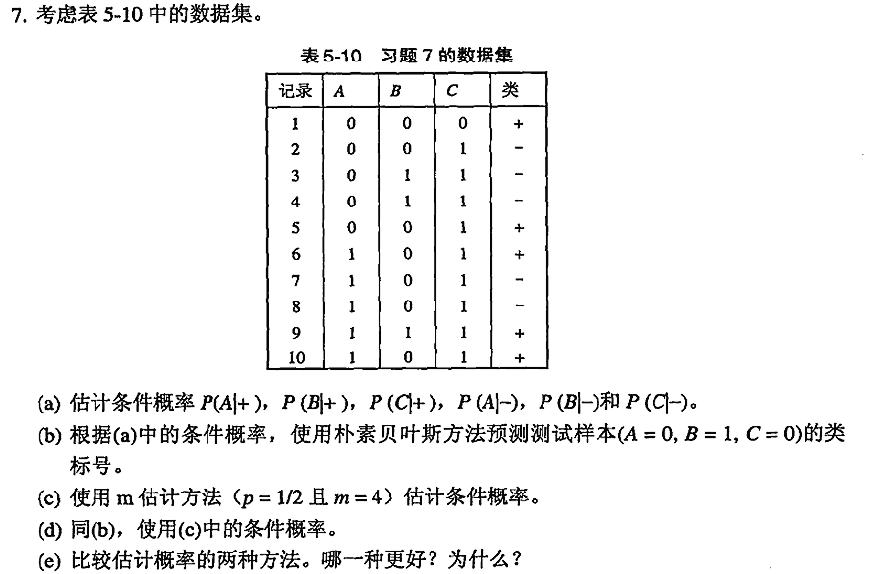
behavior, their respective gains, Δ, which are scaled differences of the

measures, do not necessarily behave in the same way, as illustrated by

the results in parts (a) and (b).

贝叶斯分类





1. *P*(*A* = 1*|−*) = 2*/*5 = 0*.*4, *P*(*B* = 1*|−*) = 2*/*5 = 0*.*4,

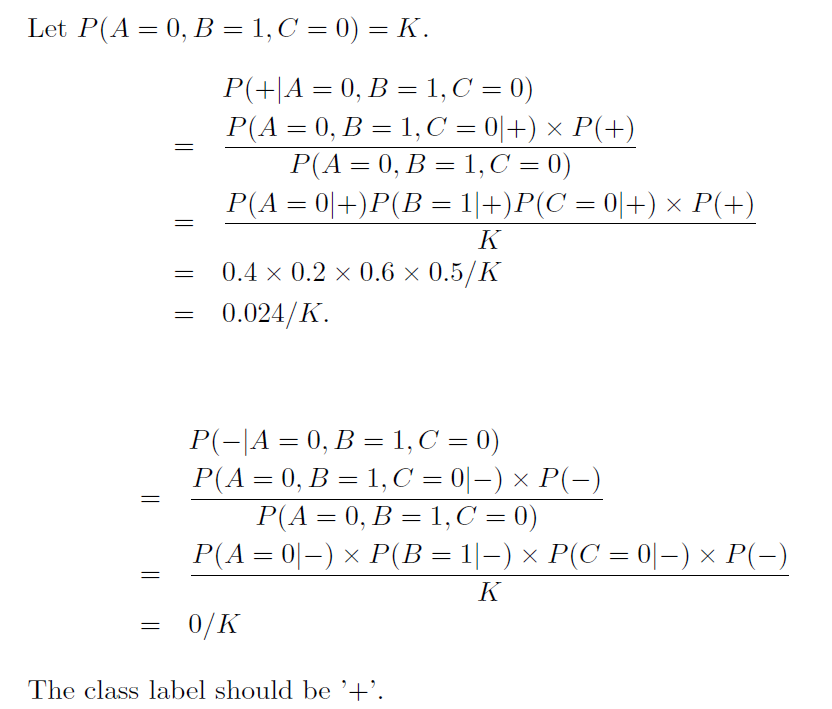
*P*(*C* = 1*|−*) = 1, *P*(*A* = 0*|−*) = 3*/*5 = 0*.*6,

*P*(*B* = 0*|−*) = 3*/*5 = 0*.*6*, P*(*C* = 0*|−*) = 0; *P*(*A* = 1*|*+) = 3*/*5 = 0*.*6,

*P*(*B* = 1*|*+) = 1*/*5 = 0*.*2, *P*(*C* = 1*|*+) = 2*/*5 = 0*.*4,

*P*(*A* = 0*|*+) = 2*/*5 = 0*.*4, *P*(*B* = 0*|*+) = 4*/*5 = 0*.*8,

*P*(*C* = 0*|*+) = 3*/*5 = 0*.*6.

1. 
2. *P*(*A* = 0*|*+) = (2 + 2)*/*(5 + 4) = 4*/*9,

*P*(*A* = 0*|−*) = (3+2)*/*(5 + 4) = 5*/*9,

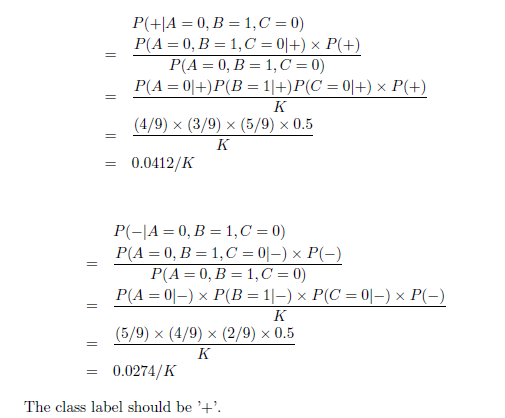
*P*(*B* = 1*|*+) = (1 + 2)*/*(5 + 4) = 3*/*9,

*P*(*B* = 1*|−*) = (2+2)*/*(5 + 4) = 4*/*9,

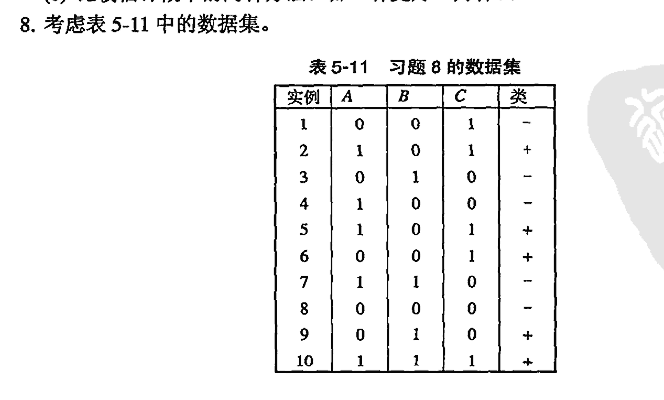
*P*(*C* = 0*|*+) = (3 + 2)*/*(5 + 4) = 5*/*9,

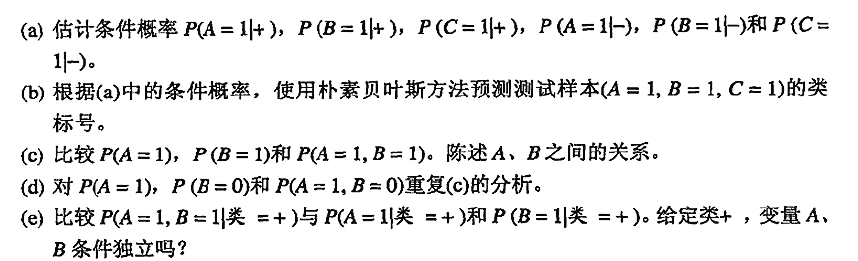
*P*(*C* = 0*|−*) = (0+2)*/*(5 + 4) = 2*/*9.

1. Let *P*(*A* = 0*,B* = 1*, C* = 0) = *K*



1. 当的条件概率之一是零，则估计为使用m-估计概率的方法的条件概率是更好的，因为我们不希望整个表达式变为零。





1. *P*(*A* = 1*|*+) = 0*.*6, *P*(*B* = 1*|*+) = 0*.*4, *P*(*C* = 1*|*+) = 0*.*8, *P*(*A* =

1*|−*) = 0*.*4, *P*(*B* = 1*|−*) = 0*.*4, and *P*(*C* = 1*|−*) = 0*.*2

2.

Let *R* : (*A* = 1*,B* = 1*, C* = 1) be the test record. To determine its

class, we need to compute *P*(+*|R*) and *P*(*−|R*). Using Bayes theorem, *P*(+*|R*) = *P*(*R|*+)*P*(+)*/P*(*R*) and *P*(*−|R*) = *P*(*R|−*)*P*(*−*)*/P*(*R*).

Since *P*(+) = *P*(*−*) = 0*.*5 and *P*(*R*) is constant, *R* can be classified by

comparing *P*(+*|R*) and *P*(*−|R*).

For this question,

*P*(*R|*+) = *P*(*A* = 1*|*+) *× P*(*B* = 1*|*+) *× P*(*C* = 1*|*+) = 0*.*192

*P*(*R|−*) = *P*(*A* = 1*|−*) *× P*(*B* = 1*|−*) *× P*(*C* = 1*|−*) = 0*.*032

Since *P*(*R|*+) is larger, the record is assigned to (+) class.

3.

*P*(*A* = 1) = 0*.*5, *P*(*B* = 1) = 0*.*4 and *P*(*A* = 1*,B* = 1) = *P*(*A*) *×*

*P*(*B*) = 0*.*2. Therefore, *A* and *B* are independent.

4.

*P*(*A* = 1) = 0*.*5, *P*(*B* = 0) = 0*.*6, and *P*(*A* = 1*,B* = 0) = *P*(*A* =1)*× P*(*B* = 0) = 0*.*3. *A* and *B* are still independent.

5.

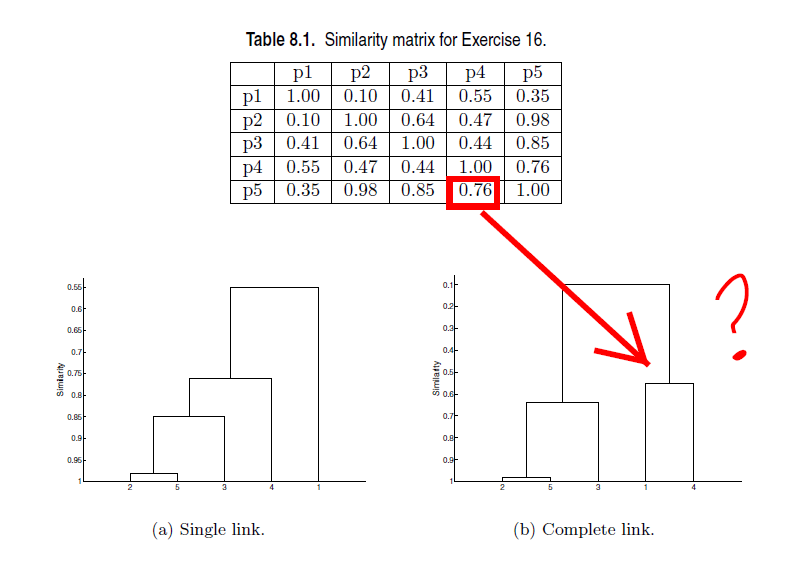
Compare *P*(*A* = 1*,B* = 1*|*+) = 0*.*2 against *P*(*A* = 1*|*+) = 0*.*6 and

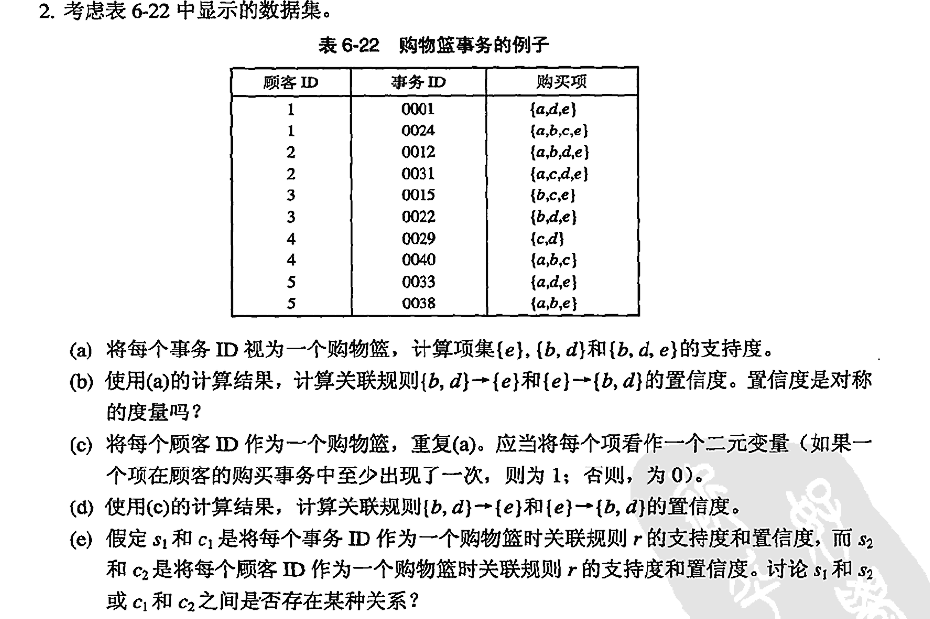
*P*(*B* = 1*|Class* = +) = 0*.*4. Since the product between *P*(*A* = 1*|*+)

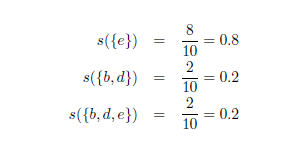
and *P*(*A* = 1*|−*) are not the same as *P*(*A* = 1*,B* = 1*|*+), *A* and *B* are

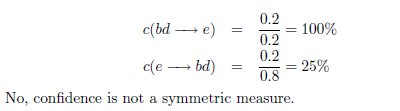
not conditionally independent given the class.

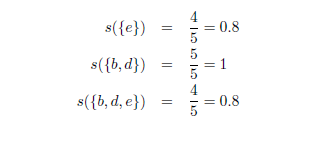
1. 使用下表中的相似度矩阵进行单链和全链层次聚类。绘制树状况显示结果，树状图应该清楚地显示合并的次序。

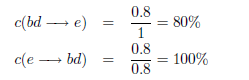


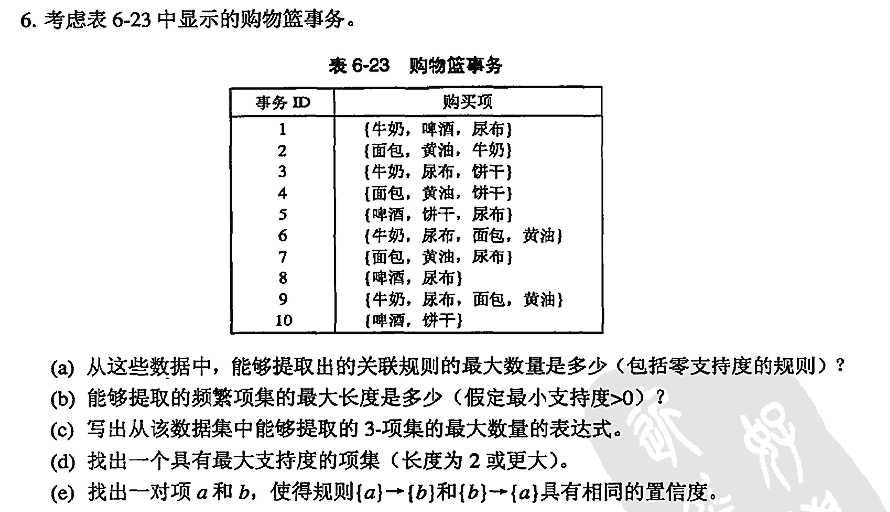


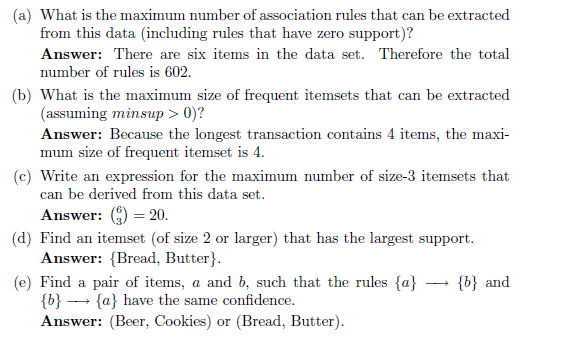


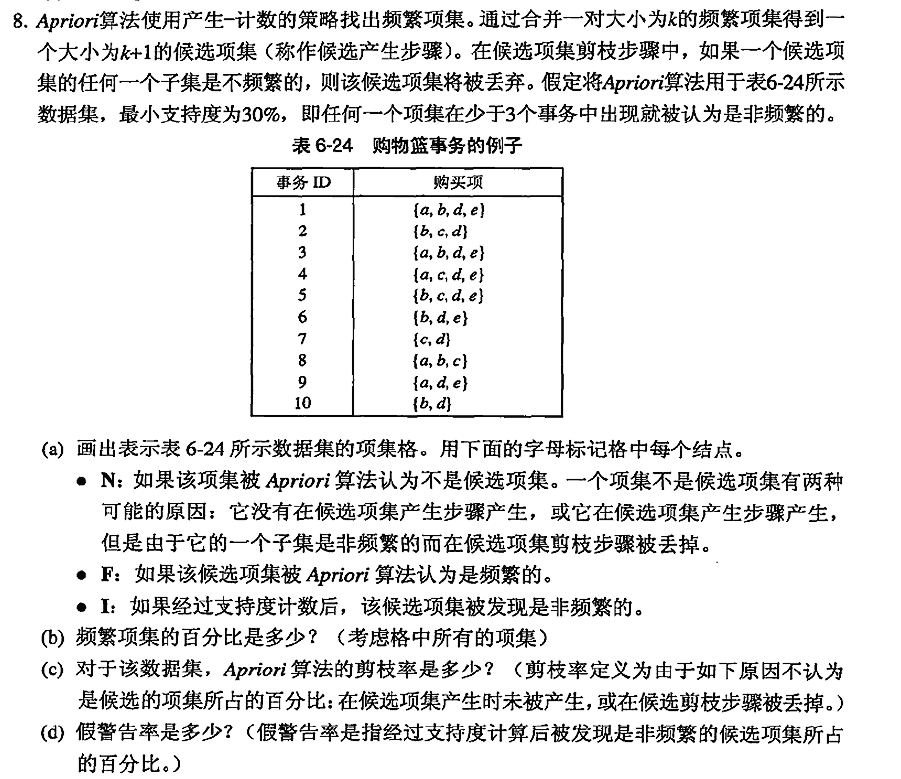


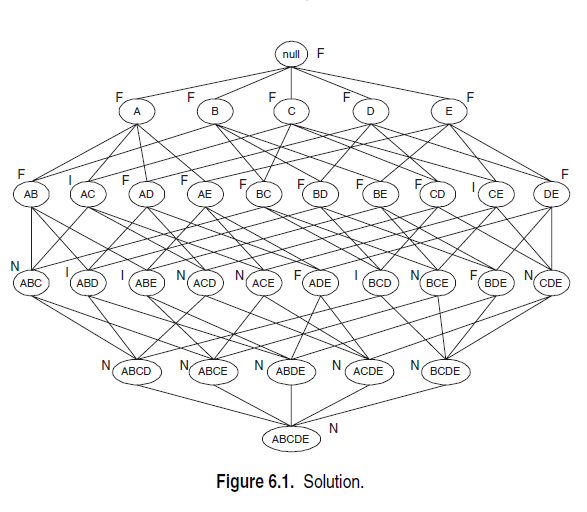


There are no apparent relationships between *s*1, *s*2, *c*1, and *c*2.



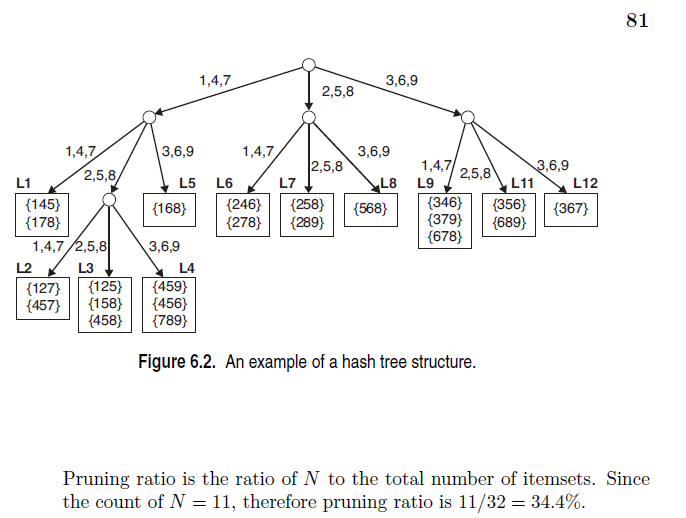






A2: Percentage of frequent itemsets = 16*/*32 = 50*.*0% (including the null

set).



A4: False alarm rate is the ratio of *I* to the total number of itemsets. Since

the count of *I* = 5, therefore the false alarm rate is 5*/*32 = 15*.*6%.