(a) 计算这两个规则的支持度和置信度。

(b) 为了使用传统的 Apriori 算法找出这些规则，我们需要离散化连续属性 A。假定我们使用等宽分箱方法离散化该数据，其中箱宽分别为 2，3，4。对于每个箱宽值，上面两个规则是否能够被 Apriori 算法发现？(注意，由于属性 A 可能具有较宽或较窄的区间，所以规则不一定与前面的规则完全相同。)对于每个与前面规则对应的规则，计算其支持度和置信度。 511

(c) 评述使用等宽分箱方法对上述数据集分类的有效性。是否有合适的箱宽度，以便很好地发现上面两个规则？如果没有，可以使用何种其他方法，以确保能够同时发现以上两个规则？

4. 考虑表 6.13 所示数据集。

表 6.13　习题 4 的数据集

年龄(A)	每周上网时数(B)				
	0～5	5～10	10～20	20～30	30～40
10～15	2	3	5	3	2
15～25	2	5	10	10	3
25～35	10	15	5	3	2
35～50	4	6	5	3	2

(a) 对于下面的每组规则，确定具有最高置信度的规则。

i. $15<A<25 \rightarrow 10<B<20$，$10<A<25 \rightarrow 10<B<20$ 和 $15<A<35 \rightarrow 10<B<20$。 512

ii. $15<A<25 \rightarrow 10<B<20$，$15<A<25 \rightarrow 5<B<20$ 和 $15<A<25 \rightarrow 5<B<30$。

iii. $15<A<25 \rightarrow 10<B<20$ 和 $10<A<35 \rightarrow 5<B<30$。

(b) 假定我们希望找出年龄在 15 岁到 25 岁之间的互联网用户每周的平均上网小时数。写一个基于统计学的关联规则，来刻画这个年龄段的用户。为了计算平均上网小时数，用中点近似值来表示每个区间(例如，使用 $B=7.5$ 来表示区间 $5<B<10$)。

(c) 通过将(b)中的平均上网小时数与不属于该年龄段的其他用户的平均上网小时数进行比较，检查(b)的量化关联规则是否具有统计意义。

5. 对于具有下面给出的属性的数据集，描述如何将它转换成适合于关联分析的二元事务数据集。具体地，指出原数据集中的每个属性：

(a) 对应于事务数据集中多少个二元属性；

(b) 原属性的值如何映射到二元属性的值？

(c) 数据属性值中是否有分层结构可以用来分组数据，形成少量二元属性。

下面是该数据集的属性列表以及它们的可能值。假定所有的属性都基于每个学生收集。

- **年级**：一年级、二年级、三年级、四年级、硕士研究生、博士研究生、专业人员。
- **邮政编码**：美国学生的家庭邮政编码，非美国学生的住处邮政编码。
- **院**：农学、建筑学、继续教育、教育、文学、工程、自然科学、商学、法律、医学、牙科、药学、护理学、兽医学。
- **住校**：如果学生住校为 1，否则为 0。
- 以下每项是一个属性，如果学生说对应的语言，则取 1，否则取 0。
  - 阿拉伯语
  - 孟加拉语