1. 为什么卡方（）检验可以用于连续型属性的离散化？请说明其原理（**要求举例说明**）。

**原理：**

卡方检验可以计算2个离散型变量之间的相关性程度，卡方值越大则相关性越高，反之则相关性越低。卡方检验在对连续型数据进行离散化时，采用自底向上合并的方法：（1）对数值属性A的每个不同属性值按从小到大排序；（2）假定排序后的不同取值个数为n，以相邻两个取值的平均值为分裂点，将n个不同属性值划分为n个不同的区间（分箱）；（3）计算每对相邻区间的卡方检验值，选择具有最小卡方检验值的一对相邻区间进行合并；（4）重复步骤（3）直到区间数量等于设定的离散化分箱数量。

选择最小卡方检验值的相邻区间进行合并的原理：给定一对相邻区间，设计两个离散型变量来衡量类别分布与所属不同区间的相关性，即一个变量*I*表示所属区间（可用取值1、0分别对应于两个相邻区间），另一个变量*C*表示每个区间上的不同类别数据对象的频数。按照这个卡方检验方法，具有最小卡方值的相邻区间将被优先合并在一起，因为低卡方值表明*I*与*C*的相关性低，也就说明这两个相邻区间具有相似的类分布（类分布情况与所属哪个区间的相关性很低）。

**现举例如下：**

例子：现有三个种类的鸢尾花A, B, C, 均具有属性“萼片长度”。要对该属性进行卡方检验离散化，具体步骤如下:

1. 首先将数据集中出现的n个不同的花萼长度值从小到大排序，并以两个相邻长度值的均值作为区间分裂点，将属性值划分为n个区间，每一个长度值对应单独一个区间，并将各个区间的对应A, B, C频数统计出来。
2. 计算两个相邻区间之间的卡方检验值，初始设置阈值为1.4，不断合并卡方值最小的两个相邻区间，并重新计算新的卡方值，直到各个相邻区间的卡方值都不小于1.4为止，此时的离散化分箱结果如下表1：

表1 三种鸢尾花的萼片长度离散化结果（阈值=1.4）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 长度（cm） | A（个） | B（个） | C（个） |  |
| [4.3, 4.9) | 16 | 0 | 0 | **4.1** |
| [4.9, 5.0) | 4 | 1 | 1 | 2.4 |
| [5.0, 5.5) | 25 | 5 | 0 | 8.6 |
| [5.5, 5.6) | 2 | 5 | 0 | 2.9 |
| [5.6, 5.7) | 0 | 5 | 1 | 1.7 |
| [5.7, 5.8) | 2 | 5 | 1 | 1.8 |
| [5.8, 5.9) | 1 | 3 | 3 | 2.2 |
| [5.9, 6.3) | 0 | 12 | 7 | 4.8 |
| [6.3, 6.6) | 0 | 6 | 15 | 4.1 |
| [6.6, 6.7) | 0 | 2 | 0 | 3.2 |
| [6.7, 7.0) | 0 | 5 | 10 | 1.5 |
| [7.0, 7.1) | 0 | 1 | 0 | 3.6 |
| [7.1, +∞) | 0 | 0 | 12 |  |

1. 如果再设置阈值为4.6，继续合并，以区间[4.3, 4.9)为例，其与相邻区间[4.9, 5.0)的值的计算公式及涉及到参数为：

其中i = 1,2,..., m, j = 1,2,…, k, 其中m = 2，表示用取值1、0分别对应于两个相邻区间；k=3，即对应三种类别的鸢尾花。

注：该计算结果与Randy Kerber1992年提出ChiMerge的论文中实例结果的“”有差异，经反复确认，以上计算步骤和结果无误，与论文中结果有差异的原因待考证。

1. 依次合并值最小的相邻区间，即合并[6.7, 7.0）和[7.0, 7.1)两个区间，并不断重复取最小值的相邻区间合并的操作，直到所有值都不小于设定的初始阈值。最终算法执行完毕之后得到的结果如下表2：

表2 三种鸢尾花的萼片长度离散化结果（阈值=4.6）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 长度（cm） | A（个） | B（个） | C（个） |  |
| [4.3, 5.5) | 45 | 6 | 1 | 30.9 |
| [5.5, 5.8) | 4 | 15 | 2 | 6.7 |
| [5.8, 6.3) | 1 | 15 | 10 | 4.9 |
| [6.3, 7.1) | 0 | 14 | 25 | 5.9 |
| [7.1, +∞) | 0 | 0 | 12 |  |

参考文献：

[1] Randy Kerber. ChiMerge: Discretization of Numeric Attributes. In Proceedings of AAAI 1992.

1. 是否可以将卡方检验用于决策树归纳（分类）算法中选择最佳属性？如果可以请说明其原理及算法的主要流程。

可以。

**原理：**

决策树将在每个节点选择分类效果最好的属性对样本进行二分类，而使用卡方检验可以从各个属性中选择最佳的属性。具体的说，卡方检验可以在各个自变量中各自寻找最显著的分割方法，并将这些自变量分割方法的显著程度相互比较，以确定要选择的自变量以及其分割法。

**算法的主要流程：**

1. 将各个属性之间的数据进行规范化。因为在处理多个属性的过程，容易出现计算结果偏向于数量级较大的属性，从而就会出现决策树构建过程中常见的偏置问题。
2. 对每个自变量和因变量之间的相关性进行卡方验证。如果存在部分自变量没有显著差异性，选取卡方值最小的一组区间进行合并，再重新进行卡方验证，直到都存在显著性差异为止。
3. 根据得到的卡方值进行分析，卡方值越大说明该属性与分类属性的相关性越强烈，在选择分类属性的时候，就选择卡方值最大的为当前分类节点的分类属性。
4. 对于每个分裂出来的子节点重复应用2、3流程，直到达到预设目标。