**Data Mining Take Home Exam**

学号:  **xxxx**  姓名: **xxx**

1. （20分）考虑下表的数据集。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 顾客ID | 性别 | 车型 | 衬衣尺码 | 类 |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20 | 男  男  男  男  男  男  女  女  女  女  男  男  男  男  女  女  女  女  女  女 | 家用  运动  运动  运动  运动  运动  运动  运动  运动  豪华  家用  家用  家用  豪华  豪华  豪华  豪华  豪华  豪华  豪华 | 小  中  中  大  加大  加大  小  小  中  大  大  加大  中  加大  小  小  中  中  中  大 | C0  C0  C0  C0  C0  C0  C0  C0  C0  C0  C1  C1  C1  C1  C1  C1  C1  C1  C1  C1 |

（1）计算整个数据集的Gini指标值。

（2）计算属性性别的Gini指标值

（3）计算使用多路划分属性车型的Gini指标值

（4）计算使用多路划分属性衬衣尺码的Gini指标值

（5）下面哪个属性更好，性别、车型还是衬衣尺码为什么

解：（1）

|  |  |
| --- | --- |
| **类** | **顾客** |
| c0 | 10 |
| c1 | 10 |

Gini=1-(10/20)^2-(10/20)^2=

(2)

|  |  |  |
| --- | --- | --- |
| **性别** | **男** | **女** |
| c0 | 6 | 4 |
| c1 | 4 | 6 |

Gini=[{1-(6/10)^2-(4/10)^2}\*1/2]\*2=

(3)

|  |  |  |  |
| --- | --- | --- | --- |
| 车型 | 家用 | 运动 | 豪华 |
| c0 | 1 | 8 | 1 |
| c1 | 3 | 0 | 7 |

Gini={1-(1/4)^2-(3/4)^2}\*4/20+{1-(8/8)^2-(0/8)^2}\*8/20+{1-(1/8)^2-(7/8)^2}\*8/20=26/160=

(4)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 尺码 | 小 | 中 | 大 | 加大 |
| c0 | 3 | 3 | 2 | 2 |
| c1 | 2 | 4 | 2 | 2 |

Gini={1-(3/5)^2-(2/5)^2}\*5/20+{1-(3/7)^2-(4/7)^2}\*7/20+[{1-(2/4)^2-(2/4)^2}\*4/20]\*2=8/25+6/35=

(5)

比较上面各属性的Gini值大小可知，车型划分Gini值最小，即使用车型属性更好。

2. （20分）考虑下表中的购物篮事务数据集。

|  |  |  |
| --- | --- | --- |
| 顾客ID | 事务ID | 购买项 |
| 1  1  2  2  3  3  4  4  5  5 | 001  0024  0012  0031  0015  0022  0029  0040  0033  0038 | {a,d,e}  {a,b,c,e}  {a,b,d,e}  {a,c,d,e}  {b,c,e}  {b,d,e}  {c,d}  {a,b,c}  {a,d,e}  {a,b,e} |

(1) 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。

（2）使用（1）的计算结果，计算关联规则 {b,d}→{e}和{e}→{b,d} 的置信度。

（3）将每个顾客ID作为一个购物篮，重复（1）。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一次，则为1，否则，为0）。

（4）使用（3）的计算结果，计算关联规则{b,d}→{e}和{e}→{b,d} 的置信度。

答：（1）由上表计数可得{ｅ}的支持度为8/10=；{ｂ，ｄ}的支持度为2/10=；｛b,d,e｝的支持度为2/10=。

（2）c[{b,d}→{e}]=2/8=; c[{e}→{b,d}]=8/2=4。

（3）同理可得：{e}的支持度为4/5=，{b,d}的支持度为5/5=1， {b,d,e}的支持度为4/5=。

（4）c[{b,d}→{e}]=5/4=，c[{e}→{b,d}]=4/5=。

3. （20分）以下是多元回归分析的部分R输出结果。

> ls1=lm(y~x1+x2)

> anova(ls1)

Df Sum Sq Mean Sq F value Pr(>F)

x1 1 \*\*\*

x2 1 \*\*

Residuals 7

> ls2<-lm(y~x2+x1)

> anova(ls2)

Df Sum Sq Mean Sq F value Pr(>F)

x2 1 \*\*

x1 1 \*\*\*

Residuals 7

（1）用F检验来检验以下假设 (α =

H0: β1 = 0

Ha: β1 ≠ 0

计算检验统计量；是否拒绝零假设，为什么

（2）用F检验来检验以下假设 (α =

H0: β2 = 0

Ha: β2 ≠ 0

计算检验统计量；是否拒绝零假设，为什么

（3）用F检验来检验以下假设 (α =

H0: β1 = β2 = 0

Ha: β1 和 β2 并不都等于零

计算检验统计量；是否拒绝零假设，为什么

解：（1）根据第一个输出结果F=>F（2，7）=，p<，所以可以拒绝原假设，即得到不等于0。



（2）同理，在α=的条件下，F=>F(2,7)=，p<，即拒绝原假设，得到不等于0。



（3）F={（+）/2}/（7）=>F=(2,7)=,即拒绝原假设，得到和并不都等于0。



4. （20分）考虑下面20个观测值：

[1]

[6]

[11]

[16]

（1）用3\*S 标准来检测该数据中是否存在异常值。

（2）用\*IQR 标准来检测该数据中是否存在异常值。

要求提供均值、标准差、IQR值和所用的区间，以及你判断的依据。

解：数据放入R运行得

> hhh=c,,,,,

+ + ,,,,,

+ + ,,,,,

+ + ,,,,

> mean(hhh)

[1]

> sqrt(var(hhh))

[1]

> jjj<-(hhh-mean(hhh))/sqrt(var(hhh))

> which(abs(jjj)>=3)

[1] 5

> Q1=quantile(hhh,

> Q3=quantile(hhh,

> IQR=Q3-Q1

> Q1

25%

> Q3

75%

> IQR

75%

> kk<\*IQR

> ll<-Q3+\*IQR

> which(hhh<kk)

[1] 1

> which(hhh>ll)

[1] 5

>

（1）由以上数据处理可得第五个数据为异常值。

（2）由数据分析可以得出第一个和第五个均为异常值。

5. 简述题 （20分）。

（1） 在这门课中你学习了几种聚类分析的算法简述这几种算法的基本工作原理。

答：（1）学习了三种聚类分析算法，分别为系统聚类分析、K均值分析和DBSCAN分析。

其中系统聚类又分为许多种计算规则，主要的有最短距离法、最长距离法、重心法、离差平方和法及类平均法等，其主要思想是在不同方法下把客观认为最相近的两类合并为一类，再次执行并类步骤，那么就可以把所有的观测值都聚为一类，对于想要知道分多少类的具体情况有很大的帮助，可以依照树状聚类图得到清晰的表示。

K-均值聚类原理是依一定观测值的均值来确定聚类中心，从而根据其他的观测值和中心的距离来判断归为哪一类中。该聚类方法事先需要确定要分的类别数据，计算量要小得多，效率比层次聚类要高。但对于想要知道其他几类的分类情况就需要重新计算分析。

DBSCAN分析算法工作原理是对一组观测值优先聚合靠的最近的两个值，然后再寻找下一个靠的最近的值，依次类推，最后得到用户想要的簇的种类。其中可能会出现噪声点，那么根据区间半径考虑是否舍弃的情况。

（2）关联规则挖掘任务主要分为哪几个主要的子任务简述你学到的算法在这些子任务中的工作原理。

答：主要分为两个子任务：频繁项集的产生和关联规则的产生。

第一种算法是Apriori算法是基于支持度的剪枝技术，系统地控制候选项集指数增长。根据先验原理，对一个系统所有的项集从最简单的集合到最复杂的集合采取逐层检验步骤，计算各项的支持度和置信度大小和阕值比较，即测试过程；对指定的非频繁项集和包含该非频繁项集的非频繁项采取舍弃操作，即剪枝过程；直到没有新的非频繁项集产生时，执行算法结束过程。