广东工业大学考试试卷（ A）

课程名称：数据挖掘　　试卷满分 100 分

考试时间：2016 年 6 月 17 日 （第 16 周 星期五 ）

题　号	一	二	三	四	五	六	七	八	九	十	总分
评卷得分											
评卷签名											
复核得分											
复核签名											

1。（10 分）对于数据：{12，9，7，6，20，100，35，21，11，18，25，37}

1) 计算它的平均值，20%的截断均值和中位数，并说明这三个统计特征在描述数据集方面的特点。

2）使用 MIN-MAX 规范方法将值其中的 6,100,35 转换到[0,1]。

3)对数据按照深度为 4 进行划分，再写出按边界值进行平滑后的结果。解释一下一般会因为什么目的对数据进行平滑处理。

2。（10 分）在现实的数据挖掘任务中，举例说明收集的数据会有什么样的噪声数据，请说明一般有哪些噪声数据处理方法？

3.（10 分）某单位想做一个基于数据挖掘的系统，该系统有一份调查问卷。希望一个高中生通过回答调查问卷中的问题，然后系统可以判断出该学生适合读大学的什么专业。如果调查问卷已经设计好了。请说明应该如何收集数据？应该使用哪一类数据挖掘任务来完成该系统？

4。（10 分）某学校对入学的新生进行性格问卷调查，没有心理学家的参与，根据学生对问题的回答，把学生的性格分成了 8 个类别。请说明该数据挖掘任务是属于分类任务还是聚类任务？为什么？并利用该例说明聚类分析和分类分析的异同点。

5.（15 分）如下表所示的数据集。请写出按属性 A 和 B 划分时的信息增益的计算表达式。不需要计算出最后结果。并回答计算信息增益在分类算法中的作用。

A	B	C	类标号
S	H	G	#
S	H	G	#
V	H	G	*
R	M	G	*
R	C	N	*
R	C	N	#
V	C	N	*
S	M	G	#
S	C	N	*
R	M	N	*
S	M	N	*
V	M	G	*
V	H	N	*
R	M	G	#

6.（15 分）请举例说明什么是聚类数据挖掘任务？

7.（15 分）假设描述一个运动员的信息包含属性：性别，籍贯，年龄。有两条记录 p，q 和 C1，C2 的信息如下，分别求出记录和簇彼此之间的距离。

p={男，湖南，17}，q={女，广东，22}

C1={男：25，女：5；湖南：18，湖北：4，广东：8；20}

C2={男：3，女：12；福建：10，广西：2，广东：3；24}

8.（15 分)画出如下数据的 FP 树，并按支持度阈值是 2 找到频繁项集。

序号	事务
1	BEER, EGG, DIAPER, KEY, BOX, WINE
2	HAM, EGG, DIAPER, KEY, BOX, WINE
3	BEER, DESK, KEY, TV
4	BEER, DESK, C, KEY, WINE
5	BOX, EGG, KEY, BEER, TV
6	WINE, BEER, KEY, BOX