选择题(40分)

一共二十个选择题，每个两分，总体上来说选择题考的都是细小的知识点基本概念什么的，中规中矩，有些题只看选项就能选出来，但是其中考了几个关于数据仓库的问题，数据仓库的相关概念课件上没有，需要上网去搜，例如数据仓库和数据库的区别，与数据集市的区别等。

有一个选择考了数据转换：最小-最大规范化。1.2~9.8映射为0~1，请给出7.6在0~1范围内映射的值。

填空题(14分)

1. 给出如下数据集{(0, 2)(1.5, 0)(0, 0)(5, 0)(5, 2)} 使用k均值算法进行聚类。其中k=2 初始节点为(0, 2)(0, 0), 写出两个簇所包含的结果。可能的算出的结果：【(0, 2) (5, 2)】【(1.5, 0)(0, 0)(5, 0)】（5分）
2. 分别用元组A（12,1,50,10）和B（30,0,36,6）表示,请算出A和B之间的欧几里得距离（3分）
3. 写出数据挖掘的六个步骤（6分）

判断题(10分)

选择题用“是”或“否”来回答，难度不大，都是概念认识题。

简答题(36分)

三个简答题每个12分

1. 假如你是一个高校的数据分析师，请使用CRISP-DM的流程，对电商平台做有用的数据挖掘，在解析流程中写出你是用什么技术做数据的聚类、分类、关联规则挖掘以及异常值检测的。
2. 写出在数据挖掘过程中为什么要做数据预处理的原因，写出数据预处理的基本流程及内容，请举一个例子来说明数据预处理的重要性。
3. 通用计算机和关系数据库系统不依赖于领域进行实现，然而有人认为通用数据挖掘系统却不能成为数据挖掘市场的主流，写出你对此观点的看法。请问构建不依赖于特定领域的数据挖掘系统好还是构建依赖于特定领域的数据挖掘系统好。

卷子总体评价

卷子大量考的都是概念题和主观题。计算题考的较少，仅考了一个k均值算法，其他的算法都没考，例如DBSCAN算法、决策数算法、ID3算法、Apriori算法、FP-Growth算法，辛辛苦苦练了算法，结果就考一个。￣へ￣，建议记一些课件后面的应用领域例子，考试的时候会用得上，例如简答题第二个， 数据挖掘每年题都不太一样，但是这次后面的开放性大题和上一届的一样。

以上试题内容都是考完试后回忆写出，会和原题所描述的内容有些出入，但是大致内容肯定是一样的。