第一大题为概念简单题，选5个

第二到第十题为问答和计算题，按之前的卷子应该是六道题目比较合适

第十一题为开放式问答题

所有题目的考察点已经在题目前标出

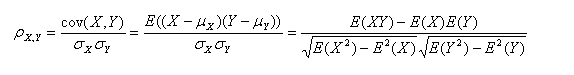
1. 简答题 （概念题选五个）
2. 两个仅包含二元属性的对象之间的相似性度量称为相似系数，简述三种(含)以上相似系数的计算方法与应用场景

**杰卡德相似系数(Jaccard Similarity Coefficient)**

**Jaccard(a,b)=**

只关心个体间的各维度值是否一致这个问题

**皮尔逊相关系数(Pearson Correlation Coefficient)**



数据存在“分数膨胀“问题

**余弦相似度(Cosine Similarity)**

**Cos(a,b)=**

数据是稀疏的，就使用余弦相似度

1. 简述关联规则中支持度和置信度的概念，并解释为什么采用这两种度量来表示关联规则的强度

**Support（支持度）：**表示同时包含 A 和 B 的事务占所有事务的比例

**Confidence（可信度）：**表示包含 A 的事务中同时包含 B 的事务的比例，即同时包含 A 和 B 的事务占包含 A 的事务的比例。

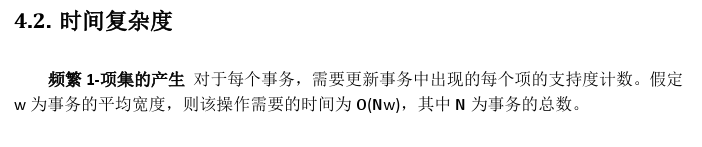
1. 简述Apriori算法的计算复杂度受哪些因素影响，并加以解释

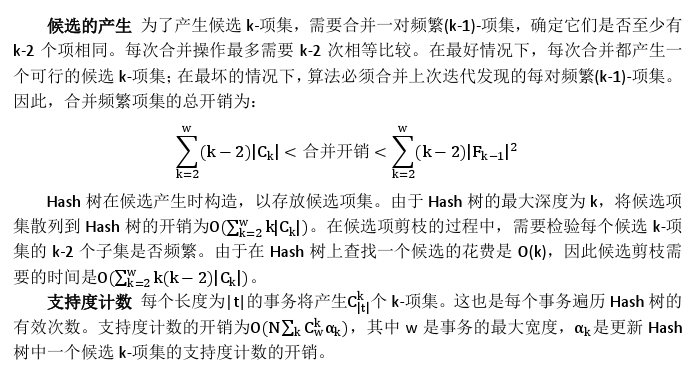
项数（维度）

事务平均宽度

事务数

支持度阀值





1. 在分类算法的评价指标中，recall和precision分别是什么含义

查全率，查准率

1. 请写出构建决策树时不纯度度量的三种指标

Ent(d)信息熵，信息增益，增益率，基尼系数

1. SVM中核函数的作用是什么？

非线性可分时，增加维度，减少计算代价

1. 介绍k-means 算法对初始点敏感的缺点（可以图示辅助分析）

朴素的初始化方法是直接在数据点中随机抽取 k 个作为聚类初始中心点。不够好的初始值可能造成收敛速度很慢或者聚类失败。

1. 传统的推荐系统算法主要是哪两种？

1. 基于用户行为数据的协同过滤算法   
2. 基于内容数据的过滤算法

9.请写出两个social network方向的研究内容，如影响力分析

静态网络的随机产生机制

研究一个静态网络和node变量的互动关系

二、（关联规则）Apriori算法使用产生—计数的策略找出频繁项集。通过合并一对大小为k的频繁项集得到一个大小为k+1的候选项集（称作候选产生步骤）。在候选项集剪枝步骤中，如果一个候选项集的任何一个子集是不频繁的，则该候选项集将被丢弃。假定将Apriori算法用于表中所示数据集，最小支持度为30%，即任何一个项集在少于3个事务中出现就被认为是非频繁的。

|  |  |
| --- | --- |
| 事务ID | 购买项 |
| 1 | {a,b,d,e} |
| 2 | {b,c,d} |
| 3 | {a,b,d,e} |
| 4 | {a,c,d,e} |
| 5 | {b,c,d,e} |
| 6 | {b,d,e} |
| 7 | {c,d} |
| 8 | {a,b,c} |
| 9 | {a,d,e} |
| 10 | {b,d} |

1. 画出表示表中所示数据的项集格，用下面的字母标记格中的每个结点。

* **N：**如果该项集被Apriori算法认为不是候选项集。一个项集不是候选项集有两种可能的原因：它没有在候选项集产生步骤产生，或它在候选项集产生步骤产生，但是由于它的一个子集是非频繁的而在候选项集简直步骤被丢掉
* **F:** 如果该候选项集被Apriori算法认为是非频繁的
* **I：**如果经过支持度计数后，该候选项集被发现是非频繁的

1. 频繁项集的百分比是多少？（考虑格中所有的项集）
2. 对于该数据集，Apriori算法的剪枝率是多少？（剪枝率定义为由于如下原因不认为是候选的项集所占的百分比：在候选项集产生时未被产生，或在候选剪枝步骤被丢掉）
3. 假警告率是多少？（假警告率是指经过支持度计算后被发现是非频繁的候选项集所占的百分比）

三、（朴素贝叶斯）试由下表的训练数据学习一个朴素贝叶斯分类器并确定x = （2，S）T 的类判别结果y。表中X（1）， X（2）为特征，Y为类标记。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| X（1） | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 3 |
| X（2） | S | M | M | S | M | M | L | M |
| Y | -1 | -1 | 1 | 1 | -1 | 1 | 1 | 1 |

四、（SVM）已知正例点x1 = (2.5, 2.5)T，x2 = (5, 2)T，负例点x3 = (1.5, 1.5)T，试用SVM对其进行分类，求最大间隔分离超平面，并指出所有的支持向量。

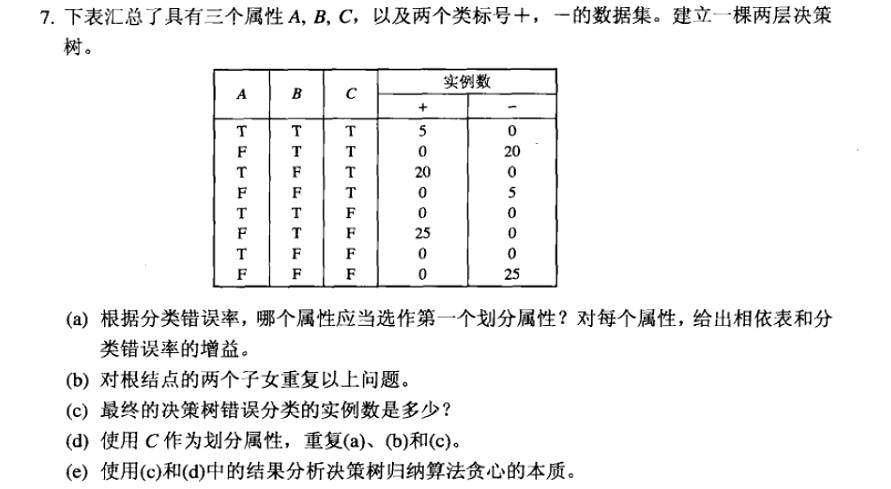
五、（决策树）下表是一个由15个贷款申请训练数据，数据包括贷款申请人的四个特征属性：分别是年龄，是否有工作，是否有自己的房子以及信贷情况，表的最后一列为类别，是否同意贷款。

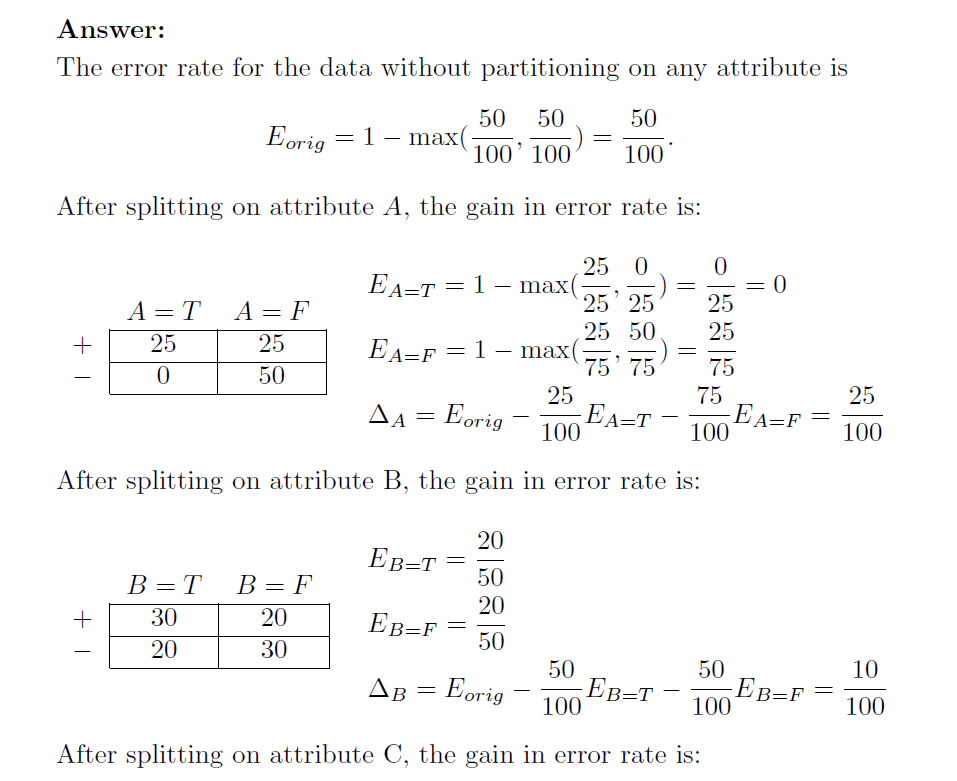
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | 年龄 | 有工作 | 有自己的房子 | 信贷情况 | 类别 |
| 1 | 青年 | 否 | 否 | 一般 | 否 |
| 2 | 青年 | 否 | 否 | 好 | 否 |
| 3 | 青年 | 是 | 否 | 好 | 是 |
| 4 | 青年 | 是 | 是 | 一般 | 是 |
| 5 | 青年 | 否 | 否 | 一般 | 否 |
| 6 | 中年 | 否 | 否 | 一般 | 否 |
| 7 | 中年 | 否 | 否 | 好 | 否 |
| 8 | 中年 | 是 | 是 | 好 | 是 |
| 9 | 中年 | 否 | 是 | 非常好 | 是 |
| 10 | 中年 | 否 | 是 | 非常好 | 是 |
| 11 | 老年 | 否 | 是 | 非常好 | 是 |
| 12 | 老年 | 否 | 是 | 好 | 是 |
| 13 | 老年 | 是 | 否 | 好 | 是 |
| 14 | 老年 | 是 | 否 | 非常好 | 是 |
| 15 | 老年 | 否 | 否 | 一般 | 否 |

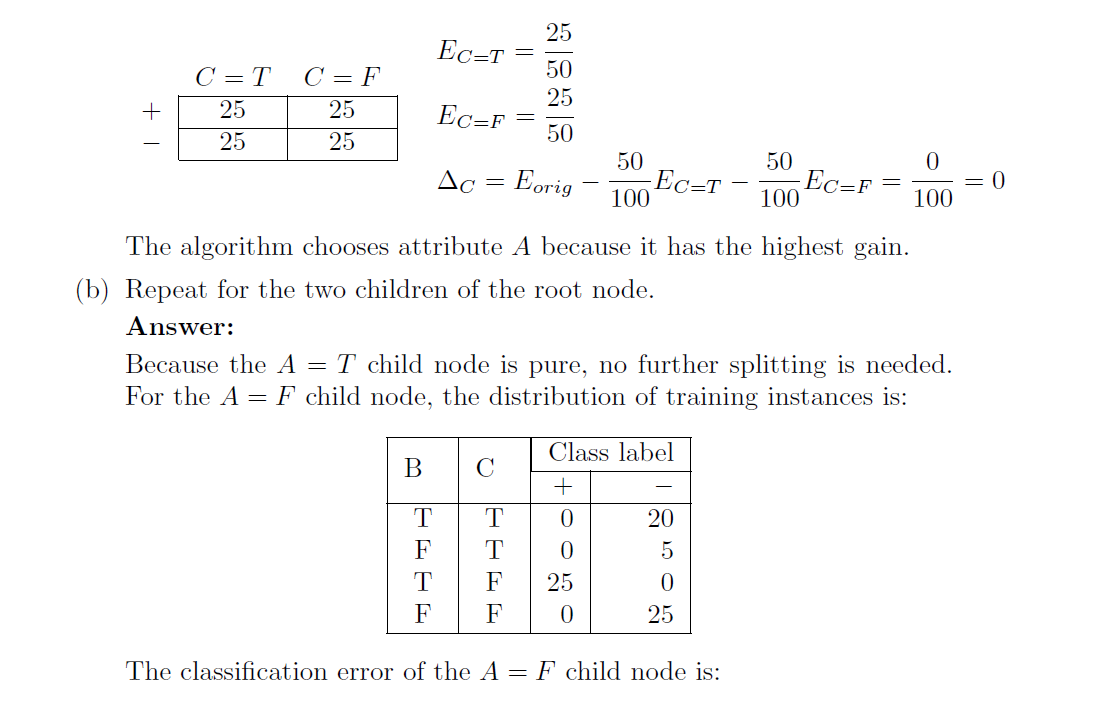
1）请根据上表的训练数据，以错误率作为划分标准来构建预测是否进行放贷的决策树。

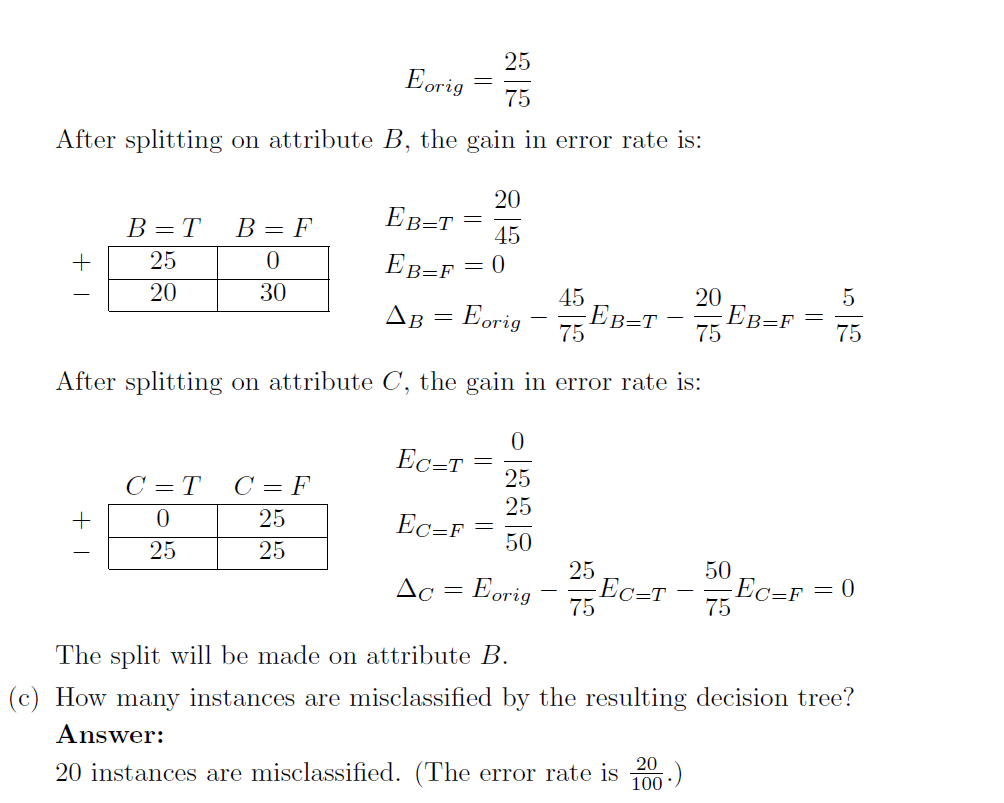
2）按照所构建的决策树，对属性值为（中年，无工作，无自己的房子，信贷情况好）的申请者是否进行放贷

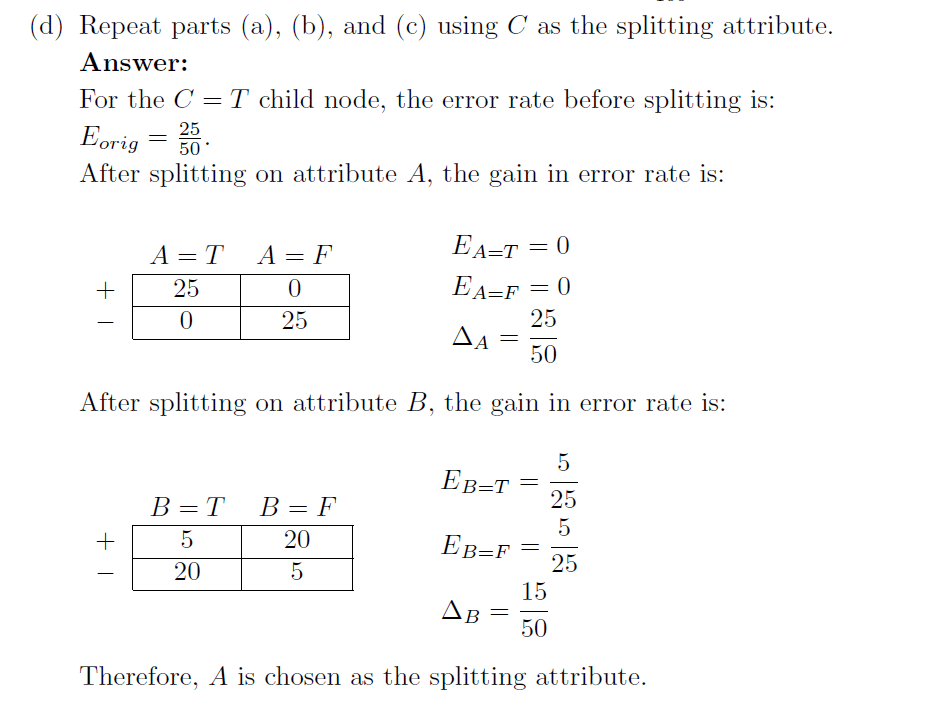
1. 在构建决策树的时候，可能会出现过拟合的问题，有哪些方法可以避免或者解决？
2. 对于含有连续型属性的样本数据，决策树有哪些处理方法？

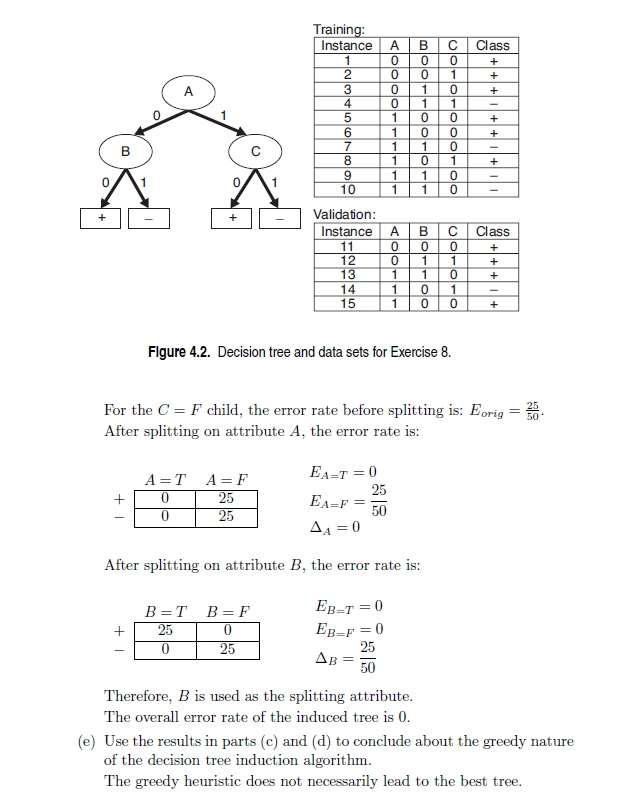












六、聚类分析（聚类）

1. 在聚类分析中，传统的K-means算法都有哪些局限性？有哪些相应的改进方法？

（1）对于离群点和孤立点敏感；   
(2)k值选择;   
(3)初始聚类中心的选择；   
(4)只能发现球状簇。

首先针对（1），对于离群点和孤立点敏感，如何解决？笔者在前面的一篇博客中，提到过离群点检测的LOF算法，通过去除离群点后再聚类，可以减少离群点和孤立点对于聚类效果的影响。

针对（2）k值的选择问题，在安徽大学李芳的硕士论文中提到了k-Means算法的k值自适应优化方法。下面将针对该方法进行总结。   
首先该算法针对K-means算法的以下主要缺点进行了改进：   
1）必须首先给出k（要生成的簇的数目），k值很难选择。事先并不知道给定的数据应该被分成什么类别才是最优的。   
2)初始聚类中心的选择是K-means的一个问题。   
李芳设计的算法思路是这样的：可以通过在一开始给定一个适合的数值给k，通过一次K-means算法得到一次聚类中心。对于得到的聚类中心，根据得到的k个聚类的距离情况，合并距离最近的类，因此聚类中心数减小，当将其用于下次聚类时，相应的聚类数目也减小了，最终得到合适数目的聚类数。可以通过一个评判值E来确定聚类数得到一个合适的位置停下来，而不继续合并聚类中心。重复上述循环，直至评判函数收敛为止，最终得到较优聚类数的聚类结果。

针对（3）对初始聚类中心的选择的优化。一句话概括为：选择批次距离尽可能远的K个点。具体选择步骤如下。首先随机选择一个点作为第一个初始类簇中心点，然后选择距离该点最远的那个点作为第二个初始类簇中心点，然后再选择距离前两个点的最近距离最大的点作为第三个初始类簇的中心点，以此类推，直至选出K个初始类簇中心点。

针对（4）只能获取球状簇的根本原因在于，距离度量的方式。在李荟娆的硕士论文K\_means聚类方法的改进及其应用中提到了基于2种测度的改进，改进后，可以去发现非负、类椭圆形的数据。但是对于这一改进，个人认为，并没有很好的解决K-means在这一缺点的问题，如果数据集中有不规则的数据，往往通过基于密度的聚类算法更加适合，比如DESCAN算法。

1. 请简要描述聚类与关联分析的主要相似点和不同点。

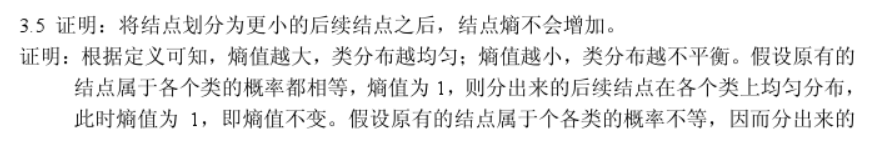
聚类。聚类类似于分类，但与分类的目的不同，是针对数据的相似性和差异性将一组数据分为几个类别。属于同一类别的数据间的相似性很大，但不同类别之间数据的相似性很小，跨类的数据关联性很低。

关联规则。关联规则是隐藏在数据项之间的关联或相互关系，即可以根据一个数据项的出现推导出其他数据项的出现。关联规则的挖掘过程主要包括两个阶段：第一阶段为从海量原始数据中找出所有的高频项目组;第二极端为从这些高频项目组产生关联规则。关联规则挖掘技术已经被广泛应用于[金融](http://www.vsharing.com/industry/1636.html)行业企业中用以预测客户的需求，各银行在自己的ATM 机上通过捆绑客户可能感兴趣的信息供用户了解并获取相应信息来改善自身的营销。

1. 请举出一个采用聚类作为主要的数据挖掘方法的实际应用例子。

图像分割；在商业上，聚类分析被用来发现不同的客户群，并且通过购买模式刻画不同的客户群的特征。聚类分析是细分市场的有效工具，同时也可用于研究消费者行为，寻找新的潜在市场、选择实验的市场，并作为多元分析的预处理。在生物上，聚类分析被用来动植物分类和对基因进行分类，获取对种群固有结构的认识。在地理上，聚类能够帮助在地球中被观察的数据库商趋于的相似性。在保险行业上，聚类分析通过一个高的平均消费来鉴定汽车保险单持有者的分组，同时根据住宅类型，价值，地理位置来鉴定一个城市的房产分组。在因特网应用上，聚类分析被用来在网上进行文档归类来修复信息。在电子商务上，聚类分析在电子商务中网站建设数据挖掘中也是很重要的一个方面，通过分组聚类出具有相似浏览行为的客户，并分析客户的共同特征，可以更好的帮助电子商务的用户了解自己的客户，向客户提供更合适的服务。

七、（决策树）证明：在决策树分类方法中，将结点划分为更小的后继结点后，结点熵不会增加





八、（效果评价ROC）请评价两个分类器M1和M2的性能。所选择的测试集包含26个二值属性，记作A到Z。

表中是模型应用到测试集时得到的后验概率（图中只显示正类的后验概率）。因为这是二类问题，所以P(-)=1-P(+),P(-|A,…,Z)=1-P(+|A,…,Z)。假设需要从正类中检测实例

1. 画出M1和M2的ROC曲线（画在一幅图中）。哪个模型更多？给出理由
2. 对模型M1，假设截止阈值t=0.5.换句话说，任何后验概率大于t的测试实例都被看作正例。计算模型在此阈值下的precision，recall和F-score
3. 对模型M2使用相同的截止阈值重复（b）的分析。比较两个模型的F-score，哪个模型更好？所得结果与从ROC曲线中得到的结论一致吗》
4. 使用阈值t=0.1对模型M2重复（b）的分析。t=0.5和t=0.1哪一个阈值更好?该结果和你从ROC曲线中得到的一致吗？

|  |  |  |  |
| --- | --- | --- | --- |
| 实例 | 真实类 | P(+|A,…,Z,M1) | P(-|A,…,Z,M2) |
| 1 | + | 0.73 | 0.61 |
| 2 | + | 0.69 | 0.03 |
| 3 | - | 0.44 | 0.68 |
| 4 | - | 0.55 | 0.31 |
| 5 | + | 0.67 | 0.45 |
| 6 | + | 0.47 | 0.09 |
| 7 | - | 0.08 | 0.38 |
| 8 | - | 0.15 | 0.05 |
| 9 | + | 0.45 | 0.01 |
| 10 | - | 0.35 | 0.04 |

九（频繁项）考虑下面的候选3-项集的集合：{1，2，3}，{1，2，5}，{1，2，6}，{1，3，4}，{2，3，4}，{2，4，5}，{3，4，6}，{4，5，6}

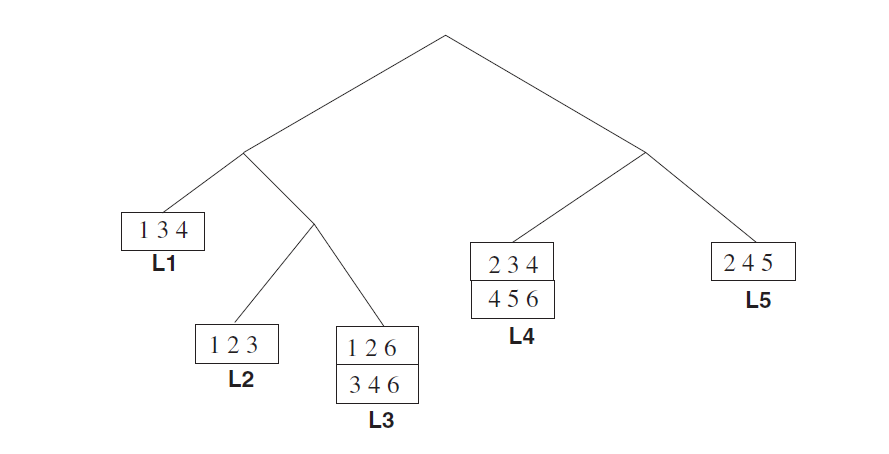
1. 构造以上候选3-项集的Hash树，假定Hash树使用这样一个Hash函数：所有奇数项都被散列到节点的左子女，所有的偶数项都被散列到右子女。一个候选k-项集按照如下方法被插入到Hash树中：散列候选项集中的每个相继项，然后再按照散列值到相应的分支。一旦到达叶节点，候选项集将按照下面的条件插入：

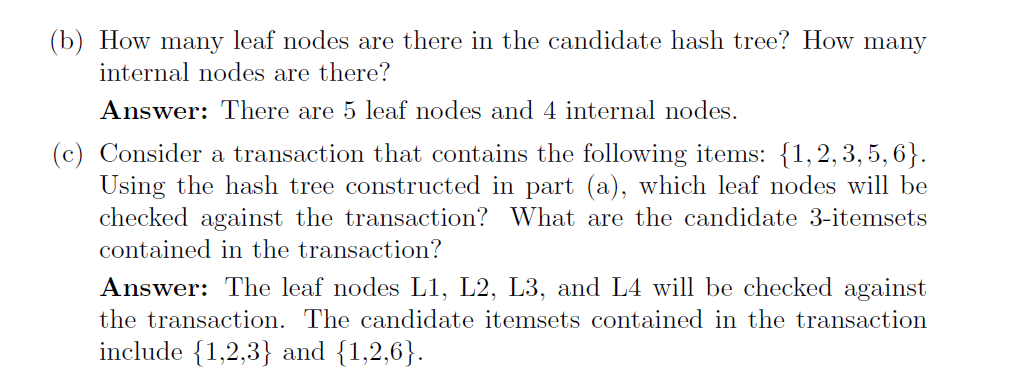
* 条件1：如果该叶节点的深度等于k（假设根节点的深度为0），则不管该节点已经存储了多少项集，将该候选插入该节点
* 条件2：如果该叶节点的深度小于k，则只要该节点存储的项集数不超过maxsize，就把它插入到该叶节点。这里，假定maxsize为2
* 条件3：如果该叶节点的深度小于k且该节点已存储的项集数量超过maxsize，则这个叶节点转变为内部节点，并创建新的叶节点作为老的叶节点的子女。先前老叶节点中存放的候选项集按照散列值分布到其子女中。新的候选项集也按照散列值存储到相应的叶节点

（b）候选Hash树中共多少个叶节点，多少个内部节点？

（c）考虑一个包含项集{1，2，3，4，5，6}的事务，使用（a）所创建的Hash树，则该事务要检查哪些叶节点？该事务包含哪些候选3-项集

习题10





十、(ensemble组合方法)请简述构建组合（集成）分类器的几种方法，并说明集成分类器能够改善分类器性能的原因。

Boosting , Bagging and 随机森林 ，组合多种方式，看书上。

十一、（开放课题）现有一个城市的数据集，包括交通卡、交通事故、出租车轨迹、公交车运行、地铁运行、空气质量、气象检测、新浪微博等（具体特征如下表）。

请利用你所学过的机器学习和数据挖掘的方法解决预测该城市空气质量的问题：

1. 哪些数据或者特征可能用到，并简要说明原因
2. 可以使用所学过的哪些机器学习方法解决该问题？
3. 请简要给出一个解决方案（最大限度地利用现有数据）。

|  |  |  |
| --- | --- | --- |
| 序号 | 数据集名称 | 具体数据项 |
| 1 | 城市道路交通指数 | 状态、区域、当前指数、参考指数、指数差值 |
| 2 | 地铁运行数据 | 线路、车站、换乘站数据、首末班车各站时刻表数据、站间运行时间数据、限流车站、封站数据、路网票价矩阵、列车实时到发站台时刻、线路拥挤及阻塞数据、出入口、厕所、残疾电梯数据 |
| 3 | 一卡通乘客刷卡数据 | 卡号、交易日期、交易时间、线路/地铁站点名称、行业名称（公交、地铁、出租、轮渡、P+R停车场）、交易金额、交易性质（非优惠、优惠、无） |
| 4 | 浦东公交车实时数据 | 设备号码,线路编码,站点编码,协议编号,进出站状态,方向,车载上报时间、编码对应表 |
| 5 | 强生出租汽车行车数据 | 车辆ID、GPS时间、经纬度、速度、卫星颗数、营运状态高架状态、制动状态 |
| 6 | 空气质量状况 | 序号，日期，PM2.5，PM10，O3，SO2，NO2，CO，AQI，质量评价，首要污染物 |
| 7 | 气象数据 | 日期、时间、监测点、天气类型、温度、风速、风向、降水量 |
| 8 | 道路事故数据 | 事故ID、事故类型、事故地点、事故时间 |
| 9 | 高架匝道关闭数据 | 匝道ID、位置信息、关闭时间、开放时间 |
| 10 | 新浪微博交通数据 | 涵盖路况、交通工具、天气等与交通相关的关键词的微博信息 |