一、 填空题（5*3=15，每空一分）

1） 关联规则的算法包括： APriori 和 FP-growth ，其中 FP-growth 的效率更高

2） 分类器设计包含三个过程：划分数据集、分类器构造，分类器测试

3） 分类中常用的评价准则有：精确度，查全率和查准率，几何均值（四个空加上 F-measure）

4） 支持向量机常用的核函数有：多项式核函数，径向基核函数，S 型核函数

5） 连续型属性的数据样本之间的距离有：欧氏距离，曼哈顿距离，明考斯基距离

6） 层次聚类方法包括：凝聚型层次聚类，分解型层次聚类

7） 聚类分析包括：连续型，二值离散型，多值离散型，混合类型。

8） 划分聚类方法对数据集进行聚类时包含三个要点：选定某种距离作为数据样本间的相似性度量、选择评价聚类性能的准则函数，选择某个初始分类，之后用迭代的方法得到聚类结果，使得评价聚类的准则函数取得最优值。

二、 简答题（5*5=25）

1、比较 PCA 和 LDA 的区别

PCA 和 LDA 都是经典的降维算法，都假设数据是符合高斯分布，也利用了矩阵特征分解的思想，但他们还是有一定的区别：

1）PCA 是无监督的，也就是训练样本不需要标签；LDA 是有监督的，也就是训练样本需要标签。

2）PCA 是去掉原始数据冗余的维度，LDA 是选择一个最佳的投影方向，使得投影后相同类别的数据分布紧凑，不同类别的数据尽量相互远离；

3）LDA 可能会过拟合数据。

2、请分析特征选择和特征提取有何区别

特征提取和特征选择是降维的两种方法，针对于维灾难，都可以达到降维的目的，但是这两个有所不同：特征选择定义为从有 N 个特征的集合中选出具有 M 个特征的子集，并满足条件 M≤N。特征选择能够为特定的应用在不失去数据原有价值的基础上选择最小的属性子集，去除不相关的和冗余的属性。

特征提取广义上指的是一种变换，将处于高维空间的样本通过映射或变换的方式转换到低维空间，达到降维的目的。它可以从一组特征中去除冗余或不相关的特征来降维

特征提取

1:特征抽取后的新特征是原来特征的一个映射

2:将机器学习算法不能识别的原始数据转化为算法可以识别的特征的过程

特征选择

1:特征选择后的特征是原来特征的一个子集

2:特征选择是从所有的特征中选择一个最好的特征子集

3、聚类和分类有什么区别和联系?

分类和聚类都是常用的数据挖掘的方法，分类可以更精确、有效的挖掘出信息，从训练集中得到模型，之后对未知类标号的数据样本进行分类，在许多实际的应用领域中，由于缺少形成类别的先验知识，收集或者存储的数据集样本没有类标号，对于这类数据集常采用聚类分析分析方法

区别：

1） 对象所属类别是否为事先。分类是把某个对象划分到某个具体的已经定义的类别当中，而聚类是把一些对象按照具体特征组织到若干个类别里

2） 分类算法的基本功能是做预测，而聚类算法的功能是降维。

3） 分类是有监督的学习，而聚类是无监督的学习。有监督的算法并不是实时的，需要给定一些数据对模型进行训练，有了模型就能预测。分类算法中，对象所属的类别取决于训练出来的模型，间接地取决于训练集中的数据。而聚类算法中，对象所属的类别，则取决于待分析的其他数据对象。

4） 典型的分类算法有：决策树，神经网络，支持向量机模型，Logistic 回归分析，以及核估计等等。聚类的方法有，基于链接关系的聚类算法，基于中心度的聚类算法，基于统计分布的聚类算法以及基于密度的聚类算法等等

## 4、TF.IDF 算法是什么，有什么实际意义?

TF—IDF 是自然语言处理中的一个简单的模型。TF 代表 term frequency，也就是词频，而 IDF 代表着 inverse document frequency，叫做逆文档频率，这两个属性都是属于单词的属性。概括来说，TF—IDF 模型是用来给文档中的每个词根据重要程度计算一个得分，这个得分就是 TF—IDF。

实际应用意义

自动提取关键词、找相似文章、自动摘要

1:首先，可以计算文档中的每个词的得分，从而选分数高的作为关键词，这就是关键词自动提取

2:搜索引擎常见的把网页上的相关文档排序的做法

3:查重：我们可以找到两篇文章，可以找到两篇文章的关键词集合并计算出词频向量，从而计算文本相似度。

4:TF-IDF 是一种加权技术,用以评估字词对于一个文件集或一个语料库中的其中一份文件的重要程度

## 5、数据挖掘与统计的区别与联系

虽然两者采用的某些分析方法是相同的，但是数据挖掘和统计学是有本质区别的：

1） 统计学主要利用概率论建立数学模型，是研究随机现象的常用数学工具之一。

2） 数据挖掘分析大量数据，发现其中的内在联系和知识，并以模型或规则表达这些知识。

一个主要差别在于处理对象（数据集）的尺度和性质。数据挖掘经常会面对尺度为 GB 甚至 TB 数量级的数据库，而用传统的统计方法很难处理这么大尺度的数据集。传统的统计处理往往是针对特定的问题采集数据（甚至通过试验设计加以优化）和分析数据来解决特定问题；而数据挖掘却往往是数据分析的次级过程，其所用的数据原本可能并非为当前研究而专门采集的，因而其适用性和针对性可能都不强，在数据挖掘的过程中，需要对异常数据及冲突字段等进行预处理，尽可能提高数据的质量，然后才经过预处理的数据进行数据挖掘。

另一个差别在于面对结构复杂的海量数据，数据挖掘往往需要采用各种相应的数学模型和应用传统统计学以外的数学工具，才能建立最适合描述对象的模型或规则。

6、请简述数据挖掘中关联规则 Apriori 算法的思想

Apriori 算法的基本思想是通过对数据库的多次扫描来计算项集的支持度，发现所有的频繁项集从而生成关联规则。Apriori 算法对数据集进行多次扫描。第一次扫描得到频繁 1-项集的集合 L1，第 K（k>1）次扫描首先利用第(k-1)次扫描的结果 Lk-1 来产生候选集 k-项集的集合 Ck，然后在扫描的过程中确定 Ck 的支持度。最后，在每次扫描结束时计算频繁 k-项集的集合 Lk，算法在候选集 k-项集的集合 Ck 为空时结束。

7、什么是关联规则？关联规则的应用有哪些

关联规则挖掘最初由 R.Agrawal 等人提出，用来发现超级市场中用户购买的商品之间的隐含关联关系，并用规则的形式表示出来，称为关联规则(Association Rule)。关联规则除了可以发现超市购物中隐含的关联关系之外，还可以应用于其他很多领域。关联规则 的应用还包括文本挖掘、商品广告邮寄分析、网络故障分析等

8、什么是分类？分类的应用领域有哪些？

分类是指把数据样本映射到一个事先定义的类中的学习过程，即给定一组输入的属性向量及其对应的类，用基于归纳的学习算法得出分类。分类问题是数据挖掘领域中研究和应用最为广泛的技术之一，许多分类算法被包含在统计分析工具的软件包中，作为专门 的分类工具来使用。分类问题在商业、银行业、医疗诊断、生物学、文本挖掘、因特网筛选 等领域都有广泛应用。例如，在银行业中，分类方法可以辅助工作人员将正常信用卡用户和 欺诈信用卡用户进行分类，从而采取有效措施减小银行的损失；在医疗诊断中，分类方法可 以帮助医疗人员将正常细胞和癌变细胞进行分类，从而及时制定救治方案，挽救病人的生命； 在因特网筛选中，分类方法可以协助网络工作人员将正常邮件和垃圾邮件进行分类，从而制定有效的垃圾邮件过滤机制，防止垃圾邮件干扰人们的正常生活

9、什么是聚类分析？聚类分析的应用领域有哪些？

聚类分析是将物理的或者抽象的数据集合划分为多个类别的过程，聚类之后的每个类别中任意两个数据样本之间具有较高的相似度，而不同类别的数据样本之间具有较低的 相似度。 聚类分析是数据挖掘应用的主要技术之一，它可以作为一个独立的工具来使用， 将未知类标号的数据集划分为多个类别之后，观察每个类别中数据样本的特点，并且对某些 特定的类别作进一步的分析。此外，聚类分析还可以作为其他数据挖掘技术（例如分类学习、 关联规则挖掘等）的预处理工作。聚类分析在科学数据分析、商业、生物学、医疗诊断、文 本挖掘、Web 数据挖掘等领域都有广泛应用。在科学数据分析中，比如对于卫星遥感照片， 聚类可以将相似的区域归类，有助于研究人员根据具体情况做进一步分析；在商业领域，聚 类可以帮助市场分析人员对客户的基本数据进行分析，发现购买模式不同的客户群，从而协 助市场调整销售计划；在生物学方面，聚类可以帮助研究人员按照基因的相似度对动物和植 物的种群进行划分，从而获得对种群中固有结构的认识；在医疗诊断中，聚类可以对细胞进 行归类，有助于医疗人员发现异常细胞的聚类，从而对病人及时采取措施；在文本挖掘和 Web 数据挖掘领域中，聚类可以将网站数据按照读者的兴趣度进行划分，从而有助于网站 内容的改进。

三、 计算题（3*15=45）

关联规则 P95；P115、P130 决策树增益的计算；贝叶斯、全概率 P152

四、 论述题（15 分）

列举几项你知道的数据挖掘应用，并论述数据挖掘在其中的作用。

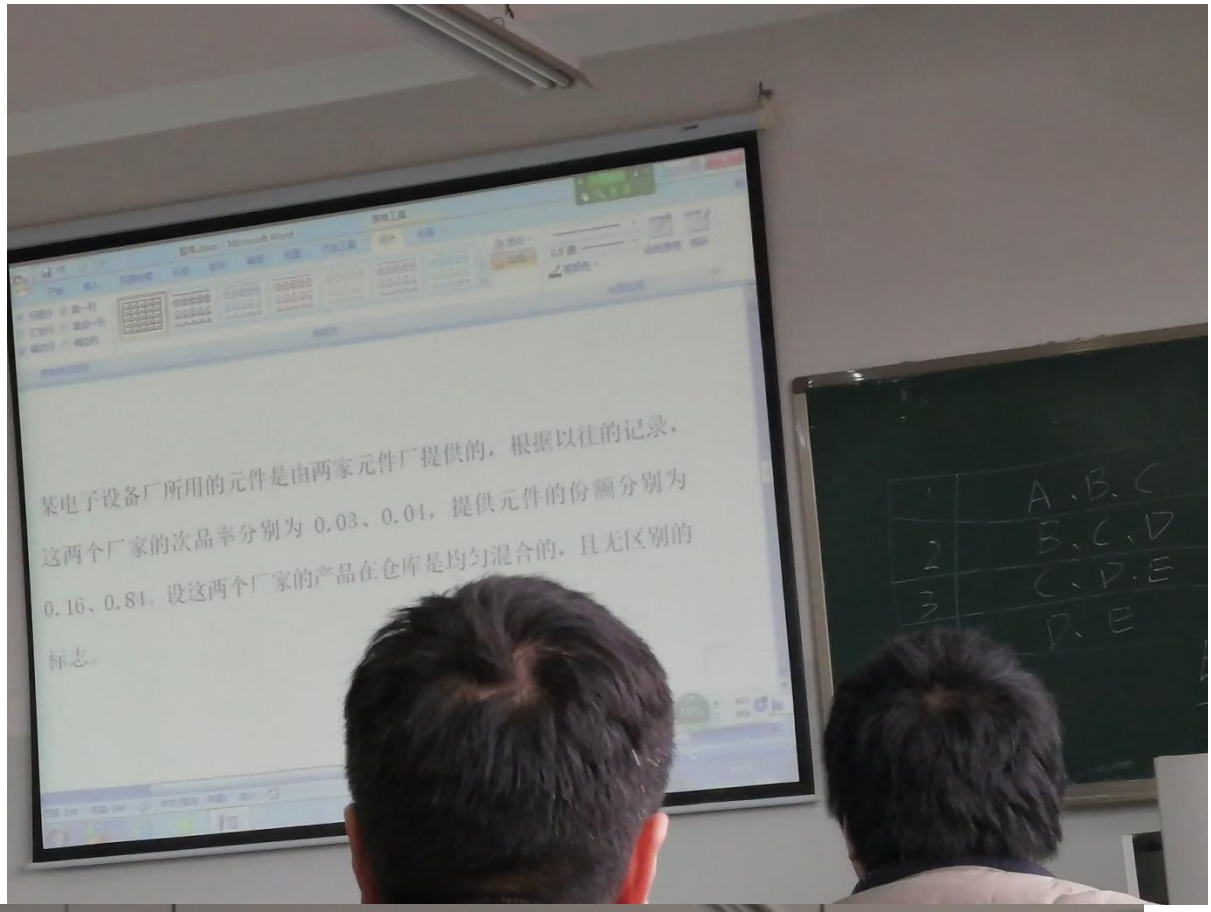
（1）分类，根据特征判断对象属于哪个类别，有指导学习。预测肿瘤细胞是良性还是恶性；识别信用卡交易是否合法还是欺诈；电信客户流失分析；图片、音频、视频标签；蛋白质结构功能分类等。

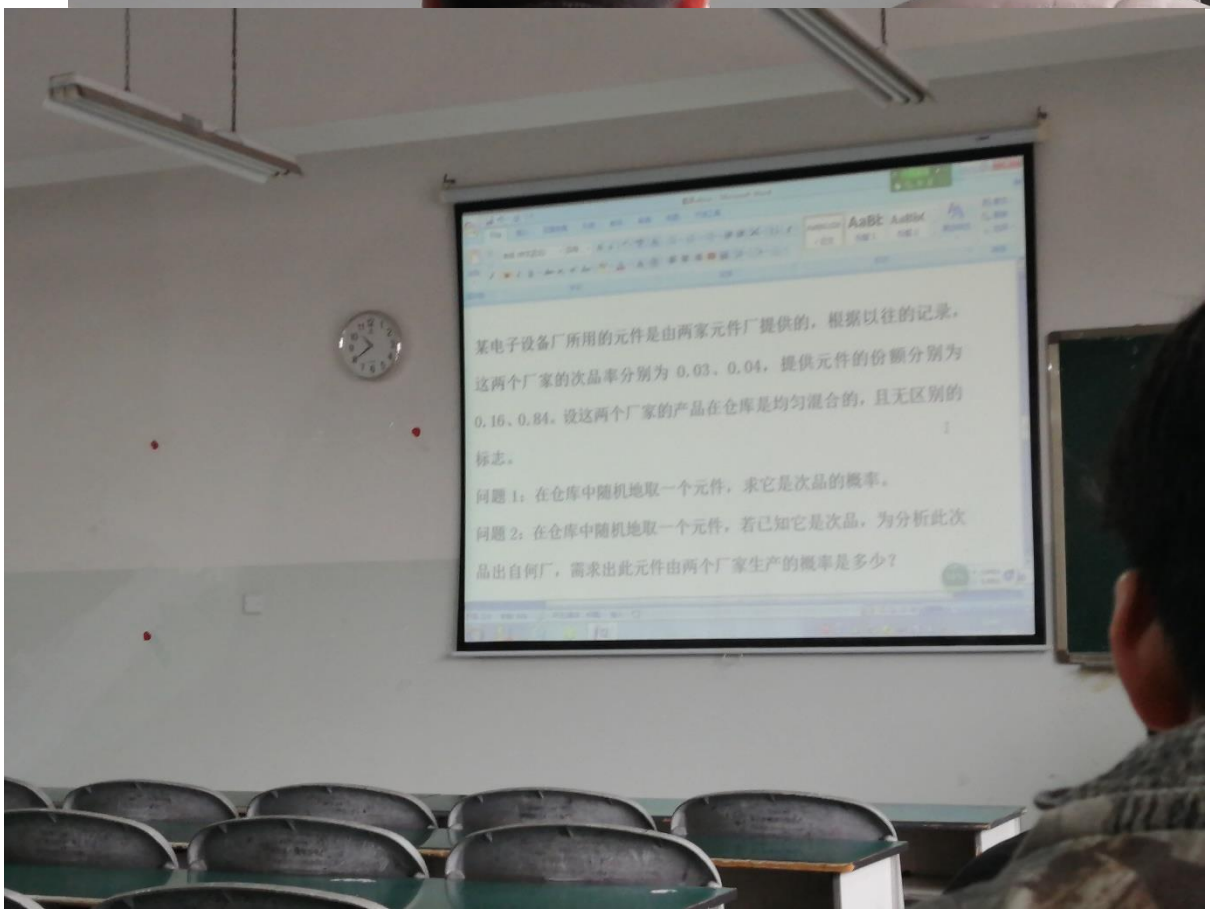
（2）聚类，给对象归类使得同组对象尽可能相似，不同组对象尽可能不相似，无指导学习。把相关文档归并方便浏览；市场分割，细分为不同的客户群；获取价格波动相似的股票有助于决策；相关案件放在一起寻找嫌疑人的特征。

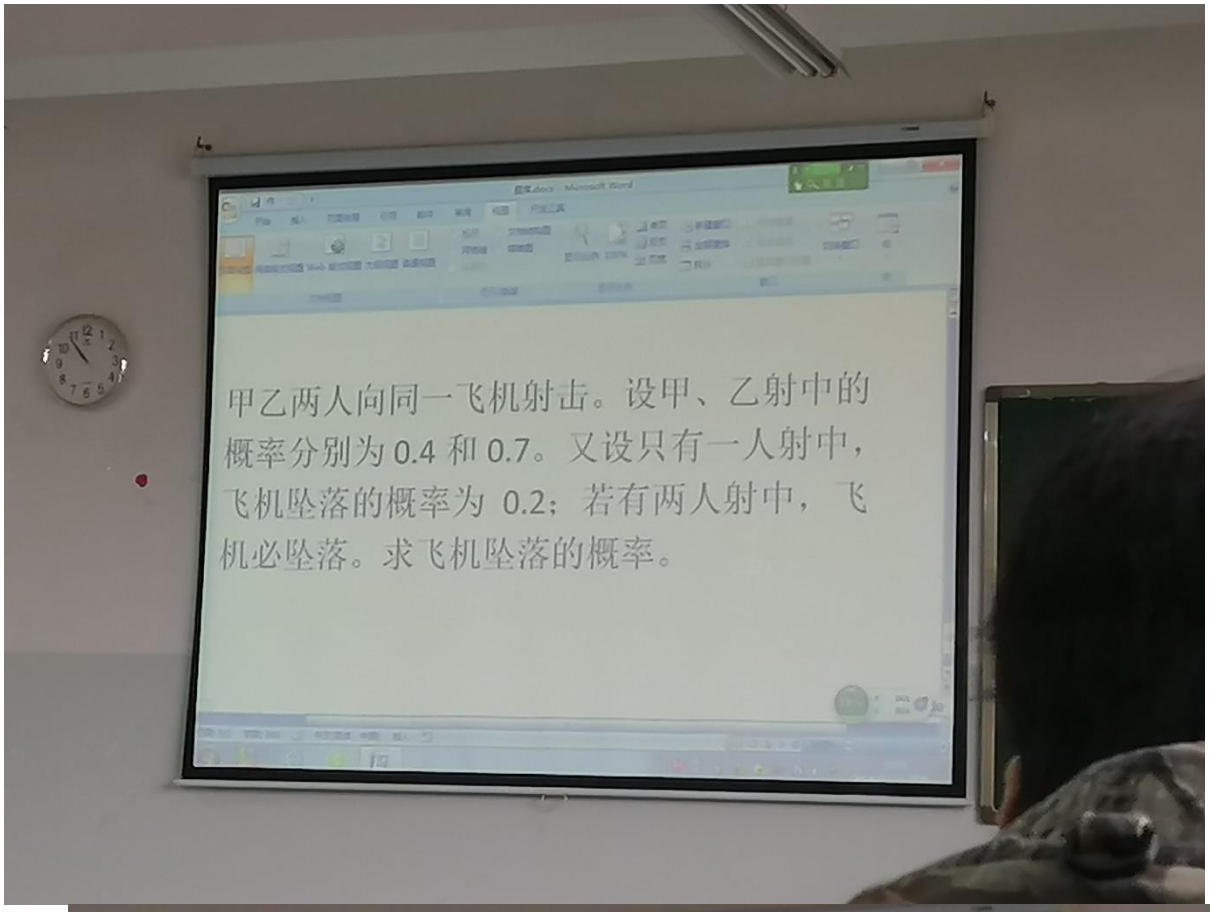
（3）关联分析，给定一组记录，分析项目之间的依赖关系。购物分析，用于促销、货价管理存货管理；医疗信息发现与某种疾病与症状的关联以便通过症状诊断病症

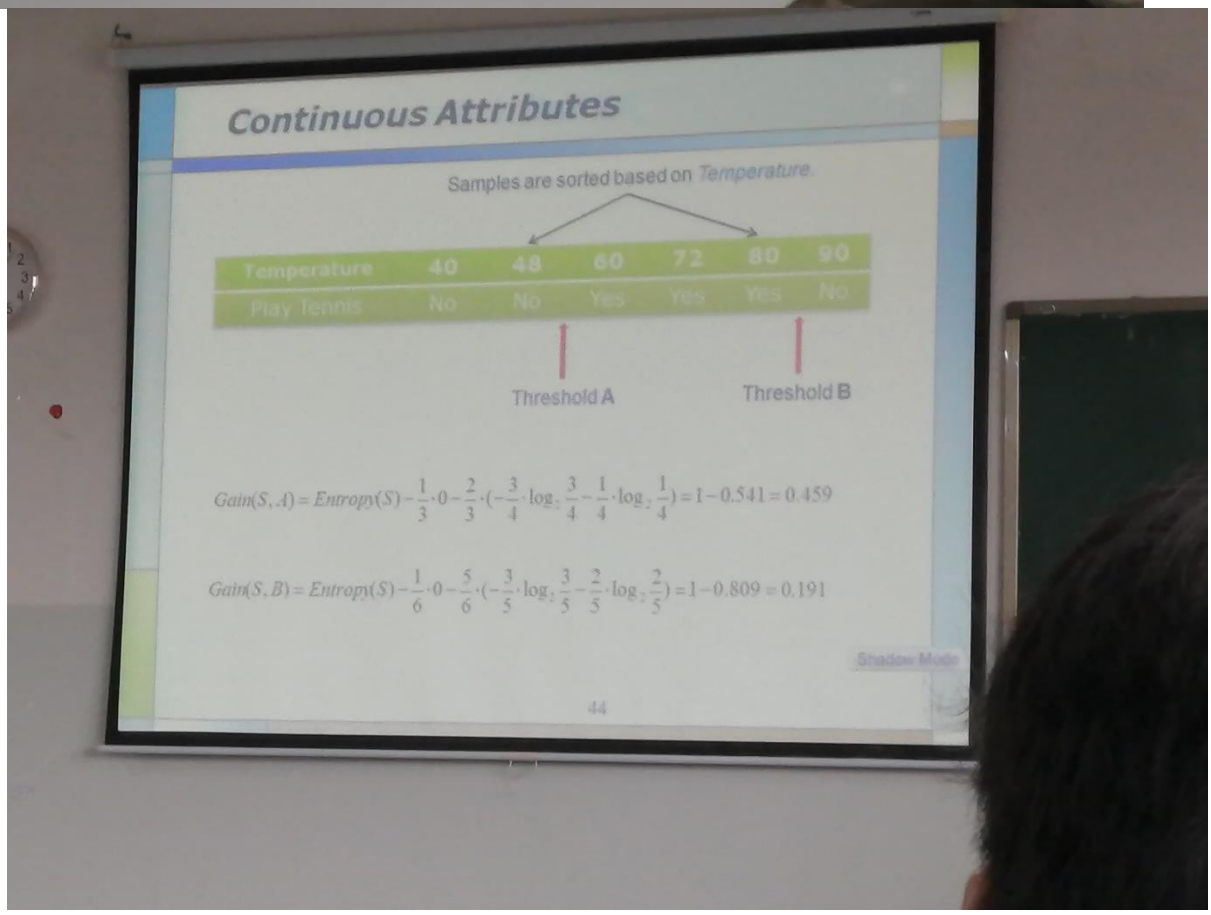
（4）顾客分类，数据挖掘能够告诉我们什么样的顾客买什么产品（聚类或分类）

识别顾客需求，对不同的顾客识别最好的产品，使用预测发现什么因素影响新顾客。汽车保险检测假造事故骗取保险赔偿的人。检测电话欺骗，通话距离、通话时间，每天或每周通话次数


某电子设备厂所用的元件是由两家元件厂提供的，根据以往的记录，
这两个厂家的次品率分别为 0.03、0.04，提供元件的份额分别为
0.16、0.84。设这两个厂家的产品在仓库是均匀混合的，且无区别的
标志。
A、B、C
B、C、D
C、D、E
D、E


某电子设备厂所用的元件是由两家元件厂提供的，根据以往的记录，
这两个厂家的次品率分别为 0.03、0.04，提供元件的份额分别为
0.16、0.84。设这两个厂家的产品在仓库是均匀混合的，且无区别的
标志。
问题 1：在仓库中随机地取一个元件，求它是次品的概率。
问题 2：在仓库中随机地取一个元件，若已知它是次品，为分析此次
品出自何厂，需求出此元件由两个厂家生产的概率是多少？


甲乙两人向同一飞机射击。设甲、乙射中的概率分别为 0.4 和 0.7。又设只有一人射中，飞机坠落的概率为 0.2；若有两人射中，飞机必坠落。求飞机坠落的概率。


Continuous Attributes
Samples are sorted based on Temperature.
Temperature 40 48 60 72 80 90
Play Tennis No No Yes Yes Yes No
Threshold A
Threshold B
$Gain(S,A) = Entropy(S) - \frac{1}{3}\cdot 0 - \frac{2}{3}\cdot(-\frac{3}{4}\cdot\log_2\frac{3}{4} - \frac{1}{4}\cdot\log_2\frac{1}{4}) = 1 - 0.541 = 0.459$
$Gain(S,B) = Entropy(S) - \frac{1}{6}\cdot 0 - \frac{5}{6}\cdot(-\frac{3}{5}\cdot\log_2\frac{3}{5} - \frac{2}{5}\cdot\log_2\frac{2}{5}) = 1 - 0.809 = 0.191$
Shadow Mode
44