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：

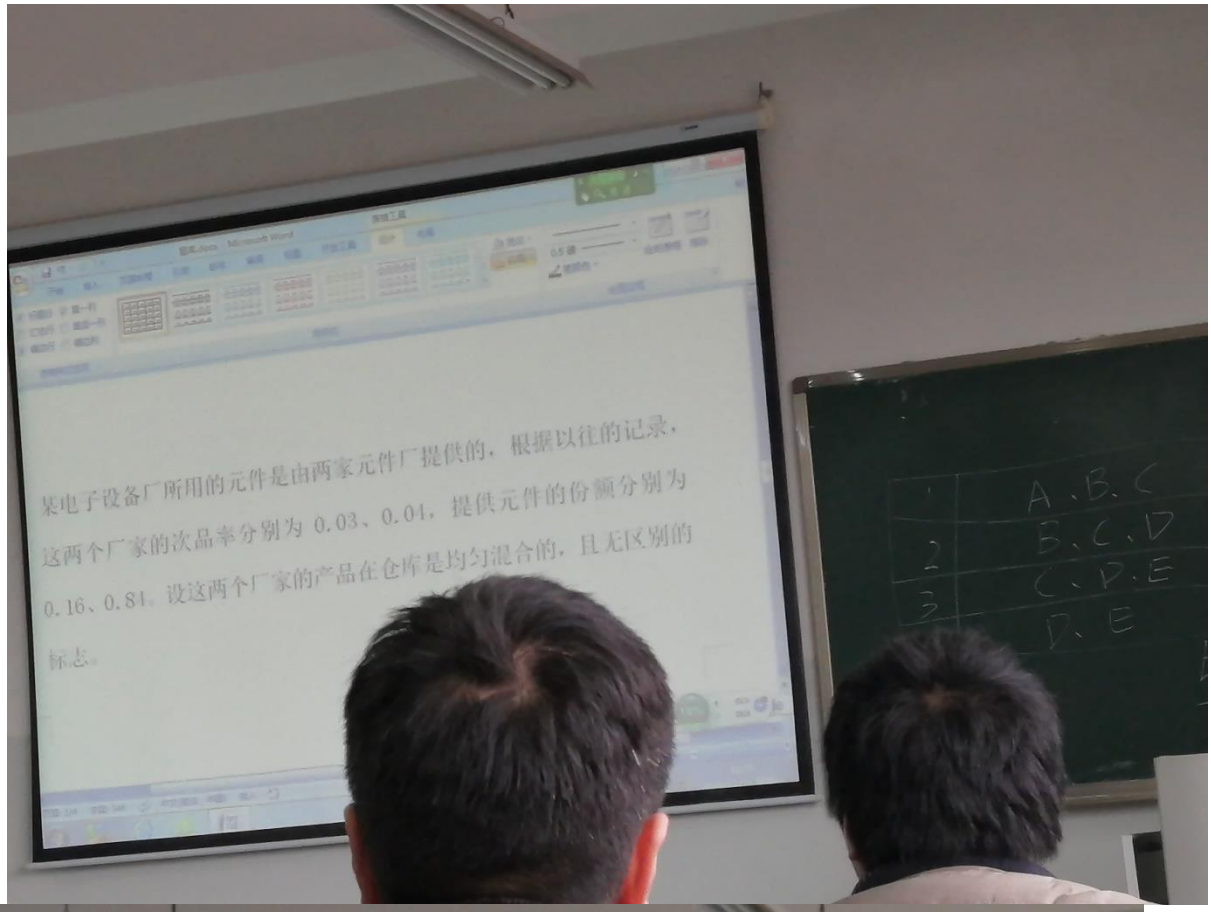
1） 统计学主要利用概率论建立数学模型，是研究随机现象的常用数学工具之一。

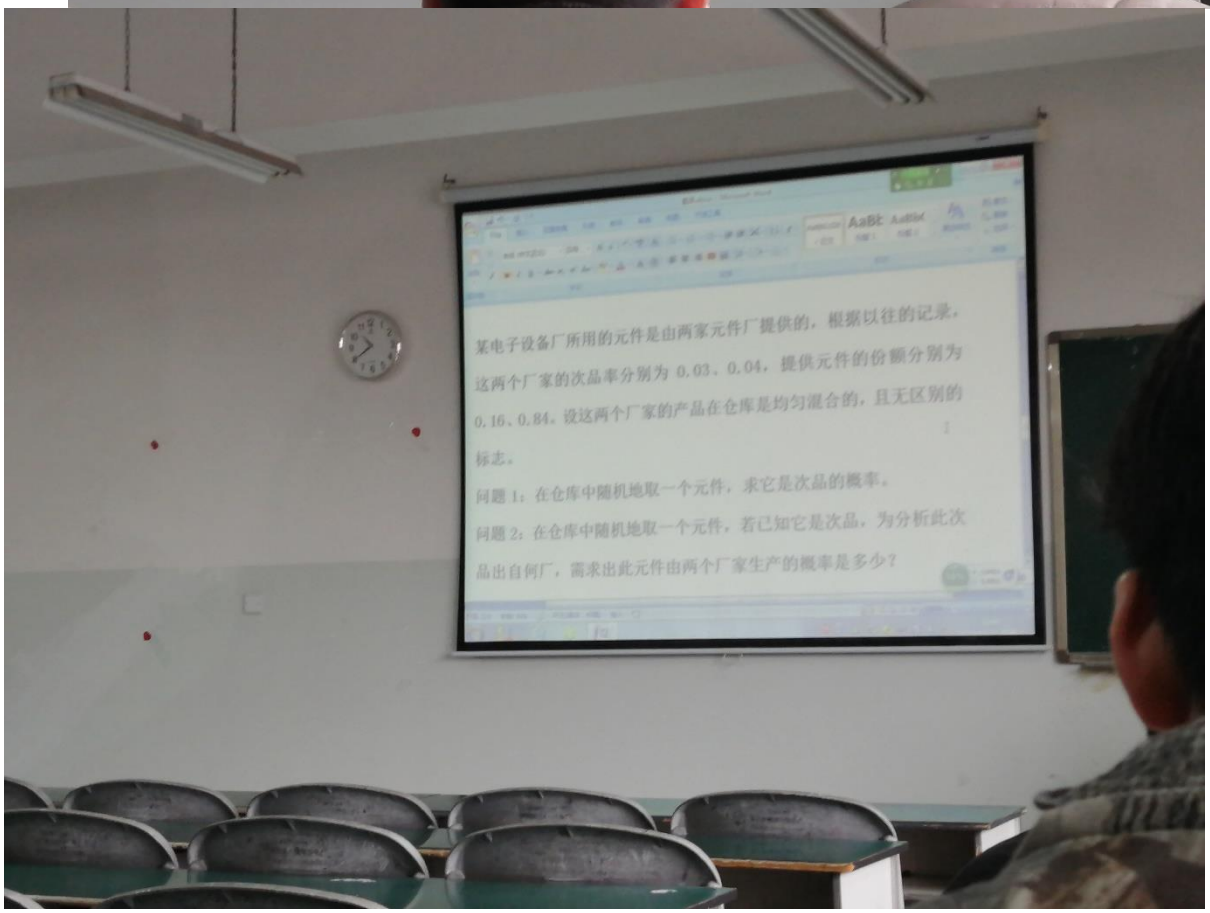
2） 数据挖掘分析大量数据，发现其中的内在联系和知识，并以模型或规则表达这些知识。

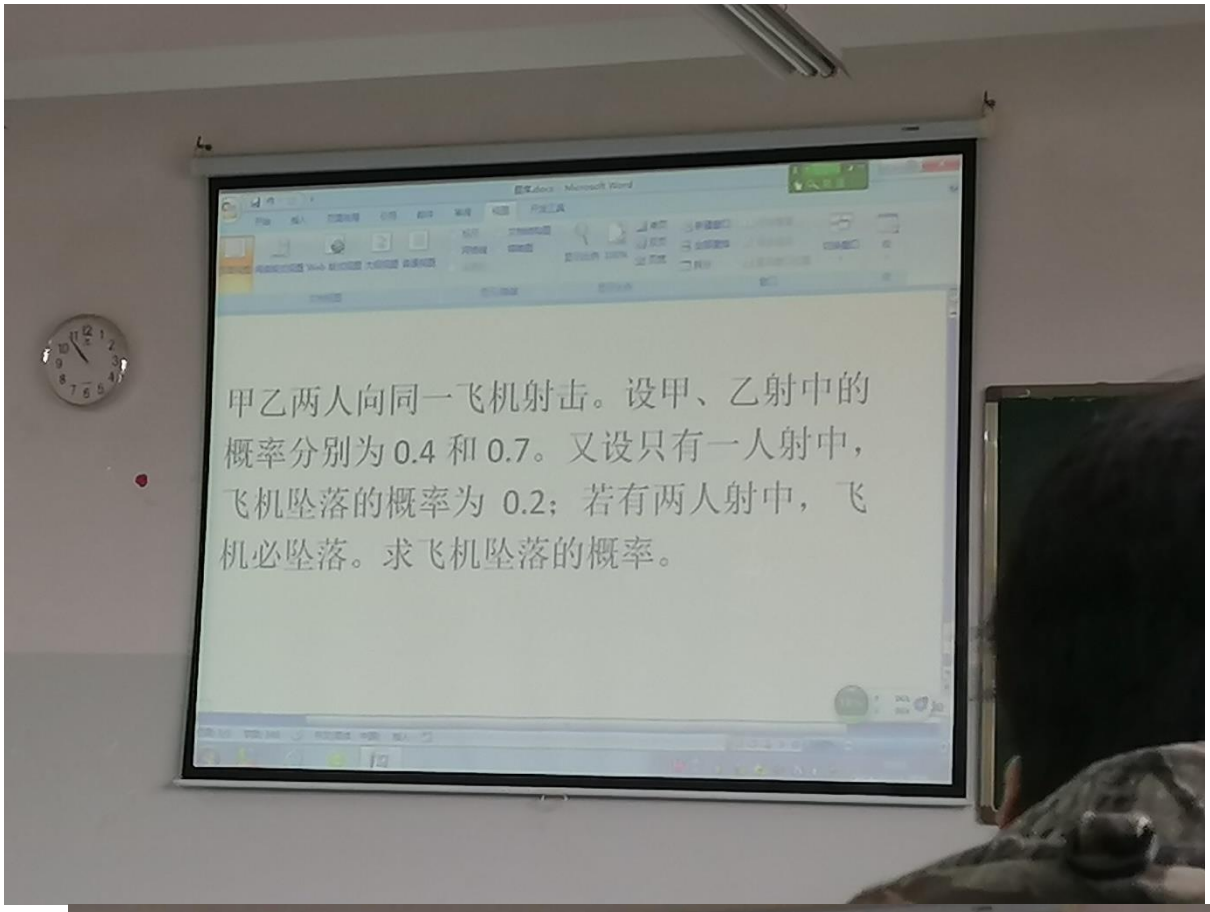
一个主要差别在于处理对象（数据集）的尺度和性质。数据挖掘经常会面对尺度为 GB 甚至 TB 数量级的数据库，而用传统的统计方法很难处理这么大尺度的数据集。传统的统计处理往往是针对特定的问题采集数据（甚至通过试验设计加以优化）和分析数据来解决特定问题；而数据挖掘却往往是数据分析的次级过程，其所用的数据原本可能并非为当前研究而专门采集的，因而其适用性和针对性可能都不强，在数据挖掘的过程中，需要对异常数据及冲突字段等进行预处理，尽可能提高数据的质量，然后才经过预处理的数据进行数据挖掘。

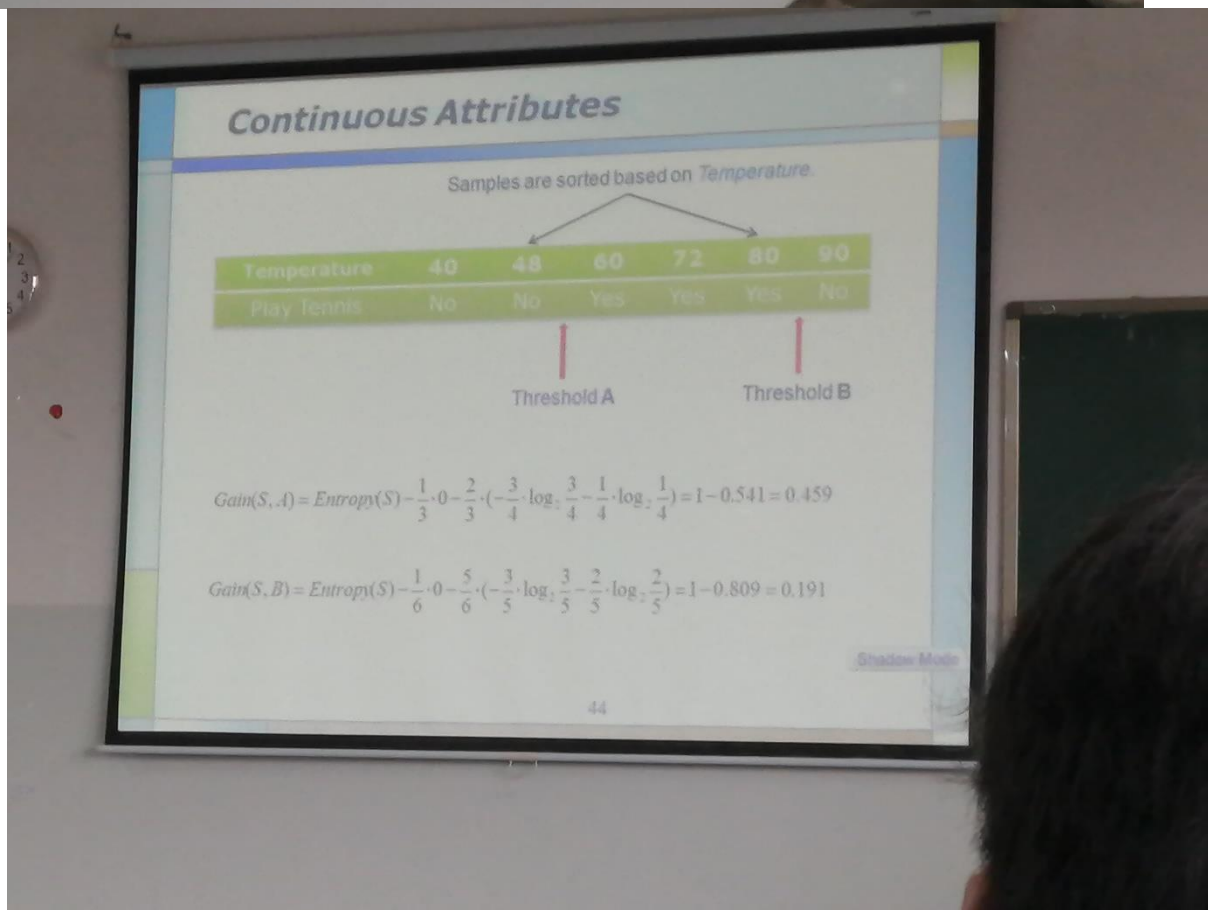
另一个差别在于面对结构复杂的海量数据，数据挖掘往往需要采用各种相应的数学模型和应用传统统计学以外的数学工具，才能建立最适合描述对象的模型或规则。

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


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？





《数据挖掘》试题一

一、辨析题：请解释并辨析以下概念。

1.什么是过拟合，泛化性？并分析两者的联系和区别。

答：为了得到一致假设而使假设变得过度复杂称为过拟合。想像某种学习算法产生了一个过拟合的分类器，这个分类器能够百分之百的正确分类样本数据（即再拿样本中的文档来给它，它绝对不会分错），但也就为了能够对样本完全正确的分类，使得它的构造如此精细复杂，规则如此严格，以至于任何与样本数据稍有不同的文档它全都认为不属于这个类别。

一个假设能够正确分类训练集之外数据（即新的，未知的数据）的能力称为该假设的泛化性。

2.请分析特征选择和特征提取有何区别？

答：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。

3.试分析回归和分类的区别？

答：分类问题和回归问题都要根据训练样本找到一个实值函数 g(x)。回归问题的要求是：给定一个新的模式，根据训练集推断它所对应的输出 y（实数）是多少。也就是使用 y=g(x)来推断任一输入 x 所对应的输出值。分类问题是：给定一个新的模式，根据训练集推断它所对应的类别（如：+1，-1）。也就是使用 y=sign(g(x))来推断任一输入 x 所对应的类别。综上，回归问题和分类问题的本质一样，不同仅在于他们的输出的取值范围不同。分类一般针对离散型数据而言的，回归是针对连续型数据的，但是其实本质上是一样的。

4.请论述 LDA 和 Fisher LDA，并辨析其区别。

答：LDA 是线性判别式分析，鉴别分析的基本思想是将高维的模式样本投影到最佳鉴别矢量空间，以达到抽取分类信息和压缩特征空间维数的效果，投影后保证模式样本在新的子空间有最大的类间距离和最小的类内距离，即模式在该空间中有最佳的可分离性。就是说，它能够保证投影后模式样本在新的空间中有最小的类内距离和最大的类间距离，即模式在该空间中有最佳的可分离性。

二、综述题：请解释并论述以下问题。

1.请描述有监督学习、无监督学习以及半监督学习的区别和联系？

答：利用一组已知类别的样本调整分类器的参数，使其达到所要求性能的过程，称为有监督学习。无监督学习：设计分类器时候，用于处理未被分类标记的样本集，目标是我们不告诉计算机怎么做，而是让它（计算机）自己去学习怎样做一些事情。非监督学习一般有两种思路。第一种思路是在指导 Agent 时不为其指定明确的分类，而是在成功时采用某种形式的激励制度。需要注意的是，这类训练通常会置于决策问题的框架里，因为它的目标不是产生一个分类系统，而是做出最大回报的决定。半监督学习（Semi-supervised Learning）是模式识别和机器学习领域研究的重点问题，是监督学习与无监督学习相结合的一种学习方法。它主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的问题。半监督学习对于减少标注代价，提高学习机器性能具有非常重大的实际意义。

2.试论述如何将聚类用于数据预处理和选择特征。

答：

数据挖掘的完整流程是什么？

答：（1）数据理解：数据理解阶段从初始的数据收集开始，通过一些活动的处理，目的是熟悉数据，识别数据的质量问题，首次发现数据的内部属性，或是探测引起兴趣的子集去形成隐含信息的假设。（2）数据准备：数据准备阶段包括从未处理数据中构造最终数据集的所有活动。这些数据将是模型工具的输入值。这个阶段的任务有个能执行多次，没有任何规定的顺序。任务包括表、记录和属性的选择，以及为模型工具转换和清洗数据。（3）建模：在这个阶段，可以选择和应用不同的模型技术，模型参数被调整到最佳的数值。一般，有些技术可以解决一类相同的数据挖掘问题。有些技术在数据形成上有特殊要求，因此需要经常跳回到数据准备阶段。（4）评估： 到项目的这个阶段，你已经从数据分析的角度建立了一个高质量显示的模型。在开始最后部署模型之前，重要的事情是彻底地评估模型，检查构造模型的步骤，确保模型可以完成业务目标。这个阶段的关键目的是确定是否有重要业务问题没有被充分的考虑。在这个阶段结束后，一个数据挖掘结果使用的决定必须达成。（5）部署：通常，模型的创建不是项目的结束。模型的作用是从数据中找到知识，获得的知识需要便于用户使用的方式重新组织和展现。根据需求，这个阶段可以产生简单的报告，或是实现一个比较复杂的、可重复的数据挖掘过程。

讨论题：（3 选 2）

1.如何改进 k-means 算法中的 k 的选取问题？

2.请描述 EM 算法原理和技术。

答：EM 算法是一种迭代算法,主要用来计算后验分布的众数或极大似然估计,广泛地应用于缺损数据、截尾数。在统计计算中，最大期望（EM）算法是在概率（probabilistic）模型中寻找参数最大似然估计或者最大后验估计的算法，其中概率模型依赖于无法观测的隐藏变量（Latent Variable）。最大期望经常用在机器学习和计算机视觉的数据聚类（Data Clustering）领域。最大期望算法经过两个步骤交替进行计算：第一步是计算期望（E），利用对隐藏变量的现有估计值，计算其最大似然估计

值；第二步是最大化（M），最大化在 E 步上求得的最大似然值来计算参数的值。M 步上找到的参数估计值被用于下一个 E 步计算中，这个过程不断交替进行。总体来说，EM 的算法流程如下：1.初始化分布参数 2.重复直到收敛：E 步骤：估计未知参数的期望值，给出当前的参数估计。M 步骤：重新估计分布参数，以使得数据的似然性最大，给出未知变量的期望估计。据、成群数据、带有讨厌参数的数据等所谓的不完全数据的统计推断问题。

3.决策树算法有哪些种类和改进？

答：决策树算法是一种逼近离散函数值的方法。它是一种典型的分类方法，首先对数据进行处理，利用归纳算法生成可读的规则和决策树，然后使用决策对新数据进行分析。本质上决策树是通过一系列规则对数据进行分类的过程决策树的典型算法有 ID3，C4.5，CART 等。现存的决策树算法也存在着很多不足之处，如计算效率低下、多值偏向等。