mahout

1、<http://blog.csdn.net/gdp12315_gu/article/details/49777797>

**聚类（Clustering）：**

将物理或抽象对象的集合分成由类似的对象组成的多个类的过程被称为聚类。

**分类（Classification）：**

在已有分类标准下，对新数据进行划分，分类。

假设有一批人的年龄的数据，大致知道其中有一堆少年儿童，一堆青年人，一堆老年人。

**聚类**就是自动发现这三堆数据，并把相似的数据聚合到同一堆中。所以对于这个例子，如果要聚成3堆的话，那么输入就是一堆年龄数据，注意，此时的年龄数据并不带有类标号，也就是说我只知道里面大致有三堆人，至于谁是哪一堆，现在是不知道的，而输出就是每个数据所属的类标号，聚类完成之后，就知道谁和谁是一堆了。

而**分类**就是，我事先告诉你，少年儿童、青年人及老年人的年龄是什么样的，现在新来了一个年龄，输出它的类标号，就是它是属于少年儿童、青年人、老年人的哪个类。一般来说，分类器是需要训练的，也就是要告诉你的算法，每个类的特征是什么样子，它才能识别新的数据。

刚才举的是一个超级简单的例子，方便大家理解。下面再举一个实际的例子。

对于聚类，比如有些搜索引擎有“查看相似网页”的功能，这个就可以用聚类来做，把网页就行聚类，在聚类的结果中，每一个类中的网页看成是相似的。

对于分类，比如手写识别就可以看到是分类问题，比如我写了10个“我”字，然后对这10个“我”字进行特征提取，就可以告诉算法，“我”字具有什么样的特征，于是来了一个新的“我”字，虽然笔画和之前的10个“我”字不完全一样，但是特征高度相似，于是就把这个手写的字分类到“我”这个类，就识别出来了。

2、回归（regression）：输出连续的

分类：输出离散的

(1)分类。分类是找出数据库中的一组数据对象的共同特点并按照分类模式将其划分为不同的类，其目的是通过分类模型，将数据库中的数据项映射到摸个给定的类别中。可以应用到涉及到应用分类、趋势预测中，如淘宝商铺将用户在一段时间内的购买情况划分成不同的类，根据情况向用户推荐关联类的商品，从而增加商铺的销售量。

(2)回归分析。回归分析反映了数据库中数据的属性值的特性，通过函数表达数据映射的关系来发现属性值之间的依赖关系。它可以应用到对数据序列的预测及相关关系的研究中去。在市场营销中，回归分析可以被应用到各个方面。如通过对本季度销售的回归分析，对下一季度的销售趋势作出预测并做出针对性的营销改变。

(3)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。

(4)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。

(5)神经网络方法。神经网络作为一种先进的人工智能技术，因其自身自行处理、分布存储和高度容错等特性非常适合处理非线性的以及那些以模糊、不完整、不严密的知识或数据为特征的处理问题，它的这一特点十分适合解决数据挖掘的问题。典型的神经网络模型主要分为三大类：第一类是以用于分类预测和模式识别的前馈式神经网络模型，其主要代表为函数型网络、感知机;第二类是用于联想记忆和优化算法的反馈式神经网络模型，以Hopfield 的离散模型和连续模型为代表。第三类是用于聚类的自组织映射方法，以ART 模型为代表。虽然神经网络有多种模型及算法，但在特定领域的数据挖掘中使用何种模型及算法并没有统一的规则，而且人们很难理解网络的学习及决策过程。