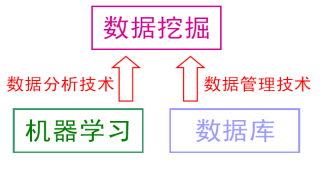
机器学习与数据挖掘



一数据挖掘与机器学习

数据挖掘和机器学习的区别和联系，数据挖掘受到很多学科领域的影响，其中数据库、机器学习、统计学无疑影响最大。对数据挖掘而言，数据库提供数据管理技术，机器学习和统计学提供数据分析技术。

统计学界提供的很多技术通常都要在机器学习界进一步研究，变成有效的机器学习算法之后才能再进入数据挖掘领域。统计学主要是通过机器学习来对数据挖掘发挥影响，而机器学习和数据库则是数据挖掘的两大支撑技术。

从数据分析的角度来看，绝大多数数据挖掘技术都来自机器学习领域，但机器学习研究往往并不把海量数据作为处理对象，因此，数据挖掘要对算法进行改造，使得算法性能和空间占用达到实用的地步。同时，数据挖掘还有自身独特的内容，即关联分析。

关联分析就是希望从数据中找出“买尿布的人很可能会买啤酒”这样看起来匪夷所思但可能很有意义的模式。实际上，在面对少量数据时关联分析并不难，可以直接使用统计学中有关相关性的知识，而关联分析的困难其实完全是由海量数据造成的，因为数据量的增加会直接造成挖掘效率的下降，当数据量增加到一定程度，问题的难度就会产生质变。

二数据挖掘

数据挖掘：从数据中提取出隐含的过去未知的有价值的潜在信息。

数据挖掘涉及六类常见的任务：

异常检测 – 识别不寻常的数据记录，错误数据需要进一步调查。

关联规则学习 – 搜索变量之间的关系。

聚类 – 是在未知数据的结构下，发现数据的类别与结构。

分类 – 是对新的数据推广已知的结构的任务。例如，一个电子邮件程序可能试图将一个电子邮件分类为“合法的”或“垃圾邮件”。

回归 – 试图找到能够以最小误差对该数据建模的函数。

汇总 – 提供了一个更紧凑的数据集表示，包括生成可视化和报表。

三机器学习

机器学习：探究和开发一系列算法来如何使计算机不需要通过外部明显的指示，而可以自己通过数据来学习，建模，并且利用建好的模型和新的输入来进行预测的学科。

机器学习步骤框架：

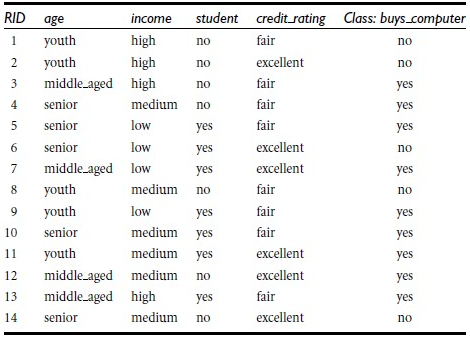
1 把数据拆分为训练集和测试集

2 用训练集和训练集的特征向量来训练算法

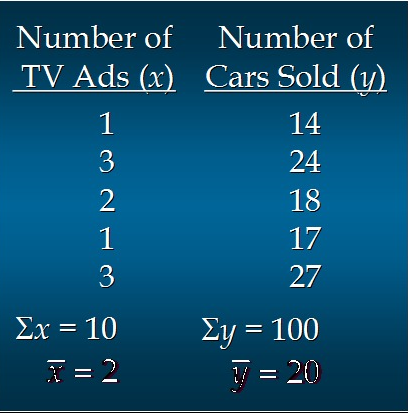
3用学习来的算法运用在测试集上来评估算法 （可能要设计到调整参数或用验证集）

四 几个概念：

分类 : 目标标记为类别型数据



 回归: 目标标记为连续性数值



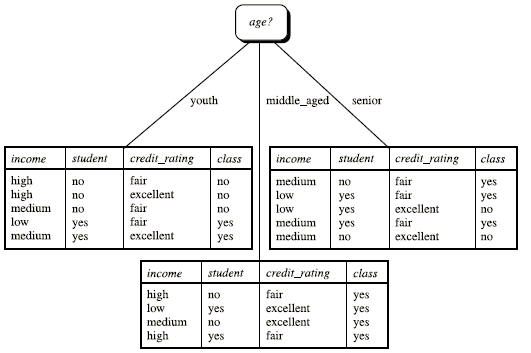
监督学习： 训练集有类别标记。

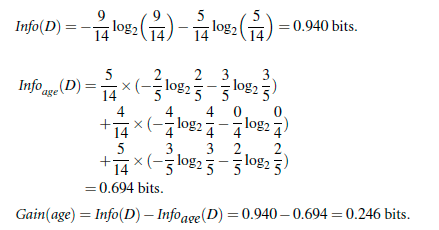
分类：决策树，临近取样（KNN），支持向量机（SVM），神经网络。

回归：线性回归，非线性回归。

决策树算法：

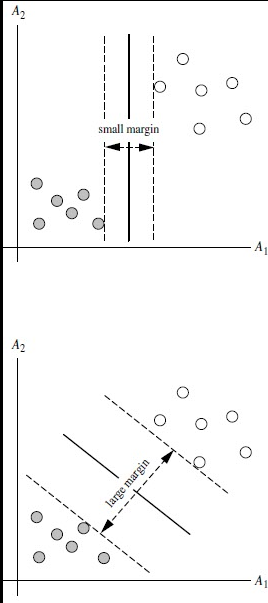
其每个非叶节点表示一个特征属性上的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶节点存放一个类别。使用决策树进行决策的过程就是从根节点开始，测试待分类项中相应的特征属性，并按照其值选择输出分支，直到到达叶子节点，将叶子节点存放的类别作为决策结果。

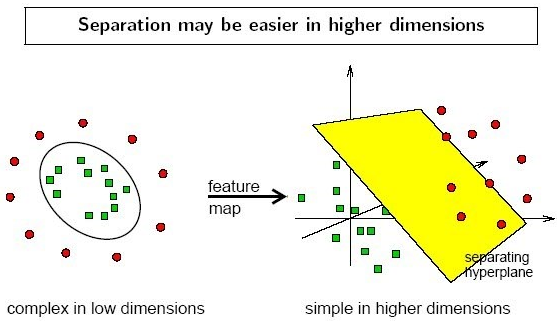




SVM:

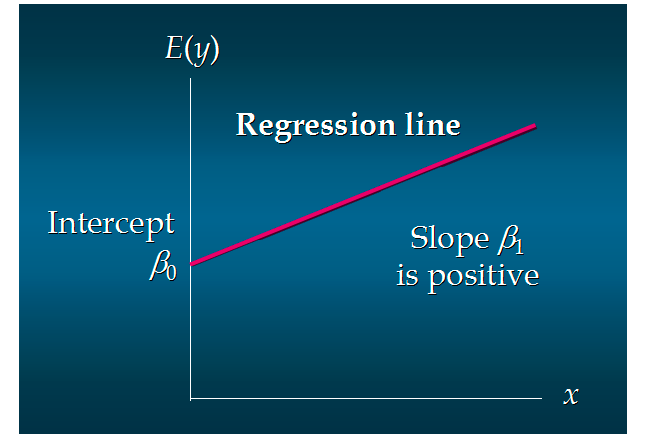
它是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器，即支持向量机的学习策略便是间隔最大化，最终可转化为一个凸二次规划问题的求解。 SVM寻找区分两类的超平面), 使边际最大。

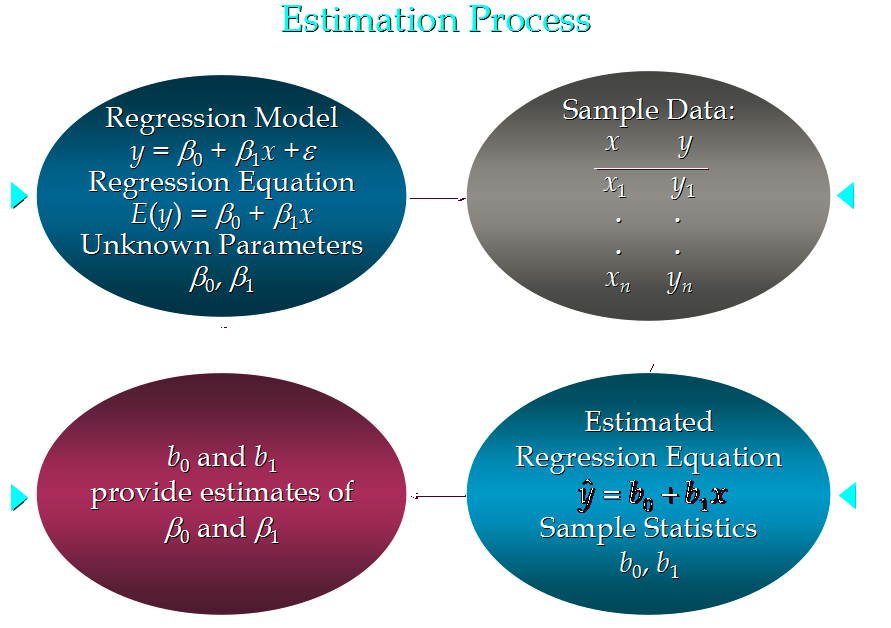




线性回归：

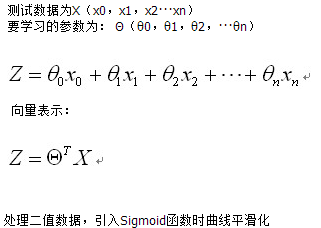
回归分析(regression analysis)用来建立方程模拟两个或者多个变量之间如何关联

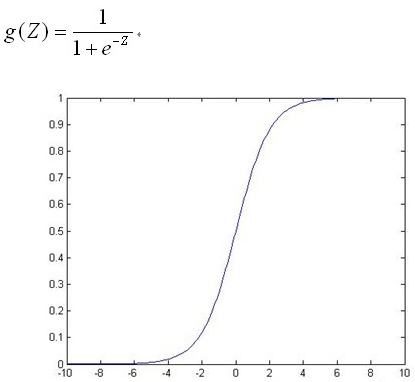


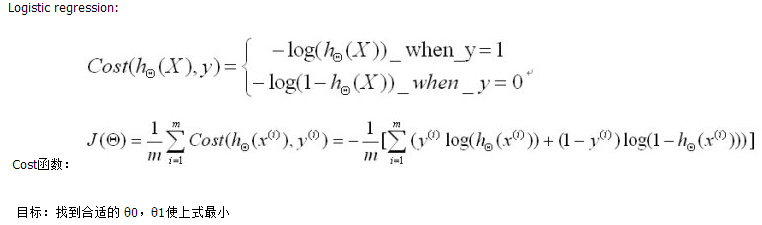




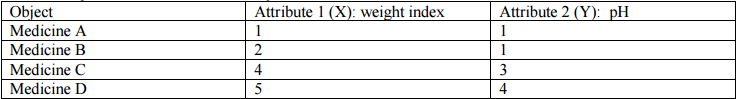
非线性回归（逻辑回归）：







无监督学习： 无类别标记。（K-mean算法聚类）。



半监督学习：有类别标记的训练集 + 无标记的训练集。

五 机器学习算法

**5.1 K-means聚类算法**

K-means 算法数据挖掘十大经典算法之一。

算法思想：算法接受参数 k ；然后将事先输入的n个数据对象划分为 k个聚类以便使得所获得的聚类满足：同一聚类中的对象相似度较高；而不同聚类中的对象相似度较小。  以空间中k个点为中心进行聚类，对最靠近他们的对象归类。通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果。

  算法描述：

1. 适当选择C个类的初始中心；

        2） 在第k次迭代中，对任意一个样本，求其到C各中心的距离，将该样本归到距离最短的中心所在的类；

        3）利用均值等方法更新该类的中心值；

        4）对于所有的C个聚类中心，如果利用（2）（3）的迭代法更新后，值保持不变，则迭代结束，否则继续迭代。

算法流程：

 输入：k, data[n];

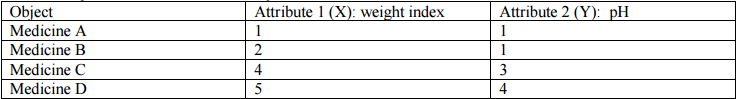
1)选择k个初始中心点，例如c[0]=data[0],…c[k-1]=data[k-1];

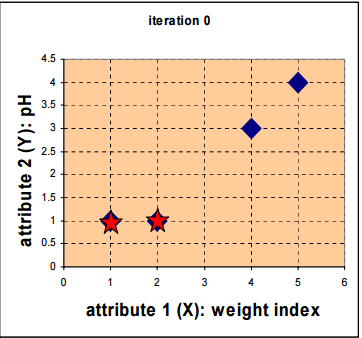
2)对于data[0]….data[n], 分别与c[0]…c[k-1]比较，假定与c[i]差值最少，就标记为i;

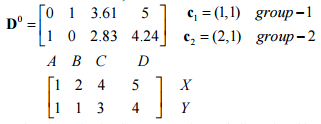
3)对于所有标记为i点，重新计r算c[i]={所有标记为i的data[j]之和}/标记为i的个

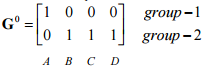
4)重复(2)(3),直到所有c[i]值的变化小于给定阈值。

例子：已知下面药属于两类药，确定每种药的类别。



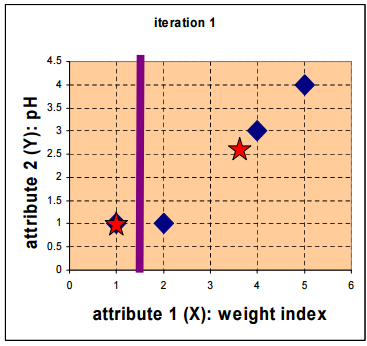


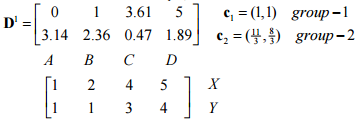


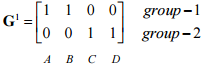






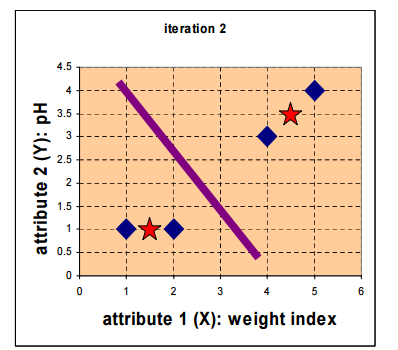


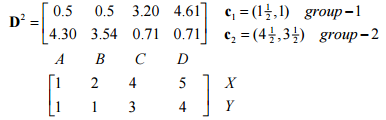


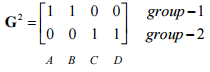




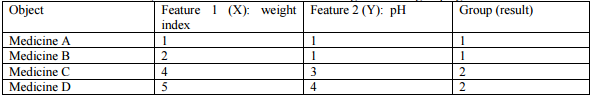












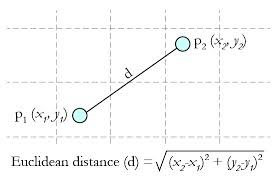
**5.2 KNN算法**

算法步骤：为了判断未知实例的类别，以所有已知类别的实例作为参照， 选择参数K，计算未知实例与所有已知实例的距离。选择最近K个已知实例，根据少数服从多数的投票法则，让未知实例归类为K个最邻近样本中最多数的类别。

细节：

 关于K。

 关于距离的衡量方法。



其他距离衡量：余弦值,曼哈顿距离。

例子：判断未知电影的类型





