

常见的数据挖掘算法：

1. **决策树：**
2. **CHAID:**卡方自动相互关系检测，利用卡法检验来选择对因变量最有影响的自变量，卡方检验适用于类别性变量的检验，如果自变量是区间型变量，CHAID改用F检验。
3. **CART:**分类与回归树，CART的分割逻辑与CHAID相同，每一层的划分都是基于对所有自变量的检验和选择上，但是，CART采用的检验标准不是卡方检验，二十基尼系数等不纯度的指标，两者最大的区别在于CHAID采用的是局部最优原则，即结合点之间不互相干，一个节点确定了之后，下面的生长过程完全在节点内进行。P19（数据万巨额与数据化运营实战）
4. **ID3:**迭代的二分器，基于信息增益的度量选择具有最高信息增益的属性作为节点的分裂属性，之后发展起来的C4.5采用信息增益率代替了ID3中的信息增益度量。

**决策树在数据化运营中的主要用途体现在：作为分类、预测问题的电影支持技术、它在用户划分、行为预测、规则梳理等方面具有广泛的应用前景、决策树甚至可以作为其他建模技术前期进行变量筛选的一种方法，即 决策树的分割来筛选有效的输入自变量。**

**决策树模型非常直观，生成一系列的如果….那么…的逻辑判断很容易让人理解和应用，这个特点是决策树赢得广泛应用的最主要原因，真正体现了简单、直观、易懂。**

**决策树搭建和应用的速度比较快，并且可以处理区间行变量和类别性变量。**

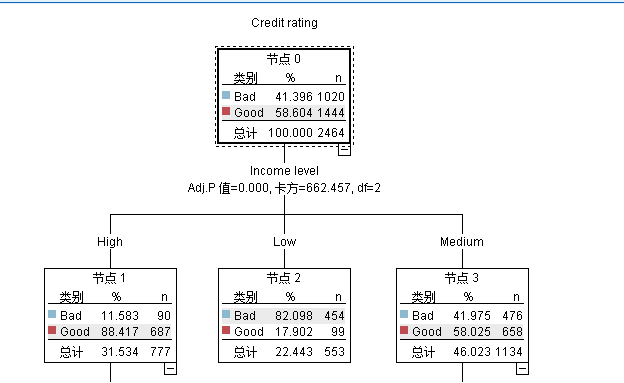
**决策树最大缺点为原理中的贪心算法，贪心算法总是做出在当前看来最好的选择，却并不从整体上思考最优的划分，因此，它所做的只能是某种意义上的局部最优算法。**

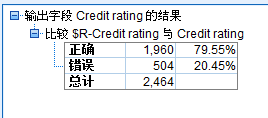
**如果目标变量是连续型变量，那么决策树就不适用了，最好改用线性回归算法解决。**

**决策树缺乏像回归或者聚类那样丰富多样的检测指标和评价方法，这也是今后努力的一个方向。**

**决策树为监督算法**

**决策树的输出结果如下：**





**决策树是可以通过对现有的监督集进行训练，可以用来预测现有的数据**

1. **神经网络**

简单的讲，神经网络就是通过输入多个非线性模型以及不同模型之间的加权互联（加权的过程在隐蔽层完成），最终得到一个输出模型，其中，隐藏层所包含的就是非线性函数。

神经网络技术的主要缺点和注意事项：

神经网络需要比较长的模型训练时间，在面对大数据量时尤其如此。

对于神经网络模型来说，少而精的变量才可以充分发挥神经网络的模型效率，但是神经网络本身是无法挑选变量的，因此，对于神经网络的实际应用来讲，之前的变量挑选环节必不可少，虽然变量的选择对于任何一个模型的搭建来说都是很重要的环节，但是必须强调的是，对于神经网络模型来说，尤其重要，这是尤其复杂的内部结构决定的。

神经网络本身对于缺失值比较敏感，所以，应用该技术时要注意针对缺失值进行适当的处理，或者赋值、或者替换、或者删除。

它具有过度拟合数据的倾向，可能导致模型应用于新数据时效率显著下降，鉴于此，针对神经网络模型的应用要仔细检查，在确保稳定的前提下才可以投入业务的落地应用。

由于其结构的复杂性和解决的难以解释性，神经网络在商业实践中远远没有回归和决策树应用得广泛。

**神经网络技术在数据化运营中的主要用途体现在：作为分类、预测问题的重要技术支持、在用户划分、行为预测、营销相应等诸多应用方面具有广泛的应用场景。**

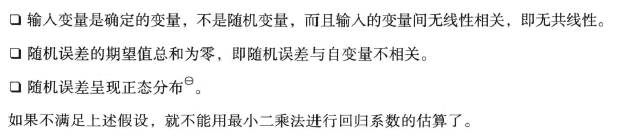
1. **回归**

回归主要指**多元线性回归**和**Logistic回归**，其中在数据化运营中更多使用的是Logistic回归，它又包括相应预测、分类划分等内容。

当目标变量是二元变量（即是与否）的时候，逻辑回归分析是一种非常成熟的、可靠的主流模型算法。

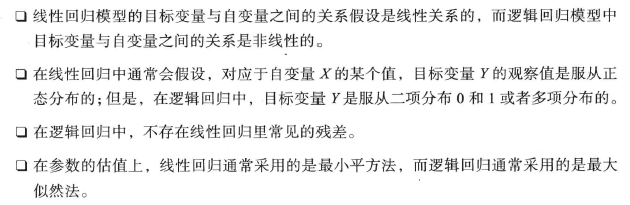
**1.多元线性回归：**描述一个因变量如何随着一批自变量的变化而变化，其回归方程就是因变量和自变量关系的数据反映。因变量的变化包括两部分：系统性变化与随机变化，其中，系统性变化是由自变量引起的（自变量可以解释的），随便变化是不能由自有变量解释的，通常也称作残值，在用来估算多元线性回归方程中自变量系数的方法中，最常用的是最小二乘法，即找出一组对应自变量的相应参数，以使因变量的实际观测值与回归方程的欲则指之间的总方差减到最小。

对于多元线性回归方程的参数估计，是基于下面假设：



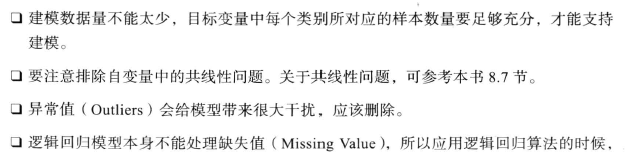
**2.Logistic回归：**相比于线性回归来说，在数据化运营中有更主流更频繁的应用，主要是因为该分析技术可以很好滴回答诸如预测、分类等数据化运营常见的分析项目主题。简单来说，凡是预测两选一的时间可能性，比如相应还是不响应，买还是不买，流失还是不流失都可以采用Logistic回归。

线性回归与逻辑回归的最大区别在于目标变量是类别型的还是区间型的，线性回归所针对的目标变量是区间型的，而逻辑回归针对的目标变量是类别型的，另外线性回归和逻辑回归模型的主要区别如下：



无论是线性回归还是逻辑回归，在回归拟合的过程中，都要进行变量的筛选，并且有各种不同的变量筛选方案，其中最常见、最著名的3中方法分别是向前引入法、向后剔除法、逐步回归法。

逻辑回归注意事项：



聚类与分类的主要区别为聚类为无监督，分类为有监督。

分类与回归的区别为，分类为离散，回归为连续

回归与时间序列的区别在于回归不强调数据间的先后顺序，而时间序列需要考虑时间特征。

1. **关联规则**

关联规则最经典的案例就是购物篮分析，通过分析顾客购物篮中商品之间的关联，可以挖掘顾客的购物习惯，从而帮助零售商更好地指定有针对性的营销策略。

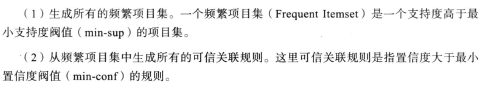
举例来说：

婴儿尿不湿--------------🡪啤酒[支持度=10%，置信度=70%]

在所有顾客中，有10%的顾客同事购买了婴儿尿不湿和啤酒，而在所有购买了婴儿尿不湿的顾客中，占70%的人同事还购买了啤酒。

事实上，支持度和置信度是衡量关联规则强度的两个重要指标，它分别反映着所发现规则的有用性和确定性。如果支持度太小，则说明相应规则只是偶发事件，偶发事件是没有商业价值的。置信度主要衡量规则的确定性，置信度的规则太低在实践应用中也没有太大的用户。

在众多关联规则和数据挖掘算法中，最著名就是Apriori算法。



关联额规则算法不仅在数值型数据集的分析中有很大用途，而且在纯文本文档和网页文件中，也有着重要用途，比如发现单词间的并发关系以及Web的使用模式等，这些都是Web数据挖掘、搜索以及推荐的基础。

1. **聚类**

针对几个特定的业务指标，可以将观察对象的群体按照相似性和相异性进行不同群组的划分，经过划分后，每个群组内部各对象间的相似度会很高，而在不同群组之间的对象，彼此之间将具有很高的相异度。

聚类分析的算法可以分为划分的方法、层次的方法、基于密度的方法、基于网格的方法、基于模型的方法等，其中，前面两种方法最为常用。

**划分的方法：**最常用的划分方法是K-Means方法。

**层次的方法：**依次让最相似的数据对象两两合并，这样不断地合并，最后就形成了一颗聚类树，层次的方法又可以分为凝聚层次聚类和分裂层次聚类。

**基于密度的方法：**该方法优势明显，最大的缺点就是需要用户确定输入参数、而且对参数十分敏感。

**基于网格的方法：**该算法的处理有点是速度快。

聚类技术主要用途表现在：为业务方的精细化运营提供具体的细分依据和相应的运营方案建议，又可在数据处理阶段用作数据探索的工具，包括发现离散点、孤立点、数据将为的手段和方法。**在数据化运营实践中，针对发规模数据集所采用的聚类方法主要是K-Means，尤其是对于大数据集，K-Means算法基本算唯一的主流算法。因为其简洁、高效、易理解、易实施。**

**K-Means聚类方法的主要缺点：**

1. 分析师需要提前制定聚类的数目k，在实践中需要不同的k指才能根据效果比较来选择最合适的k值，这个过程可能比较耗时。
2. 算法对噪声和异常值比较敏感。由于K-Means算法是采用均值作为每个聚类的聚类中心，所以异常值会严重干扰聚类中心的计算，造成聚类失真。

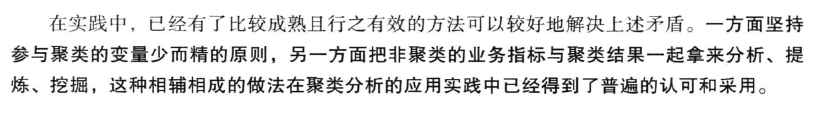
**聚类评价指标：**

**RMSSTD：**群体中所有变量的综合标准差。

**R-Square：**聚类后群体间差异的大小，一个好的聚类结果，应该R-square在[0,1]之间，并且R-Square越接近1越好。

**SPR：**该指标适用于层次方法中的凝聚层次聚类算法，它表示当原来两个群里合并成新的群体后，其所损失的群内相似性的比例。

**DBC：**该指标适用于层次方法中的凝聚层次聚类算法，它表示在要合并的两个细分群体时，分别计算两个群体的中心，以求得两个群体的距离。



1. **贝叶斯分类方法**

贝叶斯分类方法是非常成熟的统计学分类方法，它主要用来预测类成员间关系的可能性。 朴素贝叶斯方法作为一种简单贝叶斯方法甚至也可跟决策树和神经网络算法相媲美。

贝叶斯分类方法再数据化运营实践中主要用于分类问题的归类等应用场景。

1. **支撑向量机**

支持向量机与传统的神经网络技术相比，支持向量机不仅结构简单，而且各项技术性能也明显提升，因此它成为当今机器学习的重点领域之一。

支持向量机的缺点是训练数据较大，但是它的优点也很明显，对于复杂的非线性的决策边界的建模能力高度准确，并且也不太容易过拟合，支持向计机主要用在预测、分类这样的实际分析需求中。

**支持向量机特别适应于分析预测变量字段非常多（数千个）的数据。**

1. **主成分分析**

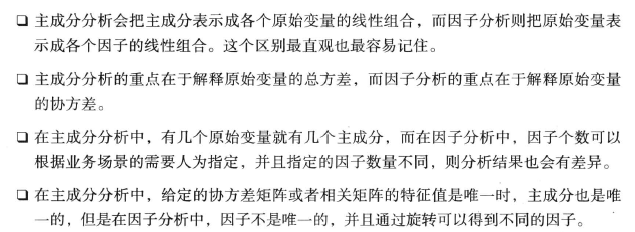
主成分分析会通过线性组合将多个原始变量合并成若干个主成分，这样每个主成分都变成了原始标量的线性组合，这种转变的目的，一方面可以大幅度降低原始数据的维度，同时也在此过程中发现原始数据属性之间的关系。

主要步骤如下：

1. 变量标准化。
2. 选择协方差阵或者相关阵计算特征根及对象的特征向量。
3. 计算方差贡献率，并根据方差贡献率的阈值选取合适的主成分个数。
4. 根据主成分载荷的大小对选择的主成分进行命名。
5. 根据主成分载荷计算各个主成分的得分。

将主成分进行推广和延伸即成为因子分析，因子分析再综合原始变量信息的基础上将会力图构建若干个意义较为明确的公因子，即采用少数几个因子描述多个指标之间的联系，将比较密切的变量归为同一类中，每类变量即一个因子。之所以称为因子，因为他们实际上是不可测量的，只能解释。

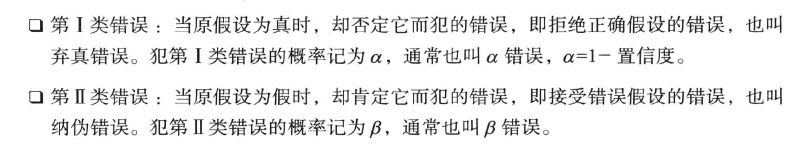
主成分分析是因子分析的一个特例，两者的区别和联系主要表现在：



主成分分析和因子分析再数据化运营实践中主要用户数据处理、降维、变量间关系的探索等方面。

1. **假设检验**

假设检验主要分为两种错误：



在数据化运营的商业实践中，假设检验最常用的场景就是用于运营效果的评估上。

探索性数据分析：从数据中找出一些可能进一步进行评估的电子。

BCG矩阵，波士顿矩阵认为一般决定产品结构的基本因素有两个：即市场引力与企业实力。

波士顿矩阵对于企业产品所处的四个象限具有不同的定义和相应的战略对策。

**（1）明星产品（stars）**

它是指处于高增长率、高市场占有率象限内的产品群，这类产品可能成为企业的现金牛产品，需要加大投资以支持其迅速发展。采用的发展战略是：积极扩大经济规模和市场机会，以长远利益为目标，提高市场占有率，加强竞争地位。发展战略以投明星产品的管理与组织最好采用事业部形式，由对生产技术和销售两方面都很内行的经营者负责。

**（2）现金牛产品（cash cow）**

又称厚利产品。因而成为企业回收资金，支持其它产品，尤其明星产品投资的后盾。

对这一象限内的大多数产品，市场占有率的下跌已成不可阻挡之势，因此可采用收获战略：即所投入资源以达到短期收益最大化为限。①把设备投资和其它投资尽量压缩；②采用榨油式方法，争取在短时间内获取更多利润，为其它产品提供资金。对于这一象限内的销售增长率仍有所增长的产品，应进一步进行市场细分，维持现存市场增长率或延缓其下降速度。对于现金牛产品，适合于用事业部制进行管理，其经营者最好是市场营销型人物。

现金牛业务指低市场成长率、高相对市场份额的业务，这是成熟市场中的领导者，它是企业现金的来源。由于市场已经成熟，因为如果市场环境一旦变化导致这项业务的市场份额下降，公司就不得不从其他业务单位中抽回现金来维持现金牛的领导地位，否则这个强壮的现金牛可能就会变弱，甚至成为瘦狗。

**（3）问题产品（question marks）**

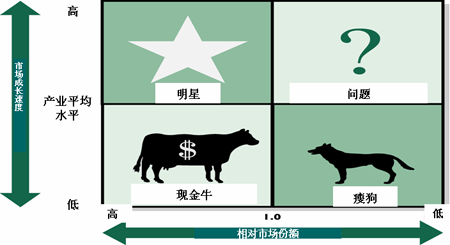
它是处于高增长率、低市场占有率象限内的产品群。前者说明市场机会大，前景好，而后者则说明在市场营销上存在问题。其财务特点是利润率较低，所需资金不足，负债比率高。例如在产品生命周期中处于引进期、因种种原因未能开拓市场局面的新产品即属此类问题的产品。

对问题产品应采取选择性投资战略。即首先确定对该象限中那些经过改进可能会成为明星的产品进行重点投资，提高市场占有率，使之转变成“明星产品”；对其它将来有希望成为明星的产品则在一段时期内采取扶持的对策。因此，对问题产品的改进与扶持方案一般均列入企业长期计划中。对问题产品的管理组织，最好是采取智囊团或项目组织等形式，选拔有规划能力，敢于冒风险、有才干的人负责。

**（4）瘦狗产品（dogs）**

也称衰退类产品。它是处在低增长率、低市场占有率象限内的产品群。其财务特点是利润率低、处于保本或亏损状态，负债比率高，无法为企业带来收益。

对这类产品应采用撤退战略：首先应减少批量，逐渐撤退，对那些销售增长率和市场占有率均极低的产品应立即淘汰。其次是将剩余资源向其它产品转移。第三是整顿产品系列，最好将瘦狗产品与其它事业部合并，统一管理。



充分了解了四种业务的特点后还须进一步明确各项业务单位在公司中的不同地位，从而进一步明确其战略目标。通常有四种战略目标分别适用于不同的业务。

**发展**

以提高经营单位的相对市场占有率为目标．甚至不惜放弃短期收益。要是金牛类业务想尽快成为“明星”，就要增加资金投入。

**保持**

投资维持现状，目标是保持业务单位现有的市场份倾、对于较大的“金牛”可以此为目标，以使它们产生更多的收益。

**收割**

这种战略主要是为了获得短期收益，目标是在短期内尽可能地得到最大限度的现金收入。对处境不佳的金牛类业务及没有发展前途的问题类业务和瘦拘类业务应视具体情况采取这种策略。

**放弃**

目标在于清理和撤销某些业务，减轻负担，以便将有限的资源用于效益较高的业务。这种目标适用于无利可图的瘦狗类和问题类业务。一个公司必须对其业务加以调整，以使其投资组合趋于合理。

云标签

支持向量机：高斯核函数进行映射

Knn（k近邻）进行分类