多元线性回归

使用前提：线性(Linear)，独立(Independence)，正态(Normality)，齐性(Equal variance)，缩写为LINE原则。

(1) 线性：各自变量xi与因变量 yi 之间存在线性关系，可以通过绘制散点图来进行判断；

(2) 独立：因变量yi的取值之间相互独立，反映到回归模型中，实际上就是要求残差（残差:观测值与拟合值的偏离）ei 之间相互独立；

(3) 正态性：构建多重线性回归模型后，残差ei服从正态分布；

(4) 方差齐性：残差ei的大小不随xi取值水平的变化而变化，即残差ei具有方差齐性。

只有准确把握了LINE核心原则，才能够保证构建符合统计学要求的多重线性回归模型。但是，由于多重线性回归模型具有一定的“抗偏倚性”，如果只是想通过构建方程来探讨自变量和因变量之间的关联性，而非对因变量进行预测，那么后面两个条件可以适当放宽。

此外，还应该注意以下几点：

(5) 因变量yi为连续性变量，而非分类变量；

(6) 自变量xi可以为连续变量，也可以为分类变量，当自变量为多分类无序变量时，则需要设置哑变量，当为有序变量时，则需要根据等级顺序进行赋值。

(7) 对于自变量xi的分布特征没有具体的限定，只要求自变量xi间相互独立，不存在多重共线性；

(8) 对于样本量的要求，根据经验一般要求样本量应当为纳入模型的自变量的20倍以上为宜，比如模型纳入5个自变量，则样本量应当在100以上为宜。

构建方法：

1. 反向淘汰：
   1. 设置一个SL的值，确定小于该值的影响因子被模型保留下来。
   2. 将所有的影响因子带入分类器进行拟合。
   3. 选择P Value最大的因子，如果P > SL则剔除该因子并进入步骤(4)，否则直接完成。
   4. 重新拟合。
2. 顺向选择：

(1)设置一个SL值，，小于该值的因子进入模型。

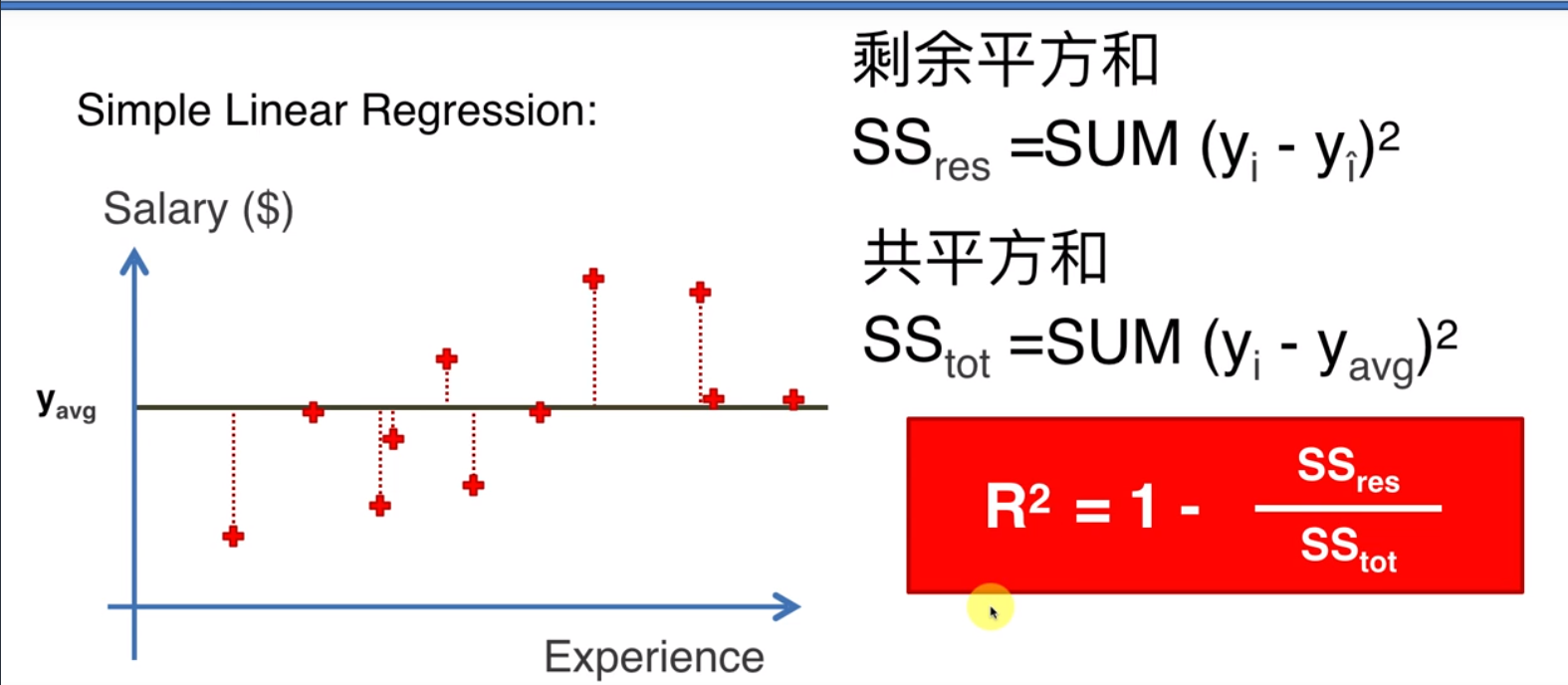
(2)对所有因子单独的拟合simple regression model，选择P Value最小的因子保留。

(3)对所有其他因子的模型，加入保留因子重新拟合。

(4)寻找(3)中P Value最小的中模型，如果 P < SL，返回(3)，否则算法结束。

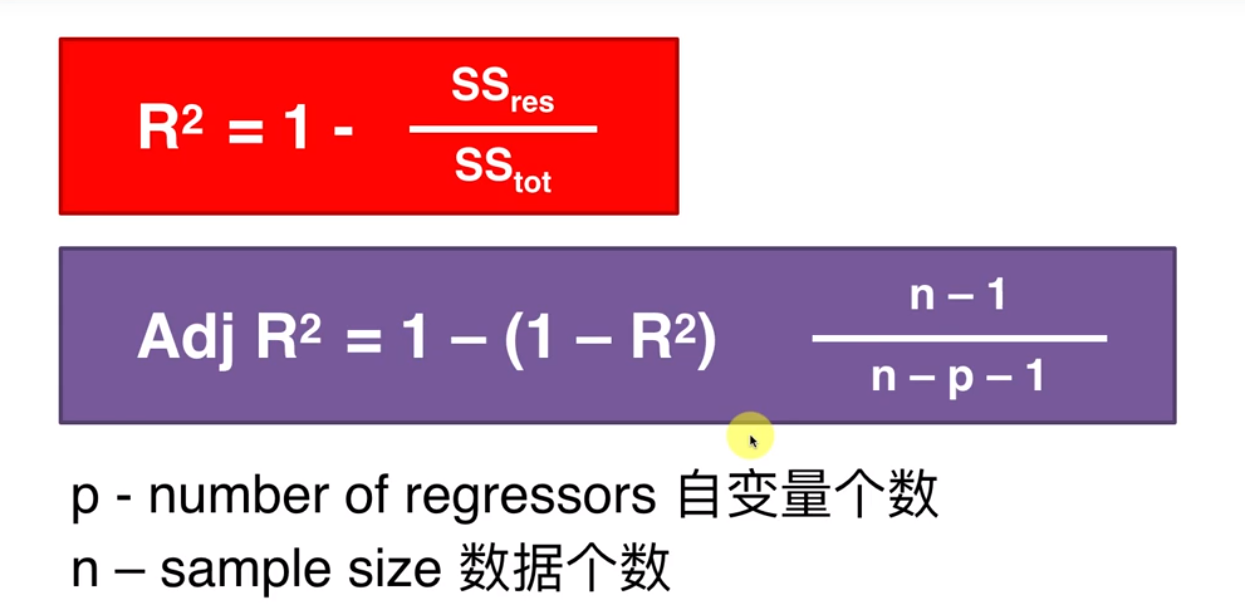
1. 双向选择：
2. 设置SL enter和SL stay的值，确保进入和剔除模型的值。
3. 进行一步顺向传播（P < SL ENTER的变量加入）
4. 进行一次完整的反向淘汰（P < SL LEAVE的变量保留）
5. 直到没有新变量可以添加且没有老变量可以剔除。
6. 信息量比较：

评价回归模型的好坏：

1. 方差：

When R^2 increase, prediction will be better, R^2 will never decrease

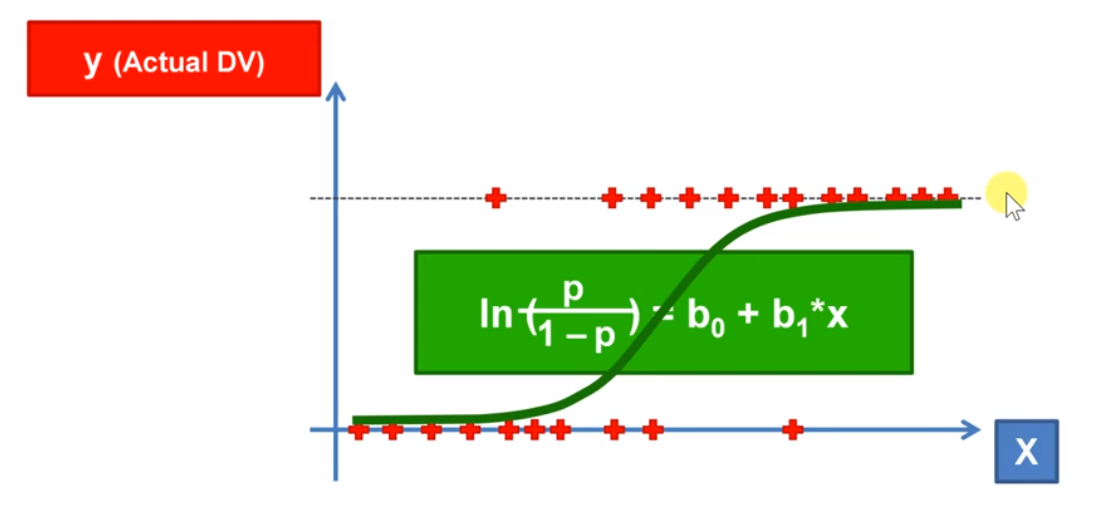
1. 广义方差：



对P（自变量个数）的上升具有惩罚作用，可以避免无限加入新自变量导致提升拟合效果的问题。

1. 回归模型系数的意义：Estimate系数的正负和大小表示该项变化对于预测结果的影响。当各自变量物理性质不同时将其相互进行直接比较说明自变量之间的数学关系（不具有现实意义的情况）。

**逻辑回归**



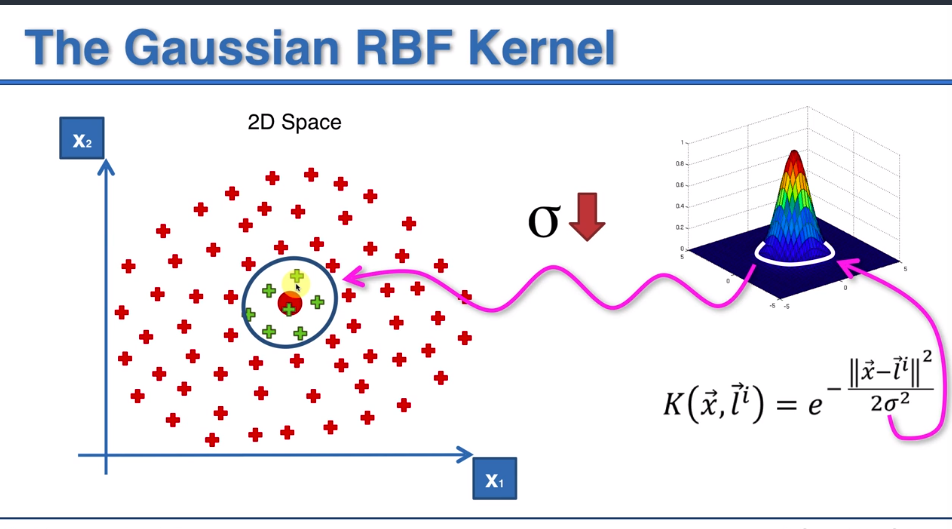
根据Y值与0.5的关系进行分类。归为1或0。

**SVM**

SVM可以对数据集中的极端特征进行分析。SVM寻找Maximum Margin Hyperplane（margin越大，可信度越高），该hyperplane由边缘的两个点构成的support vector指明位置。

**核函数SVM**

没有线性解的时候，需改进SVM。常见方法如下。

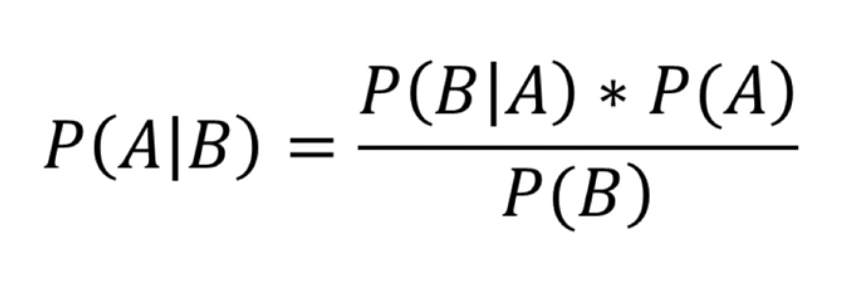
1. 高维投射：将原有数据的维度拉高寻找分割解。但是可能的运算量是巨大的，并不能随意采用。
2. 套用Kernel（以高斯RBF核为例）：

图右侧的变幻函数将平面上的点变换成3维空间中的图象。在右侧的函数图像中，绿色特征的数据的高度值趋向于Z>0, 红色特征的数据的高度值趋向于Z = 0，该图像的中心点就是绿色数据特征的中心点，函数式指数项的分母为控制函数图像的直径的项。

补充：还有其他的核函数可以使http://crsouza.com/2010/03/17/kernel-functions-for-machine-learning-applications/

**朴素贝叶斯**

基于贝叶斯定理。



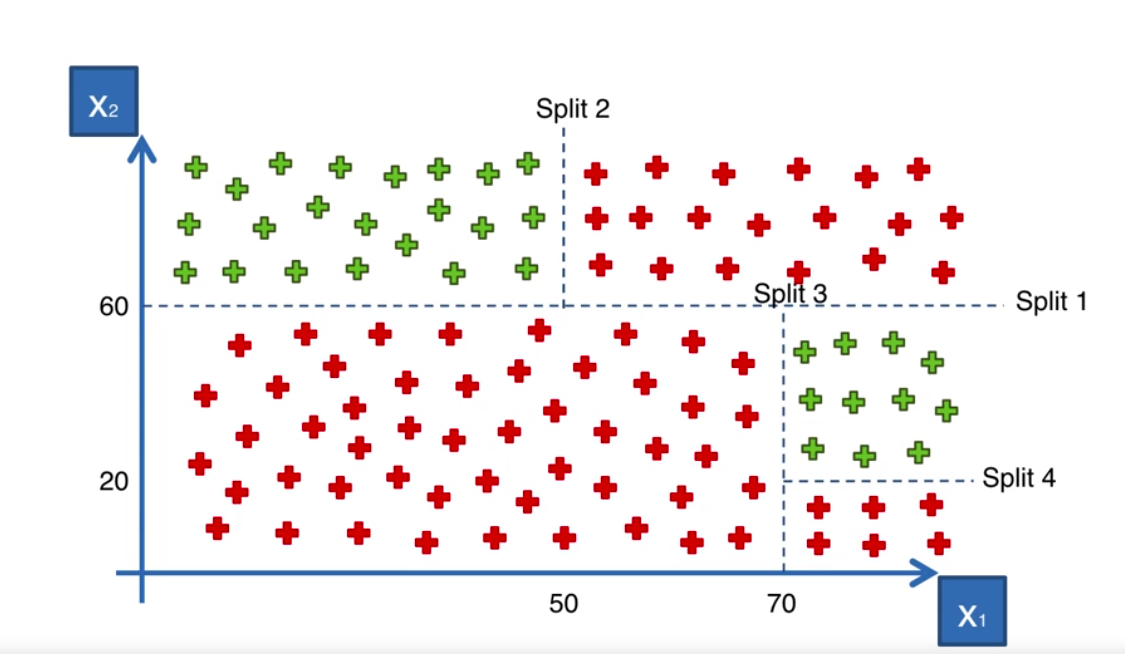
A,B均为数据的特征。

朴素贝叶斯定理假设数据分类仅有两种特征且互相独立，这并不是通常的情况。所以称为朴素贝叶斯。

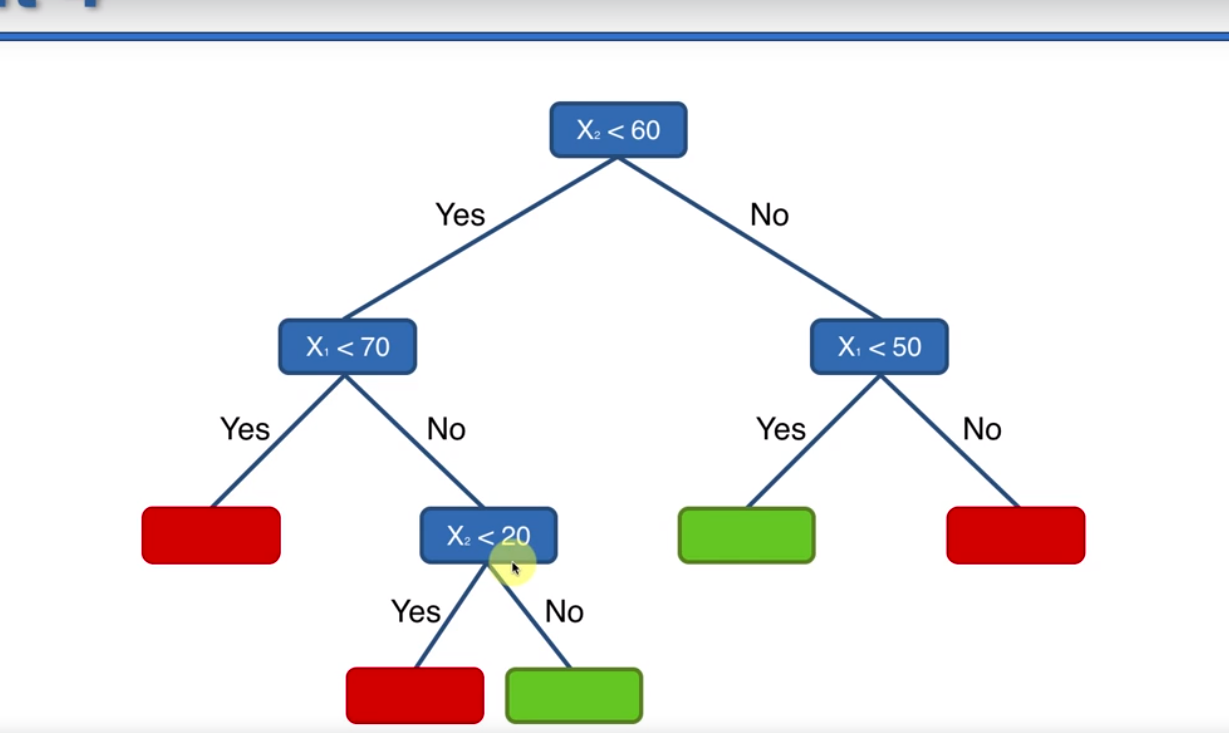
如果有第三种特征则需要三个特征都相互独立，计算三个后验概率来进行比较，归类进入最大的后验概率的那一项。

**决策树**

使用水平或垂直的直线反复的切割数据集。



具体思维流程如下



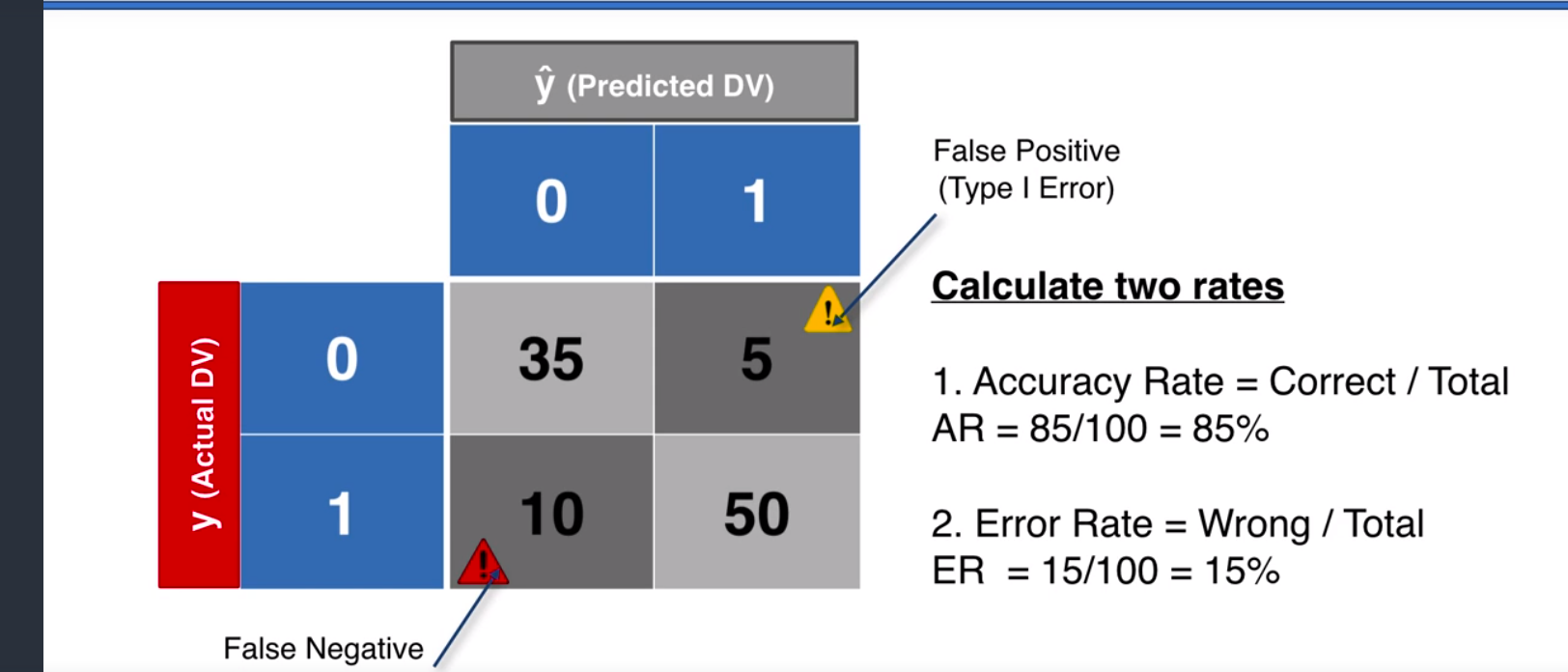
需要注意的是，上图中X2<20的部分应当被去除以免针对训练集的过度拟合。这是对决策树的修剪。

**随机森林**

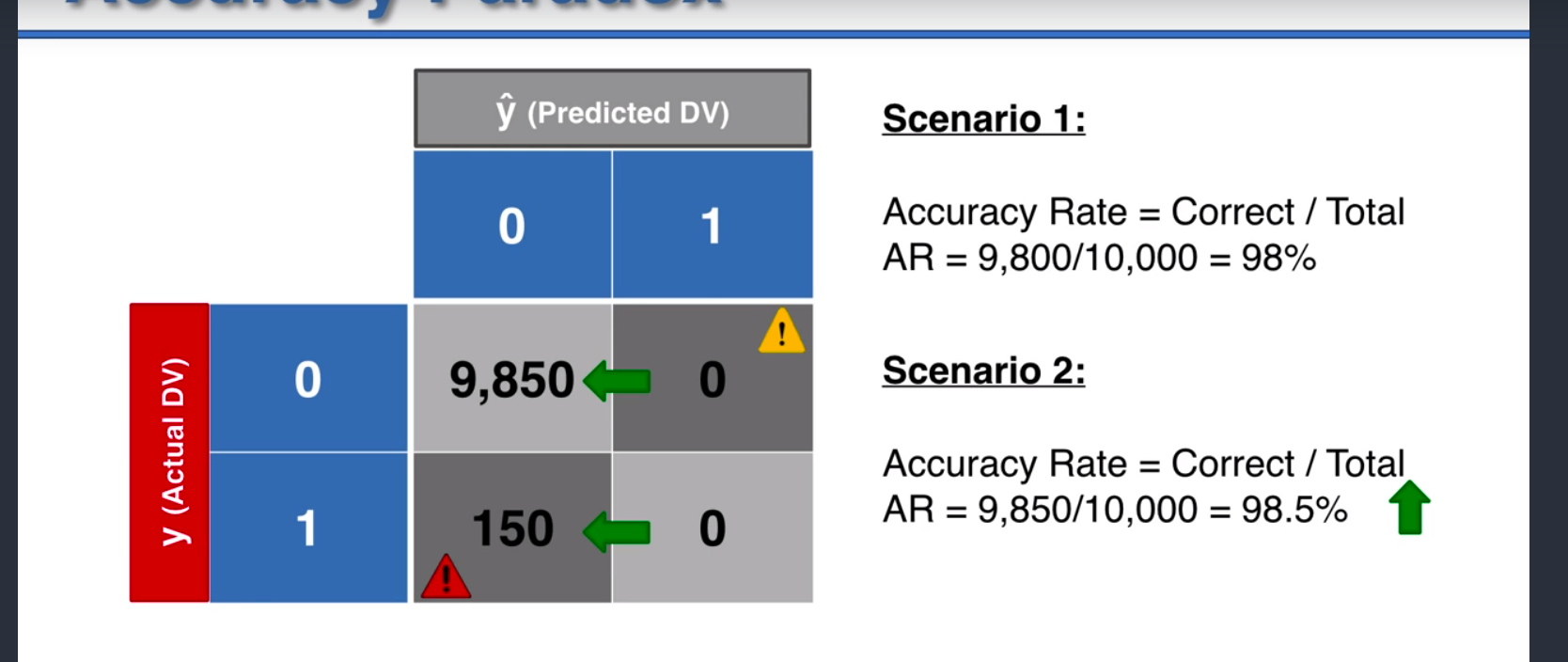
使用使用多个决策树集合起来进行投票决定最终结果。

**分类器的效果评估指标**

1. False Positive:又称伪阳性，Type I error，将实际生活中没有发生的事情判断为发生。
2. False Negative:又称伪阴性，Type II error，将实际生活中发生的事情判断为没有发生。
3. 混淆矩阵：

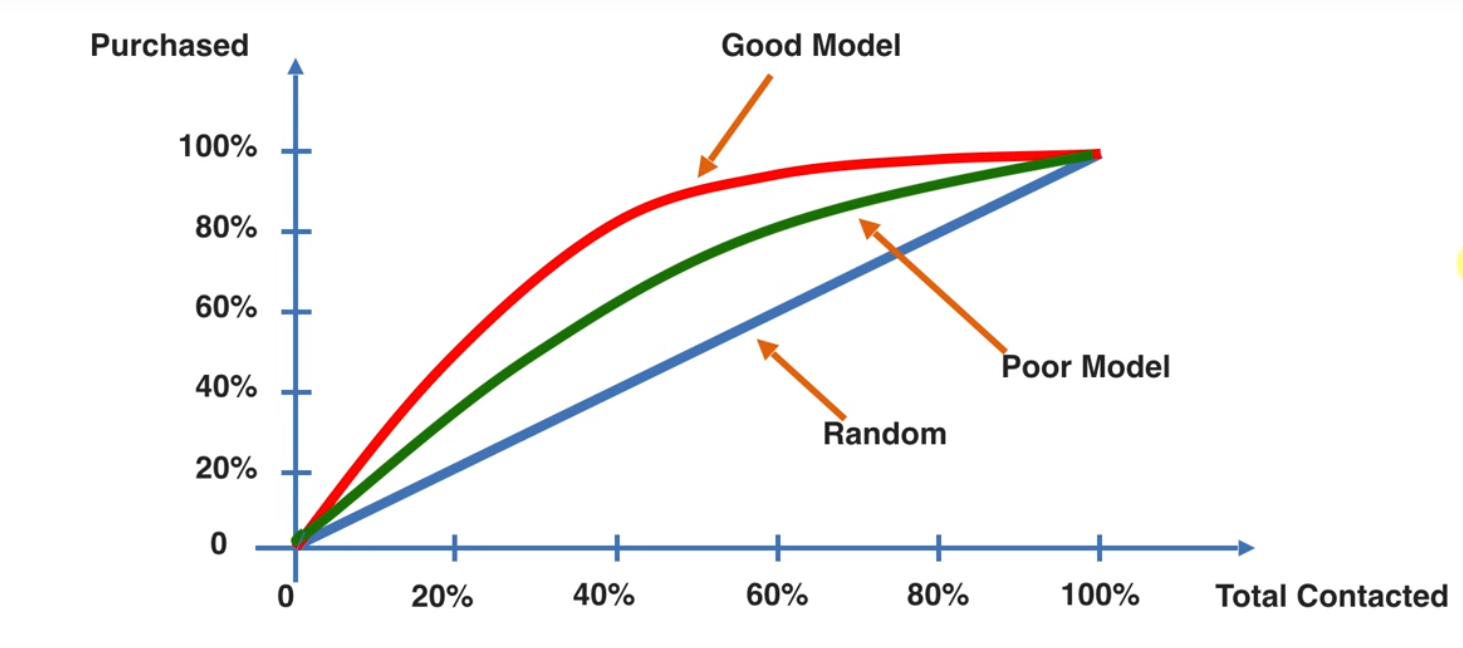


混淆矩阵可能会陷入准确率悖论（Accuracy paradox）在某一个单元出现极端个数的时候会失效，如下图，为此我们引入累计准确曲线。

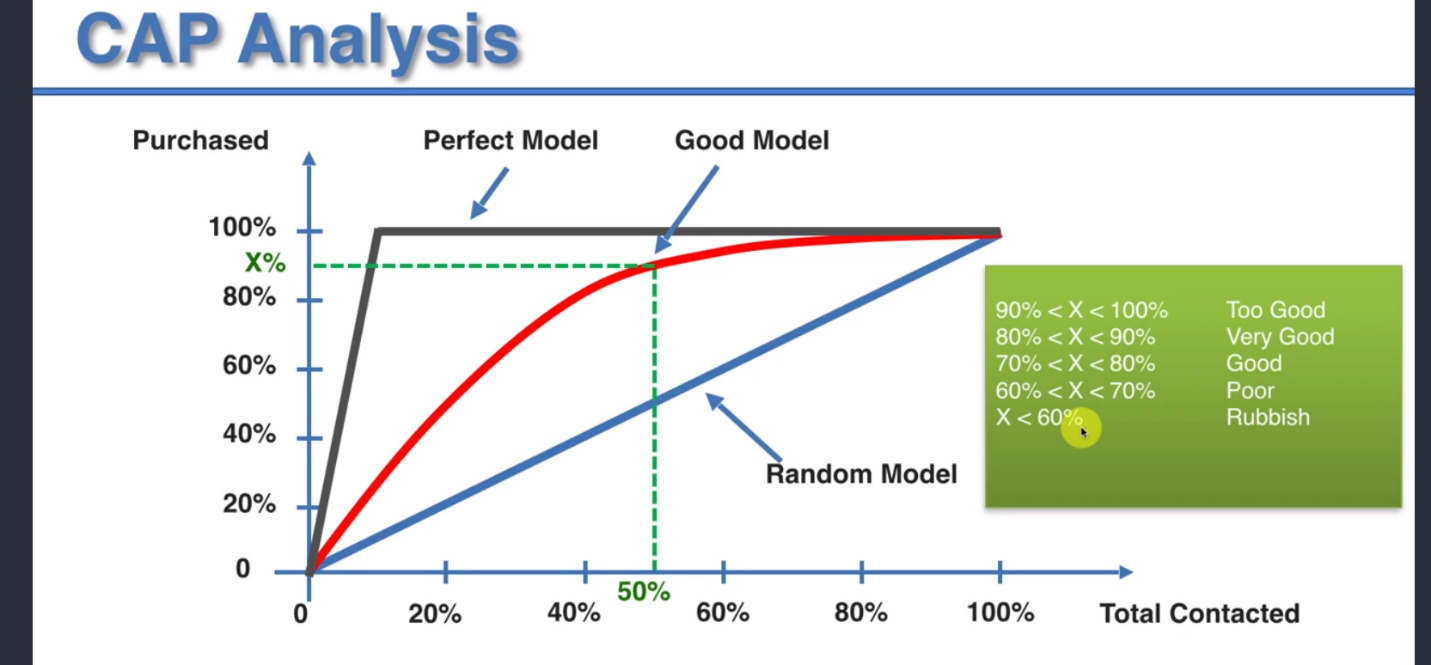


1. 累计准确曲线：

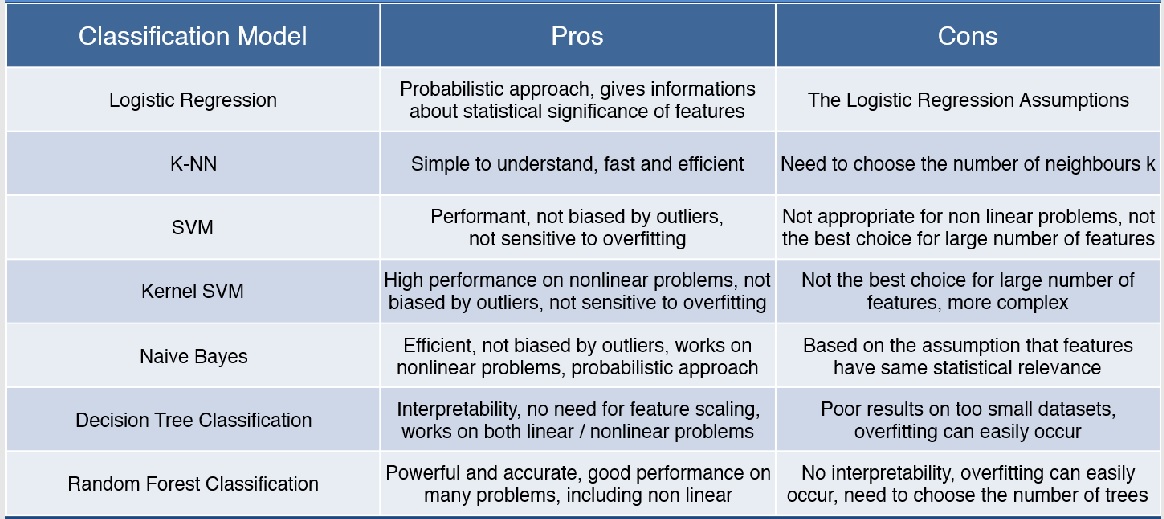
这个曲线曲度越大表示模型越准确，直线为随机预测结果。所有机器学习模型都应该好于直线结果。

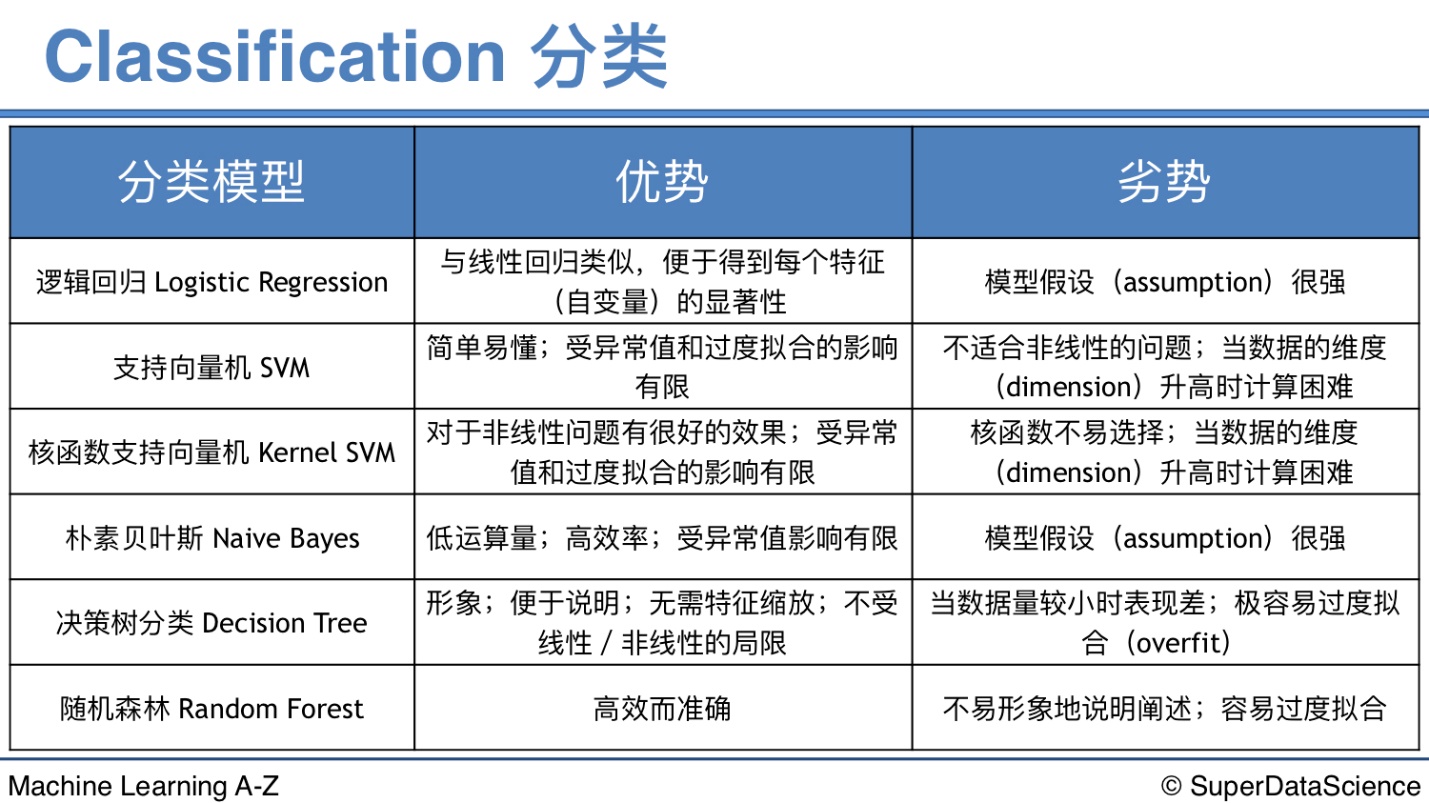


可以利用X值为50%处的Y值分布来量化的分析CAP。我们应该避免90%以上的过度拟合或者隐藏的自变量和因变量之间的关系。



1. 目前的模型优劣势汇总





首先，您需要判断您面对的是一个线性还是非线性的问题。您可以在第10部分- 模型选择中得到更多的答案。接下来：

* 假如是线性的问题，您应该选择逻辑回归（logistic regression）或者支持向量机SVM。
* 假如是非线性的问题，您应该选择朴素贝叶斯（naive bayes），决策树（decision tree）或者是随机森林（random forest）。在接下来的课程中我们会讲到神经网络（neural network），也是一个十分强大的方法。

然后，在每种情况下，应该如何选择分类模型呢？您会在第10部分的K次交叉验证（k-fold cross validation）中学到。

现在，从实际操作的角度来说，您可以大致遵循以下的规则：

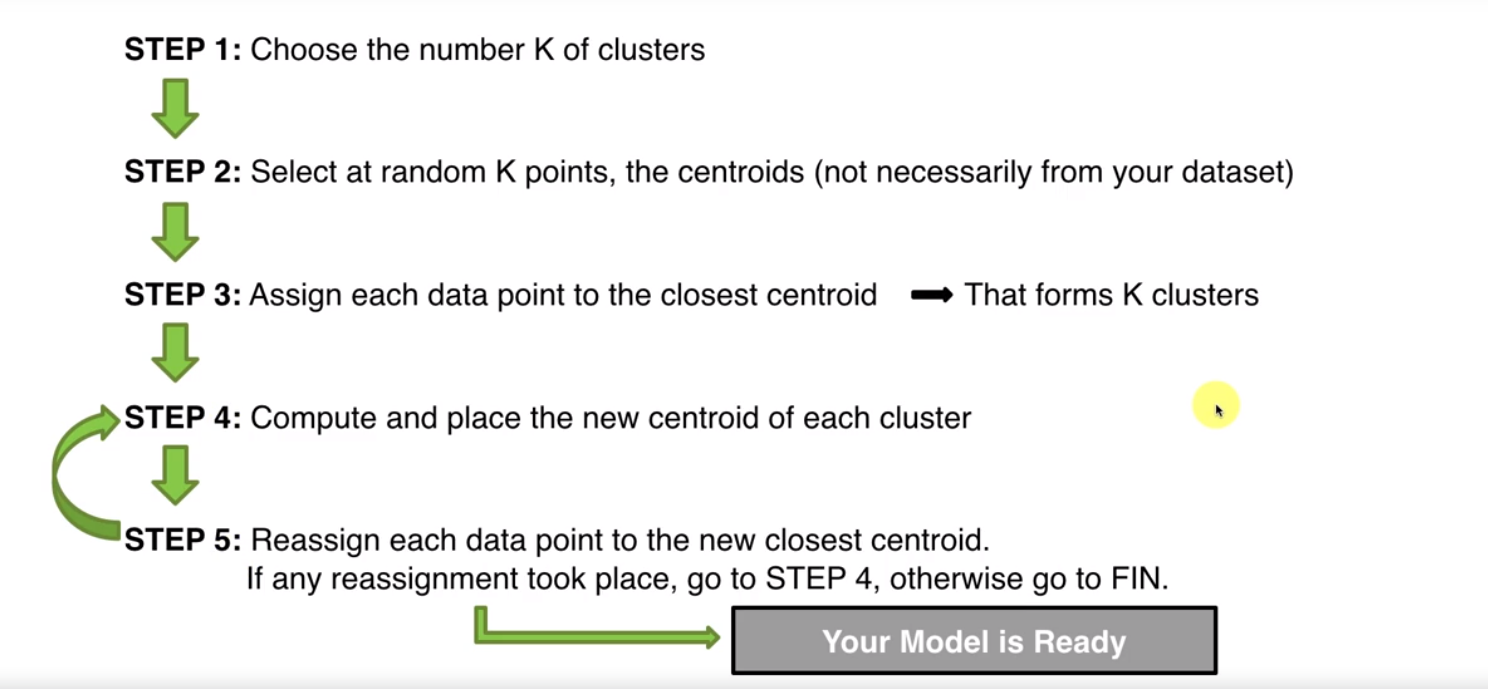
* 假如您想要给最终预测概率进行排序，您应该选择逻辑回归（logistic regression）或是朴素贝叶斯（Naive Bayes）。举个例子：您想要预测不同客户购买某项产品的概率，并将这些概率从大到小进行排序，以便锁定目标客户群。在这样的情形下，如果您的问题是线性的，您应该运用逻辑回归（logistic regression）；假如您的问题是非线性的，您应该选择朴素贝叶斯（naive bayes）模型。
* 假如您想要预测每一个客户属于哪一个划分（segment），您应该选择SVM。市场和客户群体的划分可以是已完成的市场调研或者集群分析（clustering）的结果。
* 假如您想要非常直观地展示／阐述模型，那么决策树（Decision Tree）是最佳选择。
* 假如您想要最好的模型的分类表现，并且不太在意模型的展示／阐述，那么随机森林（random forest）是不错的选择。

**集群（Clustering）**

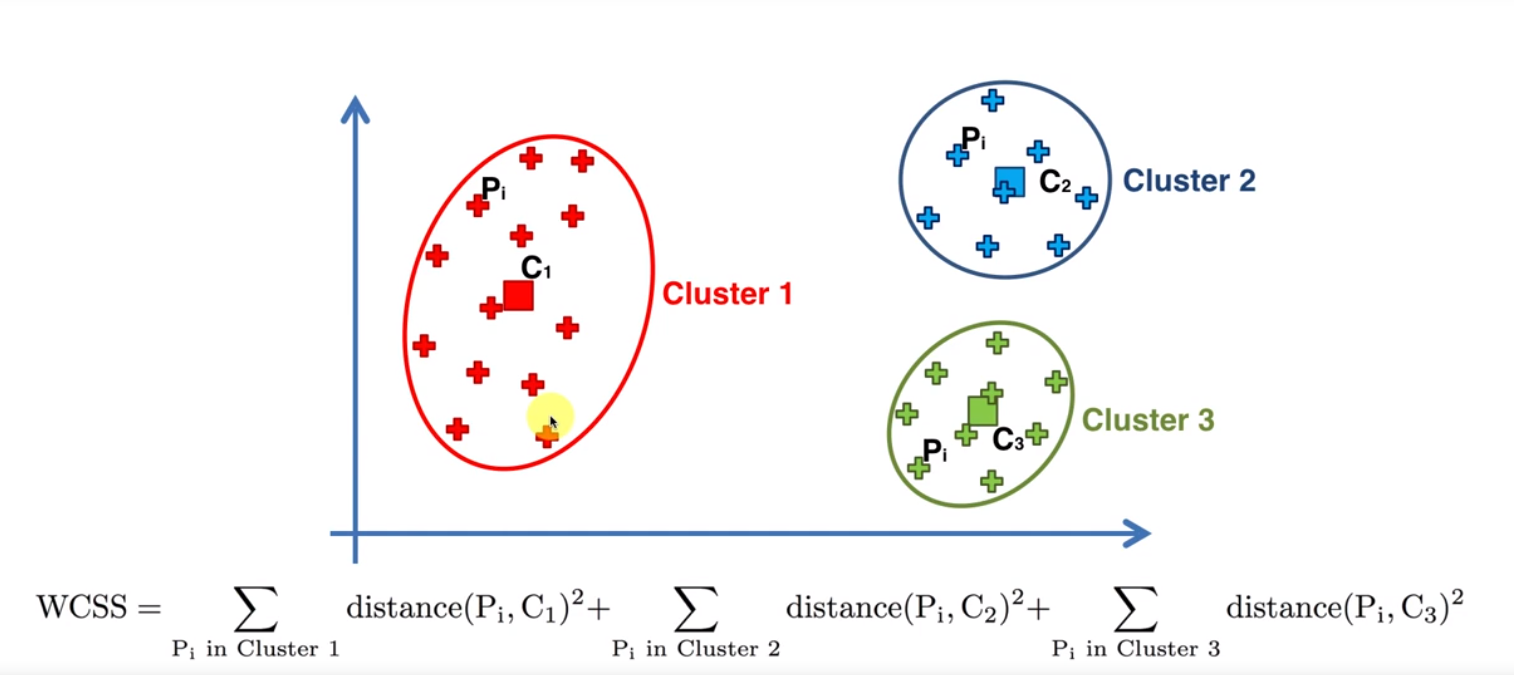
K means

该算法的初始中心点不可以随意选择，会陷入局部最小值的陷阱。可采用K-means ++来规避这个错误。

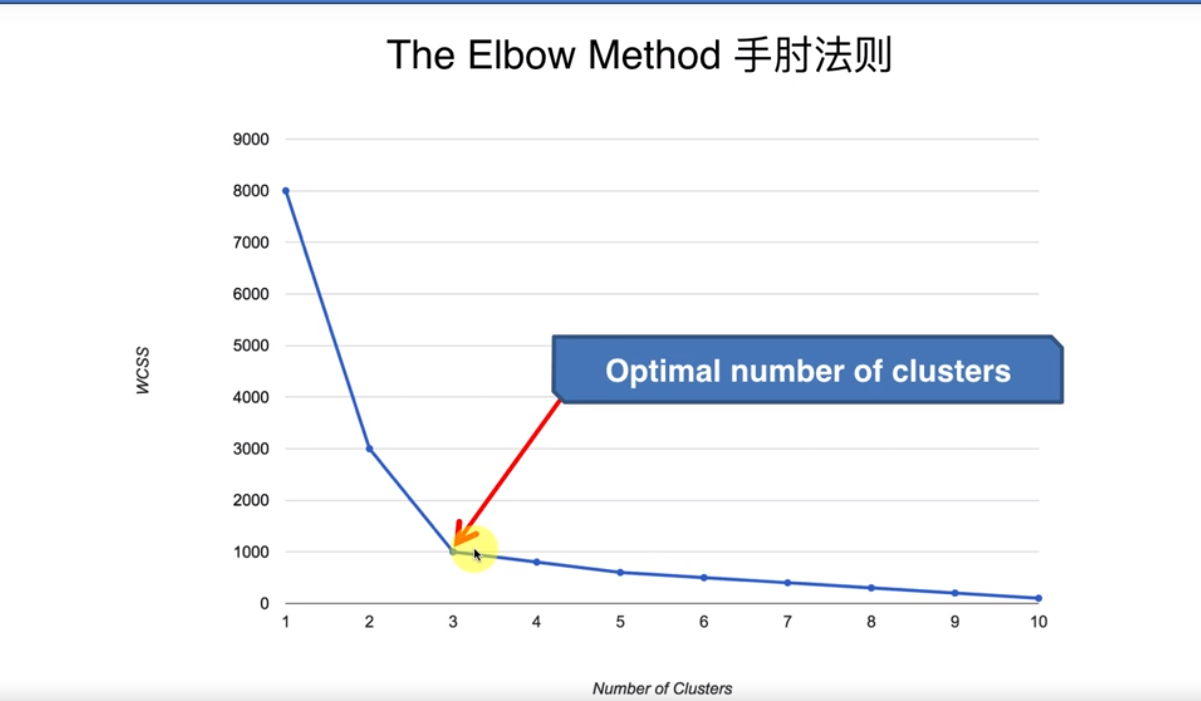
具体步骤如下



SETP 1中的选取K clusters可以使用组内平方和来决定。具体示例如下图。



对于上图中的公式，我们可以发现分越多的组，则组内平方和的值就会越小，而我们不期望进行太多的分组，所以可以根据手肘法则选择合适的分组个数，参考下图。

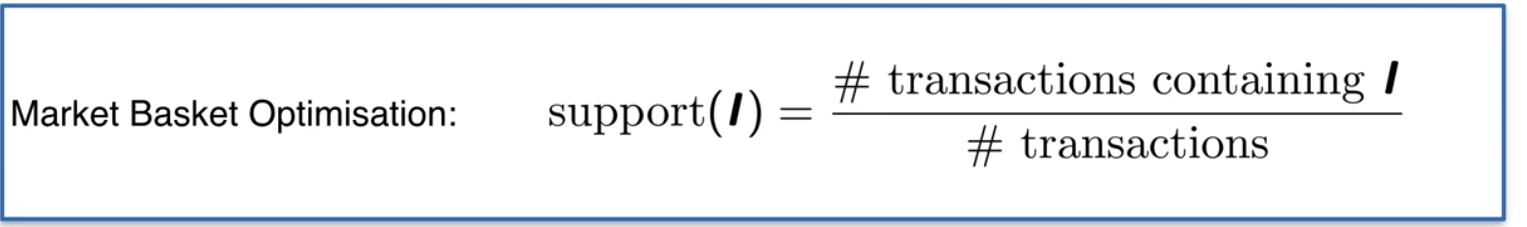


**关联规则学习——先验算法（apyori）**

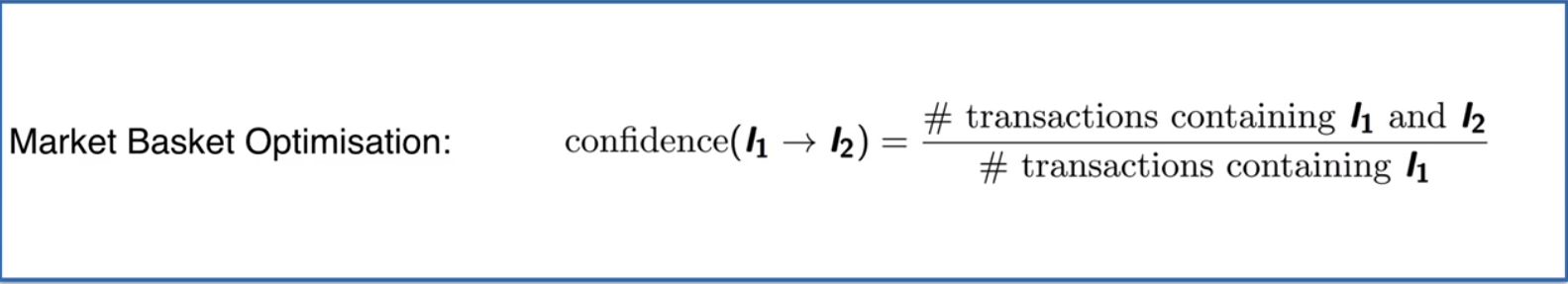
为了分辨数据集中的一个数据内部关联强度而发明的算法。比如对于顾客的一次购买商品之间的联系规律进行分析。

算法主要依靠三个函数实现。Support, Confidence, lift。

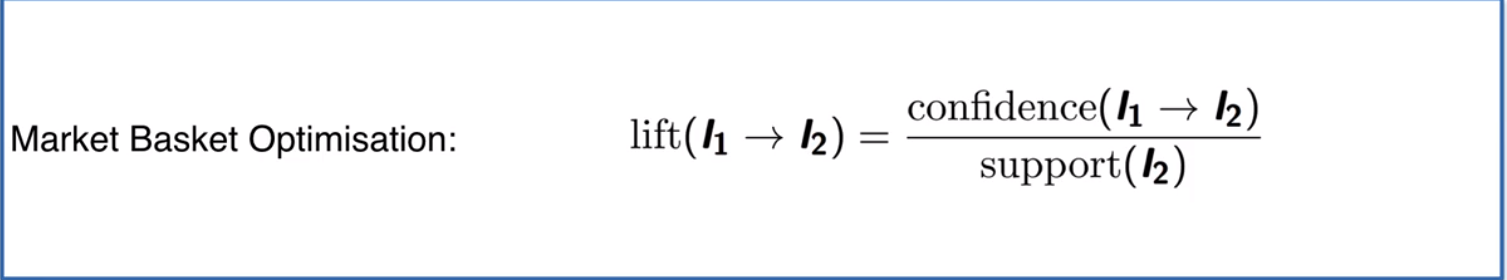
以商品销售为例，support表明了在所有销售记录中包含销售L的数据



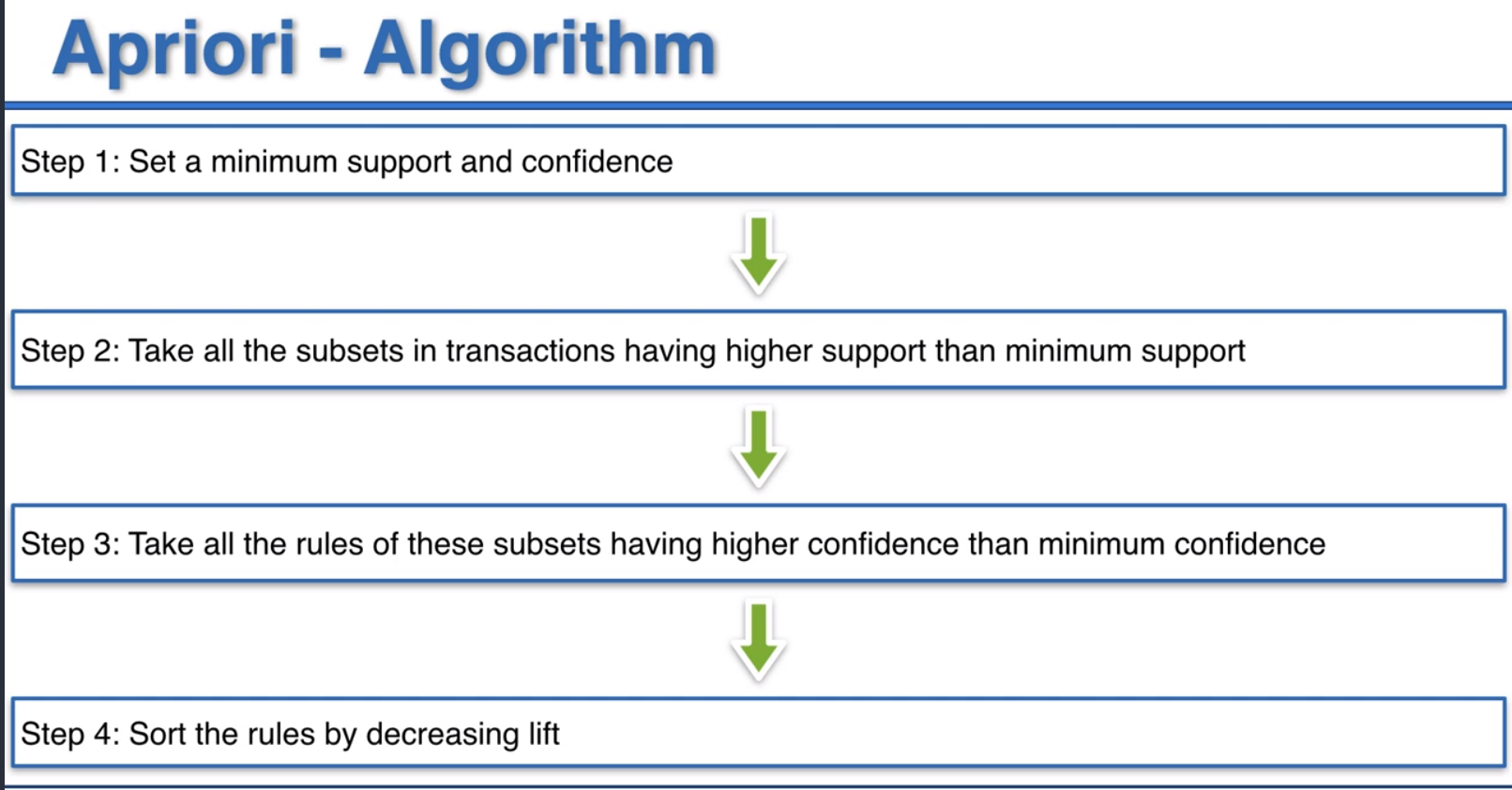
Confidence表明了销售中同时包含了L1和L2的除以L1的销售记录，肯定是一个<1的比值。



提升度表明了购买L1对于购买L2可能性的提升程度。



具体流程如下：先手动的设置合理的最小support和confidence值，后进行分析，可反复调整来获得有意义的结果。



**强化学习**

强化学习是尽可能的利用每一次对问题探索的结果，来更快地接近目标。

以多臂老虎机问题为例，对于很多台摆在面前的老虎机，假设有一台机器的回报率最佳，赌徒们想要找出这个最好的老虎机赌博。

如果针对每一个老虎机都进行100次尝试，最后统计每一老虎机的回报概率，对然可以但是在尝试过程中浪费了很多资源。因为针对除了最佳结果以外的机器每一次尝试事实上都会产生一个遗憾，也就是资源的浪费。

强化学习则是在每一次尝试之后都根据结果对下一轮尝试的方向进行修订，这样可以节省资源，减少遗憾。

**UCB算法（置信区间上界）**