P5：分类问题

混淆矩阵(Confusion Matrix)

数据挖掘：数据必须有规律可挖，像彩票中奖号码是随机生成，无法挖掘。

两条曲线有一定的相关性不能说明有因果关系。

观看数据关系时，必须考虑时间维度。

# 数据预处理(缺失、不完整、冗余)

## 缺失数据

原因：未提供、不适用、设备故障

**对策：忽略：缺失数据较少时可忽略**

**手动填充数据：利用领域知识推测/猜**

**自动填充数据：填充固定值、样本均值等**

**离群点：**最小二乘法时离群点对曲线影响很大/聚类算法时离群点是孤立的点

**异常点**：

距离：计算离群点应该计算相对距离进行比较；计算点B与其K紧邻点的平均距离的d1、B的近邻点的K近邻平均距离d2，若d1>>d2，说明B可能为离散点

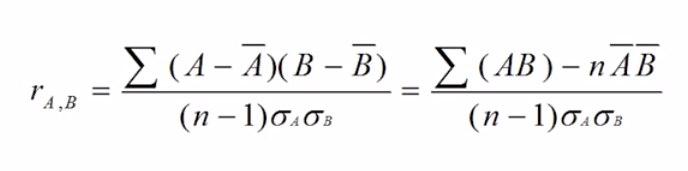
## 2.4.1 数据描述与可视化

**数据标准化**：

1. min-max标准化：最简单的是将数据映射到[0,1]之间(前提：有明确的上下界)
2. z-score 标准化：利用高斯均值和标准差，计算其与均值偏差多少个标准差

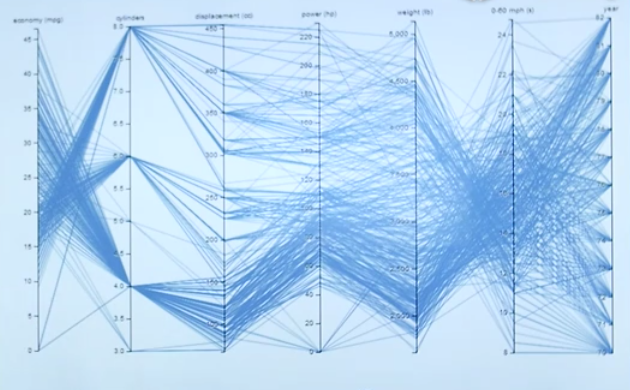
**数据描述**：均值、中位数、MODE(某一数组或数据区域中出现频率最多的数值)、方差（variance）

**两组数据相关性：**



1. 正相关 r(A,B) >0时
2. 负相关 r(A,B) <0时
3. 不线性相关 r(A,B) =0时

**多维数据可视化：**每条竖线代表一个维度，线与竖线的交点为在该维的取值。



可视化工具软件：CiteSpace--->使用可视化文本

Gephi

特征选择：

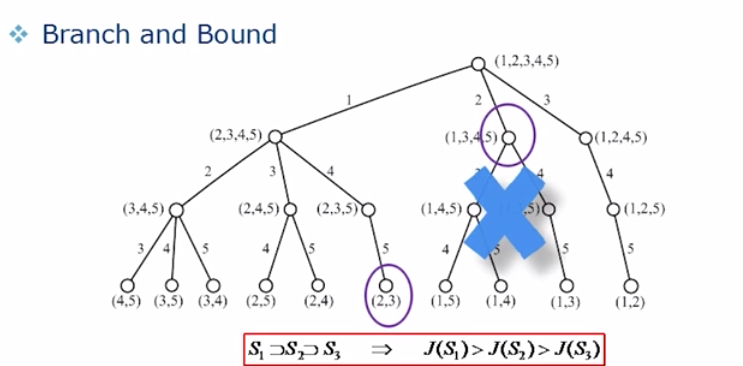
1. 熵(Entropy)：描述不确定性，当概率为0,5时，熵最大为1；当概率为0或1时，熵最小为0。
2. 信息增益：额外属性的价值（对不确定性的降值：原来的熵值-添加该属性后的熵值）

**解决属性选择问题**：从100个属性中选择最优的n个属性的一个组合

Branch and Bound(分支定界)：

**单调性假设**：**随着属性越来越少，属性组的效率（J值）越来越低**

如下图：(2,3)节点的效率大于(1,3,4,5)节点的效率，则(1,3,4,5)节点下面的树节点无需计算。

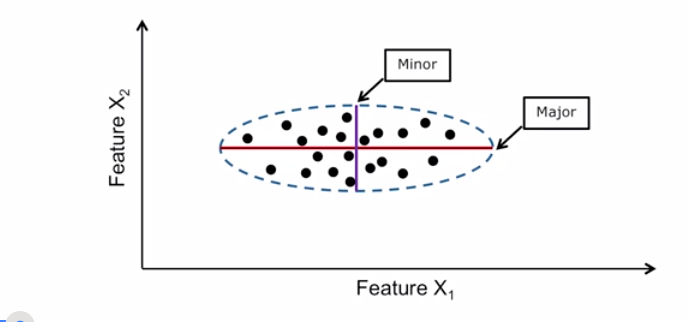


另外的方法：

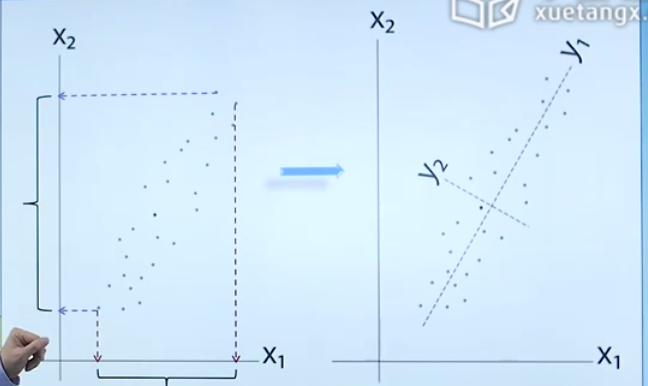
1. 遍历每个选择TOP N个
2. 在最优的n个属性基础上，遍历每个属性加入先前最优的n组成最优的n+1个属性。
3. 在最优的n+1个属性基础上，去掉每一个属性得到一个最优的n个属性组合。

### 主成分分析(PCA-图像处理中常用降维算法)

1. 方差(Variance):若沿着某个属性方差越大(分布的越开)，说明分布差别越大，该属性越重要(蕴含的信息量越大)，如下图：选择属性X1，在X1上投影成一维数据。



1. 2D-->变换坐标轴，将联系移除，如下图，可选择属性y1



**优化方法：拉格朗日乘数法**--->求带条件约束的式子最值