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

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