# 3.6类别不平衡问题

## 原理

如果不同类别的训练样例数目差别很大，如998个正例，2个反例，会对学习过程造成困扰，称为类别不平衡(class-imbalance)问题。

## 处理方法

### e.g logisitic regression

**① 当正反例可能性相同时：**

若$\frac{y}{1-y}>1$，则预测为正例，否则判断为反例。（$y>0.5$）

**② 假设训练集是样本总体的无偏采样（即总体样本是什么比例，采集过来就是什么比例）时，设正例数目$m^+$反例数目$m^-$：**

若$\frac{y}{1-y}>\frac{m^+}{m^-}$，则预测为正例。（$y<0.5$）

再缩放(rescaling):如果分类器默认基于①进行预测，只需将上式两端同乘$\frac{m^-}{m^+}$，得到$\frac{y}{1-y}*\frac{m^-}{m^+}>1$，即$\frac{y'}{1-y'}=\frac{y}{1-y}*\frac{m^-}{m^+}$

**③若上述假设均不成立，无法基于训练集类别数量推断真实几率：**

再缩放

1. 欠采样(undersampling)
   删除一些使得正反例数目接近，但可能会丢失一些重要信息。其中EasyEnsemble算法是将反例划分为若干个集合供不同的学习器使用，每个学习器进行了欠采样，全局来看是否会丢失重要性息。
2. 过采样(oversampling)
   增加一些样例，但不能简单地对初始样本进行重复采样，否则会招致严重过拟合。
3. 阈值移动(threshold-moving)
   原始数据集训练，嵌入$\frac{y'}{1-y'}=\frac{y}{1-y}*\frac{m^-}{m^+}$式中