数据不平衡，又称样本比例失衡。对于二分类问题，在正常情况下，正负样本的比例应该是较为接近的，很多现有的分类模型也正是基于这一假设。但是在某些特定的场景下，正负样本的比例却可能相差悬殊，如社交网络中的大V判断、电商领域的恶意差评检测、金融领域的欺诈用户判断、风控领域的异常行为检测、医疗领域的肿瘤诊断等。

根据**数据量的多少**和**数据不平衡程度**，可以将数据不平衡问题分为以几类：

（1）**大数据**+**轻微**数据不平衡（正负样本数量相差在一个数量级内）

如**豆瓣上的电影评分数据分布**等，此类问题常见的机器学习模型即可解决，无须额外处理。

（2） **小数据** + **轻微数据不平衡**（正负样本数量相差在一个数量级内）

此类问题可以通过加随机扰动的上采样方法处理。

（3） **大数据** + **中度数据不平衡**（正负样本数量相差在两个个数量级内）

此类问题可以通过**下采样**的方法处理。

（4） **小数据** + **中微数据不平衡**（正负样本数量相差在两个数量级内）

如**医学图像分析**。此类问题可以用**数据合成**的方法来解决。

（5） **大数据** + **重度数据不平衡**（正负样本数量相差超过两个数量级）

如**金融领域的欺诈用户判断**,**风控领域的异常行为检测**。此类问题可以通过**一分类（one-class Learning）**或**离群点异常检测**方法来解决。

（6） **小数据** + **重度度数据不平衡**（正负样本数量相差超过两个数量级）

如**医疗领域的肿瘤诊断**。此类问题可以通过**一分类（one-class Learning）**或**离群点异常检测**方法来解决。

Learning from class-imbalanced data-Review of methods and applications

重采样方法用来对非平衡数据集在数据空间再平衡，这样可以消除在学习过程中倾斜的类别分布所造成的影响。重采样方法非常通用，因为它不依赖于分类器。重采样