3.不均衡学习

1.不均衡学习原理

不均衡分类，指的是样本不同类别的数量差异越来越大的情况下，模型越来越偏向于预测大类样本的现象，因此，模型分类性能越来越差。

单纯从样本不均衡的角度出发（不考虑分布变化，小样本学习，分类问题的困难程度等其它问题），不均衡的类别对模型造成影响的原因：

1.目标函数优化的方法，使用梯度下降法优化目标函数的模型对于不均衡问题更敏感；而tree模型纯粹基于贪心策略进行分裂的方法则对此并不敏感；

2.目标函数的使用，hinge loss和交叉熵对于不均衡的敏感度不同；

2.不均衡学习的处理方法?

1.改变损失函数（代价敏感性学习）

2.生成样本

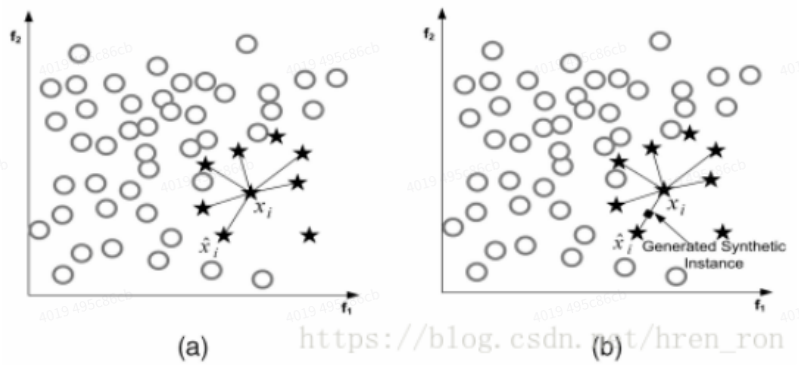
2.1smote算法的原理?

SMOTE（The Synthetic Minority Oversampling Technique）

SMOTE是一种**合成采样**的一种解决不平衡学习的方法，它已经被证明在很多领域都比较有效。它主要是基于现存的少数类样本，计算样本特征空间之间的相似度，然后创建人工合成样本。

1. 对于少数类$S_{min} \in S$中的样本，即，$x_i \in S_{min}$，计算它的K个近邻；
2. 通过计算 n 维空间的欧式距离，得到距离x_i最近的 K 个S_{min}中的样本数据；
3. 然后从 K 个近邻中，随机选择一个样本，产生人工合成的数据；具体的方法如下：

$x_{new} = x_i + (\hat{x}_i - x_i) \times \delta$，$x_i \in S_{min}$是一个少数类的样本，$\hat{x}_i$是$x_i$的其中一个近邻，$\delta \in [0, 1]$是一个随机数。



上图展示了SMOTE的具体过程。（a）图展示了一个典型的不平衡的数据，SMOTE中的 K 取值为6。（b）图中展示了一个随机产生的合成样本，这个样本是沿着和的直线产生的。

2.2为什么平常很少使用smote这类基于样本生成的方法?

SMOTE方法是一种过采样的方法，它克服了过采样的一些缺点，而且加强了原始数据。但是，**SMOTE方法可能会造成一定的过拟合**。

2.3过采样（上采样）和生成样本的区别?

上采样不一定是生成样本，例如简单的repeat式的上采样，通过repeat不涉及样本生成的过程，但生成样本一定是一种上采样；

3. 降采样平衡后的AUC值和预测概率值有怎样的变化?

roc曲线对类别数量的变化不敏感，因此auc的计算结果整体不会发生明显变化；

通过下采样平衡后，变相增大了正样本数量，分类决策边界远离正样本，预测概率整体变大；

4.class_weight的思想是什么?

class_weight对应的简单加权法是代价敏感学习最简单的一种方法，思想就是**小类样本加权，使其在loss中比重变大；**

而且有很多研究表明，代价敏感学习和样本不平衡问题有很强的联系，并且使用代价敏感学习的方法解决不平衡学习问题要优于使用随机采样的方法。