8.提升方法

提升方法得前提是，有学者在概率近似正确（probably approximately correct，PAC）学习框架下提出 强可学习（strongly learnable） 和 弱可学习（weakly learnable）是等价的，即是强可学习的充要条件是它是弱可学习的。

> 强可学习：在概率近似正确学习的框架下，一个任务可以用一个多项式学习并且最后准确率很高

相对应的：

> 弱可学习：在概率近似正确学习的框架下，一个任务可以用一个多项式学习，但是结果仅比随机略好

所以强可学习问题难找，但是弱可学习问题不难找，提升方法就是用多个 弱 提升（boost）到 强。

Adaboost 算法

大多数的提升方法都是改变训练数据的概率分布（训练数据的权值分布），针对不同训练数据分布学习一系列的弱分类器，所以这就**面临 2 个问题**：
（1） 如何改变数据的权值分布/概率分布
（2） 如何将弱分类器组合成一个强分类器

Adaboost 算法解决上述问题的方法：

问题1： 每一轮加大没有正确分类样本的权值，降低正确分类样本的权值
问题2： 加权多数表决，加大差错率小的权重，使其起较大的表决作用，减小差错率大的权重，使其起较小的表决作用。

对 Adaboost算法 的另一个解释就是，其**模型是加法模型、损失函数未指数函数、学习算法未前向分步算法**的二类学习方法。

> AdaBoost是AdaptiveBoost的缩写，表明该算法是具有适应性的提升算法。
>
> 算法的步骤如下：
>
> 1）给每个训练样本（ x1,x2,….,xN ）分配权重，初始权重 w1 均为1/N 2）针对带有权值的样本进行训练，得到模型 Gm （初始模型为G1）
> 3）计算模型 Gm 的误分率 e_m
> 4）计算模型 Gm 的系数 $a_m = 0.5\log[(1-e_m)/e_m]$
> 5）根据误分率e和当前权重向量 w_m 更新权重向量 w_{m+1}
> 6）计算组合模型 f(x) 的误分率
> 7）当组合模型的误分率或迭代次数低于一定阈值，停止迭代；否则，回到步骤2）

提升树

以决策树为基函数的提升方法，对分类问题是二叉分类树，对回归问题是二叉回归树。对于二分类问题，提升树是 Adaboost 算法的特殊情况。

梯度提升

梯度提升（gradient boosting）是解决损失函数除了 MSE 与指数损失函数外的损失函数的学习问题，因为损失函数为 MSE 或者指数损失函数时，每一步优化都很简单，但是对于一般函数就未必如此，所以梯度提升算法利用最速下降的近似方法。