链滴

熟悉算法 --bagging & boosting

作者：moloee

原文链接：https://ld246.com/article/1514475070410

来源网站：链滴

相同点：
<blockquote>
都是将已有的分类器或回归算法通过一定的方式组合起来，形成一个功能更强大的分类器。说白就是组装弱分类器，形成强分类器。
</blockquote>
下面先分别介绍
<h3 id="Bagging--Boostrap-aggregating-">Bagging (Boostrap aggregating)</h3>
套袋法，过程如下：

从原始的 N 个样本中，有放回的抽取 <code>K轮</code>，每轮抽取 n(n<=N)个样本（因有放回，所以可能重复），得到 k 个训练样本集
用 k 个训练集可以训练 k 个独立的分类器 (模型)
整体的分类结果由 k 个模型投票决定，回归问题取平均

<h3 id="Boosting">Boosting</h3>
主要思想：将弱分类器组合成强分类器，在 PAC（概率近似正确）的学习框架下 <code>[?!]</code>，一定可以将弱分类器组合成强分类器。
两个问题：

如何改变训练数据的权值或概率分布？
在每一轮训练中，提高被弱分类器误分类的样本的权值，降低正确分类的样本的权值，从而提高模型误分样本的识别效果？<code>这里的误分应该指的是实际上是 TP/TN，却被识别错了？所以要增强值，使其能够被正确分类？以提高后续分类过程中的识别效果。</code>

怎么组合弱分类器？
加法模型线性组合
AdaBoost 通过加权表决的方式，增大错误率小的分类器的权值，减小错误率大的分类器权值
提升树 通过拟合残差的方式逐步减小残差，将每一步生成的模型叠加得到最终模型

整体来说二者的区别如下：
<table>
<thead>
<tr>
<th align="left"></th>
<th align="left">Bagging</th>
<th align="left">Boosting</th>
</tr>
</thead>
<tbody>
<tr>
<td align="left">样本选择</td>
<td align="left">在原始集中有放回的选取训练集，训练集相互独立</td>
<td align="left">每一轮的训练集不变，只是训练集中每个样例在分类器中的权重发生变化。而权值根据上一轮的分类结果进行调整。</td>
</tr>
<tr>
<td align="left">样例权重</td>
<td align="left">均匀取样，每个样例的权重相等</td>
<td align="left">根据错误率不断调整样例的权值，错误率越大则权重越大</td>
</tr>
<tr>

<td align="left">预测函数</td>
<td align="left">所有预测函数的权重相等</td>
<td align="left">每个弱分类器都有相应的权重，对于分类误差小的分类器会有更大的权重</td>
</tr>
<tr>
<td align="left">并行计算</td>
<td align="left">各个预测函数可以并行生成</td>
<td align="left">各个预测函数只能顺序生成，因为后一个模型参数需要前一轮模型的结果</td>
</tr>
</tbody>
</table>
all:
整合的方式不一样，结果也不一样。总体来说会提高准确率，但是计算量会更大。
几个例子：

Bagging + 决策树 = 随机森林
AdaBoost + 决策树 = 提升树
Gradient Boosting + 决策树 = GBDT
