三、随机森林

随机森林也是为了解决决策树的过拟合问题。

3.1 Bootstrap

假设有一个大小为N的样本，我们希望从中得到m个大小为N的样本用来训练。bootstrap的思想是：首先，在N个样本里随机抽出一个样本x1，然后记下来，放回去，再抽出一个x2，...，这样重复N次，即可得到N个新样本，**这个新样本里可能有重复的**。重复m次，就得到了m个这样的样本。实际上就是一个有放回的随机抽样问题。每一个样本在每一次抽的时候有同样的概率（1/N）被抽中。

3.2 bagging策略

bagging的名称来源于： Bootstrap Aggregating，意为自助抽样集成。既然出现了Bootstrap那么肯定就会使用到Bootstrap方法，其基本策略是：

1. 利用Bootstrap得到m个样本大小为N的样本集。
2. 在所有属性上，对每一个样本集建立分类器。
3. 将数据放在这m个分类器上，最后根据m个分类器的投票结果，决定数据最终属于哪一类。如果是回归问题，就采用均值。

什么时候用bagging？当模型过于复杂容易产生过拟合时，才使用bagging，决策树就容易产生过拟合。

3.3 out of bag estimate（包外估计）

在使用bootstrap来生成样本集时，由于我们是有放回抽样，那么可能有些样本会被抽到多次，而有的样本一次也抽不到。我们来做个计算：假设有N个样本，每个样本被抽中的概率都是1/N，没被选中的概率就是1-1/N，重复N次都没被选中的概率就是(1−1/N)^N，当N趋于无穷时，这个概率就是1/e，大概为36.8%。也就是说样本足够多的时候，一个样本没被选上的概率有36.8%，那么这些没被选中的数据可以留作**验证集**。每一次利用Bootstrap生成样本集时，其验证集都是不同的。

以这些没被选中的样本作为验证集的方法称为包外估计。

3.4 样本随机与特征随机

在我们使用Bootstrap生成m个样本集时，每一个样本集的样本数目不一定要等于原始样本集的样本数目，比如我们可以生成一个含有0.75N个样本的样本集，此处0.75就称为采样率。

同样，我们在利用0.75N个样本生成决策树时，假设我们采用ID3算法，生成结点时以信息增益作为判断依据。我们的具体做法是把每一个特征都拿来试一试，最终信息增益最大的特征就是我们要选的特征。但是，我们在选择特征的过程中，也可以只选择一部分特征，比如20个里面我只选择16个特征。 那可能有的人就要问了，假设你没选的4个特征里面刚好有一个是最好的呢？这种情况是完全可能出现的，但是我们在下一次的分叉过程中，该特征是有可能被重新捡回来的，另外别的决策树当中也可能会出现那些在另一颗决策树中没有用到的特征。

随机森林的定义就出来了，**利用bagging策略生成一群决策树的过程中，如果我们又满足了样本随机和特征随机，那么构建好的这一批决策树，我们就称为随机森林(Random Forest)。**

实际上，我们也可以使用SVM，逻辑回归等作为分类器，这些分类器组成的总分类器，我们习惯上依旧称为**随机森林**。