16. 随机森林常见问题

1．什么是随机森林

随机森林（Random Forest）使用多个**CART决策树作为弱学习器**，不同决策树之间没有关联。当我们进行分类任务时，新的输入样本进入，就让森林中的每一棵决策树分别进行判断和分类，每个决策树会得到一个自己的分类结果，决策树的分类结果中哪一个分类最多，那么随机森林就会把这个结果当做最终的结果。随机森林算法是Bagging集成框架下的一种算法，它同时对训练数据和特征采用随机抽样的方法来构建更加多样化的模型。

2．随机森林为什么采用树结构

3．随机森林的随机性体现在哪里

- 随机采样：随机森林在计算每棵树时，从全部训练样本（样本数为N）中选取一个可能有重复的、大小同样为N的数据集进行训练（即Booststrap采样）。
- 特征选取的随机性：在节点分裂计算时，随机地选取所有特征的一个子集，用来计算最佳的分割方式。

4．随机森林需要归一化吗

随机森林是基于树的bagging算法，归一化对基于树的算法是没有提升效果的，不需要进行归一化处理。

5．为什么要有放回的抽样?

保证样本集间有重叠，若不放回，每个训练样本集及其**分布**都不一样，可能导致训练的各决策树差异性很大，最终多数表决无法 “求同”，即最终多数表决相当于“求同”过程。

6．为什么RF的训练效率优于bagging?

因为在个体决策树的构建过程中，Bagging使用的是“确定型”决策树，bagging在选择划分属性时要对每棵树是对所有特征进行考察；而随机森林仅仅考虑一个特征子集。

7．随机森林需要剪枝吗?

不需要，后剪枝是为了避免过拟合，随机森林随机选择变量与树的数量，已经避免了过拟合，没必要去剪枝了。一般rf要控制的是树的规模，而不是树的置信度，剩下的每棵树需要做的就是尽可能的在自己所对应的数据(特征)集情况下尽可能的做到最好的预测结果。剪枝的作用其实被集成方法消解了，所以用处不大。

8．随机森林如何处理缺失值?

根据随机森林创建和训练的特点，随机森林对缺失值的处理还是比较特殊的。

- 首先，给缺失值预设一些估计值，比如数值型特征，选择其余数据的中位数或众数作为当前的估计值。
- 然后，根据估计的数值，建立随机森林，把所有的数据放进随机森林里面跑一遍。记录每一组数据在决策树中一步一步分类的路径。
- 判断哪组数据和缺失数据路径最相似，引入一个相似度矩阵，来记录数据之间的相似度，比如有N组数据，相似度矩阵大小就是N*N。
- 如果缺失值是类别变量，通过权重投票得到新估计值，如果是数值型变量，通过加权平均得到新的估计值，如此迭代，直到得到稳定的估计值。

9．随机森林分类效果的影响因素

- 森林中每棵树的分类能力：每棵树的分类能力越强，整个森林的错误率越低。
- 森林中任意两棵树的相关性：相关性越大，错误率越大；

10．随机森林有什么优缺点

优点:

- 在当前的很多数据集上，相对其他算法有着很大的优势，表现良好。
- 它能够处理很高维度（feature很多）的数据，并且不用做特征选择(因为特征子集是随机选择的)。
- 在训练完后，它能够给出哪些feature比较重要。
- 训练速度快，容易做成并行化方法(训练时树与树之间是相互独立的)。
- 在训练过程中，能够检测到feature间的互相影响。
- 对于不平衡的数据集来说，它可以平衡误差，因为随机森林是基于CART树的。
- 如果有很大一部分的特征遗失，仍可以维持准确度。

缺点:

- 随机森林已经被证明在某些噪音较大的分类或回归问题上会过拟合。
- 对于有不同取值的属性的数据，取值划分较多的属性会对随机森林产生更大的影响，所以随机森林在这种数据上产出的属性权值是不可信的

11．什么是OOB？随机森林中OOB是如何计算的，它有什么优缺点?

假设有N个样本，每次对每个样本有放回的采样，采一个样本的概率为1/N，未被采的概率为1-1/N，N次抽样始终都未被采的概率为(1-1/N)^N，N趋向于无穷大时，结果可近似为1/e，即总有1/3的数据自始至终都未被使用过。

$$(1-\frac{1}{N})^N=\frac{1}{(\frac{N}{N-1})^N}=\frac{1}{(1+\frac{1}{N})^N}\approx\frac{1}{e}\approx 0.368$$

上面我们提到，构建随机森林的关键问题就是如何选择最优的m（特征子集），要解决这个问题主要依据计算袋外错误率oob error（out-of-bag error）。对于已经生成的随机森林,用袋外数据测试其性能,假设袋外数据总数为O,用这O个袋外数据作为输入,带进之前已经生成的随机森林分类器,分类器会给出O个数据相应的分类。因为这O条数据的类型是已知的,则用正确的分类与随机森林分类器的结果进行比较,统计随机森林分类器分类错误的数目,设为X,则袋外数据误差大小=X/O。

优点：这已经经过证明是无偏估计的,所以在随机森林算法中不需要再进行交叉验证或者单独的测试集来获取测试集误差的无偏估计。

12．随机森林的过拟合问题

要用交叉验证来调整树的数量

- https://blog.csdn.net/Roaddd/article/details/114093838?utm_medium=distribute.pc_aggpage_search_result.none-task-blog-2~aggregatepage~first_rank_ecpm_v1~rank_aggregation-4-114093838.pc_agg_rank_aggregation&utm_term=%E9%9A%8F%E6%9C%BA%E6%A3%AE%E6%9E%97+%E9%9D%A2%E8%AF%95%E9%A2%98&spm=1000.2123.3001.4430 <https://blog.csdn.net/Roaddd/article/details/114093838?utm_medium=distribute.pc_aggpage_search_result.none-task-blog-2~aggregatepage~first_rank_ecpm_v1~rank_aggregation-4-114093838.pc_agg_rank_aggregation&utm_term=%E9%9A%8F%E6%9C%BA%E6%A3%AE%E6%9E%97+%E9%9D%A2%E8%AF%95%E9%A2%98&spm=1000.2123.3001.4430>
- https://blog.csdn.net/jaffe507/article/details/105088940?utm_medium=distribute.pc_aggpage_search_result.none-task-blog-2~aggregatepage~first_rank_ecpm_v1~rank_aggregation-2-105088940.pc_agg_rank_aggregation&utm_term=%E9%9A%8F%E6%9C%BA%E6%A3%AE%E6%9E%97+%E9%9D%A2%E8%AF%95%E9%A2%98&spm=1000.2123.3001.4430 <https://blog.csdn.net/jaffe507/article/details/105088940?utm_medium=distribute.pc_aggpage_search_result.none-task-blog-2~aggregatepage~first_rank_ecpm_v1~rank_aggregation-2-105088940.pc_agg_rank_aggregation&utm_term=%E9%9A%8F%E6%9C%BA%E6%A3%AE%E6%9E%97+%E9%9D%A2%E8%AF%95%E9%A2%98&spm=1000.2123.3001.4430>
- https://zhuanlan.zhihu.com/p/385260652 <https://zhuanlan.zhihu.com/p/385260652>