1.随机森林如何处理缺失值?

方法1 暴力填补

python中的na,roughfix包提供简单的缺失值填补策略：对于训练集中处于同一个类别下的数据，如果是类别类别变量缺失，则用众数补全；如果是连续变量，则用中位数。

方法2 相似度矩阵填补

RF的python实现中，有一个rfImpute包，可以提供更加高层的缺失值填补。

1. 首先用暴力填补法进行粗粒度填充。
2. 然后使用上述填补后的训练集来训练随机森林模型，并统计相似度矩阵（proximity matrix)，然后再看之前缺失值的地方，如果是分类变量，则用没有缺失的观测实例的相似度中的权重进行投票；如果是连续性变量，则用相似度矩阵进行加权求平均值。
3. 上述投票方案迭代进行4-6次。

相似度矩阵解释：

相似度矩阵就是任意两个观测实例间的相似度矩阵，原理是如果两个观测实例落在同一棵树的相同节点次数越多，则这两个观测实例的相似度越高。

详细来说：

Proximity 用来衡量两个样本之间的相似性。原理就是如果两个样本落在树的同一个叶子节点的次数越多，则这两个样本的相似度越高。当一棵树生成后，让数据集通过这棵树，落在同一个叶子节点的”样本对(xi,xj)” proximity 值 P(i,j) 加 1 。所有的树生成之后，利用树的数量来归一化 proximity matrix。继而，我们得到缺失值所在样本的权重值，权重值相近的可以用于缺失值的填补参考。

2.随机森林如何评估特征重要性?

集成学习模型的一大特点是可以输出特征重要性，特征重要性能够在一定程度上辅助我们对特征进行筛选，从而使得模型的鲁棒性更好。

随机森林中进行特征重要性的评估思想为：

判断每个特征在随机森林中的每棵树上做了多大的贡献，然后取平均值，最后比一比特征之间的贡献大小。

有两种评价方法：

1.基于基尼指数

将变量重要性评分（variable importance measures)用VIM来表示，将Gini指数用GI来表示，假设有m个特征X1，X2，X3,...,Xc，现在要计算出每个特征Xj的Gini指数评分VIMj，亦即第j个特征在所有RF决策树中节点分裂不纯度的平均该变量。

Gini指数的计算公式为：

$$GI_m = \sum_{k=1}^{|K|} \sum_{k' \neq k} p_{mk} p_{mk'} = 1 - \sum_{k=1}^{|K|} p_{mk}^2$$

其中，K表示有K个类别，pmk表示节点m在类别k所占的比例。

特征Xj在节点m的重要性，即节点m分枝前后的Gini指数变化量为：

$$VIM_{jm}^{(Gini)} = GI_m - GI_l - GI_r$$

其中，GI_l和GI_r分别表示分枝后两个新节点的$Gini$指数。

如果特征Xj在决策树i中出现的节点在集合M中，那么Xj在第i棵树的重要性为：

$$VIM_{ij}^{(Gini)} = \sum_{m \in M} VIM_{jm}^{(Gini)}$$

假设RF中共有n棵树，那么

$$VIM_{j}^{(Gini)} = \sum_{i=1}^{n} VIM_{ij}^{(Gini)}$$

最后，把所有求得的重要性评分做一个归一化处理即可。

$$VIM_j = \frac{VIM_j}{\sum_{i=1}^{c} VIM_i}$$

2.基于袋外数据

对于一棵树 T_i，用OOB样本可以得到误差 e1，然后随机改变OOB中的第 j 列，保持其他列不变，对第 j 列进行随机的上下置换，得到误差 e2。至此，可以用 e1-e2 来刻画特征 j 的重要性。其依据就是，如果一个特征很重要，那么其变动后会非常影响测试误差，如果测试误差没有怎么改变，则说明特征j不重要。

而该方法中涉及到的对数据进行打乱的方法通常有两种：

1）是使用uniform或者gaussian抽取随机值替换原特征；

2）是通过permutation的方式将原来的所有N个样本的第 i 个特征值重新打乱分布（相当于重新洗牌）。

> 比较而言，第二种方法更加科学，保证了特征替代值与原特征的分布是近似的（只是重新洗牌而已）。这种方法叫做permutation test（随机排序测试），即在计算第 i 个特征的重要性的时候，将N个样本的第 i 个特征重新洗牌

3.什么是OOB?随机森林中OOB是如何计算的，它有什么优缺点?

bagging方法中Booststrap每次约有1/3的样本不会出现在Bootstrap所采集的样本集合中，当然也就没有参加决策树的建立，把这1/3的数据称为袋外数据oob(out of bag)，它可以用于取代测试集误差估计方法。

袋外数据(oob)误差的计算方法如下：

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0,用这0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0个数据相应的分类,因为这0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/0;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。