问题

随机森林算法思想，怎么增加随机性，如何评估特征重要性，为什么不容易过拟合

随机森林思想怎么添加的随机性

随机森林 (RF) 是 Bagging 的一个变体。RF在以决策树为基学习器构建 Bagging 集成的基础上，进一步在决策树的训练过程中引入随机性：

> 传统决策树在选择划分属性时，是在当前结点的属性集合（假定有 d 个属性）中选择一个最优属性；而在 RF 中，对基决策树的每一个结点，**先从结点的属性集合中随机选择一个包含 k 个属性的子集**，然后再从这个子集当中选择一个最优属性用于划分。

这里的参数 k 控制了随机性的引入程度。若令 $k = d$ ，则基决策树的构建与传统决策树相同，一般情况下，推荐值为 $k = log_2 d$ 。

为什么不容易过拟合

因为随机森林中每棵树的训练样本是随机的，每棵树中的每个结点的分裂属性也是随机选择的。这两个随机性的引入，使得随机森林不容易陷入过拟合。且树的数量越多，随机森林通常会收敛到更低的泛化误差。理论上当树的数目趋于无穷时，随机森林便不会出现过拟合，但是现实当做做不到训练无穷多棵树。

如何评估特征的重要性

这个问题是决策树的核心问题，而随机森林是以决策树为基学习器的，所以这里大概提提，详细的可以去看看决策树模型。

决策树中，根节点包含样本全集，其他非叶子结点包含的样本集合根据选择的属性被划分到子节点中，叶节点对应于分类结果。决策树的关键是在非叶子结点中怎么选择最优的属性特征以对该结点中的样本进行划分，方法主要有信息增益、增益率以及基尼系数 3 种，下面分别叙述。

信息增益 (ID3决策树中采用)

“信息熵”是度量样本集合纯度最常用的一种指标，假定当前样本结合 D 中第 k 类样本所占的比例为 $p_k (k = 1, 2, \dots, c)$ ，则 D 的信息熵定义为：

$$Ent(D) = -\sum_{k=1}^{c} p_k log_2 p_k \tag{1}$$

$Ent(D)$ 的值越小，则 D 的纯度越高。注意因为 $p_k \leq 1$ ，因此 $Ent(D)$ 也是一个大于等于 0 小于 1 的值。

假定离散属性 a 有 V 个可能的取值 $\{a^1, a^2, \dots, a^V\}$ ，若使用 a 来对样本集合 D 进行划分的话，则会产生 V 个分支结点，其中第 v 个分支结点包含了 D 中所有在属性 a 上取值为 a^v 的样本，记为 D^v 。同样可以根据上式计算出 D^v 的信息熵，再考虑到不同的分支结点所包含的样本数不同，给分支结点赋予权重 $\frac{|D^v|}{|D|}$ ，即样本数越多的分支结点的影响越大，于是可以计算出使用属性 a 对样本集 D 进行划分时所获得的“信息增益”：

$$Gain(D, a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) \tag{2}$$

一般而言，信息增益越大越好，因为其代表着选择该属性进行划分所带来的纯度提升，因此全部计算当前样本集合 D 中存在不同取值的那些属性的信息增益后，取信息增益最大的那个所对应的属性作为划分属性即可。

缺点：对可取值数目多的属性有所偏好

增益率 (C4.5决策树中采用)

从信息增益的表达式很容易看出，信息增益准则对可取值数目多的属性有所偏好，为减少这种偏好带来的影响，大佬们提出了增益率准则，定义如下：

$$Gain\_ratio(D,a) = \frac{Gain(D,a)}{IV(a)} \tag{3}$$
$$IV(a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} log_2 \frac{|D^v|}{|D|}$$

$IV(a)$ 称为属性 a 的“固有值”。属性 a 的可能取值数目越多，则 $IV(a)$ 的值通常会越大，因此一定程度上抵消了信息增益对可取值数目多的属性的偏好。

缺点：增益率对可取值数目少的属性有所偏好

> 因为增益率存在以上缺点，因此C4.5算法并不是直接选择增益率最大的候选划分属性，而是使用了一个启发式：**先从候选划分属性中找出信息增益高于平均水平的属性，再从中选择增益率最高的。**

基尼指数 (CART决策树中采用)

这里改用基尼值来度量数据集 D 的纯度，而不是上面的信息熵。基尼值定义如下：

$$Gini(D) = \sum_{k=1}^{c} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{c} p_k^2 = 1 - \sum_{k=1}^{c} (\frac{D^k}{D})^2 \tag{4}$$

直观来看，$Gini(D)$ 反映了从数据集 D 中随机抽取两个样本，其类别标记不一致的概率，因此 $Gini(D)$ 越小，则数据集 D 的纯度越高。

对于样本D，个数为|D|，根据特征A的某个值a，把D分成|D1|和|D2|，则在特征A的条件下，样本D的基尼系数表达式为：

$$Gini\_index(D,A) = \frac{|D^1|}{|D|} Gini(D^1) + \frac{|D^2|}{|D|} Gini(D^2) \tag{5}$$

于是，我们在候选属性集合A中，选择那个使得划分后基尼系数最小的属性作为最优划分属性即可。

参考资料

《机器学习》周志华
决策树算法原理 (CART决策树)