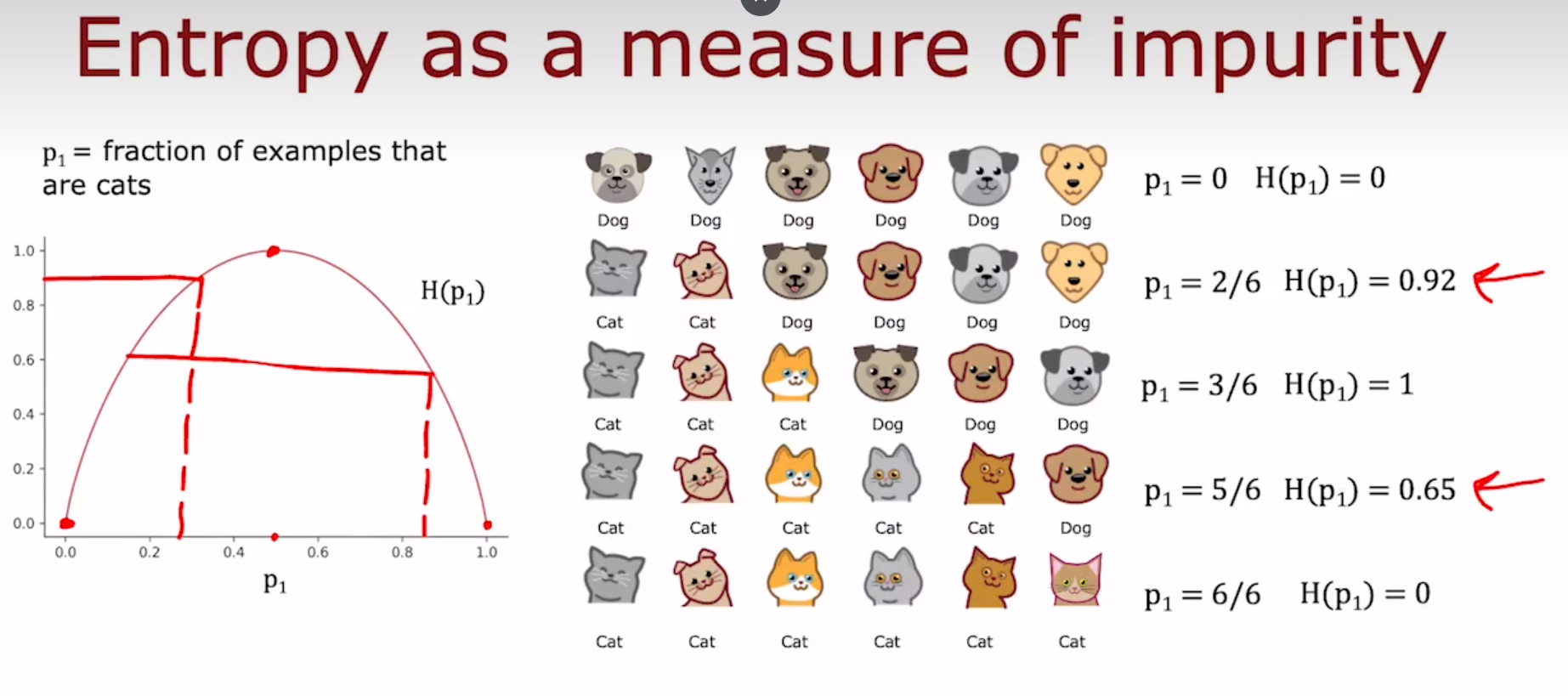
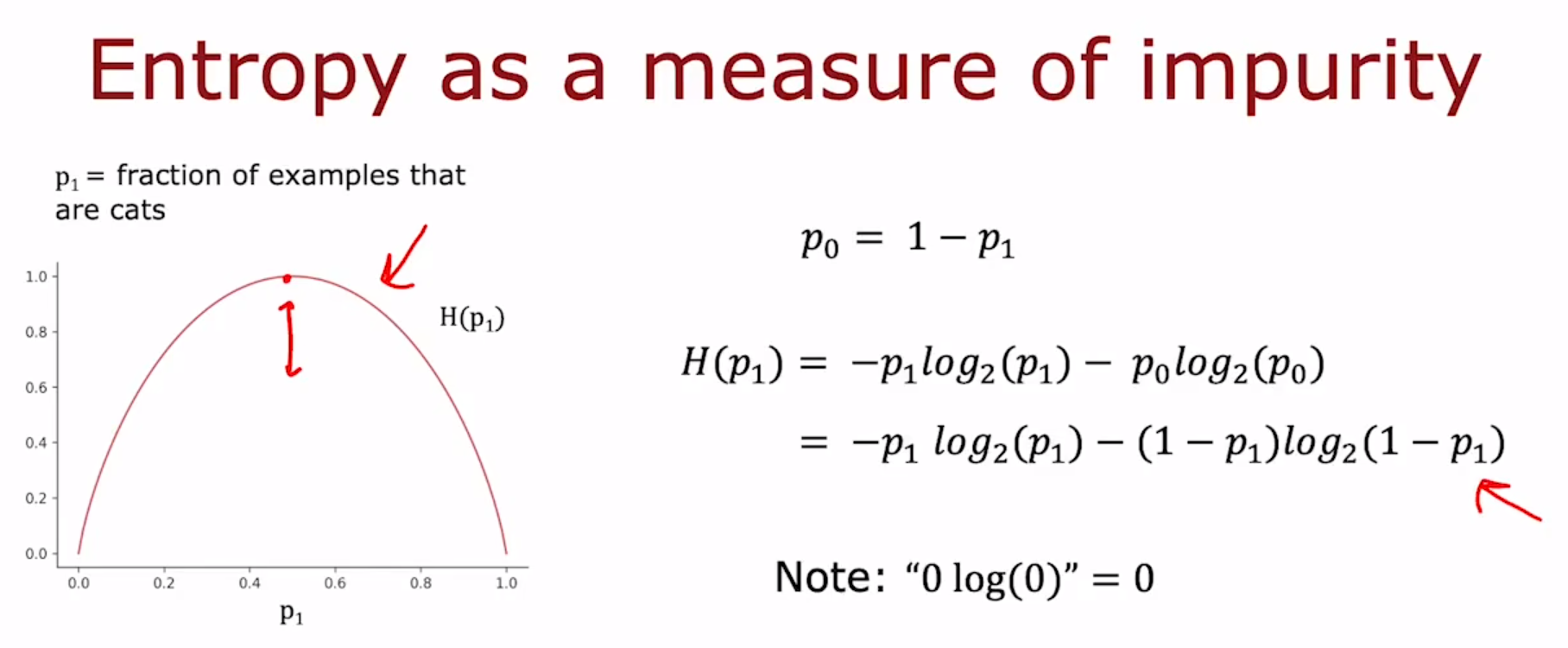
0决策树构建

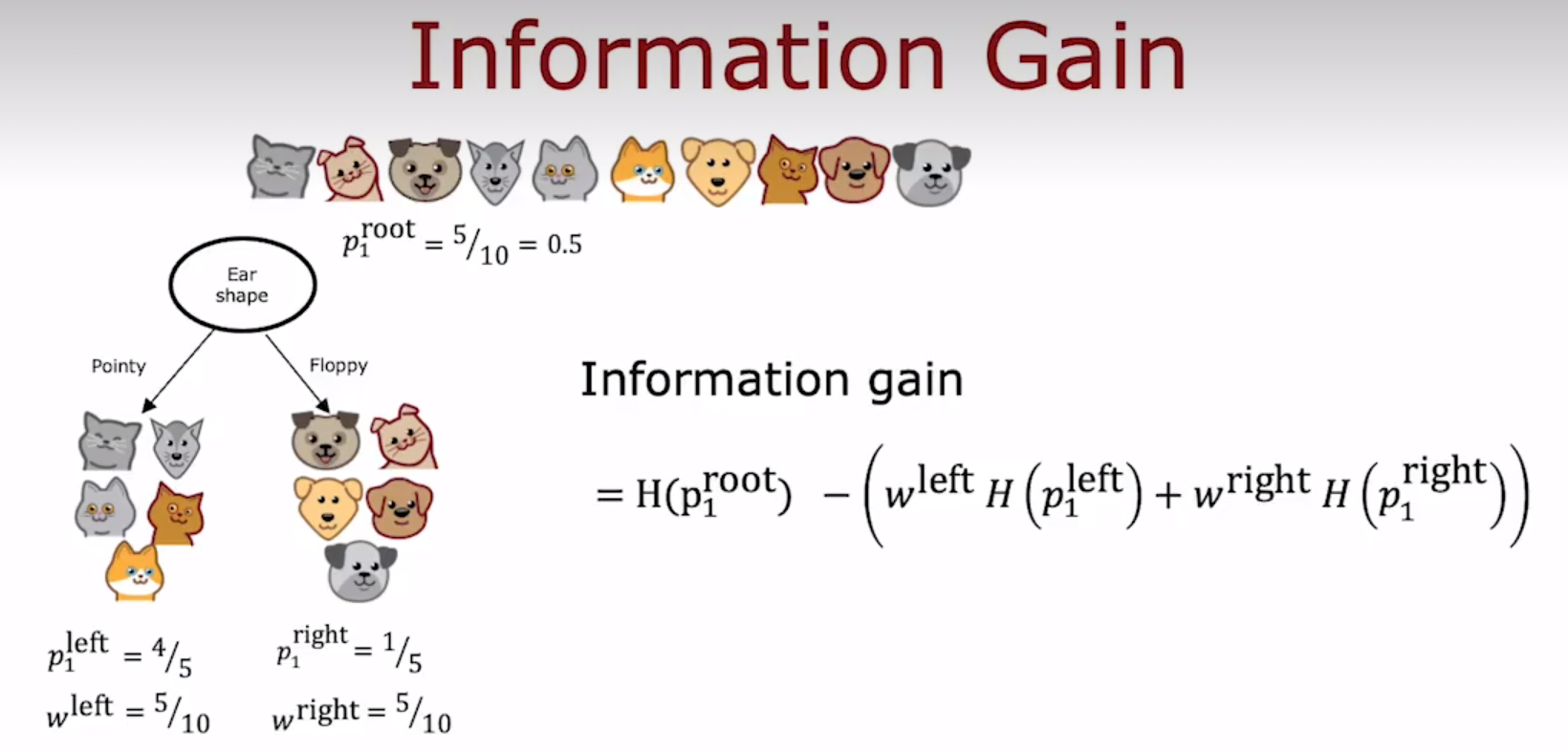
信息熵函数图像



信息熵计算公式

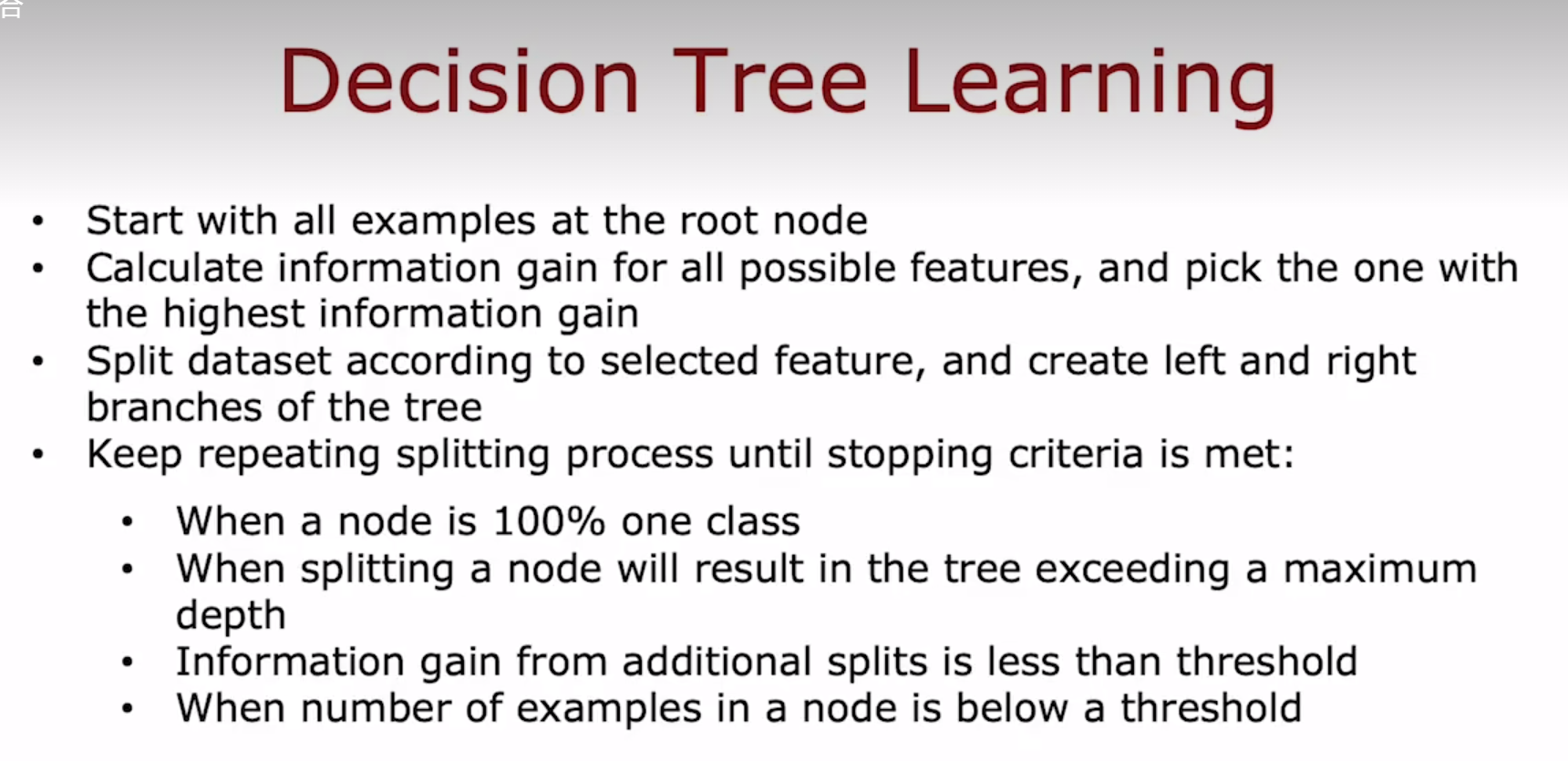


信息增益

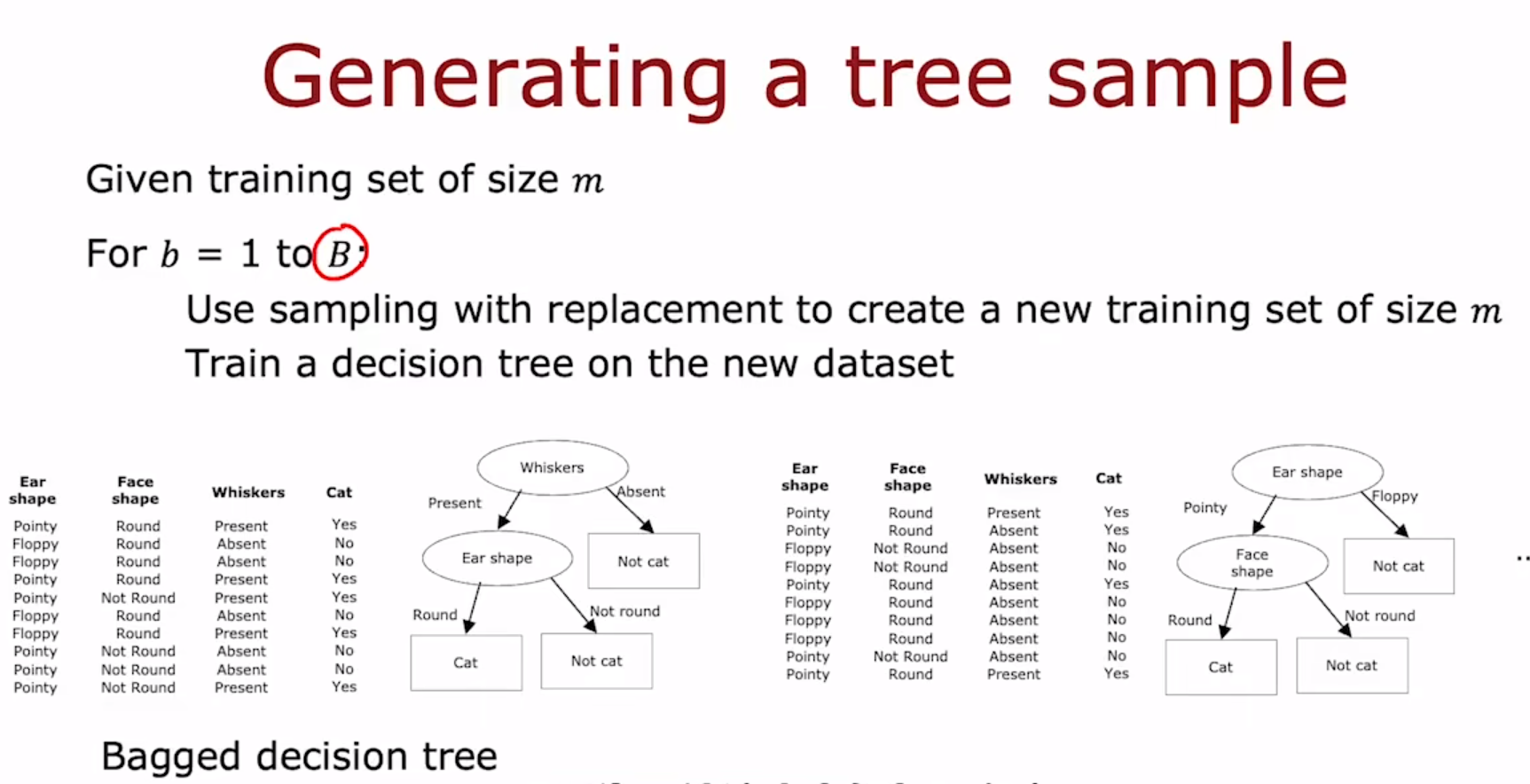


终止决策树的一个终止判断条件是如果当此决策树划分所获得的最大信息增益小于一个指定的数值，那么就决定不再进行决策树划分，防止在信息增益效果不大的情况下出现又过拟合情况，得不偿失。

构建决策树步骤



1袋装决策树算法



1.1 构建决策森林

将原所有的n个训练样本放入一个虚拟袋子，每次从袋子中有放回抽样摸出n个样本，很显然n个样本中可能有重复样本。对这n个样本构建一个决策树并将其保存到决策树集合中。重复上述过程B次获得含有B个决策树的决策森林(B一般情况下小于等于100)。

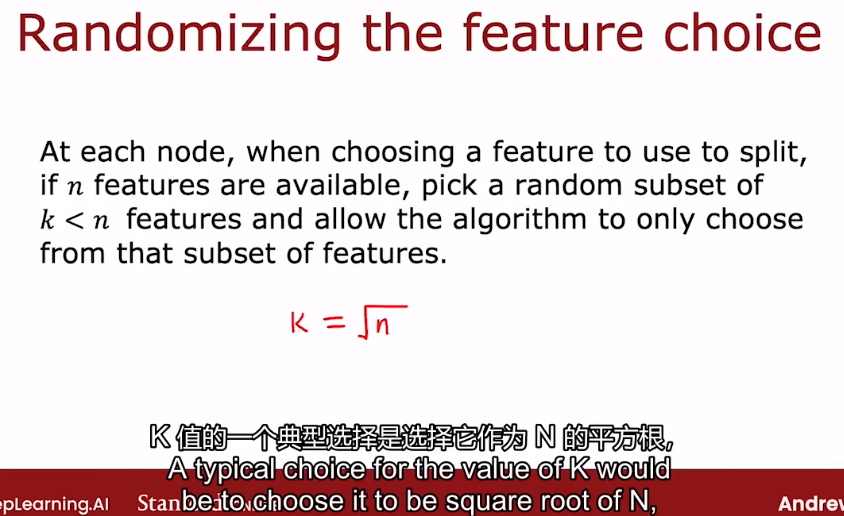
1.2 利用决策森林进行预测

森林中的每个决策树最后对测试样本进行预测，其预测权重为1/B（相当于每个决策树对同一个测试样本进行投票，总共投出B票，那么每个决策树的话语权就是1/B）。

1.3 缺点

对于比较大的训练集，构建决策森林时往往会得到相似度很高的决策树甚至一模一样的决策树，即决策森林内大部分决策树相似度高，那么这个森林的效果就不那么好。举个极端的例子，如果最后森林中所有决策树一模一样，则相当于还是使用单个决策树来进行决策，单一决策树对测试样本的细微变化非常敏感，因此其鲁棒性弱。因此解决方法就是增加决策森林中决策树结构的多样性。

2 算法改进

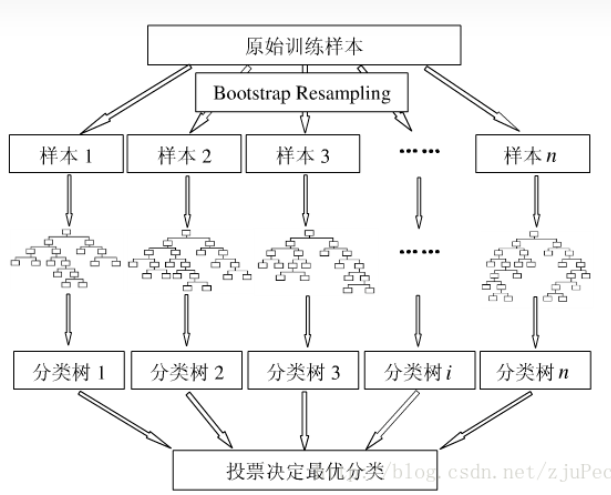


2.1 随机森林算法

构建任一决策树时，对于该决策树的任意节点，若当前节点有n个可选择特征进行划分训练集，那么从这n个特征中选取k个特征作为待选集，再从待选集中选取1个特征作为最终用于划分的特征。当n比较大时常常取

。该算法往往用于具有大量特征的训练集。由于其随机性，随机森林中的决策树探索了更多可能性，因此对于可能出现新类型的测试集来说随机森林的鲁棒性更强，其搜索范围更广。

由此可见随机森林的随机性体现在，每棵树的训练样本是随机抽取的，每棵树的特征也是随机选取的。随机性可以使得随机森林不容易过拟合，提高了鲁棒性。减小决策树的特征个数k，树之间的相关性和树的分类能力会相应的降低；增大k，两者会随之增大。所以选择最优的k（或者是范围）很重要，这也是随机森林唯一的一个参数。



2.2 随机森林核心特点

随机森林核心点是「随机」和「森林」，也是给它带来良好性能的最大支撑。

「随机」主要体现在两个方面：

样本扰动：直接基于自助采样法（Bootstrap Sampling），使得初始训练集中约63.2%的样本出现在一个采样集中。并带来数据集的差异化。

属性扰动：在随机森林中，对基决策树的每个结点，先在该结点的特征属性集合中随机选择k个属性，然后再从这k个属性中选择一个最优属性进行划分。这一重随机性也会带来基模型的差异性。

「集成」体现在：根据多个（差异化）采样集，训练得到多个（差异化）决策树，采用简单投票或者平均法来提高模型稳定性和泛化能力。