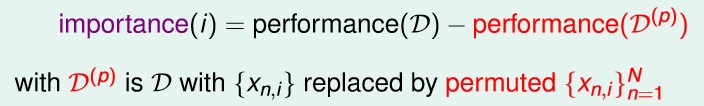
1. 随机森林简介、

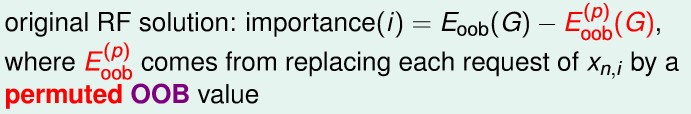
在机器学习中，随机森林是一个包含多个决策树的分类器，并且其输出的类别是由个别树输出的类别的众数而定。 Leo Breiman和Adele Cutler发展出推论出随机森林的算法。而"Random Forests"是他们的商标。这个术语是1995年由贝尔实验室的Tin Kam Ho所提出的随机决策森林（random decision forests）而来的。这个方法则是结合Breimans的"Bootstrap aggregating"想法和Ho的"random subspace method" 以建造决策树的集合。

1. Feature Importances

在实践中，数据中会有很多多余甚至无关特征，这些特征会严重影响模型的分类/回归效果。在随机森林中我们可以根据它自带的feature importance筛选特征。 如果你了解决策树的话，会知道决策树会根据信息熵逐一选取重要的特征。那么在随机森林中如何计算feature importance呢？核心idea：如果特征i对模型是有利的，那么将第i维特征置换为随机值，将会降低模型的性能。



将完整模型的性能减去置换第i维特征后的模型，就得到了第i维特征的重要性。那么问题来了，~~挖掘技术哪家强~~，要评估置换第i维特征后的模型性能，我们岂不是要重新训练并用验证集来评估性能？当然不用，忘了我们的OOB error吗？在随机森林中我们可以用OOB error来衡量模型性能，此处同样可以引入OOB error。



一、KMeans算法

K-平均算法源于信号处理中的一种向量量化方法，现在则更多地作为一种聚类分析方法流行于数据挖掘领域。k-平均聚类的目的是：把 {\displaystyle n} n个点（可以是样本的一次观察或一个实例）划分到k个聚类中，使得每个点都属于离他最近的均值（此即聚类中心）对应的聚类，以之作为聚类的标准。

1. KMeans算法描述

已知观测集，其中每个观测都是一个-维实向量，k-平均聚类要把这在个观测划分到k个集合中(k≤n),使得组内平方和（WCSS within-cluster sum of squares）最小。换句话说，它的目标是找到使得下式满足的聚类，



其中是在中所有点的均值。

5.2 问题二的模型建立与求解

问题二要求设计或采用一个方法，找出某种疾病最有可能的一个或几个致病位点，并给出相关的理论依据。为了解决这个问题，本文将使用两个方案找出相关的致病位点。然后通过分析两种方案的对比，选取比较合适的模型。下面将对两种方案模型应用与本问题的过程、步骤、结果等进行详细的阐述。

5.2.1 方案一： 随机森林模型的应用

5.2.2 方案二： 卡方检验模型的应用