# 第九章，集成学习

1. 集成学习 (ensemble 通过构建并结合多个 学习 器来完成学习任务 。有时 也被称为多分类器系统 (multi-classifier system) 、基于委员会的学习 (committee based learning)等。先学习产生一组个体学习期，然后再用某种策略将他们结合起来。
2. 同质集成中的个体学习器由相同的学习算法生成，个体学习器称为基学习器
3. 异质集成，个体学习器由不同的算法组成，个体学习器称为组件学习器。
4. 集成学习要显著优于单一个体学习器必须满足两 个必要条件：
    1. 个体学习器之间应该是相互独立的
    2. 个体学习器应当好于随机猜测学习器
5. AdaBoost算法，先从初始训练集中训练处一个基学习器，再根据基学习器的表现对训练的样本的权值进行调整，使得先前基学习器做错的训练样本在后续受到更多的关注。最后再进行基学习器的加权组合。***各个基学习器之间存在着强依赖关系***
6. 自助采样法：给定包含 m 个样本的数据集，我们先随机取出一个样本放入采样集中，再把 该样本放回初始数据集，使得下次采样时该样本仍有可能被选中。经过 m 次 随机采样操作，我们得到含 m 个样本的采样集。
7. Bagging算法，按照自助来样 法， 我们可采样出 T 个含 m 个训练样本的采样集，然后基于每个采样集训 练出一个基学习器，再将这些基学习器进行结合，通常对于分类问题采用投票的结合策略，对于回归问题采用简单均值的策略。
8. 随机森林，Bagging+决策树（效果很好）
9. 集成学习方法，为了在保持个体学习器足够好的前提下，尽量增加学习器的多样性，一般的思路是在学习的过程中引入随机性，常用的有：
    1. 训练样本扰动
    2. 输入属性扰动
    3. 输出标记扰动
    4. 算法参数扰动
    5. 混合扰动