- 集成学习
  - 概述
    - 个体学习器同质 称为基学习器
    - 个体学习器异质称为 组件学习器个体学习器应“好而不同“
  - 集成学习优势
    - 1.从统计方面，由于假设空间往往很大，单个学习器使用不同假设可能有同样性能，使用集成学习可能会降低风险
    - 2.从计算方面，学习算法往往易陷入局部极小，多次运行结合可降低陷入糟糕局部极小的风险
    - 3.从表示方面，由于真实假设可能不在当前学习算法所考虑的假设空间中，通过多个学习器，由于假设空间扩大，则有可能效果更优
  - Boosting
    - 个体学习器间存在强依赖关系、必须串行生成的序列化方法Boosting主要关注降低偏差
    - 工作机制
      - 1.根据初始训练集训练出一个基学习器
      - 2.根据基学习器的表现对训练样本分布进行调整，对分错的训练样本给予更多的关注
      - 3.基于调整后的样本分布训练训练下一个基学习器
      - 4.重复进行，直到基学习器数目达到指定T值
      - 5.将T个基学习器进行加权结合
    - 代表 AdaBoost
      - 标准AdaBoost只适用于二分类任务
      - 比较易理解的AdaBoost模型为“加性模型”，即基学习器的线性组合：其所用的优化机制为最小化指数损失函数
      - 基学习器取样方法
        - “重赋权法” 在每一轮训练中，根据样本分布为每一个训练样本重新赋予权重
        - “重采样法” 在每一轮训练中，对训练数据重新采样，可获得“重启动”机会避免训练过程过早停止
  - Bagging与随机森林
    - Bagging
      - 并行式集成学习方法最著名代表
      - 可以不经修改的用于多分类、回归任务
      - 工作机制
        - 1.利用自助采样方法采取T个含有m个训练样本的采样集
        - 2.利用每个采样集训练出一个基学习器，再将这些基学习器结合
        - 3.输出通常对于分类任务采用简单投票法，对于回归任务采用简单平均法
      - 自助采样给Bagging带来的优势
        - 由于初始训练集中约有36.8%样本未被使用，这些样本课本用作验证集来进行泛化性能的“外包估计”
        - 当基学习器为决策树时，可使用外包样本来辅助剪
        - 当基学习器为神经网络时，可使用外包样本来辅助早期停止以防过拟合
    - 随机森林（RF）
      - Bagging的一个扩展变体，且训练效率通常优于Bagging
      - RF在一决策树为基学习器构建Bagging继承的基础上，进一步在训练过程中引入了随机属性选择
      - 工作机制 传统决策树在训练划分属性时是在当前节点引入了最优属性选择； 而RF中对于及决策树的每个节点，先从该节点的属性集合中随机选择一个包含k个属性的子集，再从这个子集中选择一个最优属性用于划分
  - 结合策略
    - 平均法
      - 对数值型输出，最常见的结合策略是使用平均法
      - 简单平均法（simple averaging）
      - 加权平均法（weighted averaging） 可视为集成学习研究的基本出发点，不同的集成方法可视为通过不同的方式来确定加权平均法中的基学习器权重
      - 在个体学习器性能相差较大时使用加权平均；在个体学习器性能相近时使用简单平均
    - 投票法
      - 对分类任务来说，最常见的结合策略是使用投票法
      - 绝对多数投票法（majority voting） 若某标记得票过半数，则预测为该标记，否则拒绝预测
      - 相对多数投票法（plurality voting） 预测为得票最多的标记，若标记或票数相同，则随机选取
      - 加权投票法（weighted voting） 权重可设定方法;估算出个体学习器误差，然后令权重大小与误差大小成反比。注意：权重选取应非负
      - 注意事项
        - 1.不同类型的h（x）值不能混用，如要使用需将类标记与类概率值转化统一
        - 2.若基学习器的类型不同，则其类概率值不能直接进行比较，通常将类概率转化为类标记输出后再投票
    - 学习法
      - 当训练数据很多时，可使用一种更强大的结合策略“学习法”
      - 代表 Stacking 工作机制
        - 1.先从初始数据集训练出初级学习器
        - 2.“生成”一个新数据集用于训练次级学习器
        - 在新数据集中，初级学习器的输出被当作样例输入特征，而初始样本的标记仍被当做样例标记
        - Stacking通常优于BMA，其鲁棒性更优，且BMA对模型近似误差非常敏感
  - 多样性
    - 误差-分歧分解 个体学习器准确性越高、多样性越大、则集成越好
    - 多样性度量
      - 多样性度量用于度量集成中的个体分类器的多样性，典型做法为考虑个体分类器的两两相似\不相似性
      - 指标
        - 不合度量 值越大多样性越大
        - 相关系数 正相关值为正，负相关值为负
        - Q-统计量 与相关系数符号相同，且值更大
        - k-统计量 若分类器h1与h2在D上完全一致，则k=1；若他们仅是偶然一致，则k=0； k通常为非负，仅在h1与h2达成一致的概率甚至低于偶然性的情况下取负值
      - 多样性增强
        - 数据样本扰动 通常基于采样法，如Bagging中使用自助采样，Adaboost中使用序列采样， 对决策树，神经网络等训练样本稍加变化就导致学习器显著变动的“不稳定基学习器”有效
        - 输入属性扰动 采用不同的子空间观察数据。随机子空间（random subspace）算法从初始属性记中抽取若干个属性子集，再基于每个属性子集训练一个基学习器。 适用于线性学习器、支持向量机、朴素贝叶斯、k近邻学习器等“稳定基学器”；不适用只包含少量属性的训练样本
        - 输出表示扰动
          - 基本思路是对输出表示进行操纵以增强多样性
          - 对类标记做变动 “翻转法”
          - 对输出表示进行转化 “输出调制法”
          - 将原任务拆解为多个可同时求解的子任务 “ECOC 法”
        - 算法参数扰动
          - 负相关法 通过正则化项来强制个体神经网络使用不同的参数
          - 对于参数较少的算法，可通过将其学习过程中的某些环节用其他类似方式代替
          - 注意：交叉验证确定参数值，实际上已经使用了不同参数训练多个学习器，只不过只对单个学习器进行使用，集成学习的区别是把这些学习器都利用起来
    - 如何理解多样性，被认为时集成学习中的圣杯问题