集成学习

集成学习가 도대체 무엇이냐?: 通过构建并结合多个学习器来完成学习任务。

그 안에서 同质集成과 异质集成으로 나뉜다.

同质集成: 안에 있는 个体学习器가 같은 학습 알고리즘으로 생성되어있음, 个体学习器를 **基学习器**라고도 부름

异质集成: 안에 있는 个体学习器가 다른 학습 알고리즘으로 생성되어있음, 个体学习器를 **组件学习器**라고도 부름

한 문제를 두고 알고리즘마다 정확하게 분류할 수 있는 데이터가 각자 다르니까, 여러 개의 个体学习器를 조합해서 오답률을 줄이는게 목적임.

常用方法:

1) 训练样本扰动（젤 자주 사용한데）: 데이터 셋에서 抽样해서 다른 样本子集를 만들고, 다시 이 다른 子集들로 서로 다른 个体学习器를 만드는 방법.
2) 输入属性扰动（随机森林）: 抽取出若干个属性子集
3) 输出标记扰动: 对训练样本的类标记稍作变动
4) 算法参数扰动: 参数임의로 설정
5) 混合扰动: 在同一个集成算法中同时使用上述多种扰动方法

드디어 나왔다 **Bagging** (**b**ootstrap自助法 **agg**regating，装袋法)시험에 나왔음

Bootsrap은 自助法：有放回的抽样方法，个体分类器之间不存在依赖关系。

Bagging算法：

1） 从原始样本集中**抽取训练集。**每轮从原始样本集中使用**Bootstrap**的方法抽取N个训练样本。工进行K轮抽取，**得到k个训练集**。

2） 每次使**用一个训练集得到一个模型**，K个训练集共得到K个模型

3） （1）对**分类问题**：将上步得到的K个模型采用**投票的方式**得到分类结果；（2）对**回归问题**，计算上述模型的**均值**作为最后的结果。

随机森林（Random Forest）是Bagging的拓展变体（升级版）

随机森林在以决策树的基学习器构建**Bagging集成的基础上**，进一步在决策树的训练过程中**引入了随机属性选择**。随机森林是从相应节点的属性中随机选择若干个， 然后基于某个度量指标选择一个最优属性。

Boosting算法（提升法，요 형님도 시험에 나왔음）

Boosting算法中，**每一个样本数据是有权重的**，每一个学习器是由先后顺序的。在PAC（概率近似正确）的学习框架下，一定可以将弱分类器组装成一个强分类器。

核心问题

1. 每一轮如何改变徐连数据的权值和概率分布？

通过提高那些在前一轮被弱学习器分错样例的权值，减小前一轮正确样例的权值，使学习器重点学习分错的样本，提高学习器的性能。

2. 通过什么方法来组合弱学习器？

通过加法模型将弱学习器进行线性组合，学习器准确率大，则相应的学习器权值大；反之，则学习器的权值小。即给学习器好的模型一个较大的确信度，提高学习器的性能。

AdaBoost算法（Adaptive Boosting， 自适应增强）

初始化训练样本的权值分布，每个样本具有相同权值；

训练弱的分类器，如果样本分类正确，则在构造下一个训练集中，它的权值就会低；反之提高。用更新过的样本集去训练下一个分类器。 → 将所有弱分类器组合成强分类器。

Bagging & Boosting 区别 (시험에 나왔어)

1) 训练样本集
   Bagging：训练集是有放回抽样，从原始集中选出的K组训练集是相互独立的。
   Boosting： 每一次迭代的训练集不变

2) 训练样本权重
   Bagging：每个训练样本的权重相等，即1/N。
   Boosting：根据学习器的错误率不断调整样例的权值，错误率越大，权值越大

3) 预测函数的权重
   Bagging：K组学习器的权重相等，即1/K
   Boosting：学习器性能好的分配较大的权重，学习器性能差的分配较小的权重。

4) 并行计算
   Bagging：K组学习器模型可以并行生成。
   Boosting：K组学习器只能顺序生成，因为后一个模型的样本权重需要前一个学习器模型的结果。

集成学习的组合策略

-绝对多数投票法：과반수 되면 그 标记로 分类한다, 아니면 분류 안함

-相对对数投票法: 가장 많은 표기로 분류, 가장 많은게 여러 개 겹치면 임의로 선택

-加权投票法：类에 权值주고 분류