第八次作业

阅读机器学习实战 第7章 7.1-7.4节

Exercise 8.1

关于集成学习算法的说法正确的是（ ）
A. 一种并行的算法框架
B.一种串行的算法框架
C.一类全新的数据挖掘算法
D.一类将已有算法进行整合的算法

Exercise 8.2

关于Bootstrap采样正确的说法是：
A.有放回的采样
B.无放回的采样
C.样本大小必须与原样本相同
D.应尽可能保证各原始数据都出现

解析：Bootstrap过程的机制是：首先有一个实际观测到的数据集(称之为原始数据集)，它含有n 个观查单位。从这个数据集中有放回地随机抽取t个组成一个新样本。

Exercise 8.3(多选题)

Bagging的主要特点有：
A.各基础分类器并行生成
B.各基础分类器权重相同
C.只需要较少的基础分类器
D.基于Bootstrap采样生成训练集

Exercise 8.4

在Baggging集成学习中，多样性是通过（）实现的。

A、数据样本扰动
B、输入属性扰动
C、输出表示扰动
D、算法参数扰动

解析：根据随机放回抽样出由不同数据特征的数据集

Exercise 8.5

以下不属于bagging的特点是（）

A、有放回抽样多个子集
B、训练多个分类器
C、最终结果为每个学习器加权后的线性组合
D、可以减少过拟合

解析：各个基础分类器权重相同

Exercise 8.6

假如你正在处理一个包含3个输入特性的二分类问题。你选择对这些数据使用bagging算法, 构造3个估计器。现在，假设每个估计器都有70%的准确率。现在基于最大投票对单个估计量的结果进行聚合，你可以得到的最小准确率是多少?

A) 大于70%
B) 大于等于70%
C) 可以小于70%
D) 都不对

解析：对各个分类器进行聚合之后并不一定会产生比原先更好的分类效果

Exercise 8.7

以下关于集成学习特性说法错误的是( )。

A. 集成学习需要各个弱分类器之间具备一定的差异性
B. 集成多个线性分类器也无法解决非线性分类问题
C. 弱分类器的错误率不能高于0.5
D. 当训练数据集较大时，可分为多个子集，分别进行训练分类器再合成

解析：可以解决线性分类问题

Exercise 8.8

在随机森林里，你生成了几百颗树($T_1, T_2, \cdots, T_n$)，然后对这些树的结果进行综合，下面关于随机森林中每颗树的说法正确的是？( )

A. 每棵树是通过数据集的子集和特征的子集构建的
B. 每棵树是通过所有的特征构建的
C. 每棵树是通过所有的数据构建的
D. 以上都不对

解析：随机森林是基于bagging的方法，是通过对数据和特征的采集来构建每一棵树。

Exercise 8.9

以下关于随机森林(Random Forest)说法正确的是( )。

A. 随机森林构建决策树时，是无放回的选取训练数据
B. 随机森林算法容易陷入过拟合
C.随机森林学习过程分为选择样本、选择特征、构建决策树、投票四个部分
D. 随机森林由若干决策树组成，决策树之间存在关联性

解析：随机森林是又放回地选取训练数据；随机森林不容易陷入过拟合；组成随机森林的决策树之间不存在关联性

Exercise 8.10

如果你已经在完全相同的训练集上训练了5个不同的模型，并且 它们都达到了95%的准确率，是否还有机会通过结合这些模型来获得更 好的结果？如果可以，该怎么做？如果不行，为什么?

解析：如果你已经训练了5个不同的模型，并且都达到了95%的精度，则可以尝试将它们组合成一个投票集成，这通常会带来更好的结果。如果模型之间非常不同（例如，一个SVM分类器、一个决策树分类器，以及一个Logistic回归分类器等），则效果更优。如果它们是在不同的训练实例（这是bagging和pasting集成的关键点）上完成训练，那就更好了，但如果不是，只要模型非常不同，这个集成仍然有效。

Exercise 8.11

包外评估的好处是什么?

解析：包外评估可以对bagging集成中的每个预测器使用其未经训练的实例（它们是被保留的）进行评估。不需要额外的验证集，就可以对集成实施相当公正的评估。所以，如果训练使用的实例越多，集成的性能可以略有提升。