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

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

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

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

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

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

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

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？

Exercise 8.11

包外评估的好处是什么？

Exercise 8.12 (实践题)

加载MNIST数据集，将其分为一个训练集、 一个验证集和一个测试集（例如，使用 50000 个实例训练、10000 个实例验证、10000 个实例测试）。然后训练多个分类器，比如一个随机森林分类器、一个极端随机树分类器和一个SVM分类器。接下来，尝试使用软投票法或者硬投票法将它们组合成一个集成，这个集成在验证集上的表现要胜过它们各自单独的表现。成功找到集成后，在测试集上测试。与单个的分类器相比，它的性能要好多少？

Note: 极端决策树：

1. 对于每个决策树的训练集，RF采用的是随机采样bootstrap来选择采样集作为每个决策树的训练集，而extra trees一般不采用随机采样，即每个决策树采用原始训练集。
2. 在选定了 K 个划分特征后，RF的决策树会基于基尼系数，均方差之类的原则，选择一个最优的特征值划分点，这和传统的决策树相同。但是extra trees比较的激进，随机选择 K 个特征，对于这

K 个特征，每个特征随机选择 1 个分裂节点，从而得到 K 个分类节点。然后计算这 K 个分裂节点的分数，选择得分最高的节点作为分裂节点。