# 第10章习题参考答案

10.1 试阐述一下为什么要使用集成学习。

回答：集成学习通过有策略地组合多个弱学习器合成一个强学习器来提升预测结果。各个弱学习器之间具有一定的差异性，这使得每个个体学习器会产生不同的决策边界，将这些弱学习器结合之后能够得到更加合理的决策边界，减少模型预测误差。另外，集成学习中部分模型的各个体学习器类型允许不同，这增强了模型的灵活性，也使得模型具有更好的拟合能力。

10.2 试阐述AdaBoost算法的步骤。

回答：可根据AdaBoost的具体步骤将其分为三个步骤：（1）确定基学习器：根据个人所使用的数据特征选择合适的基学习器，一般情况下没有特别严格要求；（2）训练基学习器：更新样本权重，基于采样的数据训练出基学习器并计算出样本误差率（用于更新样本权重）；（3）基学习器合并：基于线性加权组合各个基学习器得到最终学习器。

10.3 试阐述AdaBoost和Bagging的主要区别。

回答：（1）模型目标：AdaBoost侧重于减少偏差，而Bagging侧重于减少方差；  
（2）样本选择：AdaBoost中每一轮训练的样本都是固定的，但是权重在发生变化；而Bagging是随机有放回的采样；

（3）样本权重：AdaBoost根据样本的预测效果动态调整样本权重，预测效果差的会被赋予更大的权重，而Bagging采取的是均匀取样，因此每个样本权重相同；

（4）计算方式：AdaBoost是串行计算，而Bagging的是并行计算。

10.4 试分析AdaBoost和GBDT之间的异同点。

回答：相同点：都是通过将多个表现一般的模型组合起来集成为一个表现性能较好的模型，并且两个模型都是基于先前模型的表现调整设置下一各个体学习器；

不同点：AdaBoost通过增加错误预测数据的权重来提升模型表现性能，而GBDT通过梯度来优化模型表现性能。

10.5 给定如下表的数据，试计算Adaboost算法在第二次训练时的数据权值分布。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| X | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Y | -1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 |

回答：首先初始化所有数据权值分布，即就是：

在权值为的训练数据上，阈值取3.5时可以获得最小的分类误差，因此，基本分类器设定为：

此时，本题目所包含的数据在上的误差率为：

计算的系数：

则新的权值分布为：

可以得到：

10.6 选择自己研究领域的数据集，试用python语言自己编写集成学习算法或使用Scikit-leran第三方包实现回归或分类。

回答：略（具体可参照10.5中的案例应用实现）。