作业6:决策树与Boosting方法

1.Bagging方法

在课堂上我们已经了解到，Bagging方法可以减小模型的误差。我们现在从理论上来证明这一点。首先我们在数据集合D中有放回抽样生成了m个数据集$\{D_i\}_{i=1}^m$，在每个数据集合D_i上训练分类器$h_i(\boldsymbol{x})$，假设我们现在有n个待预测的新样本$\{\boldsymbol{x_j}\}_{j=1}^n$。对于其中的任意一个样本$\boldsymbol{x}$，Bagging方法的预测值可以定义为多个分类器的预测平均值：

$$h_B = \frac{1}{m}\sum_{i=1}^{m} h_i(\boldsymbol{x}) \tag{1}$$

若对于样本$\boldsymbol{x}$真实的预测值为$y(\boldsymbol{x})$，则可知道每个分类器$h_i(\boldsymbol{x})$的误差ϵ为:

$$\epsilon_i(\boldsymbol{x}) = h_i(\boldsymbol{x}) - y(\boldsymbol{x}) \tag{2}$$

对于m个单独的分类器，它们的平均均方误差可以定义为：

$$E_h = \frac{1}{m}\sum_{i=1}^{m}\{\frac{1}{n}\sum_{j=1}^{n}[\epsilon_i(\boldsymbol{x}_j)]^2\} \tag{3}$$

对于Bagging分类器的均方误差可以定义为：

$$E_{h_B} = \frac{1}{n}\sum_{j=1}^{n}[\epsilon_B(\boldsymbol{x}_j)]^2 = \frac{1}{n}\sum_{j=1}^{n}[h_B(\boldsymbol{x}_j) - y(\boldsymbol{x}_j)]^2 \tag{4}$$

(1)假设所有分类器的误差均值为零，而且互不相关，即：

$$\frac{1}{n}\sum_{j=1}^{n}\epsilon_i(\boldsymbol{x}_j) = 0, (i \in \{1, 2, \ldots, m\}) \tag{5}$$

$$\frac{1}{n}\sum_{j=1}^{n}\epsilon_i(\boldsymbol{x}_j)\epsilon_k(\boldsymbol{x}_j) = 0, (i, k \in \{1, 2, \ldots, m\}) \tag{6}$$

请证明：

$$E_{h_B} = \frac{1}{m} E_h \tag{7}$$

(2)但在实际情况中，往往它们的误差是高度相关的，请在(1)中条件不满足的情况下证明：

$$E_{h_B} \leq E_h \tag{8}$$

2.Adaboost、决策树与随机森林

本题中，我们将实现一个基于“决策树桩”的Adaboost算法，并对所提到的这三种算法进行比较。我们仍然使用的是MNIST数据集作为样本，使用的是其中的数字“4”和“9”，从数据集的训练样本中抽样500个作为本题的训练样本，测试样本与数据集中的保持一致，预处理方法仍然按照作业4中所提到的方法即可。请根据提示，完成以下题目：

2.1实现基于决策树桩的Adaboost

样本定义如下：$\boldsymbol{X} \in \Re^{n \times p}$ 是一个矩阵，每一行表示一个样本，每一列是一维特征。而样本标签$\boldsymbol{y} \in \{-1,+1\}^n$是一个向量，每一个样本对应一个类别标签($-1$ 或者$+1$)。对于任意一个特征维度j，我们可以定义“决策树桩”：

$$h_{(a,d,j)}(\boldsymbol{x}) = \begin{cases} d & , \ x_j \leq a \\ -d & , \ x_j > a \end{cases}$$

其中$a \in \Re, j \in \{1,2,...,p\}$, $d \in \{-1,+1\}$，在这里，样本$\boldsymbol{x} \in \Re^p$是一个向量，x_j是样本$\boldsymbol{x}$的第j个特征

要求：请提交程序运行所需要的所有代码，并保持原来的文件结构。

(1)**请编写决策树桩分类器，完成代码框架decision_stump.m**

提示：

这个程序的输入为：所有样本、标签以及对应的权重($\{(\boldsymbol{x}_i, y_i, w_i)\}_{i=1}^n$，其中$w_i \leq 1, \sum_{i=1}^n w_i = 1$)。返回值为：使训练误差最小化的决策树桩。这需要为决策树桩筛选最优化的参数a, d以及特征维度j。

输出是一个三元组$(a^\star, d^\star, j^\star)$，满足：

$$a^\star, d^\star, j^\star = \arg\min_{a,d,j} \sum_{i=1}^n w_i 1\{h_{a,d,j}(\boldsymbol{x}_i) \neq y_i\} \tag{9}$$

其中$1\{h_{a,d,j}(\boldsymbol{x}_i) \neq y_i\}$表示当$h_{a,d,j}(\boldsymbol{x}_i) \neq y_i$时取1，否则取0。

请注意优化代码，尽量减少循环，实现向量化，以提高程序效率。

(2)**请编写Adaboost算法中权值更新的步骤，完成代码框架update_weights.m**

(3)**请计算Adaboost的分类错误率，完成代码框架adaboost_error.m**

(4)**设置迭代次数为200次，调用adaboost.m，完成训练和测试**

以mat格式保存200次迭代的训练错误率和测试错误率(需要提交)，在报告中说明当迭代次数为30,100,200时，训练错误率和测试错误率。

2.2训练决策树

请将训练集合中80%的样本作为训练集,其余作为验证集。在Matlab工具箱中选择“All Trees”进行训练和验证(CART算法)，它会分别在最大分裂数为4,20,100三种情况下训练决策树，老版本的同学请自行在“Advance”选项中设定，记录训练模型的错误率。分别导出模型，并在测试集合上进行预测，记录测试错误率。

2.3训练随机森林

请将训练集合中80%的样本作为训练集,其余作为验证集。在Matlab工具箱中选择“Bagged Tree”以训练随机森林，设置“Advance”中的分类器数目为30,100,300，分别进行训练，记录训练模型的错误率。分别导出模型，并在测试集合上进行预测，记录测试错误率。

2.4结果讨论

请基于上述结果，说说三种分类器之间的联系，言之有理即可。