《大数据机器学习》第 6 次作业

姓名：刘培源　　学号：2023214278

题目 1：某公司招聘职员考查身体、业务能力、发展潜力这 3 项。身体分为合格 1、不合格 0 两级，业务能力和发展潜力分为上 1、中 2、下 3 三级。分类为合格 1、不合格 −1 两类。已知 10 个人的数据，如表1所示。假设弱分类器为决策树桩。试用 AdaBoost 算法学习一个强分类器。

表 1: 应聘人员情况数据表

	1	2	3	4	5	6	7	8	9	10
身体	0	0	1	1	1	0	1	1	1	0
业务能力	1	3	2	1	2	1	1	1	3	2
发展潜力	3	1	2	3	3	2	2	1	1	1
分类	-1	-1	-1	-1	-1	-1	1	1	-1	-1

答：用 x_0、x_1 和 x_2 分别代表身体、业务能力和发展潜力，Adaboost 算法迭代如下：

1. 第 1 轮:
   - 初始权重分布:
     $D_1 = [0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1]$
   - 选取的弱分类器规则: $h_1(x) = \begin{cases} +1 & \text{if } x_0 > 0.0 \\ -1 & \text{otherwise} \end{cases}$
   - $\alpha_1 \approx 0.693$
   - 更新 D
2. 第 2 轮:
   - 初始权重分布:
     $D_2 = [0.0625, 0.0625, 0.0625, 0.0625, 0.0625, 0.0625, 0.25, 0.25, 0.0625, 0.0625]$
   - 选取的弱分类器规则: $h_2(x) = \begin{cases} +1 & \text{if } x_1 < 1.5 \\ -1 & \text{otherwise} \end{cases}$
   - $\alpha_2 \approx 0.733$

- 更新 D

3. 第 3 轮:

   - 初始权重分布:

     $D_3 = [0.1667, 0.0385, 0.0385, 0.1667, 0.0385, 0.1667, 0.1538, 0.1538, 0.0385, 0.0385]$
   - 选取的弱分类器规则: $h_3(x) = \begin{cases} +1 & \text{if } x_2 < 1.5 \\ -1 & \text{otherwise} \end{cases}$
   - $\alpha_3 \approx 0.499$
   - 更新 D

4. 第 4 轮:

   - 初始权重分布:

     $D_4 = [0.114, 0.0714, 0.0263, 0.114, 0.0263, 0.114, 0.2857, 0.1053, 0.0714, 0.0714]$
   - 选取的弱分类器规则: $h_4(x) = \begin{cases} +1 & \text{if } x_0 > 0.5 \\ -1 & \text{otherwise} \end{cases}$
   - $\alpha_4 \approx 0.582$
   - 更新 D

5. 第 5 轮:

   - 初始权重分布:

     $D_5 = [0.0748, 0.0469, 0.0553, 0.2395, 0.0553, 0.0748, 0.1875, 0.0691, 0.15, 0.0469]$
   - 选取的弱分类器规则: $h_5(x) = \begin{cases} +1 & \text{if } x_0 > 0 \\ -1 & \text{otherwise} \end{cases}$
   - $\alpha_5 \approx 0.532$
   - 更新 D

6. 第 6 轮:

   - 初始权重分布:

     $D_6 = [0.0503, 0.0315, 0.0372, 0.1611, 0.0372, 0.0503, 0.3654, 0.1346, 0.1009, 0.0315]$
   - 选取的弱分类器规则: $h_6(x) = \begin{cases} +1 & \text{if } x_2 < 2.5 \\ -1 & \text{otherwise} \end{cases}$
   - $\alpha_6 \approx 0.545$
   - 更新 D

7. 第 7 轮:

- 初始权重分布:

  $D_7 = [0.0336, 0.0627, 0.0739, 0.1076, 0.0248, 0.1001, 0.2441, 0.0899, 0.2006, 0.0627]$
- 选取的弱分类器规则: $h_7(x) = \begin{cases} +1 & \text{if } x_1 < 1.5 \\ -1 & \text{otherwise} \end{cases}$
- $\alpha_7 \approx 0.573$
- 更新 D

8. 第 8 轮:

   - 初始权重分布:

     $D_8 = [0.0697, 0.0413, 0.0487, 0.2229, 0.0164, 0.2074, 0.1608, 0.0593, 0.1322, 0.0413]$
   - 选取的弱分类器规则: $h_8(x) = \begin{cases} +1 & \text{if } x_0 > 0 \\ -1 & \text{otherwise} \end{cases}$
   - $\alpha_8 \approx 0.633$
   - 更新 D

9. 第 9 轮:

   - 初始权重分布:

     $D_9 = [0.0447, 0.0265, 0.0312, 0.1429, 0.0105, 0.133, 0.3654, 0.1346, 0.0848, 0.0265]$
   - 选取的弱分类器规则: $h_9(x) = \begin{cases} +1 & \text{if } x_0 > 0.5 \\ -1 & \text{otherwise} \end{cases}$
   - $\alpha_9 \approx 0.499$
   - 更新 D

10. 第 10 轮:

    - 初始权重分布:

      $D_{10} = [0.0306, 0.0181, 0.058, 0.2653, 0.0195, 0.091, 0.2501, 0.0921, 0.1573, 0.0181]$
    - 选取的弱分类器规则: $h_{10}(x) = \begin{cases} +1 & \text{if } x_0 < 0 \\ -1 & \text{otherwise} \end{cases}$
    - $\alpha_{10} \approx 0.327$
    - 更新 D

11. 第 11 轮:

    - 初始权重分布:

      $D_{10} = [0.0232, 0.0138, 0.0441, 0.2016, 0.0148, 0.0692, 0.3654, 0.1346, 0.0848, 0.0265]$
    - 发现预测准确率已经为 1，停止迭代

题目 2：比较支持向量机、AdaBoost、逻辑斯蒂回归模型的学习策略与算法。

答：

1. **支持向量机：**

   - **学习策略：**SVM 的主要目标是最大化分类间隔，同时允许一定程度的分类错误。这通过最小化一个包含正则化项和合页损失（hinge loss）的目标函数来实现：

     $$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

     其中，$\mathbf{w}$ 是权重向量，b 是偏置项，ξ_i 是松弛变量，用于处理不完全可分的情况，C 是正则化参数。

   - **学习算法：**序列最小最优化算法（SMO）是一种高效的算法，用于训练 SVM。SMO 通过分解大的二次规划问题为一系列最小的二次规划问题来工作。

2. **AdaBoost：**

   - **学习策略：**AdaBoost 通过迭代地增加难以分类的样本的权重，构建一个强分类器。其目标是最小化加法模型的指数损失函数：

     $$\min_{\alpha,\mathbf{G}} \exp\left(-\sum_{i=1}^{n} y_i \sum_{m=1}^{M} \alpha_m G_m(\mathbf{x}_i)\right)$$

     其中，G_m 是弱分类器，α_m 是分类器的权重。

   - **学习算法：**前向分步加法算法，该算法逐步添加弱分类器，每次添加都是为了最小化当前的指数损失函数。

3. **逻辑斯蒂回归：**

   - **学习策略：**在 Logistic 回归中，通常使用极大似然估计来找到最佳的参数，这可以通过最小化对数损失来实现。在有正则化的情况下，目标函数变为：

     $$\min_{\mathbf{w},b} -\sum_{i=1}^{n}\left[y_i \log(\sigma(\mathbf{w}^T\mathbf{x}_i + b)) + (1-y_i)\log(1-\sigma(\mathbf{w}^T\mathbf{x}_i + b))\right] + \lambda\|\mathbf{w}\|^2$$

     其中，$\sigma(\cdot)$ 是 Sigmoid 函数，λ 是正则化系数。

   - **学习算法：**包括改进的迭代尺度算法、梯度下降和拟牛顿法等。这些算法通过不断更新参数来最小化损失函数。