2023–2024 学年第 2 学期《机器学习基础》期中试卷

- **考试时间**：2024年4月16日 13:00–14:50
- **说明**：共六道大题，请将答案写在答题纸上。

---

1. （20分）信息增益、基尼指数、朴素贝叶斯估计

考虑具有三个二值离散特征的二类分类问题训练数据集：

$$D = \{((a, y, c), +1), ((a, n, c), +1), ((b, y, d), +1), ((b, n, c), -1), ((b, n, c), -1), ((a, y, d), +1)\}$$

计算：

1. 第一维特征对训练数据集 D 的信息增益（**注：对数不用化为小数**）；
2. 第二维特征的基尼指数；
3. 朴素贝叶斯方法中 $P(y = 1)$ 的极大似然估计和贝叶斯估计（$\lambda = 1$），并说明为什么要引入贝叶斯估计。

---

2. （15分）岭回归解析解

给定已经中心化的训练数据集：

$$D = \{(x_i, y_i)\}_{i=1}^{N}, \quad x_i \in \mathbb{R}^R, \quad y_i \in \mathbb{R}, \quad 1 \le i \le N$$

给出**岭回归模型**，并求其解析解。

---

3. （20分）支持向量机的原始问题与对偶问题

给定训练数据集：

$$D = \{(x_i, y_i)\}_{i=1}^{N}, \quad x_i \in \mathcal{X} \subseteq \mathbb{R}^R, \quad y_i \in \{+1, -1\}$$

考虑基于构建线性支持向量机模型，在惩罚项中以 $\sum_{i=1}^{N} \xi_i^2$ 代替总松弛幅度$\sum_{i=1}^{N} \xi_i$：

1. 基于 D 给出线性支持向量机的原始最优化问题；
2. 利用拉格朗日乘子法推导出其对偶问题。

---

4. （10分）Bagging 与随机森林的比较

简要比较 **Bagging 方法** 与 **随机森林模型** 的异同。

---

5. （10分）交叉验证法与自助法的比较

简要比较 k **折交叉验证法** 与 **自助法（Bootstrap 方法）** 的异同。

---

6. （25分）AdaBoost 算法分析

1. 简述 **AdaBoost 算法** 的算法框架；
2. 如果我们假设 AdaBoost 算法中每轮学到的弱分类器训练误差都小于 0.5，训练能够持续的情况下，第 t 轮学习的弱分类器 h_t 与第 $t+1$ 轮学习的弱分类器 h_{t+1} 是否一定不同？简要证明你的结论；
3. 令 γ 为使得对任一 $1 \leq t \leq T$，满足：

   $$\left(\frac{1}{2} - \epsilon_t\right) \geq \gamma > 0$$

   成立的正数。考虑使用一固定值 $\alpha > 0$ 来替换 AdaBoost 算法中的 α_t，即对 $t \in \{1, 2, \ldots, T\}$，将 α_t 都设定为 α。通过分析训练误差，求以 γ 的函数表示的你认为最合理的 α；
4. 采用你选择的 α 在每轮权值调整中分配给当前被预测错误的样本的权值所占比例相对采用 α_t 的情形是**更多**、**更少**还是**保持不变**？简要论证你的结论。