第一次编程作业

1 任务说明

本次作业旨在引导同学们深入探索机器学习中若干基础算法，不同于简单地调用现有库函数，同学们需要从算法原理出发，从零开始实现这些机器学习算法。请从以下 4 种算法中选择 **2** 种实现并进行汇报:

- K 近邻算法 (K-Nearest Neighbors, K-NN)
- 支持向量机 (Support Vector Machine, SVM)
- 决策树 (Decision Tree, e.g., ID3, C4.5, or CART)
- 逻辑回归 (Logistic Regression)

对于你选择的**每一种算法**，都需要在以下两类数据集上进行实验和汇报。

1.1 合成数据集实验

你需要自行编写代码生成至少 **3** 个合成数据集，并通过控制数据的特性，展示你的算法效果由差变好的过程。

（提示：可通过调整数据分布、样本数量、噪声水平等控制合成数据集的特性）

1.2 真实数据集实验

对于你选择的每一种算法，至少选择 **1** 个以下的真实数据集进行训练和测试。

- **Gisette 数据集**

  这是一个源自 NIPS 2003 特征选择挑战赛的二分类数据集。它的目标是区分手写数字中的 “4” 和 “9”。由于比赛不开放测试集标签数据，本次作业中将验证集作为测试集使用。

  - **特征与样本：**数据集包含 **5,000** 个特征，训练集有 6,000 个样本，测试集（验证集）有 1,000 个样本。
  - **详细信息链接：** Gisette Dataset @ UCI Machine Learning Repository
  - **网盘下载链接：** `https://cloud.tsinghua.edu.cn/d/0fa331dd22d74bc0bc76/`

- **HIGGS 数据集**

  这是一个源自高能物理领域的二分类数据集。它的任务是从粒子加速器实验产生的背景噪声中，区分出由希格斯玻色子（Higgs boson）产生的信号。

- **特征与样本：**完整数据集包含 **1,100 万**个样本，每个样本有 **28** 个特征。在本次作业中，将使用其子集（前 50 万个样本）进行实验，请自行划分训练集/验证集/测试集。
- **详细信息链接：** HIGGS Dataset @ UCI Machine Learning Repository
- **网盘下载链接：** `https://cloud.tsinghua.edu.cn/d/f8d7e695c7144a618db3/`

2 任务和报告要求

a) 请勿直接调用现有的机器学习库（如 sklearn 等）完成算法实现。你需要从零开始完成算法实现、合成数据集实验和真实数据集实验。算法的实现可参考课程 PPT 或其他资料（请勿直接复制），真实数据集的加载方式可参见附件。

b) 在合成数据集实验部分，展示数据特性是如何影响算法效果的，并给出你的分析。

c) 在真实数据集实验部分，展示超参数调优的过程，如有 training loss 和 validation loss，请展示超参对应的 loss 曲线，说明 loss 曲线随超参变化的原因。

d) 通过以上的实验，比较你所实现的两种算法的适用性和优劣。

e) 对于每种算法，选择一组实验画出对应的决策边界（decision boundary），并解释其形状。(提示：绘制决策边界可用“网格法”，即在我们数据的二维平面上铺上一层密集的、看不见的“网格点”，然后让我们训练好的模型去预测每一个网格点的类别，最后根据预测结果为不同的区域上色。参考：决策边界绘制)

f) 将完成后的实验报告 pdf（不超过 8 页）、代码文件打包提交，建议使用 python 进行实现。