# 《机器学习》课程实验一导论

1 实验名称

基于 kNN 的手写数字识别实验

**2 实验目的**

1. 实现 k 近邻（k-Nearest Neighbor, kNN）分类器；

2. 掌握留一法（Leave-One-Out, LOO）交叉验证流程；

3. 对比自实现结果与 Weka 内置 kNN 精度；

4. 学会用图表展示实验过程与结论。

3 数据集下载与快速检查

3.1 下载地址（官网可访问）

• 官网：http://archive.ics.uci.edu/ml/datasets/Semeion+Handwritten+Digit

**3.2 数据集说明**

解压后得到 semeion.data.txt（或 semeion.data），纯文本，空格分隔：

• 前 256 列 = 16×16 手写图像展开成 0/1 像素；

• 后 10 列 = 独热编码标签，第 i 列为 1 表示数字 i-1（如第 1 列为 1 表示数字 0）。

**3.3 快速检查（Python 3 代码，实验前运行）**

|  |
| --- |
| import numpy as np  import matplotlib.pyplot as plt #注意依赖库的安装  raw = np.loadtxt('semeion.data.txt')  X, y = raw[:, :256], raw[:, 256:]  y = np.argmax(y, axis=1) # 独热 → 0~9 整数  print('样本数:', len(X), '像素数:', X.shape[1])  # 随机画 6 张图  for i in range(6):  plt.subplot(2,3,i+1)  plt.imshow(X[i].reshape(16,16), cmap='gray')  plt.title(y[i]); plt.axis('off')  plt.tight\_layout(); plt.show() |

运行后应看到 6 张清晰数字，若无图请检查文件路径。

**4 实验环境推荐**

|  |  |
| --- | --- |
| 推荐使用 | 依赖库安装命令（非唯一） |
| Python | pip install numpy matplotlib scikit-learn |

**5 基础任务：自写 kNN + LOO**

**5.1 算法提示**

1. 距离：欧氏距离 √(∑(xi − xj)²)
2. 投票：找出 k 个最近邻居后，标签多数表决（可平分时随机）
3. LOO：共 1593 轮，每次留 1 行做测试，其余 1592 行做训练

**5.2 引导代码（仅供参考）**

|  |
| --- |
| import numpy as np  ## import... 安装所需要的依赖库  def loo\_eval(X, y, k):  # 这部分为实验的核心代码  return acc  # 主流程  raw = np.loadtxt('semeion.data.txt')  X, y = raw[:, :256], np.argmax(raw[:, 256:], 1)  for k in [1, 3, 5]:  acc = loo\_eval(X, y, k)  print(f'k={k} LOO 准确率 = {acc:.4f}') |

预期输出：

k=1 LOO 准确率 = 0.9855

k=3 LOO 准确率 = 0.9868

k=5 LOO 准确率 = 0.9849

**6 Weka 对比（中级要求）**

1 打开 Weka → Explorer → Preprocess → Open file → 选 semeion.data.txt

2切换至 Classify 面板：

* 选择 classifier: lazy → IBk
* 点击 IBk 行，设置 KNN=1 → OK
* Test options: Leave-One-Out (注意 Weka 显示为 “Cross-validation folds = 1593”)
* Start，记录 “Correctly Classified Instances” 百分比

3 重复 6.2 步，把 KNN 改为 3、5，填入下表：

|  |  |  |  |
| --- | --- | --- | --- |
| k | 自写 LOO 精度 | Weka LOO 精度 | 差异（绝对值） |
| 1 |  |  |  |
| 3 |  |  |  |
| 5 |  |  |  |

若差异 >0.5%，请给出 1-2 条可能原因（如距离定义、投票平局处理等）。

**7 提交清单**

（按“学号\_姓名\_班级\_实验1.zip”命名，发送至邮箱**13031090911@163.com**）

├─ exp\_1.py （基本要求,完成5基础任务的**完整源码**）

└─ 实验报告 （文档格式，基本要求：参考本模版，对各部分**实验结果截图**、**表格**、**关键代码**，以及所**提出问题**进行补充回答。进阶要求：各实验板块思考过程、具体实现步骤、踩坑记录）

**8 学术诚信提醒**

• 允许讨论思路，禁止直接复制他人源码；

• 引用网络代码请加注释并给出 URL；

**祝实验顺利！**