单细胞转录组特征基因筛选实验

机器学习大作业

2025 年 12 月 23 日

目录

实验说明

实验选择

本实验提供两个难度相当的单细胞转录组特征筛选题目，学生可任选其一完成。两个实验均基于相同的乳腺癌单细胞转录组模拟数据集，但采用不同的技术路线和方法论。

实验一：基于多模态机器学习的特征基因筛选
侧重于传统统计方法、机器学习方法和深度学习方法的集成应用，遵循经典分析路线。

实验二：基于深度自监督学习的特征基因筛选
侧重于深度学习和自监督学习技术，探索人工智能在生物医学数据分析中的前沿应用。

评分结构：每个实验分为初级要求（30 分）、中级要求（50 分）和高级要求（20 分），总计 100 分。初级和中级要求为基础部分，高级要求为拓展部分。

实验背景

什么是单细胞转录组测序？

单细胞转录组测序（scRNA-seq）是一种革命性的生物技术，它能够在单个细胞水平上测量基因的表达量。与传统的批量测序不同，scRNA-seq 可以揭示细胞群体中的异质性，识别稀有细胞类型，并理解细胞状态转变的动态过程。

表达矩阵是什么？

单细胞转录组数据通常表示为**表达矩阵**，其维度为 $n \times m$，其中：

- n：细胞数量（本实验：约 10,000 个细胞）
- m：基因数量（本实验：约 15,000 个基因）
- 每个元素：表示特定细胞中特定基因的表达量（通常为经过标准化处理的计数）

乳腺癌及其亚型简介

乳腺癌是女性最常见的恶性肿瘤之一，具有高度异质性。根据分子标志物的表达，乳腺癌主要分为四种亚型：

亚型	特征描述
ER+（Luminal A/B）	雌激素受体阳性，占乳腺癌的 60-70%，预后相对较好，对内分泌治疗敏感
HER2+	人表皮生长因子受体 2 过表达，占 15-20%，侵袭性强，但靶向治疗（如赫赛汀）有效
TN（三阴性）	雌激素受体、孕激素受体、HER2 均为阴性，占 10-15%，侵袭性最强，缺乏靶向治疗，预后最差
ER+HER2+（双阳性）	同时表达雌激素受体和 HER2，占 5-10%，兼具两种亚型特点，需联合治疗

表 1: 乳腺癌主要亚型特征

不同亚型在基因表达模式、治疗反应和预后方面存在显著差异，因此准确识别亚型对于个性化治疗至关重要。

特征基因筛选的重要性

在高维单细胞数据中，大多数基因是噪声或不相关的。特征基因筛选的目标是：

1. **降维：**从数万个基因中筛选出数百个关键基因
2. **分类性能：**提高下游分析（如细胞亚型分类）的准确性和效率
3. **计算效率：**减少计算资源和时间消耗
4. **方法学验证：**比较不同特征筛选算法的有效性

关于模拟数据的说明

本实验使用的是乳腺癌单细胞转录组**模拟数据**，并非真实生物数据。数据中故意添加了噪声细胞（非四种标准亚型）和噪声基因（全零表达）。数据清洗是实验的重要环节，必须严格按照要求进行。由于是模拟数据，**不要求生物学意义解释**，重点关注方法学比较。

1 实验一：基于多模态机器学习的特征基因筛选

1.1 考察意义

本实验旨在通过系统比较不同方法论框架下的特征基因筛选效果，培养学生以下能力：

1. **方法掌握能力：**系统掌握统计检验、机器学习、深度学习三种方法论的特征筛选技术
2. **算法设计能力：**通过自行设计基于聚类一致性的特征筛选算法，提升数学建模和算法创新能力
3. **评估比较能力：**构建统一的评估框架，客观比较不同方法的优劣，培养科学评估思维
4. **工程实践能力：**完成从数据清洗、特征筛选到结果评估的完整数据分析流程
5. **问题解决能力：**针对模拟数据中的噪声问题，设计有效的数据清洗策略

通过本实验，学生将深入理解不同特征筛选方法的原理、适用场景和局限性，为后续的科研和工程实践奠定坚实基础。

1.2 实验核心目标

1. 系统实现并比较统计检验、机器学习、深度学习三种方法论的特征筛选效果
2. 自行设计基于聚类一致性的特征筛选算法，体现数学建模能力
3. 构建统一的评估框架，客观比较不同方法的优劣
4. 掌握单细胞模拟数据清洗的基本方法

1.3 数据集信息

乳腺癌单细胞转录组模拟数据集

项目	描述
数据格式	`.h5ad` (AnnData 格式)
原始细胞数量	约 10,000 个细胞
原始基因数量	约 15,000 个基因
乳腺癌亚型	TN, ER+, HER2+, ER+HER2+
数据特点	**模拟数据**，包含非四种亚型的噪声细胞和全零表达的噪声基因
数据质量	需要清洗：去除无效细胞和基因

表 2: 乳腺癌单细胞转录组模拟数据集基本信息

1.4 实验内容与要求

1.4.1 初级要求：数据清洗与基础特征筛选方法实现（30 分）

核心目标：掌握单细胞数据清洗、预处理和基础统计方法。

任务 1：数据清洗与预处理

完成数据清洗与预处理工作，具体包括：使用 scanpy 加载.h5ad 格式数据，统计原始数据的细胞数、基因数、亚型分布以及数据稀疏度。在数据清洗环节，需要只保留四种标准亚型的细胞，移除全零表达的基因以及低表达基因，同时过滤掉总表达量异常的低质量细胞。数据预处理阶段需进行归一化处理，包括 CTPM 方法和对数转换，并筛选出 500 个高变异基因。此外，要求绘制清洗前后的亚型分布对比条形图、基因表达量分布直方图，并使用 PCA 降维展示细胞分布。

任务 2：基于统计检验的特征筛选

完成基于统计检验的特征筛选任务，主要内容包括：使用 Wilcoxon 秩和检验识别各亚型的差异表达基因，针对每个亚型进行一对多比较，并使用 FDR 方法进行多重检验校正。要求每个亚型筛选出前 50 个差异表达基因，按 log2 Fold Change 排序，绘制火山图展示差异表达结果，并提供包含基因名、校正后 p 值、log2FC 的基因列表。

1.4.2 中级要求：多方法特征筛选集成（50 分）

核心目标：集成机器学习和深度学习方法，构建多模态特征筛选框架。

任务 3：基于机器学习的特征筛选

实现基于机器学习的特征筛选方法，包括：训练多分类随机森林模型并使用 Gini 重要性评估特征重要性，配置 XGBoost 多分类模型并使用 gain、cover 和 weight 三种重要

性类型进行评估。需要设计多方法集成策略，结合随机森林和 XGBoost 的重要性评分，通过加权集成策略筛选出前 100 个特征基因。

任务 4：基于深度学习的特征筛选

实现基于深度学习的特征筛选方法，包括：构建多层感知机分类模型并应用 Dropout、BatchNorm 等正则化技术；实现基于梯度的特征重要性评估，计算输入特征对模型输出的梯度，使用梯度绝对值或平方作为特征重要性指标。要求比较深度学习方法与机器学习方法的筛选结果重叠率，并分析不同方法的优缺点。

1.4.3 高级要求：基于聚类一致性的创新算法设计（20 分）

核心目标：自行设计数学建模方法，基于聚类一致性原理筛选特征基因。

任务 5：聚类一致性算法设计与实现

自行设计并实现基于聚类一致性的特征基因筛选算法。算法设计思想基于以下核心假设：优秀的特征基因应在不同聚类算法中稳定地展现出区分能力。具体实现时需要通过多种聚类算法验证基因的表达模式一致性，并设计评分函数结合基因表达模式与聚类稳定性。需要实现聚类一致性评分函数，使用多种聚类算法计算基因表达值的组间/组内方差比，评估基因在不同亚型中的表达特异性，并加权综合各维度得到最终一致性得分。要求将新方法与传统方法在独立验证集上比较，分析算法的时间复杂度和空间复杂度，并讨论算法的创新点和潜在改进方向。

1.5 评估与提交

1.5.1 统一评估框架

1. **评估指标**
    - 分类性能：准确率、F1 分数、精确率、召回率
    - 特征稳定性：多次运行结果的 Jaccard 相似度
2. **评估流程**
    - 使用 5 折交叉验证评估分类性能
    - 比较不同方法筛选的特征基因集合
    - 分析特征数量与分类性能的关系
3. **可视化要求**

- 绘制各方法分类性能对比柱状图
- 绘制特征重叠 Venn 图
- 绘制学习曲线（深度学习）
- 绘制基因表达热图（top 特征基因）

1.5.2 提交要求

- **实验报告（PDF 格式）**
  - 实验设计与方法描述（需包含算法设计思路）
  - 完整的结果分析与可视化
  - 方法比较与讨论
  - 创新点总结与改进方向
- **源代码**
  - 完整可执行的 Python 代码
  - 良好的代码结构和注释
  - README 文件说明运行环境和方法
- **提交格式**
  - 压缩包文件名：组长姓名 _ 实验一.zip
  - 提交至指定邮箱，附小组成员信息（学号 + 姓名）

2 实验二：基于深度自监督学习的特征基因筛选

2.1 考察意义

本实验旨在探索自监督学习在特征基因筛选中的应用，培养学生以下能力：

1. **前沿技术掌握：**掌握自监督学习、对比学习、提示学习等深度学习前沿技术
2. **表征学习能力：**理解如何从无标签数据中学习有意义的特征表示
3. **少样本学习能力：**设计提示学习框架，实现在有限标签数据下的高效微调
4. **端到端建模能力：**构建完整的深度学习工作流，从数据预处理到模型评估
5. **创新应用能力：**将计算机视觉领域的先进方法迁移到生物信息学问题

通过本实验，学生将掌握深度自监督学习在生物医学数据分析中的最新应用，了解如何利用无标签数据提升特征学习效果，为处理实际生物医学数据中的标签稀缺问题提供解决方案。

2.2 数据集说明

项目	描述
主数据文件	`breast_cancer_scRNA_simulated.h5ad`
少样本划分	`few_shot_splits/` 目录下
细胞总数	10,000 个细胞
基因总数	15,000 个基因
有标签细胞	约 9,000 个（四种标准亚型）
无标签细胞	约 1,000 个（"Unknown" 亚型）

表 3: 实验二数据集配置

少样本学习划分：

- `10_percent.npy`：10% 有标签数据（约 4,750 个细胞）
- `30_percent.npy`：30% 有标签数据（约 14,250 个细胞）
- `50_percent.npy`：50% 有标签数据（约 23,750 个细胞）
- `all.npy`：全部有标签数据（约 47,500 个细胞）

2.3 实验核心目标

1. 掌握自监督学习在单细胞数据分析中的应用原理
2. 实现基于对比学习和自编码器的无监督特征提取方法
3. 设计端到端的深度学习特征筛选框架
4. 理解表征学习对特征发现的增强作用
5. 设计提示学习框架，实现在少量标签数据下的高效微调

2.4 实验内容与要求

2.4.1 初级要求：自监督预训练模型构建（30 分）

核心目标：构建自编码器模型，学习单细胞数据的压缩表示。

任务 1：自编码器设计与实现

构建标准的自编码器模型，设计合适的编码器和解码器结构，实现重构损失函数并训练稳定的编码器-解码器架构，保存训练好的编码器模型用于特征提取。同时实现基因表达数据的增强方法，包括高斯噪声添加和随机掩码，并设计增强参数的可调节性。

2.4.2 中级要求：对比学习与特征重要性评估（40 分）

核心目标：实现对比学习框架，并通过注意力机制评估特征重要性。

任务 2：SimCLR 对比学习实现

实现 SimCLR 对比学习框架，包括数据增强生成正样本对、共享编码器结构、投影头将特征映射到对比空间，使用 InfoNCE 损失函数优化表征空间并设置温度参数。

任务 3：特征重要性评估模块

在预训练编码器后添加自注意力层，计算基因级注意力权重作为重要性分数，实现注意力可视化并基于注意力权重筛选前 50 个重要基因。同时实现集成梯度方法，计算输入特征对表征空间的贡献度，基于梯度重要性筛选前 50 个重要基因，并比较注意力与梯度方法筛选结果的重叠率。

2.4.3 高级要求：提示学习与端到端评估（30 分）

核心目标：设计提示学习框架，实现在少量标签数据下的高效微调，并进行全面评估。

任务 4：提示学习框架设计

设计提示学习框架，包括冻结预训练编码器的所有参数，设计可学习的提示向量，实现提示向量与编码特征的拼接操作，添加轻量级分类头，并仅训练提示向量和分类头参数。

任务 5：全面性能评估与解释

建立多维度评估体系，包括分类性能、聚类质量、特征稳定性和计算效率的评估。需要实现交叉验证评估分类性能，计算调整兰德指数和标准化互信息评估聚类质量，通过不同随机种子下的 Jaccard 相似度评估特征稳定性。

2.5 评估与提交

2.5.1 评估标准

1. **评估说明**：由于所使用的数据集为虚拟数据集，因此不再考虑相关指标优劣，重点在于实现完整的工作流程；

2.5.2 可视化要求

- **学习曲线：**预训练和微调过程中的损失变化
- **特征重要性热图：**展示 top 特征基因的重要性分数
- **降维可视化：** UMAP/t-SNE 展示学习到的特征表示
- **注意力可视化：**展示注意力权重在基因上的分布

2.5.3 提交要求

- **实验报告（PDF 格式）**
  - 完整的技术路线和方法描述
  - 详细的实验结果
  - 创新点总结和未来改进方向
- **源代码**
  - 完整的 PyTorch/TensorFlow 实现
  - 模块化设计，包含数据加载、模型定义、训练、评估等模块
  - 详细的 README 说明文件

- **提交格式**
  - 压缩包文件名：组长姓名 _ 实验二.zip
  - 提交至指定邮箱，附小组成员信息（学号 + 姓名）

通用要求与注意事项

数据清洗要点

1. **细胞过滤：**
   - 移除无亚型标签或标签不在四种标准亚型中的细胞
   - 移除总表达量过低或过高的细胞（技术异常）
2. **基因过滤：**
   - 移除在所有细胞中表达量为零的基因
   - 移除在少于 10 个细胞中表达的基因（低表达基因）
3. **质量控制：**
   - 记录清洗前后的数据维度变化
   - 可视化清洗效果（亚型分布、表达量分布等）

实验环境建议

基础环境配置

```
# 核心数据科学库
numpy >= 1.21.0
pandas >= 1.4.0
scipy >= 1.8.0
scikit-learn >= 1.0.0

# 单细胞分析
scanpy >= 1.9.0
anndata >= 0.9.0

```

```
# 可视化
matplotlib >= 3.5.0
seaborn >= 0.11.0
```

Listing 1: 基础环境配置

实验一专用环境

```
# 机器学习库
xgboost >= 1.6.0

# 统计和可视化
statsmodels >= 0.13.0

# 深度学习（可选）
tensorflow >= 2.8.0  # 或 pytorch >= 1.11.0
```

Listing 2: 实验一环境配置

实验二专用环境

```
# 深度学习核心
torch >= 2.0.0

# 模型解释
captum >= 0.6.0

# GPU支持
CUDA >= 11.7（推荐）
```

Listing 3: 实验二环境配置

祝同学们实验顺利！