课程作业：scRNA-seq 数据表征学习与细胞聚类

2025 年 11 月 5 日

目录

1 作业目标

单细胞 RNA 测序 (scRNA-seq) 技术使我们能够在单个细胞的分辨率上研究基因表达，但其产生的数据具有**高维性**、**稀疏性**和**高噪音**的特点。本作业的目标是探索和比较不同的特征表征学习（representation learning）算法，将高维的基因表达数据压缩到低维的“嵌入”（embedding）空间中。

你们将使用这些算法学习到的嵌入向量进行无监督聚类，并使用提供的真实细胞类型注释来评估聚类效果。通过本作业，你将：

- 掌握处理 scRNA-seq 数据的基本预处理流程。
- 实现并应用 PCA、自动编码器 (AE)、变分自动编码器 (VAE) 以及一个专门为 scRNA-seq 数据设计的 ZINB-VAE 模型。
- 学习如何使用下游任务（聚类）来评估上游表征学习算法的质量。
- 深入理解不同算法背后的数学原理及其在生物数据上的优缺点。

2 数据集

你将获得一个 `AnnData` 格式（`.h5ad` 文件）的 scRNA-seq 数据集。该数据包含：

- `adata.X`：原始的基因表达计数矩阵（稀疏矩阵）。
- `adata.obs['cell_type']`：真实的细胞类型注释。

重要提示：`adata.obs['cell_type']` 标签**严禁**用于模型训练（这是一个无监督学习任务）。它**只能**在最后一步用于评估聚类结果。

3 任务要求

3.1 任务一：数据预处理

针对不同的模型，你可能需要不同的预处理策略。请为每个模型明确说明你的预处理步骤。数据预处理请参考 Scanpy 教程（参考文献），需要和 Scanpy 教程完全一致。

1. **用于 PCA, AE, VAE**：
    - 对细胞和基因进行基本过滤。
    - 进行标准化（例如，log1p 转换后的 CPM/TPM）。
    - （可选）选择高可变基因 (Highly Variable Genes, HVGs)。

- （可选）对数据进行缩放 (Scaling)。
- **请在报告中详细说明并论证你的选择**。

2. **用于 ZINB-VAE**：
   - ZINB 类模型通常直接在**原始计数矩阵 (raw counts)** 上进行训练。
   - 你可能仍需要进行细胞和基因过滤。

3.2 任务二：模型实现与表征学习

你需要实现以下四种算法来学习细胞的低维嵌入（例如，统一学习一个 32 维的嵌入向量）。

1. **基线模型：主成分分析 (PCA)**
   - 使用 `sklearn.decomposition.PCA`。
   - 在预处理后的数据上训练 PCA。
   - 提取前 k (例如 $k = 32$) 个主成分作为细胞的嵌入。

2. **自动编码器 (Autoencoder, AE)**
   - 设计并实现一个多层感知机 (MLP) 结构的 AE（编码器-解码器结构）。
   - 编码器将输入数据压缩到 k 维的潜在空间 (bottleneck)。
   - 解码器尝试从 k 维嵌入中重建原始输入。
   - **损失函数**：使用均方误差 (Mean Squared Error, MSE) 作为重建损失。
   - 训练模型，并提取编码器输出的 k 维嵌入。

3. **变分自动编码器 (Variational Autoencoder, VAE)**
   - 实现一个 VAE。与 AE 的主要区别在于：
     - 编码器输出一个概率分布的参数（均值 z_μ 和对数方差 $z_{\log \sigma^2}$）。
     - 从该分布 $N(z_\mu, z_\sigma)$ 中采样得到潜在嵌入 z。
   - **损失函数 (ELBO)**：
     - **重建损失**：同样，先使用 MSE。
     - **KL 散度**：作为正则化项，惩罚潜在空间分布与标准正态分布 $N(0, I)$ 之间的差异。
     - $Loss = \text{MSE\_Loss} + \text{KL\_Loss}$
   - 训练模型，并使用编码器输出的**均值** z_μ 作为最终的确定性嵌入。

4. **零膨胀负二项 VAE (ZINB-VAE)**

   - 编码器：与标准 VAE 相同，输出 z_μ 和 $z_{\log \sigma^2}$。
   - 解码器：
     - 解码器不再是简单地重建输入，而是输出 **ZINB 分布**的三个参数：
       (a) π (pi)：零膨胀的概率（dropout rate）。
       (b) μ (mu)：负二项分布的均值。
       (c) θ (theta)：负二项分布的离散度 (dispersion)。
   - **损失函数**：
     - **重建损失**：不再是 MSE，而是 **ZINB 分布的负对数似然 (Negative Log-Likelihood)**。
     - **KL 散度**：与标准 VAE 相同。
     - $Loss = \text{ZINB\_NLL\_Loss} + \text{KL\_Loss}$
   - **注意**：该模型应在**原始计数数据**上训练。

3.3 任务三：聚类与评估

对于从上述四种方法中获得的每一种嵌入（PCA, AE, VAE, ZINB-VAE）：

1. **聚类**：

   - 在 k 维嵌入上运行一个标准的聚类算法（例如 `KMeans`）。
   - **提示**：你可以通过 `adata.obs['cell_type'].nunique()` 获取真实细胞类型的数量 N_{clusters}，并将 `KMeans` 的簇数 `n_clusters` 设置为该值。

2. **评估**：

   - 比较聚类算法得到的簇标签 (Predicted labels) 与真实的细胞类型标签 (True labels)。
   - 计算并报告以下两个指标：
     - **调整兰德系数 (Adjusted Rand Index, ARI)**
     - **标准化互信息 (Normalized Mutual Information, NMI)**

4 分析报告要求

请提交一份完整的报告（PDF 格式），必须包含以下内容：

1. **预处理策略**：详细说明你为不同模型（特别是 PCA/AE/VAE vs ZINB-VAE）选择的预处理步骤及理由。

2. **模型架构**：清晰描述你的 AE, VAE, 和 ZINB-VAE 的网络结构（层数、激活函数等）。

3. **结果比较**：

   - 创建一个表格（使用 `booktabs` 包），汇总四种方法在 ARI 和 NMI 指标上的得分。
   - （推荐）使用 UMAP 或 t-SNE 将四种方法学习到的 k 维嵌入降至 2 维进行可视化，并分别用**真实细胞类型**和**聚类标签**进行着色。

4. **算法分析（重点）**：

   - **PCA vs. AE**：解释线性（PCA）与非线性（AE）降维的区别。为什么 AE 在复杂生物数据上有可能比 PCA 表现更好?
   - **AE vs. VAE**：解释确定性嵌入（AE）与概率性嵌入（VAE）的区别。VAE 中的 KL 散度项起到了什么作用？它如何影响潜在空间的结构?
   - **VAE vs. ZINB-VAE**：**这是本作业的核心**。详细解释为什么在 scRNA-seq 数据上，使用 ZINB 分布作为重建损失（如 ZINB-VAE）通常远优于使用 MSE（如标准 VAE）?（提示：请围绕 scRNA-seq 数据的统计特性，如稀疏性/零膨胀、高方差/过离散等进行讨论）。

5 提交内容

请将以下文件打包提交：

1. **代码**：包含所有步骤的 Jupyter Notebook 或 Python 脚本。

2. **分析报告**：包含上述所有分析内容的 PDF 文件。

6 附录：推荐资源

6.1 核心参考文献（模型理论基础）

参考文献

[1] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). **Deep generative modeling for single-cell transcriptomics.** *Nature Methods*. (ZINB-VAE 的首选参考)

[2] Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., & Theis, F. J. (2019). **Single-cell RNA-seq denoising using a deep count autoencoder.** *Nature Communications*. (ZINB-AE 的清晰论述)

6.2 数据预处理教程（实战指南）

推荐使用 **Scanpy** (Python) 库完成预处理。

- **Scanpy 官方标准流程（推荐）：**
  - **链接:** https://scanpy.readthedocs.io/en/latest/tutorials/basics/clustering.html
  - **简介:** 涵盖了从过滤、标准化 (`normalize_total`, `log1p`)、寻找高可变基因 (`highly_variable_genes`) 到缩放 (`scale`) 和 PCA (`pca`) 的完整流程。
  - **注意: PCA, AE, VAE** 应在 标准化、Log 转换并缩放 的数据上训练（通常使用 HVGs）。**ZINB-VAE** 应在 原始计数矩阵 上训练（仅进行基本过滤）。

- **单细胞分析最佳实践:**
  - **链接:** https://www.sc-best-practices.org/
  - **简介:** 由 Scanpy 开发者维护的网站，提供了每一步预处理（QC、标准化等）背后的“为什么”和最新共识。