想用单细胞表达信息预测患者分组，训练集和测试集该怎么拆?

从同一份模拟表达表按细胞和按患者分别做交叉验证，逐折导出重叠患者、训练特征和每个被留出患者的概率。



运行教学示例

Python 3.10 或以上、pandas、numpy、matplotlib、scikit-learn，见 requirements.txt。

```
python run.py --input example --out results
```

在解压后的资料包根目录运行。example 是教学输入，results 保存本次输出，expected 保存可供核对的参考结果。

输入字段

cells.tsv.gz、make_example.py 模拟计数、患者身份、真实组别和可修改的个体指纹生成器。

核对输出，再换自己的数据

split_audit.tsv、training_features.tsv 逐折患者交集、训练折内选择的 35 个基因。

cell_oof.tsv、patient_oof.tsv、metrics.tsv 折外细胞与患者概率、患者层 AUC 和均衡准确率。

用示例验算

1 随机按细胞拆的四折，每折训练与测试共享全部 24 名患者。

2 按患者拆每折有 18 名训练、6 名测试，交集为零。

3 本次按患者拆分的 AUC 约 0.2986，按细胞拆分为 1。

换用真实数据

按真实预测对象建立分组键，在每个训练折内部完成选基因、缩放、拟合和调参，留出样本只参与最终打分。

弱信号模拟和有限患者数产生的具体 AUC 不是任何疾病模型的性能估计。

方法资料

scikit-learn 分组交叉验证

预处理与数据泄漏