气体光谱 PLS 回归建模项目说明文档

1. 项目背景

本项目基于偏最小二乘回归（PLS）方法，对气体红外吸收光谱数据进行浓度预测分析。目标气体包括 NO2 和 SO2，使用人工合成的混合光谱样本进行建模测试，验证 PLS 在气体定量分析中的适用性。

2. 脚本及功能说明

01\_preprocess.py

功能：

读取原始 NO2 和 SO2 的光谱数据（CSV 文件）

在统一的波数轴（587~3689 cm⁻¹，步长 1 cm⁻¹）上做线性插值

为后续混合和建模打基础，保证相同波数点的一一对应

输出：

interpolated\_spectra.csv

列：wavenumber，NO2，SO2

02\_generate\_mixture.py

功能：

使用 01 步插值后的光谱

按 NO2=40%、SO2=60% 比例生成一个混合样本

同时输出混合比例的标签 [0.4, 0.6]

输出：

mixed\_spectrum.csv（波数+混合谱）

mixed\_concentration.csv（混合比例标签）

03\_generate\_dataset.py

功能：

批量生成多个比例的混合样本（NO2：30%50%，SO2：70%50%，步长0.05）

每种比例生成 10 个带 1%高斯噪声扰动的样本

构成一个可用于回归建模的完整数据集

输出：

X\_dataset.csv（样本矩阵）

Y\_labels.csv（每行一个比例标签）

04a\_pls\_model\_default.py

功能：

从 03 生成的训练集读取数据

使用 PLSRegression 模型，固定 5 个成分，拆分 80%训练+20%测试

输出 RMSE、R²

绘制真实 vs 预测散点图

保存预测结果与评估指标

输出：

Y\_test\_default.csv（真实浓度）

Y\_pred\_default.csv（预测浓度）

evaluation\_default.txt（RMSE 和 R²）

pls\_prediction\_default.png（可视化散点图）

04b\_pls\_model\_custom.py

功能：

同样从 03 的数据集中读取

可配置训练比例（默认70%），以及最大成分数（10），并通过5折交叉验证自动选取最优成分

输出 RMSE、R²、最优成分数

绘制并保存可视化散点图

输出：

Y\_test\_custom.csv（真实浓度）

Y\_pred\_custom.csv（预测浓度）

evaluation\_custom.txt（RMSE、R²、最优成分数）

pls\_prediction\_custom.png（可视化散点图）

3. 项目产出总结

以上五个脚本整体完成： 1. 从原始气体光谱到插值的预处理 2. 生成可控比例的混合样本 3. 形成训练/验证集 4. PLS 模型的拟合、预测及验证

实现对 NO2、SO2 混合比例的定量预测。当前是基于人工合成光谱，未来若用于真实采样，还需结合标气进行实验验证。