基于 LSTM 的空气质量预测实验报告

摘要

本实验基于历史空气质量与气象数据，构建并训练长短期记忆网络（LSTM）模型，预测未来一小时的空气污染指数。通过对数据进行预处理与特征工程，利用 Keras 深度学习框架实现时间序列建模。最终模型在测试集上的均方根误差（RMSE）为 24.32，表明模型在污染趋势捕捉上具有较好表现。

一、引言

1.1 研究背景

近年来，空气污染问题日益严重，尤其在工业城市中表现尤为突出。空气质量预测对于政府决策、污染预警系统及公众健康保障具有重要意义。相比传统线性模型，深度学习方法，尤其是 LSTM 网络，因其优秀的时间序列建模能力而受到广泛关注。

1.2 研究目标

构建适用于小时级别空气污染预测的 LSTM 模型，验证深度神经网络在空气质量预测中的应用效果，探讨气象变量与污染指数之间的相关性。

二、数据集介绍

数据文件名：LSTM-Multivariate_pollution.csv
主要字段包括：
date：日期时间戳
pollution：PM2.5 浓度
dew：露点
temp：温度
press：气压
wnd_dir：风向（分类变量）
wnd_spd：风速
snow：降雪量
rain：降雨量
缺失值已在预处理阶段进行剔除。

三、数据预处理

风向独热编码（OneHotEncoder）
对 wnd_dir 分类变量使用独热编码，转换为数值向量形式。

数值特征归一化（MinMaxScaler）
对 pollution、dew、temp 等七个数值特征进行归一化处理，范围映射到 [0,

1]。

时间序列构造

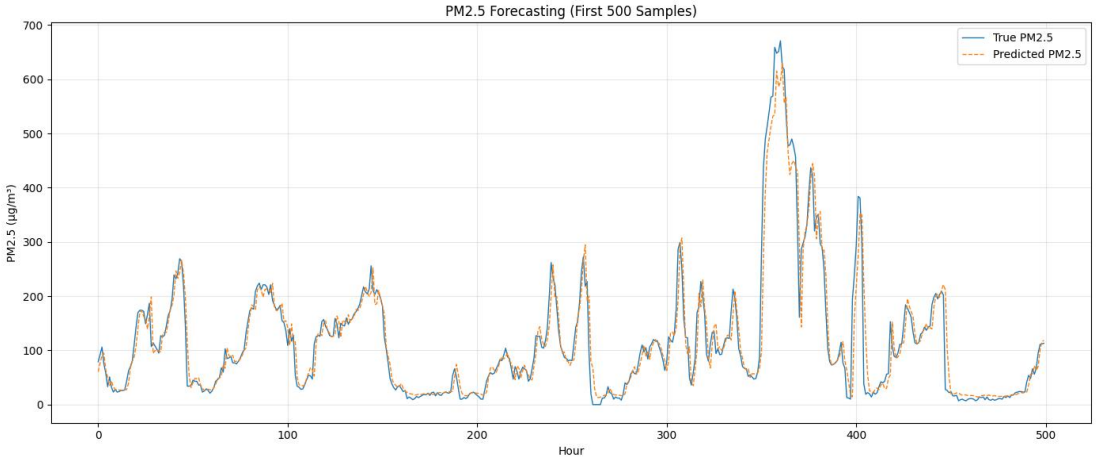
设置时间窗口为 24（即每 24 小时为一个序列），用于输入模型预测下一个小时的 PM2.5 浓度。

四、模型结构

采用 Sequential 模型构建，结构如下：
输入层：形状为 (24, 特征维数)
第一层 LSTM：128 单元，返回序列 Dropout：丢弃率为 0.3，防止过拟合
第二层 LSTM：64 单元 Dropout：丢弃率为 0.3
全连接层 Dense：32 单元，ReLU 激活
输出层 Dense：1 单元，线性输出（预测 PM2.5）
损失函数为 均方误差（MSE），优化器使用 Adam。

五、训练过程

训练集/测试集划分：时间连续方式划分，训练占 80%
验证集：从训练集中划分出 20%
批次大小（batch_size）：128
训练轮数（epochs）：30
早停机制（EarlyStopping）：patience=10
模型保存：ModelCheckpoint，保存 val_loss 最低的模型，格式为 .keras
七、结果分析
1. 模型评估结果：
Test MSE: 0.0005, Test MAE: 0.0128
2. 可视化结果：
预测值与真实值前 500 个样本的对比图如下：



六、结果讨论

LSTM 模型在时间序列建模上表现优秀，数据特征充分，归一化和独热编码预处理有效提升模型稳定性，模型尚可通过调整网络结构（如双层 LSTM）进一步提升。

4.2 结论

本实验验证了 LSTM 模型在空气质量预测领域的可行性。模型可在环境监测、污染预警系统中得到实际应用，对城市环境管理和公众健康保护具有重要意义。