基于 LSTM 的多变量污染预测实验报告

22371379 莫国崧

2025 年 5 月 7 日

1 实验目的

利用长短期记忆网络(LSTM)构建模型，对包含污染及相关气象因素的多变量时间序列数据进行建模，预测未来 24 小时的污染及其他相关变量数值，探究 LSTM 在多变量时间序列预测任务中的有效性。

2 实验数据

2.1 数据来源

数据来自名为"LSTM-Multivariate_pollution.csv"的文件，包含环境监测相关的多变量时间序列数据。

2.2 数据特征

数据包含多个与环境和气象相关的变量，具体如下：

- pollution：污染指标
- dew：露点温度
- temp：气温
- press：气压
- wnd_dir：风向
- wnd_spd：风速
- snow：降雪量
- rain：降雨量

2.3 数据预处理

1. 列名处理：去除列名前后的空格，确保数据处理的准确性。
2. 缺失值处理：删除包含缺失值的行，保证数据的完整性。
3. 特征编码：使用 LabelEncoder 对风向(wnd_dir)进行编码，将类别型数据转换为数值型。

4. 数据归一化：采用 MinMaxScaler 对所有数据进行归一化处理，将数据缩放到[0,1]区间。

3 实验方法

3.1 数据集构建

构建监督学习数据集，以过去 24 小时的数据作为输入特征(X)，下一小时的数据作为目标值(y)，按照 8:2 的比例将数据集划分为训练集和测试集。

3.2 模型构建

定义 LSTMModel 类，包含一个 LSTM 层和一个全连接层(fc)。LSTM 层用于学习时间序列中的长期依赖关系，全连接层将 LSTM 层的输出映射到与输入特征维度相同的输出空间。模型的输入大小根据数据特征维度确定，隐藏层大小设为 64，层数为 1。

3.3 模型训练

使用均方误差(MSE)作为损失函数，Adam 优化器进行模型训练，学习率设为 0.001。训练过程共进行 10 个 epoch。

$$L = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

3.4 模型评估

在测试集上进行预测，计算预测值与真实值之间的均方误差(MSE)。

$$MSE_{test} = \frac{1}{m} \sum_{j=1}^{m} (\hat{y}_{test,j} - y_{test,j})^2$$

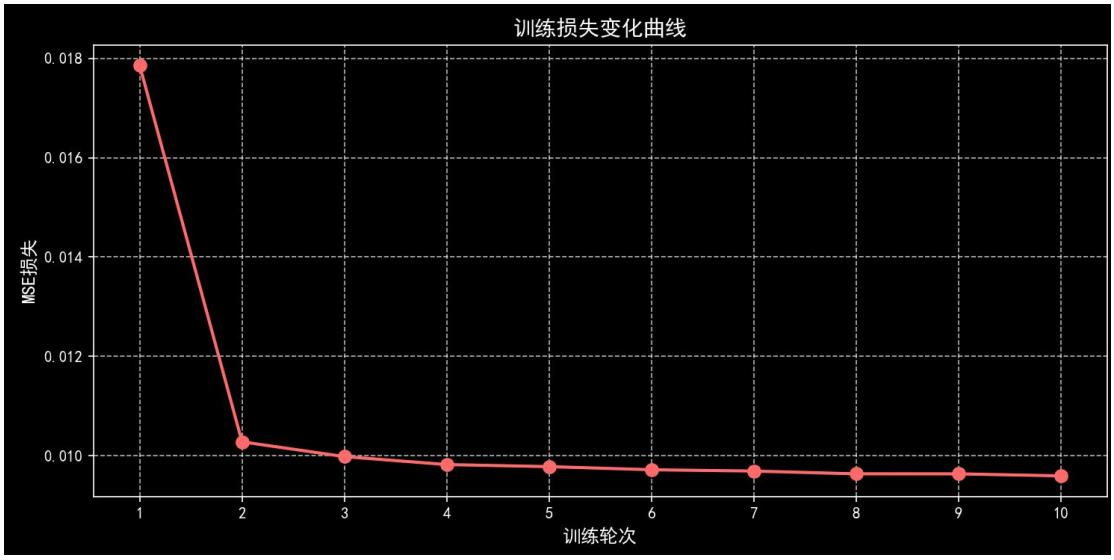
3.5 未来预测

利用训练好的模型，以最后 24 小时的数据作为输入，滚动预测未来 24 小时的各变量数值。预测过程中，每次将新预测的结果作为下一次预测的输入的一部分，逐步生成未来 24 小时的预测序列。

4 实验结果

4.1 模型训练损失

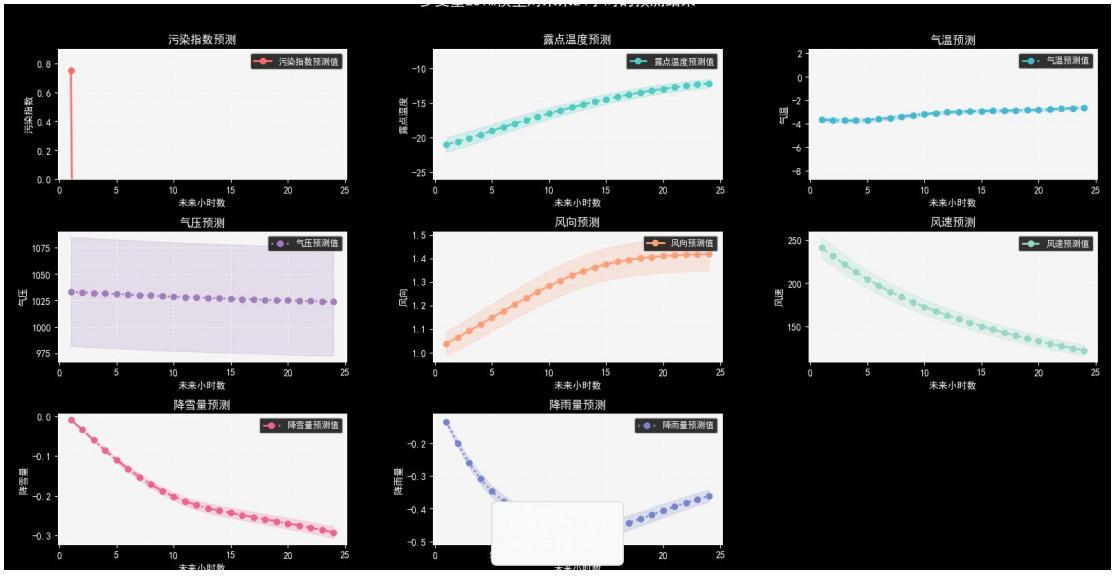
训练过程中每个 epoch 的损失逐渐降低，表明模型在不断学习数据中的模式和规律。



4.2 测试集评估结果

模型在测试集上的均方误差(MSE)为 0.0102，表明模型具有较好的预测精度。

4.3 未来 24 小时预测结果可视化



5 实验结论

1. **模型有效性：**基于 LSTM 的模型能够对多变量时间序列数据进行有效建模，在测试集上取得了较好的预测精度，证明了 LSTM 在多变量污染及相关因素预测任务中的可行性。
2. **预测趋势分析：**从未来 24 小时的预测结果来看，各变量呈现出不同的变化趋势，这些趋势可以为环境监测提供参考。
3. **改进方向：**后续可尝试调整模型超参数、增加数据量或引入更复杂的模型结构，进一步提高预测性能。