Assignment 3

Abstract

本实验利用来自 Kaggle 的北京空气质量数据集，预测未来一小时的 PM2.5 浓度。该数据集包含了五年间逐小时记录的天气和污染信息，包括温度、露点、气压、风速风向、降雪与降雨时长等变量。该问题建模为一个多变量时间序列预测任务，并使用长短期记忆网络（**LSTM**）来捕捉时间依赖性与变量之间的关系，从而提高预测的准确性。

Introduction

本实验使用的空气质量数据集记录了北京地区每小时的气象和污染信息。数据集包含多个特征，包括 PM2.5 浓度（污染值）、露点、气温、大气压力、风速风向、降雪与降雨时长等。为了进行有效的预测，首先需要对数据进行预处理。

在数据预处理过程中，对输入的数值特征进行标准化处理，并对风向进行 one-hot 编码，对其它数据进行数值归一化。

数值特征归一化：

数值特征包括 PM2.5 浓度（`pollution`）、气温（`temp`）、露点（`dew`）、大气压力（`press`）、风速（`wnd_spd`）以及降雪降雨时长（`snow`, `rain`）。采用 **MinMaxScaler** 将这些特征归一化到 [0, 1] 区间。

$$x_{\text{norm}} = \frac{x_{\text{raw}} - x_{\min}}{x_{\max} - x_{\min}}$$

其中，x_min 和 x_max 分别为该特征的最小值和最大值。

风向的 one-hot 编码：

风向（`wnd_dir`）是一个类别特征，通常采用 one-hot 编码将其转换为向量形式。如果风向有 KK 个不同的类别，需要通过对每个风向类别生成一个长度为类别数量的，其中对应类别位为 1，其余位置为 0：

$$\mathbf{v}_k = [0, 0, \ldots, 1, \ldots, 0]$$

通过以上两步处理，数据集的特征矩阵将是一个数值特征和编码后的风向特征的拼接矩阵，设为：

$$X = [X_{\text{numeric}}\ X_{\text{wind}}]$$

其中，前面为归一化后的矩阵，后面为经过 one-hot 编码后的风向矩阵。

时间序列构建

由于使用的是 LSTM 模型进行时间序列预测，因此需要将数据转换为多个时间步的序列样本。人为设定每个输入序列的长度，则每个输入样本包含过去对应个时间步的数据。

$$\mathbf{X}_t = [x_{t-\text{SEQ\_LEN}+1}, \ldots, x_t]$$

对应的输出是下一个时间步的 PM2.5 浓度

$$pollution\mathbf{y}_t = x_{t+1,\text{pollution}}$$

Methodology

LSTM

LSTM（Long Short-Term Memory）是一种特殊的**循环神经网络（RNN）**，它通过引入“记忆单元”和“门控机制”解决了传统 RNN 中的长期依赖和梯度消失问题，适用于时间序列预测任务。

LSTM 的核心思想是通过**三个门结构**（遗忘门、输入门、输出门）来控制信息在时间序列中的传递。

遗忘门（Forget Gate）

决定当前时间步是否“遗忘”上一时间步的记忆。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

输入门（Input Gate）

决定当前时间步输入的哪些信息用于更新“细胞状态”。

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

更新细胞状态（Cell State Update）

将遗忘部分旧的状态并加入新的信息。

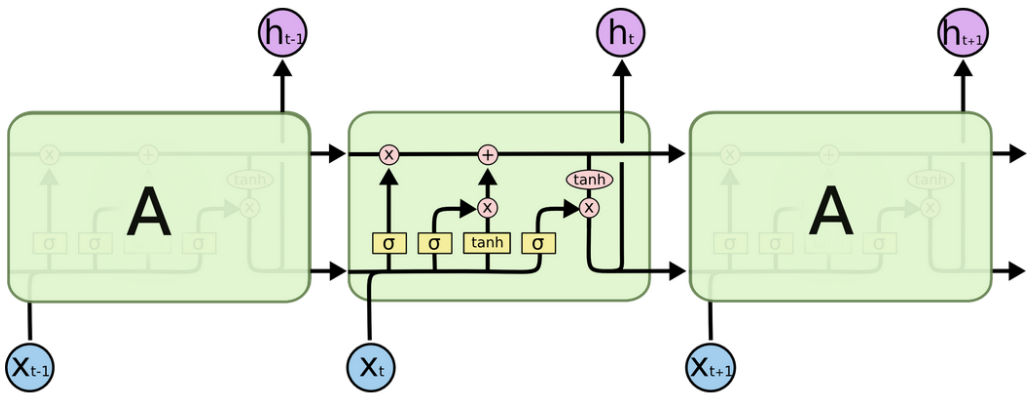
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

输出门（Output Gate）

控制当前时间步的输出内容：

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

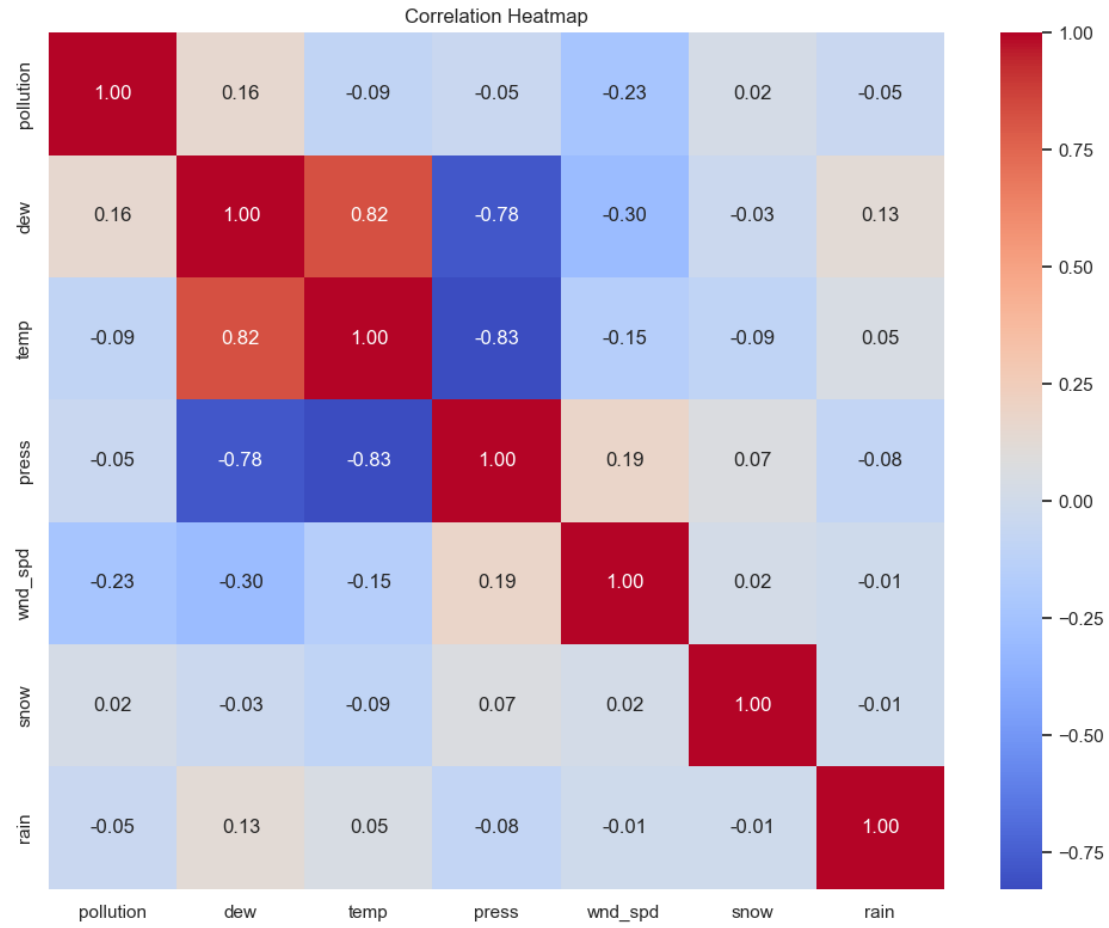


本次实验采用两层LSTM结构，且输出过程仅取最后一个输出，作为预测值。

Experimental Studies

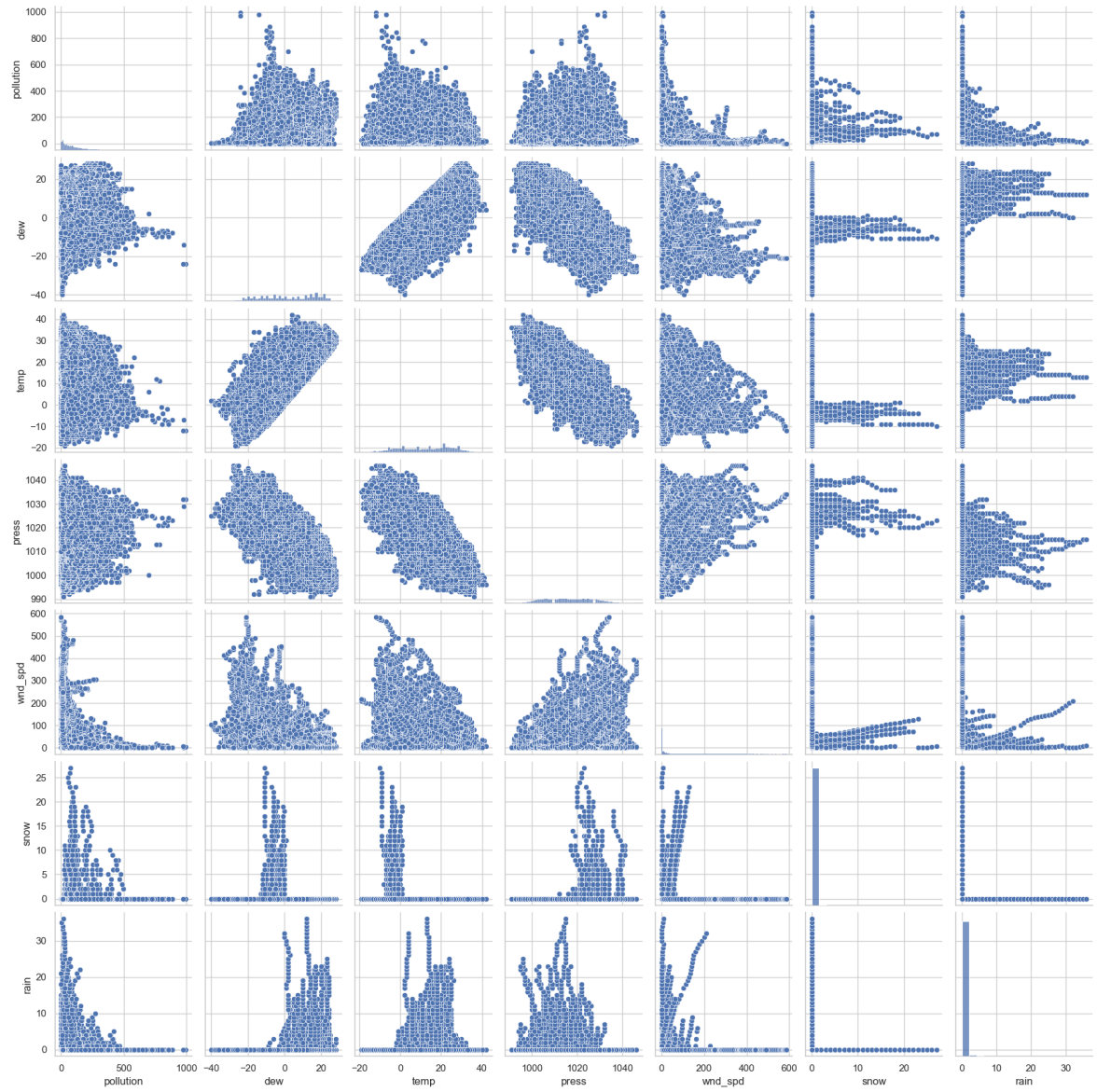
数据分析

对于高维度的数据首先通过相关性计算的得到相关性矩阵。



图中数值越接近1表明正相关相关性越强，越接近-1表示负相关相关性越强。可以看出pollution数值和dew数据正相关性相对较高，和风速数据呈负相关。

再对各种是数据分别两两做散点图观察数据分布情况。

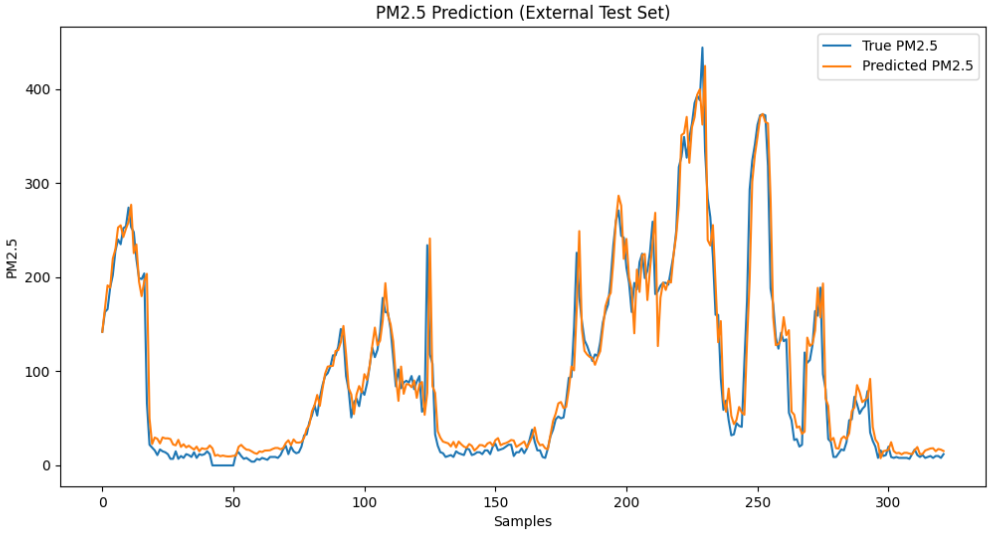


LSTM预测结果

使用两层LSTM网络，利用前一天即24小时数据预测后一小时的pollution程度，训练20轮次，并用训练结果在测试集上进行测试，得到下图结果。可以看到在训练过程中从第4轮开始，损失值基本就降低到一个较小水平，最终在测试数据集上达到

$$MSE = 0.07$$

的较小水平。



Conclusion

本实验围绕空气污染数据展开，利用LSTM神经网络对未来的PM2.5浓度进行时间序列预测。通过对原始数据集进行清洗、归一化和特征编码，构建了适合模型输入格式的多维时间序列数据。特别地，对风向这一类别变量引入了One-Hot编码，有效地提升了模型对风向信息的表达能力。在构建序列样本时，通过固定长度的滑动窗口生成输入-输出对，使得模型能够捕捉过去一段时间内的变化趋势，并据此预测未来的污染水平。

从实验结果来看，LSTM模型在该类任务中表现出良好的预测能力，能够学习出污染值与气象变量之间复杂的非线性关系。相较于传统线性模型，LSTM能够有效捕捉时间上的长期依赖关系，从而更准确地反映污染趋势。这一点在测试集上的表现中得到了验证，说明模型具有一定的泛化能力。

实验也存在一些限制。例如，模型预测的精度在某些突变天气条件下有所下降，说明模型可能尚未充分捕捉极端天气与污染之间的耦合关系。此外，风速、降雨、风向等特征在不同时间段对污染的影响权重可能并不一致，当前的模型还没有引入注意力机制或动态特征加权策略来加以区分，这为进一步优化模型提供了方向。

总体而言，本实验验证了基于LSTM的深度学习模型在空气污染预测任务中的可行性与有效性。可能增加LSTM层数可以使得模型表达能力更强。