# 基于 LSTM 的气温数据时间序列预测

## 1.问题描述

气温预测是气象学领域中的一项重要工作，它在保障社会经济活动正常运行、提高人们的生活质量、保护生态环境等方面都发挥着重要作用。气温预测不仅是一项重要的科学工作，更是社会运行的关键支撑。它帮助我们提前了解天气变化，从而在农业生产中合理安排播种和收获，保障粮食安全；在能源管理方面优化电力和供暖供应，减少资源浪费；在交通运输中提前采取措施保障交通安全。通过精准的气温预测，我们能够更好地保护人类健康，提高生活质量，促进社会和谐稳定。

数据预测研究以 Weather Undergroud API 的天气数据集为基础进行模型学习与数据预测，该数据集提供了 2013 年 1 月 1 日至 2017 年 4 月 24 日每天印度德里市的天气数据，包括气温，湿度，风速，气压 4 组数据。本研究将对其中的气温数据进行预测，模型输入 12 天的历史温度数据，输出未来 1 天的温度预测数据。

## 2.模型介绍

固定地理位置每天的气温数据预测任务具有很强的时序特性，预测模型是否能够高效地捕捉数据中隐藏的时序特征直接影响其预测结果的准确性。与传统的机器学习方法不同，循环神经网络(Recurrent Neural Networks, RNN)在设计之初就用以解决时序问题，RNN 将数据信息分时间步储存与输入，并利用隐藏状态和输入来确定输出。最简单的循环神经网络可以看作是两层全连接神经网络的扩展，在每个序列步骤中，RNN 处理当前时间步输入数据的同时处理过去步的记忆数据，记忆数据包含在上一步隐藏层的神经元中。原则上，具有足够数量隐藏单元的 RNN 可以学习近似任何序列到序列的映射。然而在实际应用中，当时间步长数量较大时，在误差反向传播的过程中，很容易出现梯度爆炸和梯度消失的现象。为解决这一问题，可以在网络的每个时间步单元中，部署一个门控结构来代替简单的完全连接的隐藏层，门控结构可以精细控制信息的流动，既能记住长期的依赖信息，也能忘记无关的细节，从而减少梯度爆炸和梯度消失情况的出现。长短期记忆(Long Short-Term Memory, LSTM)网络是最经典的门控 RNN，也是最常用的时序信息预测模型。

LSTM 网络是于 1997 年提出的 RNN 结构的一种特殊变体。LSTM 结构以 RNN 为基础，采用遗忘门$f_t$、输入门$i_t$和输出门$o_t$来选择性地记忆或遗忘信息。遗忘门会读取上一个时间步的输出和当前时间步的输入，并通过一个 sigmoid 激活函数生成一个 0 到 1 之间的值。这个值用于决定上一个时间步的单元状态中有多少信息需要被遗忘。输入门包括一个 sigmoid 层和一个 tanh 层。sigmoid 层决定哪些值将被更新，而 tanh 层创建一个新的候选值向量，这些值将被加入到状

态中。单元状态通过遗忘门和输入门的组合来更新。遗忘门决定遗忘多少旧信息，输入门决定添加多少新信息。输出门决定最终输出的值。它通过一个 sigmoid 层决定输出值的哪些部分将被保留，并通过一个 tanh 层生成最终的输出值，如下式 1 所示：

$$\begin{cases} i_t = \sigma(W_{ui}u_t + W_{hi}h_{t-1} + b_i) \\ f_t = \sigma(W_{uf}u_t + W_{hf}h_{t-1} + b_f) \\ o_t = \sigma(W_{uo}u_t + W_{ho}h_{t-1} + b_o) \\ C_t = f_tC_{t-1} + i_t * \tanh(W_{uc}u_t + W_{hc}h_{t-1} + b_c) \\ h_t = o_t * \tanh(C_t) \end{cases} \tag{1}$$

其中$u_t$为 t 时刻的输入状态，$W_{xy}$和$b_x$表示 LSTM 结构的权重与偏置参数，$\sigma$表示 sigmoid 函数，$h_t$表示隐藏状态。

在模型建立过程中，可以基于 Pytorch 建立 LSTM 模型对数据特征进行学习并进行未来数据预测。通过定义 LSTM 层的输入特征维度、输出特征维度，隐藏状态维度和网络层数构建基于 LSTM 的时间序列预测模型，并初始化隐藏状态和记忆状态，将原始数据集以 8：2 的比例分为训练集与测试集对模型进行训练与预测精度验证。本研究为了神经网络能够更好地提取数据中的时序特征，在 LSTM 结构前增加了一层一维卷积神经网络(Convolutional Neural Network, CNN)，模型通过一个全连接层进行输出(Fully Connected Layer, FC)，具体模型结构如下图 1 所示。

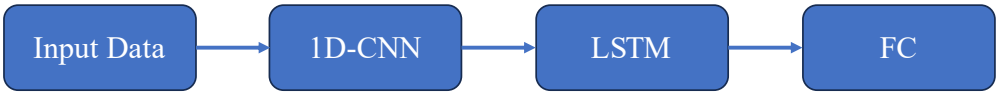



图 1 时间序列预测模型结构

3.预测结果

预测模型采用 Adam 优化器，训练过程中设置最大训练周期为 500，批量大小设置为 64，学习率设置为 0.0001，并采用均方误差(Mean Squared error, MSE)函数计算模型损失。训练过程中的损失记录如下表 1。

表 1 时间序列预测模型训练损失

| Epoch | 0 | 100 | 200 | 300 | 400 |
|---|---|---|---|---|---|
| Train Loss | 0.383 | 0.051 | 0.048 | 0.023 | 0.006 |
| Test Loss | 0.521 | 0.049 | 0.041 | 0.018 | 0.005 |

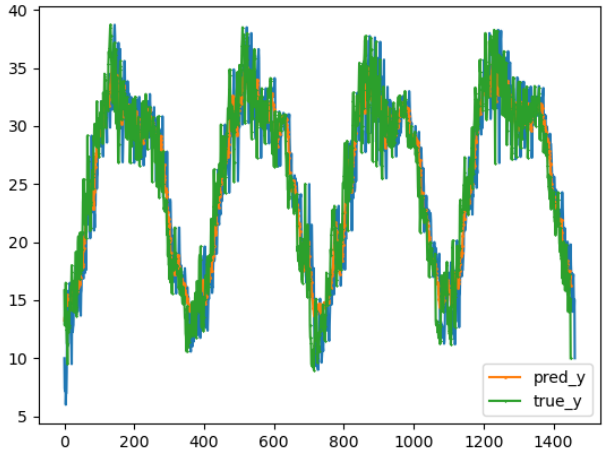
预测后的模型在测试集上计算预测误差为 5.537，误差公式如下式 2 所示：

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

其中 n 表示样本数量，$y_i$与$\hat{y}_i$分别为预测值与真实值。

模型预测值与真实值的对比可见下图 2。

图 2 预测值与真实值对比曲线




基于 LSTM 的气温数据预测模型代码见文件 Prediction_LSTM.py，数据集见文件 DailyDelhiClimateTrain.csv。