Homework 5:
股票价格预测

大数据原理与技术 (Spring 2025)

22336226 王泓沣
Lectured by: Changdong Wang
Sun Yat-sen University

1 问题描述

- 下载某股票历史数据（CSV 格式）
- 用 ARIMA 模型预测未来 7 天价格
- 用 LSTM 模型实现相同任务，对比 MAE/RMSE 指标

2 数据描述

使用 `yfinance` 库下载苹果公司 2020 年 1 月 1 日至 2025 年 3 月 20 日的股票日线价格数据。使用 30 日作为测试集，其余为训练集。

3 Method

3.1 ARIMA 模型

ARIMA（Autoregressive Integrated Moving Average）模型是一种常用且经典的时间序列分析与预测方法。它通过自回归（AR）、差分（I）以及滑动平均（MA）三个部分来对时间序列进行拟合和预测。ARIMA 模型常被记为 ARIMA(p,d,q)

1. 自回归
   AR 部分用过去的自身滞后值来解释序列的变化；若自回归部分的阶数为 p，则说明本期值会受到前 p 期值的影响。
   $$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t$$
   其中 ϕ_i 表示第 i 阶滞后自回归系数,ϵ_t 为白噪声，代表随机扰动

2. 差分
   有些时间序列并不是静止（即方差和均值不随时间变化），而是随时间呈现趋势性或其他非平稳特征。ARIMA 要求时间序列满足弱平稳（stationary），即均值、方差恒定且自相关不随时间变化。因此，需要通过对数据做差分来去除趋势等非平稳成分。若差分次数为 d，表示对序列做了 d 次差分，写作 $\nabla^d X_t$。

3. 滑动平均
   MA 部分用过去的随机误差（噪声）来解释序列当前值的变化；若滑动平均部分的阶数为 q，则表示模型中包含了前 q 个时刻的误差项。
   $$X_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$
   其中：θ_i 为第 i 阶滑动平均系数，ϵ_t 为白噪声项

3.2 LSTM

在时间步 t，LSTM 的更新流程可表示为以下公式（其中符号“*”表示元素逐元素相乘）：

$$
\begin{aligned}
f_t &= \sigma\big(W_f[h_{t-1}, x_t] + b_f\big), \quad \text{(遗忘门)} \\
i_t &= \sigma\big(W_i[h_{t-1}, x_t] + b_i\big), \quad \text{(输入门)} \\
\tilde{C}t &= \tanh\big(W_C[ht-1, x_t] + b_C\big), \quad \text{(候选记忆)} \\
C_t &= f_t * C_{t-1} + i_t * \tilde{C}t, \quad \text{(更新细胞状态)} \\
o_t &= \sigma\big(W_o[ht-1, x_t] + b_o\big), \quad \text{(输出门)} \\
h_t &= o_t * \tanh(C_t). \quad \text{(更新隐藏状态)}
\end{aligned}
$$

x_t：第 t 个时间步的输入向量 h_{t-1}：前一个时间步的隐藏状态。C_t：当前时间步的细胞状态，核心存储。$\tilde{C}_t$：当前时间步新的候选记忆，通过 tanh 函数得到。f_t, i_t, o_t：分别为遗忘门、输入门、输出门的门控向量，取值范围在 [0,1]。σ：sigmoid 函数，tanh：双曲正切函数。

4 Result

Method	MAE	RMSE
ARIMA	10.196397478154942	11.738271779572411
LSTM	8.492959594726566	10.220722256795856

Table 1: 测试集表现

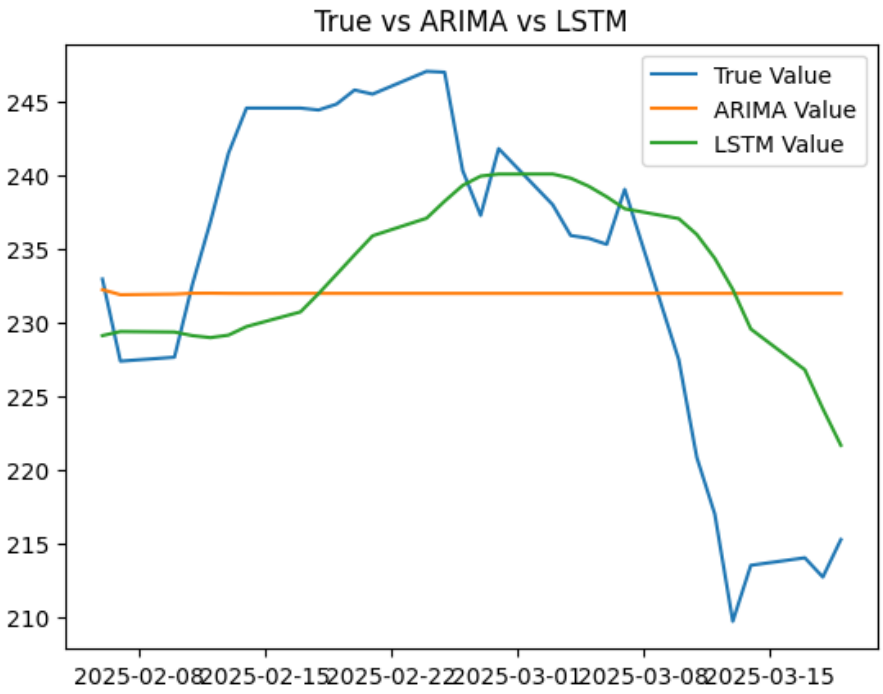

Figure 1: 未来七日预测结果