

# 基于神经网络的股票预测-最终报告

小组成员：常明 2120160580

李京松 2120160586

李杰 2120160915

## 1. 简介

股票市场作为风险与收益都较高投资市场，一直受到投资者的密切关注。各个国家的股票交易所每天都要产生巨量的交易数据，投资者与投资机构等在交易证券和投资股票时也越来越多将数据作为首要参考依据，实时交易数据，比如 K 线、分时线等等常常作为市场的客观反映。投资者和投资机构通常选取这些历史数据来进行分析和预测，以期获得更高利润。

本项目采用了一种基于循环神经网络 LSTM 的股票预测方法，根据历史交易数据对第二天股价进行预测。

## 2. 问题描述

本课题通过训练神经网络模型，使它可以根据历史交易数据预测第二天的股价，通过准确率（后文定义）来评价预测结果。

数据集采用上证综值用于模型训练，然后实时随机抓取个股来，来检验模型性能。

上证指数：

网易财经 [http://quotes.money.163.com/trade/ljsysj\\_zhishu\\_000001.html](http://quotes.money.163.com/trade/ljsysj_zhishu_000001.html)

抓取网址：

[http://quotes.money.163.com/trade/ljsysj\\_股票代码.html](http://quotes.money.163.com/trade/ljsysj_股票代码.html)

### 3. 技术方案

#### 3.1 理论基础

股票具有多个技术指标，但最为常用的是开盘价、收盘价、最高价、最低价、成交量。前 4 个指标反应了股市中卖家和买家的力量对比，成交量则反应了股市的活跃程度。因此，根据历史数据的这 5 个指标就可以对未来的走势做出个预测。

RNN 相比传统的神经网络，具备了时序观念，可以实现时间序列依次多个输入输出，且隐层之间实现自连接，展开后相当于时序之间相互影响。LSTM 是一种优化后的 RNN 神经元，它解决了 RNN 网络中梯度消失或梯度爆炸的问题。

因此可以通过将股票历史数据喂入 LSTM 网络的方式，使网络学会“理解”股票数据，从而做出对第二天价格的预测。

#### 3.2 软件平台

语言：Python

深度学习框架：Keras

爬虫模块：requests, BeautifulSoup4

#### 3.3 具体方案

##### 3.3.1 数据处理

1. 填补日期，并对非交易日的数据进行剔除；
2. 除去股票中不需要的属性
3. 将数据处理为适合传入网络的形式（分块、标准化、划分训练集与测试集等）

标准化公式：

$$n_i = \frac{p_i}{p_0} - 1$$

$n_i$  为处理后数据

$p_i$  为处理前数据

$p_0$ 为处理窗口的第一个数据

训练集：测试集 = 9：1

### 3.3.2 构建网络

网络结构



第一层 LSTM 输出维度 50；第二层输出维度 100

Dropout 为 0.2

损失评价函数：MSE（均方误差）

优化器：RMSProp

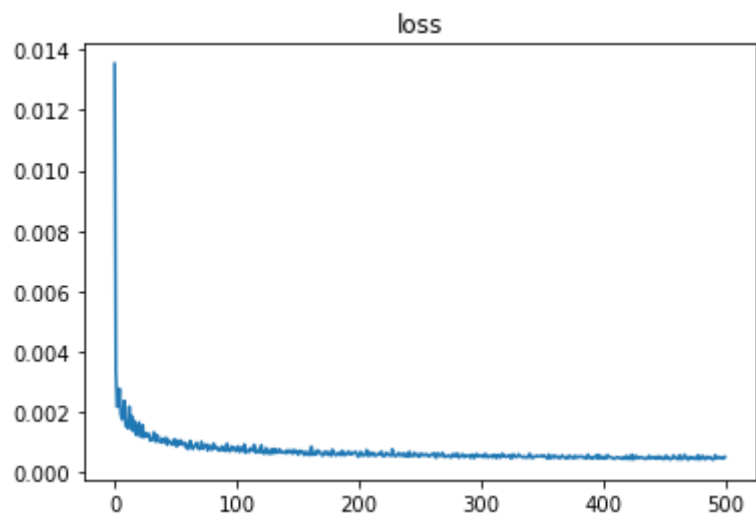
准确度计算公式：

$$\text{acc} = 1 - \left| \frac{y_t - y_p}{y_t + 1} \right|$$

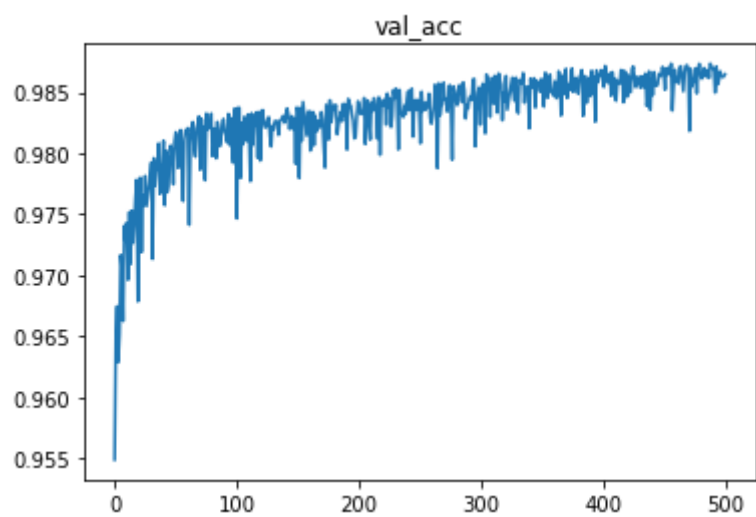
### 3.3.3 训练结果

迭代次数 500 次后

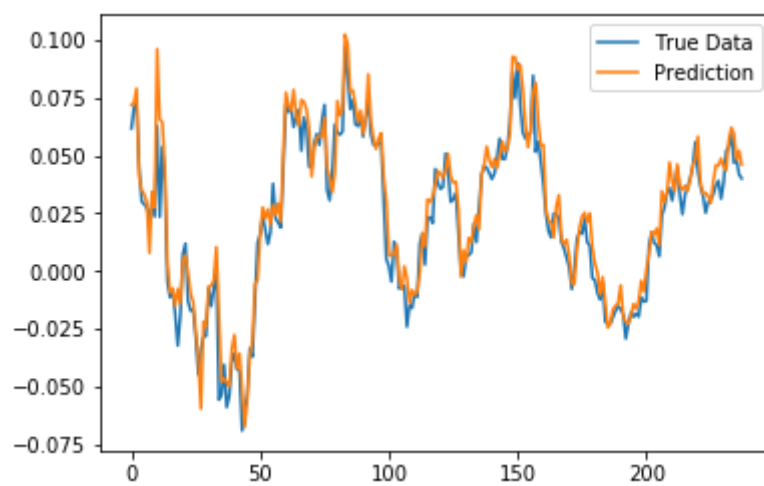
训练集损失图：



验证集准确率图：



测试集拟合结果：



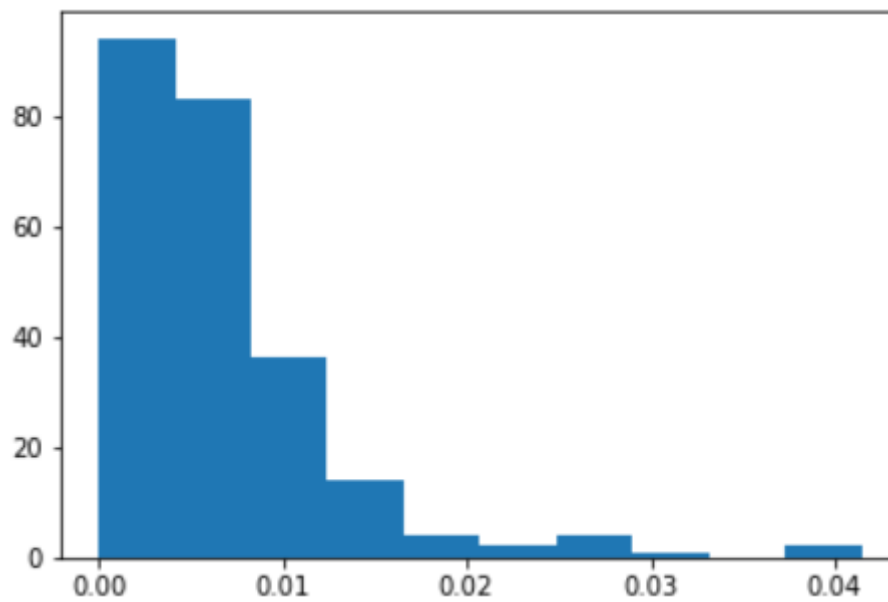
测试集误差 (1-acc) 统计:

误差 0.01 以内:81.5126%

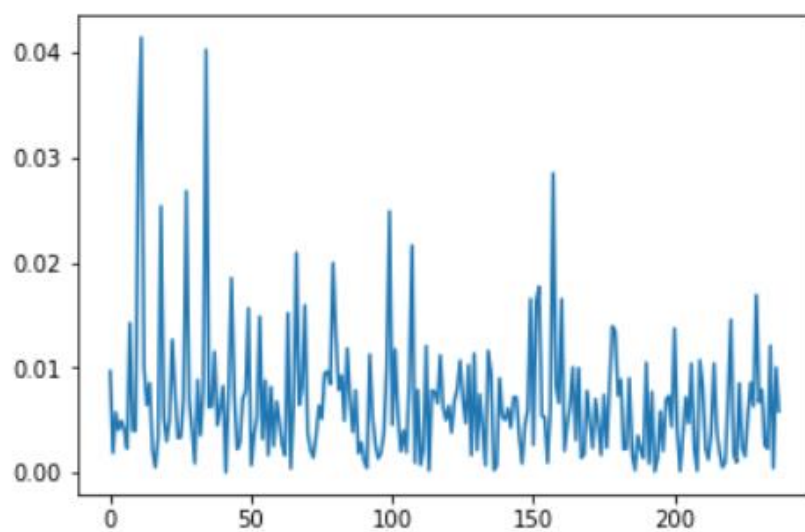
误差 0.02 以内:96.2185%

误差 0.05 以内:100.0000%

误差统计直方图



误差折线图



### 3.4 个股验证

通过爬虫随机抓取 50 只股票，误差统计结果如下：

股票代码:600329	误差 0.01 以内:58.0%	误差 0.02 以内:70.0%	误差 0.05 以内:100.0%
股票代码:600031	误差 0.01 以内:58.0%	误差 0.02 以内:78.0%	误差 0.05 以内:98.0%
股票代码:600085	误差 0.01 以内:56.0%	误差 0.02 以内:84.0%	误差 0.05 以内:100.0%
股票代码:600438	误差 0.01 以内:54.0%	误差 0.02 以内:74.0%	误差 0.05 以内:92.0%
股票代码:600218	误差 0.01 以内:52.0%	误差 0.02 以内:76.0%	误差 0.05 以内:98.0%
股票代码:600461	误差 0.01 以内:50.0%	误差 0.02 以内:76.0%	误差 0.05 以内:100.0%
股票代码:600280	误差 0.01 以内:50.0%	误差 0.02 以内:68.0%	误差 0.05 以内:100.0%
股票代码:600231	误差 0.01 以内:50.0%	误差 0.02 以内:70.0%	误差 0.05 以内:96.0%
股票代码:600240	误差 0.01 以内:48.0%	误差 0.02 以内:70.0%	误差 0.05 以内:96.0%
股票代码:600056	误差 0.01 以内:48.0%	误差 0.02 以内:78.0%	误差 0.05 以内:100.0%
股票代码:600436	误差 0.01 以内:46.0%	误差 0.02 以内:64.0%	误差 0.05 以内:98.0%
股票代码:600498	误差 0.01 以内:44.0%	误差 0.02 以内:72.0%	误差 0.05 以内:98.0%
股票代码:600265	误差 0.01 以内:44.0%	误差 0.02 以内:68.0%	误差 0.05 以内:88.0%
股票代码:600141	误差 0.01 以内:42.0%	误差 0.02 以内:72.0%	误差 0.05 以内:100.0%
股票代码:600267	误差 0.01 以内:42.0%	误差 0.02 以内:74.0%	误差 0.05 以内:100.0%
股票代码:600360	误差 0.01 以内:42.0%	误差 0.02 以内:72.0%	误差 0.05 以内:94.0%
股票代码:600121	误差 0.01 以内:42.0%	误差 0.02 以内:76.0%	误差 0.05 以内:98.0%
股票代码:600539	误差 0.01 以内:42.0%	误差 0.02 以内:66.0%	误差 0.05 以内:96.0%
股票代码:600348	误差 0.01 以内:40.0%	误差 0.02 以内:84.0%	误差 0.05 以内:98.0%
股票代码:600118	误差 0.01 以内:40.0%	误差 0.02 以内:78.0%	误差 0.05 以内:100.0%
股票代码:600193	误差 0.01 以内:38.0%	误差 0.02 以内:62.0%	误差 0.05 以内:94.0%
股票代码:600351	误差 0.01 以内:38.0%	误差 0.02 以内:76.0%	误差 0.05 以内:94.0%
股票代码:600398	误差 0.01 以内:36.0%	误差 0.02 以内:66.0%	误差 0.05 以内:98.0%
股票代码:600099	误差 0.01 以内:36.0%	误差 0.02 以内:56.0%	误差 0.05 以内:90.0%

股票代码:600313	误差 0.01 以内:34.0%	误差 0.02 以内:50.0%	误差 0.05 以内:74.0%
股票代码:600272	误差 0.01 以内:34.0%	误差 0.02 以内:62.0%	误差 0.05 以内:92.0%
股票代码:600579	误差 0.01 以内:32.0%	误差 0.02 以内:56.0%	误差 0.05 以内:90.0%
股票代码:600466	误差 0.01 以内:32.0%	误差 0.02 以内:66.0%	误差 0.05 以内:98.0%
股票代码:600235	误差 0.01 以内:32.0%	误差 0.02 以内:66.0%	误差 0.05 以内:94.0%
股票代码:600365	误差 0.01 以内:32.0%	误差 0.02 以内:54.0%	误差 0.05 以内:80.0%
股票代码:600513	误差 0.01 以内:30.0%	误差 0.02 以内:50.0%	误差 0.05 以内:82.0%
股票代码:600075	误差 0.01 以内:30.0%	误差 0.02 以内:54.0%	误差 0.05 以内:84.0%
股票代码:600361	误差 0.01 以内:30.0%	误差 0.02 以内:46.0%	误差 0.05 以内:80.0%
股票代码:600143	误差 0.01 以内:28.0%	误差 0.02 以内:68.0%	误差 0.05 以内:88.0%
股票代码:600491	误差 0.01 以内:28.0%	误差 0.02 以内:60.0%	误差 0.05 以内:96.0%
股票代码:600321	误差 0.01 以内:26.0%	误差 0.02 以内:58.0%	误差 0.05 以内:86.0%
股票代码:600388	误差 0.01 以内:26.0%	误差 0.02 以内:52.0%	误差 0.05 以内:76.0%
股票代码:600017	误差 0.01 以内:24.0%	误差 0.02 以内:76.0%	误差 0.05 以内:92.0%

注：因为抓取时有的股票交易记录不足 100 条，所以数量略少于 50

可以看出：在个股上模型的准确率有所下降，而且层次不齐，这是因为，个股更容易被操纵，被政策影响。但不难发现，模型在一些股票上的准确率还是可以接受的，因为这些股票比较稳定。

## 4. 结论

由于个股股价受多方因素影响较大，单纯利用历史交易记录来预测未来股价是极其困难的，但是本模型任有一定意义。它可以通过大量预测股票来筛选稳定的个股（这些个股的预测误差往往比较低），同时还可以利用模型对这些稳定股票第二天预测的价格来制定初步的交易策略（如预测价格涨幅 2%以上，就可以适当买入）。

注：相关代码 [https://github.com/lucifer443/stock\\_predict](https://github.com/lucifer443/stock_predict)