基于机器学习的多因子A股市场量化选股策略研究

2204

**摘要**：

**关键词**：A股市场；因子筛选；LSTM神经网络；回归分析

**模型的选择**

我们最终选择了长短期记忆人工神经网络（Long Short-Term Memory, LSTM）作为股价趋势预测的模型。在做出选择之前，我们比对了大量的模型。普通神经网络因为其“黑箱”的特点，很大概率会给模型的调参与优化造成困难，并且调试与排错的成本高昂，故被放弃。卷积神经网络缺少记忆功能，并且全连模式显得冗余且低效，故也被淘汰。而循环神经网络以其优秀的记忆能力与对非线性特征进行机器学习的能力脱颖而出。对于两种最为常用的循环神经网络，标准循环神经网络（Recurrent Neural Network, RNN）对于长短期记忆人工神经网络（LSTM）有一个重要的功能缺陷，相对严重的梯度消失。“Standard RNN cannot bridge more than 5-10 time steps … Blown-up error signals lead straight to oscillating weights, whereas with a vanishing error”(Staudemeyer & Morris, 2019)。标准循环神经网络会尝试记住所有的信息，A股市场的股价变化与海量的因子数据会带来沉重的记忆负担，导致最后保留大量的噪音数据，给进一步的分析带来困扰。而长短期记忆人工神经网络拥有记忆细胞，可以对股价变化、因子特征等信息进行筛选，并通过“遗忘门”函数清理噪音数据。“当遗忘门𝑓𝑡被打开时，𝐶𝑡的梯度可以有效地反向传递给𝐶𝑡−1……通过引入另一个隐藏状态𝐶𝑡和3个门控结构，LSTM 缓解了神经网络训练中的梯度消失问题”（林晓明，2017）保障了最终数据的可靠性。

**长短期记忆人工神经网络（LSTM）的搭建**

在决定了以长短期记忆人工神经网络为基础构建模型之后，Python语言扩展包丰富，编程自由度高，数据处理能力好的优点我们选择使用其进行编程来构建模型以及之后的模型优化与数据处理。模型以交易日的各因子数据为自变量，当日收盘价为因变量作回归分析。股票的因子数据与收盘价均从“new\_factor\_data”数据表直接读取。自变量数据表与因变量数据列受参数“mem\_day”调节。自变量与因变量搭建的步骤被封装进“LSTM\_stock”函数中，输入参数为数据表“stock”与记忆天数“mem\_day”为了避免自己搭建LSTM神经网络，降低时间成本，我们从tensorflow扩展包中的keras附属包中直接调用LSTM模块。

**模型调参与优化**

LSTM模型的调参与优化工作由代码文件“LSTM\_Test.py”执行。模型循环运行的步骤被封装进“opt\_model”函数中，输入参数为记忆天数列表“mem\_days”，神经网络层数列表“lstm\_layers”，全连接层层数列表“dense\_layers”，神经元个数列表“units”。该函数通过for循环更改参数组合，使用预设的训练集（分割自“new\_factor\_data”数据表）运行LSTM模型，每一种参数组合进行50次学习，输出模型文件并标注拟合误差率。考虑数据缺失和拟合失败等问题，一轮测试将生成不超过2,700个模型，其中拟合度最高，误差率最小的模型会被保存并进入回测阶段。

**模型效果检验**

通过调参获取的最优模型将会被读取进入代码文件“LSTM\_Formal.py”，在这里代码会展示模型的具体信息。利用sklearn扩展包的“train\_test\_split()”函数，我们对检验股票的因子数据与收盘价数据划分训练集与测试集，参数test\_size=0.1。调用读取的模型使用“.evaluate()”函数对测试集进行评估，获得模型准确率与损失数据。将这两项数据与模型拟合度标签进行对比，检验模型的可靠性。

同时，模型对因子数据测试集使用“.predict()”函数，预测检验股票的股价变化，并与真实股价变化一并绘制折线图，将拟合度数据可视化。绘图工具由matplotlib扩展包导入。

**预测数据输出**

通过运行代码文件“LSTM\_Run.py”，我们对“沪深300”股票池中的每一支股票进行模型预测。预测所得的收盘价数据与股票实际收盘价数据被录入表格中，以“file\_name.gz”的压缩包格式进行数据储存。所有数据表格将用于下一步的回测分析工作。

**参考文献：**

Staudemeyer, R. C., & Morris, E. R. (2019). *Understanding LSTM--a tutorial into long short-term memory recurrent neural networks*. 17-18.

<https://doi.org/10.48550/arXiv.1909.09586>

林晓明.（2017）. *华泰人工智能系列之九：人工智能选股之循环神经网络模型*. 金工研究/深度研究. 华泰证券.