

基于深度学习的证券价格预测

课程项目大纲

【背景】

以深度学习为代表的人工智能技术，已在语音识别、机器翻译、自然语言理解等序列预测领域，取得突破性进展。证券价格预测，属于典型的时间序列预测问题，是量化交易的关键所在。但因证券数据噪音大、信息不全面等因素，历年来传统的量化分析方法，都难以取得较高的预测精准度。本课程项目，旨在通过研究分析证券的历史交易数据，基于深度学习中的核心技术（包括但不限于：DNN / CNN / RNN / LSTM等），挖掘数据中的深层规律，探索证券价格涨跌趋势，提升预测精确度。

希望各位同学在课程项目的学习和实践中，透过数据研究证券价格的涨跌规律，学习**数据分析、数据可视化、数据预处理、特征工程**的常用方法，了解深度学习的相关技术原理，掌握 **TensorFlow/Keras/TFLearn** 等深度学习编程库的使用方法，完成深度学习**证券价格预测模型**的定义、训练、评估、预测，为构建精准量化交易策略提供决策支持。

【目标】

根据上证 A 股沪深 300 指数成分股中 10 只股票的历史交易数据，构建两类深度学习模型：

1. 预测未来 1 天的股价**涨跌幅度**（比如+1.99%、-1.99%）
2. 预测未来 1 天的股价**是涨是跌**（涨为 1、不涨为 0）

注：理论上，根据第 1 类模型的涨跌幅度，就可知道第 2 类模型的预测结果（是涨是跌）；但在深度学习模型的训练过程中，第 1、2 类模型会有区别，因此建议各自单独建模。

验收时，第 1 类模型以 10 支股票的平均涨跌幅误差为评分标准，第 2 类模型以 10 支股票的平均准确率为评分标准。误差越小越好，准确率越高越好。

【实践指引】

以下内容如有不明白的地方，建议先百度、Google，再咨询队友；如果仍未解决，可在微信群提出或问老师。

1. 数据获取

a) 才云科技深度学习平台(<http://capstone.caicloud.io>)，登录后选择左侧边栏“数据存储”>> datasets/stock

b) 百度网盘: <https://pan.baidu.com/s/1bpKvm0n> (提取码: n78w)

一共 10 支股票的历史每日交易数据，从 2012-10-08 到 2017-09-29，各列分别表示日期、开盘价、收盘价、最高价、最低价、交易量、股票代码。

2. 示例代码

使用 keras/tflearn 分别编写的 LSTM 股票价格预测模型，代码放在百度网盘（微信群里也会提供）。

一共 4 个代码文件：tflearn-lstm.py、tflearn-lstm.ipynb、keras-lstm.py、keras-lstm.ipynb。 .py 文件可用 python 直接运行，.ipynb 文件需安装 Jupyter 后运行。

3. 环境配置

操作系统：推荐使用 Linux/Ubuntu/Mac

(如果是 Windows，推荐使用 <http://capstone.caicloud.io>)。

编程语言：Python 2.7（如果使用 Python3，示例代码可能需做修改）

软件工具：Jupyter、pip(使用 pip 来安装下面的最新版本的库)

数据处理库：numpy、pandas、matplotlib、sklearn

深度学习工具库：tensorflow、keras、tflearn((Windows 下只能用 theano、keras)

4. 技术框架

4.1 开发环境配置

Python 是机器学习领域最主流的编程语言，pip 可以很方便地安装 python 工具包，jupyter 可以很方便地开发、调试 python 源代码，。

因此我们先安装 Python、pip，再使用 pip 安装 jupyter、numpy、pandas、matplotlib、sklearn、tensorflow、keras、tflearn 等工具库。安装方式示例：

```
pip install numpy
```

4.2 数据分析、数据可视化

numpy/pandas 是 python 提供的数据处理高级库，可以很方便地对 csv/excel 等行列格式的结构化数据进行读写、计算、分析。Matplotlib 是数据可视化库，能方便地将结构化数据画成图形展示。

建议大家先对这 3 个库进行学习了解，用来做数据分析与数据可视化。比如找出各支股票涨停或跌停的日期并分析原因，将数据显示成蜡烛形 K 线图。

4.3 数据预处理、特征工程

数据中常有缺失、错误、异常点等情况出现，为此我们需对数据进行预处理，比如验证有无缺失值、涨跌幅是否超过 10%，对数据做标准化/归一化等。

虽然深度学习某种程度上实现了 end-2-end，降低了对人工特征工程的依赖，但良好的人工特征往往能使深度学习事半功倍。如有时间，可提取形如前 N 日的平均成交量、N 日内的最高价与最低价之差等特征；或者从技术分析的角度出发，提取 MACD/KDJ/RSI 等技术指标类特征。

4.4 模型构建

a) 模型选择

选择 DNN、RNN(LSTM/Bi-LSTM/GRU)、CNN 等深度学习模型，或者选择 SVM、XGBoost 等机器学习模型。

b) 模型定义

使用 keras/tflearn 高层抽象代码或 tensorflow/theano 原始代码来定义深度学习模型，或者使用 sklearn/xgboost 来定义机器学习模型。

c) 参数调优

比如神经网络用多少层、每层多少个神经元、激活函数、优化方法、权重初始化方法、Dropout、正则化、学习率衰减、Batch Normalization、Attention 等。参数调优可能比较费时间，可考虑使用 <http://capstone.caicloud.io> 来训练，降低训练时间、提高训练效率。

d) 数据标签

以未来 1 天的涨跌幅度、是涨是跌作为数据标签。如有兴趣，可尝试用未来 N 天内的(最高价-开盘价)/开盘价作为数据标签。

4.5 交易策略 (*可选)

如有兴趣，可利用模型的预测结果，开发交易策略（比如预测为涨时开盘后买入，预测为跌时收盘前卖出），并统计策略收益及最大回撤。

5. 组内分工协作

建议每组推选出 1 名队长，自荐或举荐；如果没有，则默认以分组表中每组第 1 位同学为队长。队长统筹组内的人员安排，推动分工协作，组织队员学习、讨论、实践。

6. 项目验收

结束阶段，各组提交**模型文件**及相关的**训练和预测代码**，我们根据 10 支股票未来 5 个交易日的数据进行评估验收。

各组总结本次项目的开发过程、踩坑经历、填坑经验等，以供相互之间借鉴学习。