

# 基于深度学习的证券价格预测

## 课程计划

	任务	说明
第一周	数据分析、 数据可视化、 数据预处理	<ol style="list-style-type: none"> <li>1. 学习使用 <b>pandas/numpy</b> 进行数据分析，包括数据读写、数据列式计算、数据统计等</li> <li>2. 学习使用 <b>matplotlib/seaborn</b> 进行数据可视化，包括画 K 线图、柱状图、折线图等</li> <li>3. 学习 <b>sklearn.preprocessing</b> 模块中的数据预处理方法，了解深度学习中的 <b>BN(Batch Normalization)</b>方法</li> <li>4. 了解交叉验证方法，划分训练集、测试集</li> </ol>
第二周	特征设计、 特征生成、 特征选择	<ol style="list-style-type: none"> <li>1. 根据数据分析结果和对业务的思考理解，设计相应的统计特征、组合特征</li> <li>2. 根据设计结果，使用 <b>pandas/numpy</b> 等工具生成相应的特征</li> <li>3. 学习特征选择的常用方法，计算特征与标签的相关性，分析特征重要性</li> </ol>
第三周	模型设计、 模型训练、 模型超参数调优	<ol style="list-style-type: none"> <li>1. 学习全连接神经网络、LSTM 循环神经网络的技术原理，设计不同的神经网络结构，并使用 <b>keras/tflearn/tensorflow</b> 来完成模型构建</li> <li>2. 学习神经网络各参数的相关概念，使用 TaaS 平台的来加速模型参数调优，提升模型效果</li> <li>3. TensorBoard 模型训练过程可视化</li> </ol>
第四周	模型预测、 模型评估、 交易策略设计	<ol style="list-style-type: none"> <li>1. 使用模型来对测试集进行预测，对比预测结果和真实结果，统计预测误差和 Loss</li> <li>2. 根据十只股票的预测结果，设计交易策略，回测验证交易策略的可行性、统计收益率</li> </ol>

## 第一周实践目标：

### 1. 数据分析

- 找出股票的历史最低价格、历史最高价格及其对应的日期
- 计算股票从第一天开始，之后每个交易日的累积涨跌幅度
- 统计从周一到周五，五个工作日的涨跌比例和平均涨跌幅度
- 找出当日开盘价相比上日收盘价，涨跌幅超过 5% 的所有日期
- 找出涨停或跌停的所有日期
- 找出成交量超过前 5 个交易日的平均成交量 3 倍的日期
- 按月对数据进行汇总，找出涨跌幅超过 20% 的月份（提示：可使用 pandas 的 `resample` 方法）

### 2. 数据可视化

- 画出股票的最近一年的日 K 线图(包含蜡烛条、成交量、MA5、MA10、MA30)，类似于下图：（提示：可使用 matplotlib 的 `candlestick2` 方法）



- 画出股票涨跌幅度的分布图(柱形图或饼图), 用于表示  $X\% \leq \text{涨跌幅} \leq (X+1)\%$  的交易日各有多少天、占比多少，其中  $X$  为整数且  $-10 \leq X \leq 9$ 。

### 3. 数据预处理

- 数据校验：检查数据中是否有缺失值，过滤缺失值；检查数据中是否有涨跌幅  $> 10.5\%$  的值。
- 数据归一化：学习 sklearn 常用的数据预处理方法，使用 `Scaler` 对数据进行归

一化和还原（反归一化）。

c) 交叉验证：学习 sklearn 常用的交叉验证方法，使用 Cross Validation 对股票时间序列数据做交叉验证，生成训练集和测试集。

## 【学习资料】

[python 廖雪峰简易教程](#)

[10 分钟快速上手 pandas](#)

[numpy 快速入门](#)

[matplotlib 画 K 线图](#)

[matplotlib 画图示例](#)

[sklearn 数据预处理](#)

[sklearn 交叉验证](#)

## 第二周实践目标：

特征工程在机器学习中占据重要地位，决定着机器学习模型效果的上限。特征工程需要深度思考与预测目标相关的因素，找出可能的重要特征并生成，先谋而后动；再优中选优，从众多特征中，找出最重要的部分特征，用以提升模型预测效果、减少过拟合风险。

### 1. 特征设计

- 思考股票涨跌的本质，从量价时空角度，分析哪些因素与涨跌最相关。
- 学习技术指标相关概念，例如 MACD、KDJ、RSI、BOLL、EMV 等，思考其优缺点及适用场景。

### 2. 特征生成

根据设计结果，使用 pandas/numpy/talib 等工具生成相应的特征：

- 前 5、10、20、40 个交易日内的平均成交量、上涨时的平均成交量、下跌时的平均成交量
- 前 5、10、20、40 个交易日区间内的(收盘价-开盘价)/开盘价、(最高价-开盘

价)/开盘价、(最低价-开盘价)/开盘价、(最高价-收盘价)/开盘价、(最低价-收盘价)/开盘价

- c) 前 1、3、5、10 个交易日内，[(最高价+最低价)-(开盘价+收盘价)] \* 成交量
- d) 使用 talib 生成 MACD、KDJ、RSI、EMV 指标

### 3. 特征选择

- a) 学习特征选择的常用方法，例如 Wrapper、Filter、Embedding、LDA/PCA、AutoEncoder、L1 正则化、决策树特征重要性等。
- b) 计算特征与标签的相关性，分析特征重要性，选择出最重要的 top-K 特征。

### 【学习资料】

如何理解 KDJ 和 MACD 的等指标的使用

<https://www.zhihu.com/question/26729743>

机器学习中，有哪些特征选择的工程方法？

<https://www.zhihu.com/question/28641663/answer/41653367>

<http://blog.csdn.net/u012328159/article/details/53954522>

[http://scikit-learn.org/stable/modules/feature\\_selection.html#feature-selection](http://scikit-learn.org/stable/modules/feature_selection.html#feature-selection)

百度网盘：capstone > 学习资料 > talib 函数手册·zw 汉化标注版 v2.pdf

## 第三周实践目标：

1. 学习 DNN 全连接神经网络、LSTM 循环神经网络的技术原理
  - a) 分别设计 DNN 和 LSTM 两种不同的神经网络结构，并使用 keras/tflearn 中的一种和 tensorflow 来完成模型构建和模型训练。
  - b) 学习神经网络各参数的相关概念，比如学习率、优化方法、激活函数、损失函数、权重初始化方法、正则化、Dropout、Batch Normalization、Batch Size、Attention 机制等。

2. 使用 TaaS 平台的来加速模型参数调优，提升模型效果。可调参数包括：神经网络层数、各层神经元数量、神经网络类型(dense/lstm)、学习率、优化方法(sgd/adam/adagrad/rmsprop 等)、激活函数(sigmoid/relu/tanh 等)、损失函数(rmse/binary\_entropy/cross\_entropy 等)、权重初始化方法(uniform/normal/truncated\_normal)、正则化系数(L1/L2)、Dropout、Batch Normalization、Batch Size 等。
3. (\*选做) 使用 GridSearchCV 来自动搜索超参数，记录不同参数下的模型效果，找出最优参数组合；实现带 Attention 机制的 LSTM 模型。
4. 模型训练过程可视化  
使用 TensorBoard 来可视化神经网络结构，并观察 Loss、Metrics、Gradients 等指标在训练过程中的变化趋势。

## 【学习资料】

网易云课堂 神经网络和深度学习 吴恩达：

<http://mooc.study.163.com/course/2001281002>

LSTM 神经网络实现：

[http://keras-cn.readthedocs.io/en/latest/layers/recurrent\\_layer/](http://keras-cn.readthedocs.io/en/latest/layers/recurrent_layer/)

<http://tflearn.org/layers/recurrent/>

GridSearchCV 参数搜索：

<https://machinelearningmastery.com/grid-search-hyperparameters-deep-learning-models-python-keras/>

<http://blog.csdn.net/lixianjun913/article/details/52216795>

利用 tflearn 快速搭建深度学习模型，并使用 TensorBoard 可视化

[https://zhuanlan.zhihu.com/p/25322066?utm\\_source=weibo&utm\\_medium=social](https://zhuanlan.zhihu.com/p/25322066?utm_source=weibo&utm_medium=social)

深度学习调参必知技巧

<http://lamda.nju.edu.cn/weixs/project/CNNTricks/CNNTricks.html>

<http://www.36dsj.com/archives/86829>

## 第四周实践目标：

### 1. 模型预测与分析

- a) 使用模型来对测试集进行预测，生成预测结果
- b) 画图展示测试集的预测结果和真实结果，统计预测误差或预测准确率

### 2. 交易策略设计

- a) 根据预测结果，设计自动交易策略（比如今日预测明天为涨时以今日收盘价买入；或今日预测明天涨幅大于 1% 时以今日收盘价买入）。
- b) 在测试集上统计交易策略的每日累积收益率，并与基准策略（第一天开盘买入，持有不动）的每日累积收益率画图对比。