价格预测 +

模式识别

40240452

April 26, 2018

数据说明

- 该数据为 2017 年 7-8 月的数种期货数据
- 每 500ms 检查行情数据。若行情变化，将新行情写入文件。
- 文件命名方法 (大类)-(日期)-(day/night).log，记载了不同种类和时间的行情记录。文件内包含的有效样例记录如下：

```
2017-08-08 09:00:00.686  [KeyedThreadPool-3-1] INFO   - Quote[lastPrice=40950000,highestPrice=41080000,lowestPrice=40560000,volume=971204,
turnover=39657599940000,bidPrice1=40950000,bidVolume1=326,askPrice1=40960000,askVolume1=70,instrumentID=B2,exchangeID=]
2017-08-08 09:00:00.694  [KeyedThreadPool-64-1] INFO   - Quote[lastPrice=39480000,highestPrice=39740000,lowestPrice=39160000,volume=1626802,
turnover=64236528480000,bidPrice1=39480000,bidVolume1=162,askPrice1=39520000,askVolume1=104,instrumentID=B3,exchangeID=]
2017-08-08 09:00:01.183  [KeyedThreadPool-3-1] INFO   - Quote[lastPrice=40950000,highestPrice=41080000,lowestPrice=40560000,volume=971458,
turnover=39668003520000,bidPrice1=40950000,bidVolume1=329,askPrice1=40960000,askVolume1=6,instrumentID=B2,exchangeID=]
2017-08-08 09:00:01.184  [KeyedThreadPool-64-1] INFO   - Quote[lastPrice=39490000,highestPrice=39740000,lowestPrice=39160000,volume=1627528,
turnover=64265204380000,bidPrice1=39490000,bidVolume1=189,askPrice1=39510000,askVolume1=156,instrumentID=B3,exchangeID=]
2017-08-08 09:00:01.679  [KeyedThreadPool-3-1] INFO   - Quote[lastPrice=40980000,highestPrice=41080000,lowestPrice=40560000,volume=971776,
turnover=39681029120000,bidPrice1=40960000,bidVolume1=6,askPrice1=40980000,askVolume1=192,instrumentID=B2,exchangeID=]
```

数据说明

- 记录起始为记录的时间。
- instrumentID: 合约 ID。
- 对应的价格
  - highestPrice: 最高价
  - lastPrice: 最新成交价
  - lowestPrice: 最低价
- turnover: 累计成交金额
- volume: 累计成交数量
- bid, ask
  - askPrice1: 卖 1 价
  - askVolumn1: 买 1 量
  - bidPrice1: 买 1 价
  - bidVolumn1: 买 1 量

作业主要内容

根据过去一段时间（长度建议在 10 几秒到几分钟，最长不超过十分钟）的行情，预测未来 10 秒内价格变化的方向。

- 对数据进行预处理，标注价格变化方向
- 划分训练集/测试集，对价格变化方向进行预测并测评
- 任选第 6 章、第 7 章及其衍生方法进行预测。如：多层神经网络、深度置信网络、玻尔兹曼机、图模型等

作业说明

- 作业需包含指定方法的数据集划分、类别标注、结果测评，以便对比不同作业预测效果。
- 需提交作业报告，说明预处理方法、分析思路和预测方法、预测结果测评等内容。
- 需提交代码源文件和对应的说明文档。
- 若组队（推荐 2 人），请注明各成员工作内容或贡献。

作业说明

以下为作业内容的具体说明。当然，在此说明的基础上，你可以做其他的调整，并说明调整理由和分析调整之后的影响。

- 数据集划分
  - 训练集：20170703 至 20170809
  - 测试集：20170810 至 20170825
- 在预测时间 t 的价格时，可以使用 t 之前的任意数据，甚至用这些数据重新训练模型。
- 但不得使用 t 之后的任何数据，这些数据在预测之前是未知的。如无监督学习/聚类等处理方法。

作业说明

- 定义时间 t 价格 $P_{0.5}(t)$:
  - 若前 0.5s 有交易, 则:

$$P_{0.5}(t) = k \cdot \frac{turnover(t) - turnover(t-0.5)}{volume(t) - volume(t-0.5)} + (1-k) \cdot \frac{bidPrice1(t) + askPrice1(t)}{2}$$

  - 若前 0.5s 无交易, 查找之前价格:

$$P_{0.5}(t) = P_{0.5}(t-0.5)$$

- k 的建议范围 $k \in [0.2, 0.5]$, 可先尝试 $k = 0.3$

作业说明

- 记录内时间戳间隔并不是准确的 500ms，0.5s 前的数据可以找临近值或者插值等方法。
- 也可提前将时间戳对齐为 500ms。若你使用 1s 间隔，2s 间隔也可以将 0.5s 换为对应时间间隔。
- 若为重新开始交易，可舍弃该段时间起始时的价格数据 $P_{0.5}(t)$，或用 lastPrice 等值填充。

作业说明

- 定义价格变化 $d_{a,b}(t)$：我们可以列出未来 $[a, b]$ 秒的价格。如

$$P_{0.5}(t + t'), t' \in \{a, a + 0.5, .., b\}$$

- 取与 $P_{0.5}(t)$ 相差最大的 t'，即

$$t_s = \arg\max_{t'} |P_{0.5}(t + t') - P_{0.5}(t)|$$

- 得到价格变化

$$d_{a,b}(t) = \frac{P_{0.5}(t + t_s) - P_{0.5}(t)}{P_{0.5}(t)}$$

作业说明

- 需选择 $[a, b]$ 的长度，建议在 10 秒,20 秒附近。
- 随着自己对市场的理解，方便均衡数据等可以增大区间长度，但建议

$$\begin{cases} a \geq 5\mathrm{s} \\ b \leq 60\mathrm{s} \end{cases}$$

- 当然，可以任意缩小 $[a, b]$ 的长度，甚至缩小为 $\{10\}$，即

$$d_{10}(t) = \frac{P_{0.5}(t+10) - P_{0.5}(t)}{P_{0.5}(t)}$$

作业说明

- 分类标准 1：$\theta_1 = 0.15\%$ 的 3 分类：

条件	类别
$d_{a,b}(t) \geq \theta_1$	上涨 $\uparrow$
$-\theta_1 < d_{a,b}(t) < \theta_1$	不变 $\rightarrow$
$d_{a,b}(t) \leq -\theta_1$	下跌 $\downarrow$

- 由于预处理方法不一，你需要给训练集、测试集，不同合约的每个类别的数量和占比。

作业说明

- 分类标准 1：需要分别给出训练集、测试集上的测评结果:
  - 不同合约、不同类别的精确率 P、召回率 R（可附表）

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}$$

  - 不同合约、$\uparrow, \downarrow$ 类别的精确率、召回率的均值 (可附表)

$$\frac{P_{\uparrow} + P_{\downarrow}}{2}, \frac{R_{\uparrow} + R_{\downarrow}}{2}$$

  - A, B 类别中各选择一个最好的合约，求两个合约在测试集上精确率、召回率的均值，方便对比结果。

$$\frac{P_{Ai} + P_{Bi}}{2}, \frac{R_{Ai} + R_{Bi}}{2}$$

作业说明

- 分类标准 2(选做)：$\theta_1 = 0.1\%$, $\theta_2 = 0.2\%$ 的 5 分类：

条件	类别
$\theta_2 \leq d_{a,b}(t)$	快速上涨 $\uparrow\uparrow$
$\theta_1 \leq d_{a,b}(t) < \theta_2$	一般上涨 $\uparrow$
$-\theta_1 < d_{a,b}(t) < \theta_1$	方向不变 $\rightarrow$
$-\theta_2 < d_{a,b}(t) \leq -\theta_1$	一般下跌 $\downarrow$
$d_{a,b}(t) \leq -\theta_2$	快速下跌 $\downarrow\downarrow$

- 由于预处理方法不一，你需要给训练集、测试集，不同合约的每个类别的数量和占比。

作业说明

- 分类标准 2：需要分别给出训练集、测试集上的测评结果:
  - 不同合约、不同类别的精确率 P、召回率 R（可附表）

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}$$

  - 不同合约、$\uparrow, \downarrow, \uparrow\uparrow, \downarrow\downarrow$ 类别的精确率、召回率的均值 (可附表)

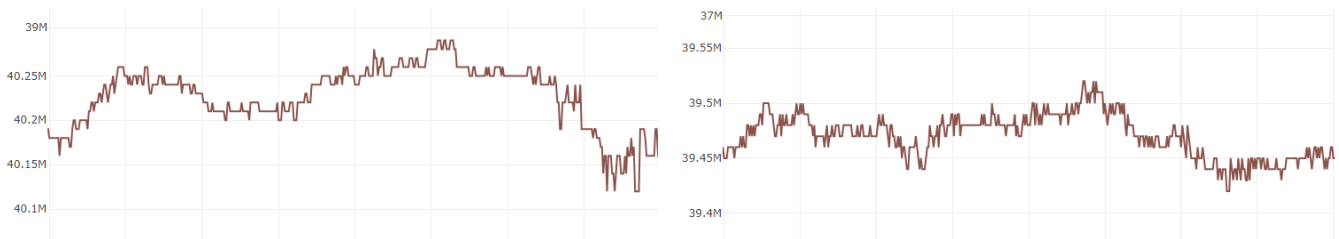
$$\frac{P_{\uparrow} + P_{\downarrow}}{2}, \frac{R_{\uparrow} + R_{\downarrow}}{2}, \frac{P_{\uparrow\uparrow} + P_{\downarrow\downarrow}}{2}, \frac{R_{\uparrow\uparrow} + R_{\downarrow\downarrow}}{2}$$

  - A, B 类别中各选择一个最好的合约，求两个合约在测试集上精确率、召回率的均值，方便对比结果。

$$\frac{P_{Ai\updownarrow} + P_{Bi\updownarrow}}{2}, \frac{R_{Ai\updownarrow} + R_{Bi\updownarrow}}{2}, \frac{P_{Ai\updownarrow\updownarrow} + P_{Bi\updownarrow\updownarrow}}{2}, \frac{R_{Ai\updownarrow\updownarrow} + R_{Bi\updownarrow\updownarrow}}{2}$$

附注

- 预测 4 类合约 A1, A3, B2, B3
- 相同类别不同合约的价格之间有较强的相关性，如 B2, B3
- 类别 A, B 之间在经济学上有一定相关性



B2, B3 在某时间内的 lastPrice，x 轴网格间距为 30s

工具

- **语言不限**: C/C++, Matlab, R, Python, ...
- **可视化**: matplotlib, plotly, ...
- **数据整理**: re, numpy, pandas, ...
- **机器学习**:
  - scikit-learn
  - Torch/PyTorch
  - Tensorflow
  - Keras, Mxnet, ...
  - Matlab Statistics and Machine Learning Toolbox