模式识别 2019 大作业

# 作业

作业主要内容根据过去一段时间的行情，预测从现在开始未来一段时间内股市中价格的变化。

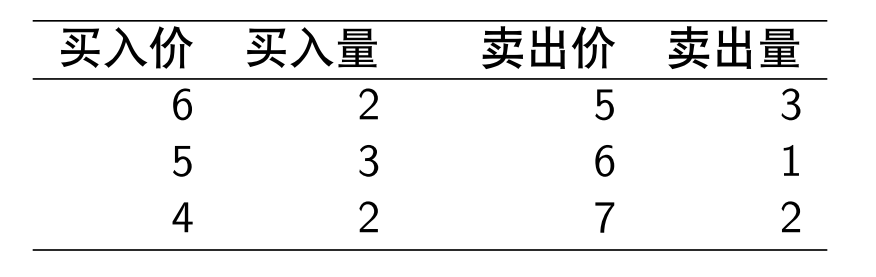
* 对数据进行预处理，选定价格变化标注方案。
* 划分训练集/测试集，对价格变化进行预测并测评。
* 任选简单的线性分类器、线性回归，曲线回归、逻辑回归、隐 markov 模型 HMM。
* 作业需包含指定方法的数据集划分、类别标注、结果测评，以便对比不同作业预测效果。
* 需提交作业报告，说明预处理方法、分析思路和预测方法、预测结果测评等内容。
* 需提交代码源文件和对应的说明文档。
* 若组队（推荐2人），请注明各成员工作内容或贡献。

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。

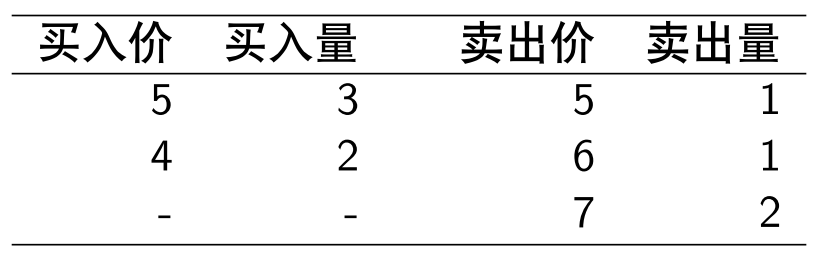
# 交易简易解释

## 连续交易：撮合

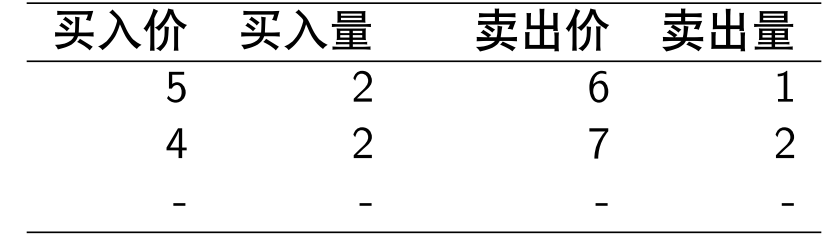
* 1.买卖双方在系统内申报价格
* 2.系统将同一合约的买卖申报指令排序，形成买单卖单列表。主要按照价格优先原则，若价格相同则按时间顺序。
  + 对于买方来讲，价格高优先成交
  + 对于卖者来讲，价格低优先成交



* 若买入价 ≥ 卖出价, 则开始交易。
* (6|2) 价格比 (5|3) 高，可以交易 2 个合约单位，即累计成交数量增加 2，累计成交金额增加 2\* 成交价格（成交价确定规则见2.2）。同时买单卖单表变为:



* (5|3) 价格等于 (5|1)，可以交易 1 个合约单位，即累计成交数量增加 1，累计成交金额增加 1\* 成交价格。同时买单卖单表变为:



* 此时买入价 < 卖出价，不继续对列表内买单卖单进行撮合交易，等待新买卖申报。剩余买单卖单根据优先顺序，分部称为买1, 买 2, ..., 卖 1, 卖 2，...



## 成交价格的确定

* 取买入价、卖出价和前一成交价（已确定成交最后一笔交易的成交价）的中值。
* 这样可以保证交易价格不比申请时的买入价高，也不比申请时的卖出价低。
  + 对之前的例子假定最初时前一成交价为 6
  + 之后，前一成交价为 6，买入价为 6，卖出价为 5，则成交价格为 6
  + 最后，前一成交价变为 6，买入价为 5，卖出价为 5，则成交价格为 5

## 特殊说明

* 交易所存在一段时间，不撮合交易，进行集合竞价。
* 每周一到周五交易，国家法定假日及周六日不能交易。
* 不是 24 小时交易，中午和半夜无法交易。
* 每天的价格变化有一定限制，当变化幅度超过限制时，不允许超过变化幅度的交易，成为涨停和跌停。当然，此时相反方向还可以交易。即涨停时可以卖出，跌停时可以买入。

# 数据说明

## 数据说明

* 数据分为3个文件夹存放：
  + Tick: 行情数据
  + Order：逐笔委托
  + OrderQueue：委托队列
* 存储格式为CSV，文件名组成方式为<TYPE>\_<ID>\_<DATE>.csv
  + TYPE：Tick / Order /OrderQueue
  + ID: 该文件对应的股票代码，ID相同为同一只股票
  + DATE：日期编号，每经过一个交易日自增1，DATE相同为同一天

## 行情数据（Tick）

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 描述 |
| nTime | int | 时间(HHMMSSmmm) |
| nPrice | int | 成交价 |
| iVolume | long long | 成交量 |
| iTurover | long long | 成交额 |
| nMatchItems | int | 成交笔数 |
| nInterest | int | IOPV(基金)、利息(债券) |
| chTradeFlag | char | 成交标志 |
| chBSFlag | char | BS标志 |
| iAccVolume | long long | 当日累计成交量 |
| iAccTurover | long long | 当日成交额(元) |
| nHigh | Int | 最高价 |
| nLow | Int | 最低价 |
| nOpen | Int | 开盘价 |
| nPreClose | Int | 前收盘 |
| nSettle | int | 结算价 |
| nPosition | int | 持仓量 |
| nCurDelta | int | 虚实度 |
| nPreSettle | int | 昨结算 |
| nPrePosition | int | 昨持仓 |
| nAskPrice | int[10] | 十档叫卖价 |
| nAskVolume | unsigned int[10] | 十档叫卖量 |
| nBidPrice | int[10] | 十档叫买价 |
| nBidVolume | unsigned int[10] | 十档叫买量 |
| nAskAvPrice | Int | 加权平均叫卖价(上海L2) |
| nBidAvPrice | Int | 加权平均叫买价(上海L2) |
| iTotalAskVolume | long long | 叫卖总量(上海L2) |
| iTotalBidVolume | long long | 叫买总量(上海L2) |
| nIndex | int | 不加权指数 |
| nStocks | int | 品种总数 |
| nUps | int | 上涨品种数 |
| nDowns | int | 下跌品种数 |
| nHoldLines | int | 持平品种数 |

## 逐笔委托数据 (Order)

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 描述 |
| nTime | int | 时间(HHMMSSmmm) |
| nIndex | int | 委托编号(从1开始递增) |
| nOrder | int | 交易所委托号 |
| chOrderKind | char | 委托类型 |
| chFunctionCode | char | 委托代码，买入B、卖出S、撤单C |
| nOrderPrice | int | 委托价格 |
| nOrderVolume | int | 委托数量 |

## 委托队列数据 (OrderQueue)

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 描述 |
| nTime | Int | 时间(HHMMSSmmm) |
| nSide | Int | 买卖方向('B':Bid,' A':Ask) |
| nPrice | int | 成交价格 |
| nOrderItems | int | 订单数量 |
| nABItems | int | 明细个数 |
| nABVolume | int[50] | 订单明细 |

# 具体说明

* 你可以自行划分数据集和测试集，并给出你的理由。
* 在预测时间的价格时，可以使用之前的任意数据，甚至用这些数据重新训练模型。
* 但建议不使用之后的任何数据，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。

# 价格定义

* 你可以自行定义价格
* 比如 =， 或者定义时间的价格 :
  + 若前有交易,则:
  + 若前无交易, 查找之前价格:
* 注意：Turnover和Price单位可能不一致。
* 建议范围, 可先尝试 。
* 记录内时间戳间隔并不是相等，可以找临近值或者插值等方法得到新数据。
* 也可提前将时间戳对齐为间隔dt。比如使用3s间隔，s间隔等。
* 若为重新开始交易，可舍弃该段时间起始时的价格数据, 或用nPrice等值填充。

## 价格变化

* 定义价格变化：我们可以列出未来秒的价格。如，
* 取与相差最大的, 即
* 得到价格变化
* 自行选择合理, 如：[0.5 min, 5 min]、 [30 min, 2h] 等。
* 随着自己对市场的理解增加，为方便均衡数据等可以选择其他区间更长长度。
* 当然，可以还任意缩小的长度，甚至缩小为, 即

# 分类方法

由于存在手续费，建议阈值至少满足 。阈值可选择调高，依据数据处理难度自行选择，但不可直接设置为0。

* 分类标准：的3分类：

|  |  |
| --- | --- |
| 条件 | 类别 |
|  | 上涨 |
|  | 不变 |
|  | 下跌 |

* 由于预处理方法不一，你需要给训练集、测试集，每个类别的数量和占比。
* 需要分别给出训练集、测试集上的测评结果:
  + 类别的精确率P、召回率R的均值

# 工具

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