

时间序列处理-股票交易实验报告 总览

By: 吴宇森 2000013137

报告总共分成两部分，一部分包括简洁的实现股票交易的综合算法 Midterm_TimeSeries.ipynb, 另一部分 details.ipynb 是为了具体描述和测试用到的机器学习相关算法性能并进行定量和图形化比较的分块展开解释文件(为了方便观察效果和加入图片, 主要内容放在这一块里面)。前一部分的函数在后者有具体呈现。

涉及的算法主要包括数据预处理、检测和学习并预测, 以及最后的决策算法四部分组成。在交易环境中表现为 judge 和 your_order 函数。其中决策所用到的环境参数只有 history_price() 传入的事件价格序列数组以及当天已有的表单。在非常有限的金融(股票)知识范围内尽可能写了一点简单的买卖策略。

一、预处理

- 1、对 DataFrame 类型 data 进行 .diff() 处理实现规整;
- 2、每隔四个数据取一个进行收缩并对得到结果用一维的卡尔曼滤波实现去除毛刺和短时干扰(这对于只有逐订单交易价格而不是收盘价, 数据间隔过短, 噪声较大的环境比较有效);
- 3、去除出现的 NAN 项。

二、数据检查

- 1、通过 adfuller 检查 data 数据规整, 反馈第二个参数较小, 证明符合要求
- 2、plot 出一部分预测数据和实际数据, 计算 MSE 并判断是否需要缩小
- 3、计算用于 ARIMA 的 p 和 q 超参数

三、序列预测

1、LR (逻辑回归) 算法进行二分类, 先用经验算法/求极值函数 signal.argrelextrema 标注各个点是否是极大/极小值点 (分别对是/不是两个分支进行标注)。选取并求取几个序列特征, 学习最后 200 项数据特征-标签匹配, 并回归预测当前是否处于极大/极小值位置, 是否进行对应的买卖。这里的特征选取的是 LinearRegression 线性回归) 得到的最后 20, 10, 7, 5 天梯度以及倒数第 9, 5, 3, 1 天股价数据, 由于没有每天的开盘/收盘价/成交量等因素 (以及对金融并不是很了解 qwq), 这些特征也许并不很能代表性质, 但也能看出一定效果。

2、ARIMA 函数等

同样基于低买高卖的想法去找极值点, 根据规整后的 data 数据和求出的 p, q 预测未来 6 天

的具体股价，并回归计算向前和向后各六天的梯度以此判断该天是否处于极值位置,最后判断是否买卖。

3、窗口化预测

把时间序列等分成多个区段，并通过比较特征找到相似度高的区段，并根据已有区段是否挣钱（处于上升区段，买入并持有一段时间后卖出可以收益），若挣钱权重高则所有相似区段都采取购买决策，若挣钱权重低则不购买。在这种方法中要求买卖间隔一定以保证结果有效性。

四、决策（贪心为主）

- 1、针对佣金，尽可能避免过于短期的交易，持有天数不小于 3 天或者处于上升期间收益率超过佣金
- 2、较有耐心的摇摆投资者策略，用优先队列保存此前买入价格，持有股票收益超过 10%即卖出获利，或损失 15%后卖出止损
- 3、朴素的想法就是极大值买入，极小值卖出；预期上升买入，预期下降卖出；根据以往相似模型经验，倾向于获利的买入并卖出
- 4、为了最后现金最大，从最后 25 天开始，每天固定抛售两股；并从最后 100 天开始提前倾向于卖出，通过更低的出价实现
- 5、对于大机会的捕捉，判定收益较大时买入较多股票，判断大风险时卖出较多股票
- 6、每日进场时机没有太多好的想法，根据股市一般特征选择 1/3 的交易时间进场，抓住机会的同时表单不易于空过
- 7、用到的数据有点少，没法加更多的特征，或者说如果画出 K 线什么的好一点

五、实现的老师上课补充内容概要

- 1、三个序列相关关系探究，并结合市场实际分析表现背后的原因
- 2、谱特征提取和分析，并进行平滑
- 3、二重聚类（分极大值/极小值两聚类）
- 4、用 ARIMA 模型预测一段时间交易价格，并和其他两个预测方式定量比较
- 5、卡尔曼滤波在预测准确率方面带来的提升，并和其他去噪方式进行比较
- 6、除了一般特征以外，实现了上课提到的分词和用两种熵（模式频繁特征）的特征提取并用于分类计算的数据集

感谢审阅！