金融机器学习算法

第二讲

金融数据结构化算法

本讲主要内容

- 金融数据结构化的动机
- 标准结构化法
- 信息驱动结构化法
- 多产品序列的结构化处理
- 对数据进行结构化抽样

金融数据结构化的动机

- 什么是结构化数据：结构化数据是以二维表结构来实现逻辑表达的数据。数据的每一行代表产生一个金融事件（或一次观测），数据的每一列代表事件（或观测）包含的变量。尽量保证每一行观测都包含独立的信息

- 非结构化的金融数据：未整理的高频数据，新闻数据，电子邮件/聊天/微信/微博, 网页搜索数据，卫星/监控数据，视屏会议，刷卡记录，App 生成数据……

- 金融数据结构化的动机：学习算法一般只接受结构化的数据，数据的独立同分布性 (i.i.d) 是机器学习算法能有效识别规律、进行预测的前提条件；数据易于交流展示、储存和关联

标准结构化法

等时长采样（Time bar）

- 标准结构法的作用是将一系列以不规律频率观测的数据，化为通过规律性采样获得的衍生均匀序列
- 等时长采样（时长把）：以固定的时长间隔（每秒、每分、每天、每周等）来记录数据的结构化方法

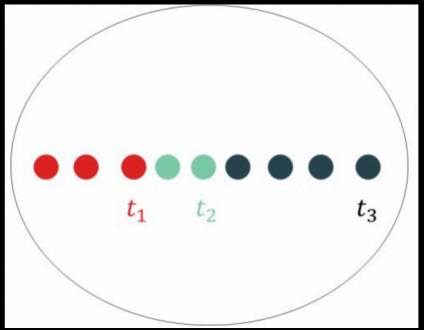
date_time	open	high	low	close	volume	turnover
2023-09-01 10:05:28	3880	3910	3855	3908	8033	0.12

- 等时长采样是将非结构化的 tick 级数据转化为结构化的周期数据

标准结构化法

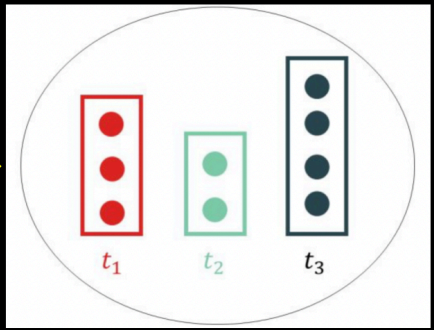
等时长采样（Time bar）

Tick 数据



结构化

时间把



标准结构化法

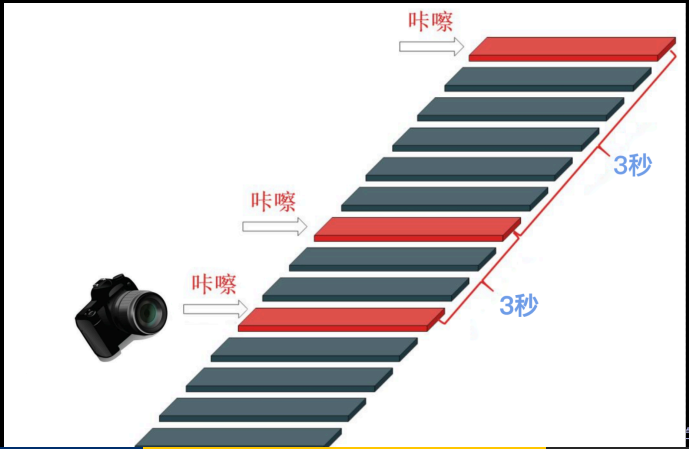
等时长采样（Time bar）

stockcode	time	price	bi	volv	vol	direction	buy1	buy2	buy3	buy4	buy5	sell1	sell2	sell3	sell4	sell5
CN600000	2020-01-02 09:32:35	12.51	34	1117028	893	B	12.50	12.49	12.48	12.47	12.46	12.51	12.52	12.53	12.54	12.55
CN600000	2020-01-02 09:32:38	12.51	32	885964	709	B	12.51	12.50	12.49	12.48	12.47	12.52	12.53	12.54	12.55	12.56
CN600000	2020-01-02 09:32:41	12.50	30	968788	775	S	12.49	12.48	12.47	12.46	12.45	12.50	12.51	12.52	12.53	12.54
CN600000	2020-01-02 09:32:44	12.52	27	1015536	812	B	12.50	12.49	12.48	12.47	12.46	12.52	12.53	12.54	12.55	12.56
CN600000	2020-01-02 09:32:47	12.52	21	343044	274	B	12.51	12.50	12.49	12.48	12.47	12.52	12.53	12.54	12.55	12.56
CN600000	2020-01-02 09:32:50	12.51	10	210160	168	S	12.50	12.49	12.48	12.47	12.46	12.51	12.52	12.53	12.54	12.55
CN600000	2020-01-02 09:32:53	12.52	19	526712	421	B	12.51	12.50	12.49	12.48	12.47	12.52	12.53	12.54	12.55	12.56
CN600000	2020-01-02 09:32:56	12.51	12	155212	124	S	12.51	12.50	12.49	12.48	12.47	12.52	12.53	12.54	12.55	12.56
CN600000	2020-01-02 09:32:59	12.51	9	125156	100	S	12.51	12.50	12.49	12.48	12.47	12.52	12.53	12.54	12.55	12.56
CN600000	2020-01-02 09:33:02	12.52	29	2322920	1856	B	12.51	12.50	12.49	12.48	12.47	12.52	12.53	12.54	12.55	12.56
CN600000	2020-01-02 09:33:05	12.52	42	848852	678	B	12.52	12.51	12.50	12.49	12.48	12.53	12.54	12.55	12.56	12.57
CN600000	2020-01-02 09:33:08	12.53	62	1120744	895	B	12.52	12.51	12.50	12.49	12.48	12.53	12.54	12.55	12.56	12.57
CN600000	2020-01-02 09:33:11	12.51	26	315684	252	S	12.51	12.50	12.49	12.48	12.47	12.52	12.53	12.54	12.55	12.56

标准结构化法

等时长采样（Time bar）

很多时候获得的数据并不是真正意义上的 tick 级，而是小时间段上的快照 (snapshot)



标准结构化法

等时长采样（Time bar）算法

等时长采样时，需要将“时间把”里的一组 tick 数据进行汇总，计算

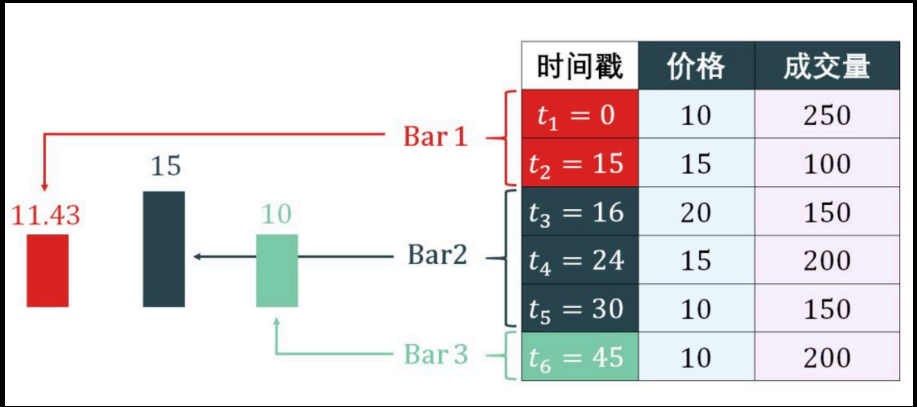
- OHLC: open, high, low, close 价格
- Volume：总成交量
- 时间加权平均价 (Time-weighted Average Price, TWAP):

$$TWAP(t_i, t_j) = \frac{1}{n} \sum_{t_{i,s}=1}^{t_{i,s}=n} price(t_{i,s})$$

- 成交量加权平均价 (Volume-weighted Average Price, VWAP):

$$VWAP(t_i, t_j) = \frac{\sum_{t_{i,s}=1}^{t_{i,s}=n} price(t_{i,s}) * volume(t_{i,s})}{\sum_{t_{i,s}=1}^{t_{i,s}=n} volume(t_{i,s})}$$

标准结构化法: 等时长采样（Time bar）算法



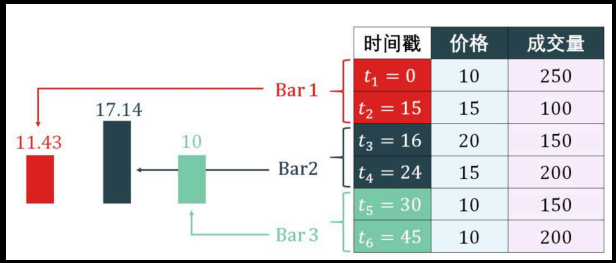
时间戳	价格	成交量
$t_1 = 0$	10	250
$t_2 = 15$	15	100
$t_3 = 16$	20	150
$t_4 = 24$	15	200
$t_5 = 30$	10	150
$t_6 = 45$	10	200

标准结构化法: 等时长采样（Time bar）的缺点

- 相同周期内金融市场的信息含量有显著不同
- 等时长采样下的金融序列的统计学特性非常差：非高斯、非 $i.i.d$

标准结构化法: 等时点采样（Tick bar）

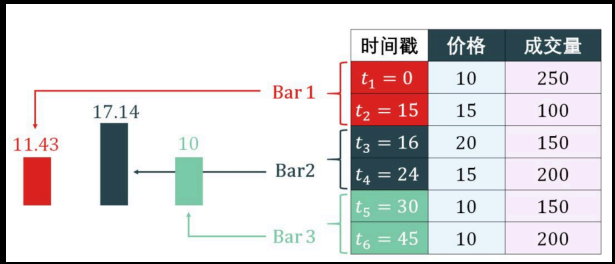
- 等时点采样（时点把）：以固定的报价变动数量（每 10 次、每 20 次、每 500 次）来记录数据的结构化方法
- 等时点采样需要将“相同时点把”里的一组 tick 数据进行汇总，计算 OHLC、总成交量、TWAP 和 VWAP



时间戳	价格	成交量
$t_1 = 0$	10	250
$t_2 = 15$	15	100
$t_3 = 16$	20	150
$t_4 = 24$	15	200
$t_5 = 30$	10	150
$t_6 = 45$	10	200

标准结构化法: 等量采样（Volume bar）

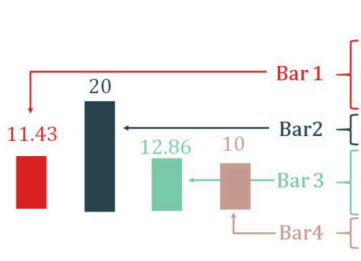
- 等量采样（成交量把）：以固定成交量（1000、5000、10000 等）来记录数据的结构化方法
- 等量采样需要将“相同成交量把”里的一组 tick 数据进行汇总，计算 OHLC、总成交量、TWAP 和 VWAP



时间戳	价格	成交量
$t_1 = 0$	10	250
$t_2 = 15$	15	100
$t_3 = 16$	20	150
$t_4 = 24$	15	200
$t_5 = 30$	10	150
$t_6 = 45$	10	200

标准结构化法: 等额采样（Value bar）

- 等额采样（成交额把）：以固定成交金额（10 万、100 万、1000 万等）来记录数据的结构化方法
- 等额采样需要将“相同成交额把”里的一组 tick 数据进行汇总，计算 OHLC、总成交量、TWAP 和 VWAP



时间戳	价格	成交量	成交额
$t_1 = 0$	10	250	2500
$t_2 = 15$	15	100	1500
$t_3 = 16$	20	150	3000
$t_4 = 24$	15	200	3000
$t_5 = 30$	10	150	1500
$t_6 = 45$	10	200	2000

标准结构化法: 等时点采样，等量采样，等额采样的优点

- 等时点采样：获得接近于 $i.i.d$ 正态分布的回报
- 等量采样：比等时点采样获得接近于 $i.i.d$ 正态分布的回报
- 等额采样：当价格有大幅波动时，采样的回报更加平稳，容易记录趋势；屏蔽股本变化带来的影响

信息驱动的结构化方法

时点平衡性采样：Tick Imbalance bars

- 信息驱动的结构化法的作用是使得新信息进入市场时能够获得更加频繁的采样，在信息匮乏时采样频率相应放慢，从而将采样和信息的流动同步，获得等信息量的衍生序列。
- 时点平衡性采样：将累积买卖平衡度在一定范围内的 tick 数据进行汇总。
- 买卖不平衡度：

$$b_t = \begin{cases} b_{t-1}, & if \Delta p_t = 0 \\ \text{sign}(\Delta p_t), & if \Delta p_t \neq 0 \end{cases}$$

- 累积买卖不平衡度：

$$\theta_{t_{i,j}} = \sum_{t_{i,s}=1}^{t_{i,j}} b_{t_{i,s}} \qquad \theta_T = \sum_{t=1}^{T} b_t$$

信息驱动的结构化方法

时点平衡性采样：Tick Imbalance bars

- 采样算法：

$$j^* = \arg\min_j \left\{ |\theta_{t_{i,j}}| \geq \mathbb{E}_{t_{i,0}}(\theta_{t_{i,j}}) \right\}$$

改写为简洁的形式：

$$T^* = \arg\min_T \left\{ |\theta_T| \geq \mathbb{E}_0(\theta_T) \right\}$$

- 可以证明：

$$\mathbb{E}_0(\theta_T) = \mathbb{E}_0(T) \cdot (P(b_t = 1) - P(b_t = -1))$$

信息驱动的结构化方法

时点平衡性采样：Tick Imbalance bars

■ 证明：

$$\mathbb{E}_0(\theta_T) = \mathbb{E}_0\left[\sum_{t=1}^{T} b_t\right]$$

$$= \mathbb{E}_0\left[\sum_{t=1}^{+\infty} b_t \cdot \phi(T > t-1)\right] = \sum_{t=1}^{+\infty} \mathbb{E}_0(b_t) \cdot \mathbb{E}_0(\phi(T > t-1))$$

$$= \mathbb{E}_0(b_t) \cdot \sum_{t=1}^{+\infty} \mathbb{E}_0(\phi(T > t-1)) = \mathbb{E}_0(b_t) \cdot \sum_{t=1}^{+\infty} P(T > t-1)$$

$$= \mathbb{E}_0(b_t) \cdot \sum_{t=0}^{+\infty} P(T > t) = \mathbb{E}_0(b_t) \cdot \mathbb{E}_0(T)$$

$$= \mathbb{E}_0(T) \cdot (P(b_t = 1) - P(b_t = -1))$$

信息驱动的结构化方法

成交量平衡性采样（Volume Imbalance bars）

- 成交量平衡性采样：将累积的成交量平衡度限定在一定范围内的 tick 数据进行汇总
- 累积成交量不平衡度:

$$\theta_T = \sum_{t=1}^{T} b_t v_t$$

- 采样算法:

$$T^* = \arg\min_T \left\{ |\theta_T| \geq \mathbb{E}_0(T) \cdot \left[P(b_t = 1)\mathbb{E}_0(v_t \mid b_t = 1) - P(b_t = -1)\mathbb{E}_0(v_t \mid b_t = -1) \right] \right\}$$

信息驱动的结构化方法

成交额平衡性采样（Value Imbalance bars）

- 成交额平衡性采样：将累积的成交额平衡度限定在一定范围内的 tick 数据进行汇总
- 累积成交额不平衡度:

$$\theta_T = \sum_{t=1}^{T} b_t q_t = \sum_{t=1}^{T} b_t v_t \, price(t)$$

- 采样算法：

$$T^* = \arg\min_T \left\{ |\theta_T| \geq \mathbb{E}_0(T) \cdot \left[P(b_t = 1)\mathbb{E}_0(q_t \mid b_t = 1) - P(b_t = -1)\mathbb{E}_0(q_t \mid b_t = -1) \right] \right\}$$

信息驱动的结构化方法

时点游程采样（Ticks Runs bars）

- 时点游程采样：将单向的买卖持续度限定在一定范围内的 tick 数据进行汇总
- 单向买卖持续度:

$$\theta_T = \max\left\{\sum_{t|b_t=1}^{T} b_t, \sum_{t|b_t=-1}^{T} -b_t\right\}$$

- 采样算法：

$$T^* = \arg\min_T\left\{|\theta_T| \geq \mathbb{E}_0(T) \cdot \max\left[P(b_t = 1), 1 - P(b_t = 1)\right]\right\}$$

信息驱动的结构化方法

交易量游程采样（Volume Runs bars）

- 时点游程采样：将单向的成交量持续度限定在一定范围内的 tick 数据进行汇总
- 单向成交量持续度:

$$\theta_T = \max\left\{\sum_{t|b_t=1}^{T} b_t v_t, \sum_{t|b_t=-1}^{T} -b_t v_t\right\}$$

- 采样算法：

$$T^* = \arg\min_T \left\{|\theta_T| \geq \mathbb{E}_0(T)\cdot\max\left[P(b_t=1)\mathbb{E}_0(v_t \mid b_t=1), 1-P(b_t=1)E_0(v_t \mid b_t=-1)\right]\right\}$$

信息驱动的结构化方法

交易额游程采样（Value Runs bars）

- 时点游程采样：将单向的成交额持续度限定在一定范围内的 tick 数据进行汇总
- 单向成交额持续度:

$$\theta_T = \max\left\{\sum_{t|b_t=1}^{T} b_t q_t, \sum_{t|b_t=-1}^{T} -b_t q_t\right\}$$

- 采样算法：

$$T^* = \arg\min_T\left\{|\theta_T| \geq \mathbb{E}_0(T)\cdot\max\left[P(b_t=1)\mathbb{E}_0(q_t \mid b_t=1), 1-P(b_t=1)E_0(q_t \mid b_t=-1)\right]\right\}$$

信息驱动的结构化方法

遗留的问题

- 如何确定 $\mathbb{E}_0(T)$?? 在实际操作中，可以采用已采样本的采样间隔 T 的指数移动平均来近似代替
- 如何计算 $P(b_t = 1)$?? 在实际操作中，可以采用已采样本中买入性 tick 所占的比例的指数移动平均来近似代替
- 【思考】如何计算 $P(v_t \mid b_t = 1)$?? $P(q_t \mid b_t = 1)$??

多产品序列的结构化处理

- 动机 1：需要对权重动态变化的资产组合进行建模
- 动机 2：需要对不定期支付息票或股息的金融产品进行处理
- 动机 3：需要考虑时间序列的结构变化（structural break）
- 目标：将任何复杂的多产品数据转化为一个回报类似 ETF 的单一数据 (ETF 技巧)

多资产序列的结构化处理

ETF 技巧

假设结构化采样的时点为 $t = 1, 2, \cdots, T$，组合中的金融资产为 $i = 1, , 2, \ldots I$。当结构化采样产生后，已知如下数据

- $o_{i,t}$: 原始的开盘价
- $p_{i,t}$: 原始的收盘价
- $\varphi_{i,t}$: 时点上一个点的价值
- $v_{i,t}$: 时点上的成交量
- $d_{i,t}$: 时点上支付的息票，息票与资金成本之差 (carry)，股利分红

多产品序列的结构化处理

ETF 技巧

w_t 表示在采样时点的子集 $B \subseteq \{1, 2, \cdots, T\}$ 中资产组合的重新平衡权重，转换为单一回报的组合价值 K_t 按照如下方式计算

$$K_t = K_{t-1} + \sum_i^I h_{i,t-1}\ \varphi_{i,t}(\delta_{i,t} + d_{i,t})$$

$$\delta_{i,t} = \begin{cases} p_{i,t} - o_{i,t}, & \text{如果 } t-1 \in B \\ \Delta p_{i,t}, & \text{否则} \end{cases} \qquad h_{i,t-1} = \begin{cases} \dfrac{w_{i,t-1} K_{t-1}}{o_{i,t} \varphi_{i,t-1} \sum_i^I |w_{i,t-1}|}, & \text{如果 } t-1 \in B \\ h_{i,t-2}, & \text{否则} \end{cases}$$

多资产序列的结构化处理

ETF 技巧

- $h_{i,t}$ 表示在时刻 t 资产 i 的持有量（合约数量）
- $\delta_{i,t}$：资产 i 从 $t-1$ 到 t 的市场价值变化
- $\frac{w_{i,t}}{\sum_i^I |w_{i,t}|}$ 表示每个资产的占比分配

多资产序列的结构化处理

ETF 技巧：w_i

- 使用主成分分析法 (Principle Component Analysis)
- $\mu_{1\times N}$: 收益均值向量
- $V_{N\times N}$: 协方差矩阵
- 第一步: 进行谱分解：$VX = X\Lambda$
- 第二步：引入 w_i，$\sigma^2 = w^T V w = w^T X \Lambda X^T w = \beta^T \Lambda \beta$
- 第三步，主成分分解：第 n 个成分的风险为 $R_n = \beta_n^2 \Lambda_{n,n} \sigma^{-2} = (X^T w)_n^2 \lambda_n \sigma^{-2}$，$R_n$ 是第 n 个主成分的风险
- 第四步，计算 β：$\beta_n = \sigma \sqrt{\frac{R_n}{\lambda_n}}$
- 第五步，计算 w_i: $w = X\beta$

对数据进行结构化抽样

数据已经结构化，为什么还要继续进行结构化抽样 (Sampling)？

- 很多机器学习算法无法支持规模很大的数据（例如 SVM）
- 只有数据具有足够的信息量时，机器学习算法才能形成具有较高预测精度的学习器。

结构化抽样

- 缩减抽样 (Sampling for reduction)
- 事件驱动抽样 (Event-based Sampling)

对数据进行结构化抽样: 缩减抽样

- 缩减抽样：抽取出符合金融机器学习算法建模要求的数据量
- 线性等分抽样 (linspace sampling)：通过恒定的步长进行序列抽样
  - 优点：简单；缺点：步长随意，结果受随机种子而变化
- 均匀抽样 (uniform sampling)：以均匀分布进行随机抽样
  - 优点：简单；缺点：数量任意，结果受随机种子而变化
- 两类抽样不一定包含那些在预测能力或信息内容方面最相关数据样本

对数据进行结构化抽样: 事件驱动抽样

- 重要的事件往往带有较强预测能力的信息
- 事件可能与某些宏观经济统计数据的发布、波动率飙升、曲线差值与平衡水平的显著偏离等有关
- 通过对特征识别来推定有信息量的事件
  - 机制转换 (Regime shift) 或结构变化 (Structural break)
  - 熵变化 (Entropy Change)
  - 市场微观结构 (Microstructural Information)

事件驱动抽样算法：CUSUM 过滤法

设结构化的数据为 $y_t,\ t = 1, 2, \cdots, T$

- 蓄势累积指标 $S_t = \max\{S_{t-1} + (y_t - \mathbb{E}_{t-1}(y_t)), 0\}$
- 如果累计指标大于阈值，对 t 进行抽样：$S_t \geqslant h$
- 变体：对称抽样

$$S_t^+ = \max\{S_{t-1}^+ + (y_t - \mathbb{E}_{t-1}(y_t)), 0\}, \quad S_0^+ = 0$$

$$S_t^- = \max\{S_{t-1}^- + (y_t - \mathbb{E}_{t-1}(y_t)), 0\}, \quad S_0^- = 0$$

- 简单应用，取 $\mathbb{E}_{t-1}(y_t) = y_{t-1}$