Financial Data Structures

ESSENTIAL TYPES OF FINANCIAL DATA

Fundamental Data	Market Data	Analytics	Alternative Data
• Assets • Liabilities • Sales • Costs/earnings • Macro variables • . . .	• Price/yield/implied volatility • Volume • Dividend/coupons • Open interest • Quotes/cancellations • Aggressor side • . . .	• Analyst recommendations • Credit ratings • Earnings expectations • News sentiment • . . .	• Satellite/CCTV images • Google searches • Twitter/chats • Metadata • . . .

Fundamental Data

- be found in regulatory filings and business analytics
- You must confirm exactly when each data point was released, so that your analysis uses that information only after it was publicly available.
- A second aspect of fundamental data is that it is often backfilled or reinstated.
    - "Backfilling" means that missing data is assigned a value, even if those values were unknown at that time.
- Fundamental data is extremely regularized and low frequency. Being so accessible to the marketplace, it is rather unlikely that there is much value left to be exploited. Still, it may be useful in combination with other data types.

Market Data

- all trading activity that takes place in an exchange (like CME) or trading venue
- Every market participant leaves a characteristic footprint in the trading records, and with enough patience, you will find a way to anticipate a competitor's next move.
- Human GUI traders often trade in round lots, and you can use this fact to estimate what percentage of the volume is coming from them at a given point in time, then associate it with a particular market behavior.

Analytics

- derivative data, based on an original source, which could be fundamental, market, alternative, or even a collection of other analytics.
- A positive aspect of analytics is that the signal has been extracted for you from a raw source. The negative aspects are that analytics may be costly, the methodology used in their production may be biased or opaque, and you will not be the sole consumer.

Alternative Data

- Alternative data offers the opportunity to work with truly unique, hard-to-process datasets. Remember, data that is hard to store, manipulate, and operate is always the most promising.

BARS

Time Bars

- should be avoided for two reasons.
    - markets do not process information at a constant time interval
- time bars oversample information during low-activity periods and undersample information during high-activity periods.
- time-sampled series often exhibit poor statistical properties,

Tick Bars

- The sample variables listed earlier (timestamp, VWAP, open price, etc.) will be extracted each time a pre-defined number of transactions takes place, e.g., 1,000 ticks.
- This allows us to synchronize sampling with a proxy of information arrival
- tick bars allow for better inference than time bars.

Volume Bars

- we could sample prices every time a futures contract exchanges 1,000 units, regardless of the number of ticks involved.
- Another reason to prefer volume bars over time bars or tick bars is that several market microstructure theories study the interaction between prices and volume.

Dollar Bars

- Dollar bars are formed by sampling an observation every time a pre-defined mar- ket value is exchanged.
- A second argument that makes dollar bars more interesting than time, tick, or volume bars is that the number of outstanding shares often changes multiple times over the course of a security's life, as a result of corporate actions.
- the bar size could be adjusted dynamically as a function of the free-floating market capitalization of a company

Information-Driven Bars

- Tick Imbalance Bars
    - sample bars whenever tick imbalances exceed our expectations
    - The so-called tick rule defines a sequence {bt}t=1,...,T where
        - $$b_t = \begin{cases} b_{t-1} & \text{if } \Delta p_t = 0 \\ \frac{|\Delta p_t|}{\Delta p_t} & \text{if } \Delta p_t \neq 0 \end{cases}$$
    - define the tick imbalance at time T
        - $$\theta_T = \sum_{t=1}^{T} b_t$$
    - estimate E0[T] as an exponentially weighted moving average of T values from prior bars
    - a tick imbalance bar (TIB) as a T∗
        - $$T^* = \arg\min_T \left\{ |\theta_T| \geq \mathrm{E}_0[T] \left| 2\mathrm{P}[b_t = 1] - 1 \right| \right\}$$
- Volume/Dollar Imbalance Bars
    - We would like to sample bars when volume or dollar imbalances diverge from our expectations.
    - define the imbalance at time T
        - $$\theta_T = \sum_{t=1}^{T} b_t v_t$$
    - compute the expected value of θT at the beginning of the bar
        - estimate E0[T] as an exponentially weighted moving average of T values from prior bars, and (2v+ − E0[vt]) as an exponentially weighted moving average of btvt values from prior bars.

- we define VIB or DIB as a T∗-contiguous subset of ticks
  - $$T^* = \arg\min_T \{|\theta_T| \geq \mathrm{E}_0[T]|2v^+ - \mathrm{E}_0[v_t]|\}$$

Tick Runs Bars

- TIBs, VIBs, and DIBs monitor order flow imbalance
- useful to monitor the sequence of buys in the overall volume, and take samples when that sequence diverges from our expectations.
  - define the length of the current run as
    - $$\theta_T = \max\left\{\sum_{t|b_t=1}^{T} b_t, -\sum_{t|b_t=-1}^{T} b_t\right\}$$
  - compute the expected value of θT at the beginning of the bar
    - estimate E0[T] as an exponentially weighted moving average of T values from prior bars, and P[bt = 1] as an exponentially weighted moving average of the proportion of buy ticks from prior bars.
  - a tick runs bar (TRB) as a T∗-contiguous subset of ticks such that the following condition is met:
    - $$T^* = \arg\min_T \{\theta_T \geq \mathrm{E}_0[T]\max\{\mathrm{P}[b_t = 1], 1 - \mathrm{P}[b_t = 1]\}\}$$

Volume/Dollar Runs Bars

- sample bars whenever the volumes or dollars traded by one side exceed our expectation for a bar.
- define the volumes or dollars associated with a run as
  - $$\theta_T = \max\left\{\sum_{t|b_t=1}^{T} b_t v_t, -\sum_{t|b_t=-1}^{T} b_t v_t\right\}$$
- compute the expected value of θT at the beginning of the bar,
  - estimate E0[T] as an exponentially weighted moving average of T values from prior bars
- define a volume runs bar (VRB) as a T∗-contiguous subset of ticks such that the following condition is met:
  - $$T^* = \arg\min_T \{\theta_T \geq \mathrm{E}_0[T]\max\{\mathrm{P}[b_t = 1]\mathrm{E}_0[v_t|b_t = 1], (1 - \mathrm{P}[b_t = 1])\mathrm{E}_0[v_t|b_t = -1]\}\}$$

Sampling Features

- several ML algorithms do not scale well with sample size
- ML algorithms achieve highest accuracy when they attempt to learn from relevant examples.

Sampling for Reduction

- one reason for sampling features from a structured dataset is to reduce the amount of data used to fit the ML algorithm.
- Downsampling: done by either sequential sampling at a constant step size (linspace sampling), or by sampling randomly using a uniform distribution (uniform sampling).

Event-Based Sampling

- Portfolio managers typically place a bet after some event takes place → struc- tural break, extracted signal, microstructural phenomena
    - events could be associated with the release of some macroe- conomic statistics, a spike in volatility, a significant departure in a spread away from its equilibrium level, etc.
- We can characterize an event as significant, and let the ML algorithm learn whether there is an accurate prediction function under those circum- stances.

The CUSUM Filter

- a quality-control method, designed to detect a shift in the mean value of a measured quantity away from a target value.
- the filter is set up to identify a sequence of upside divergences from any reset level zero.