STAT430: Machine Learning for Financial Data

LarryHua.com/teaching
Spring 2019

Financial machine learning applications

- Price prediction / algo trading
- Anomaly detection / Risk analysis
- Portfolio construction
- Credit Ratings

- See more examples: [Ten Financial Applications of Machine Learning]

Financial data types

- Essential Types of Financial Data
- TABLE 2.1 of AFML

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 • . . .

Structured bars

- Standard bars
  - Time bars
  - Tick bars
  - Volume bars
  - Dollar bars
- Information-driven bars: to sample more frequently when new information arrives
  - Tick imbalance bars
  - Volume/dollar imbalance bars
  - TIBs, VIBs, and DIBs monitor order flow imbalance, as measured in terms of ticks, volumes, and dollar values exchanged

Structured bars

- More information-driven bars:
  - Tick runs bars
  - Volume/dollar runs bars
  - Monitor the sequence of buys in the overall volume, and take samples when that sequence diverges from our expectations

Time bars

- Sampling information at fixed time intervals, e.g., once every minute
- Timestamp / Open / Close / High / Low / Volume
- Limitations:
  - Oversample / undersample
  - Poor statistical properties: serial correlation, heteroscedasticity, and non-normality
- Try R

Tick bars

- Sampling information at a pre-defined number of transactions, e.g., once every 1000 ticks
- Order fragmentation introduces some arbitrariness in the number of ticks
- Be aware of outliers due to auctions at open/close
- Try R

Volume bars

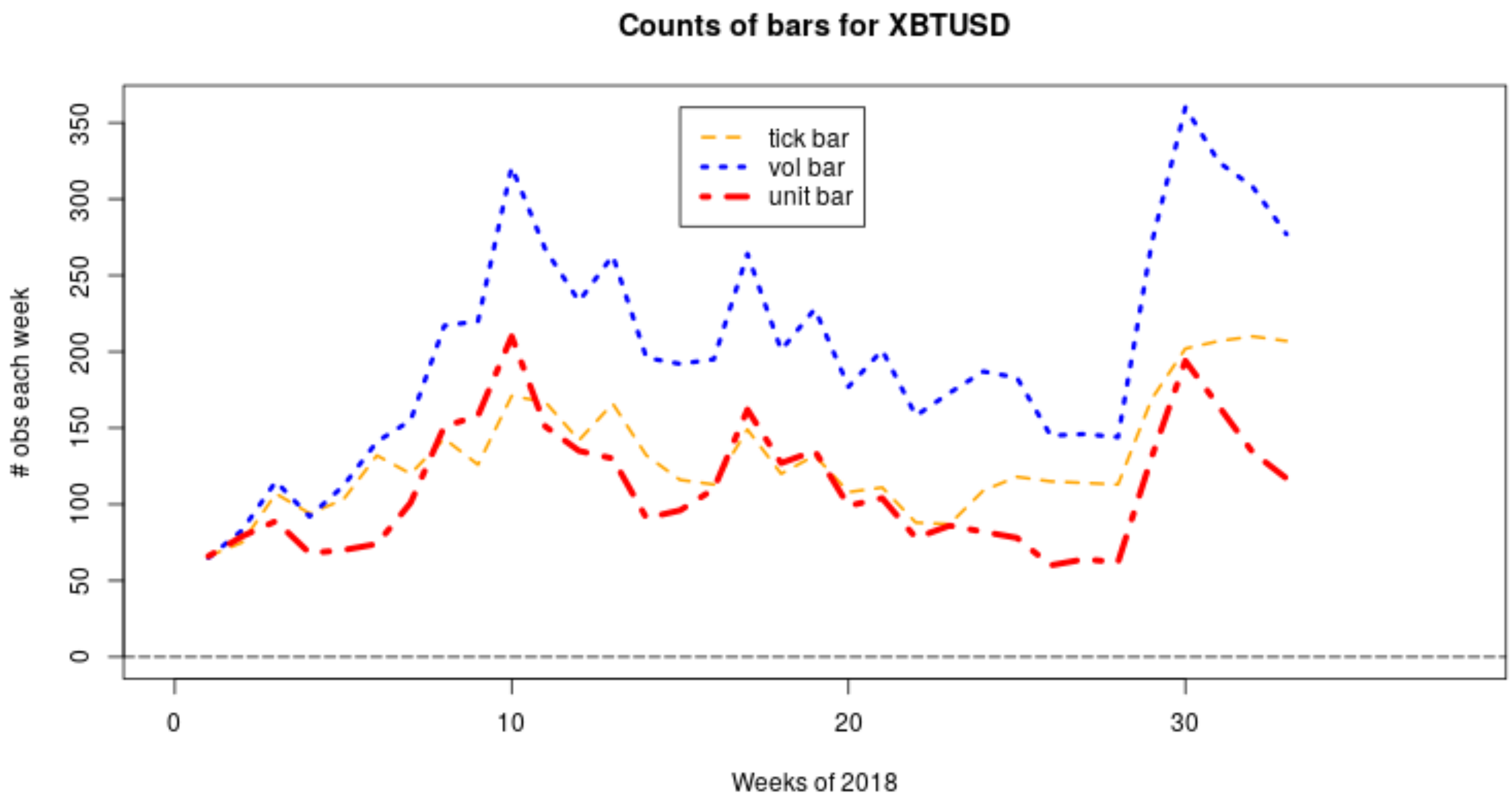
- Sampling information when a pre-defined amount of the security's units have been exchanged

Dollar bars / Unit bars

- Sampling information every time a pre-defined market value is exchanged
- More robust than volume/tick bars
- Amount of ticks and volumes may be affected by corporate actions: splits, buy-back, etc.
- bar size can be fixed over time, or linked to other factors, e.g., free-floating market capitalization of a company

Some comparisons

- Counts of different bars - E-mini S&P 500 futures
  - See [FIGURE 2.1 of AFML]
- Counts of different bars - XBTUSD



Tick imbalance bars

- Calculate a b_t sequence:

$$b_t = \begin{cases} b_{t-1} & \text{if } \Delta p_t = 0 \\ \dfrac{|\Delta p_t|}{\Delta p_t} & \text{if } \Delta p_t \neq 0 \end{cases}$$

- Find tick imbalance at T

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

- $E_0[\theta_T] = E_0[T](P[b_t = 1] - P[b_t = -1])$

- Sample information at T^*

$$T^* = \arg\min_T \left\{ |\theta_T| \geq \mathrm{E}_0[T] \left| 2\mathrm{P}\left[b_t = 1\right] - 1 \right| \right\}$$

Tick imbalance bars

In practice:

- Estimate $E_0[T]$ as an exponentially weighted moving average of T values from prior bars
- Estimate $2P[b_t = 1] - 1$ as an exponentially weighted moving average of bt values from prior bars.
- Try R
- Question: any potential problems for approximating $E_0[T]$??

Volume/dollar imbalance bars

- Find imbalance at T

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

- $$\mathrm{E}_0[\theta_T] = \mathrm{E}_0\left[\sum_{t|b_t=1}^{T} v_t\right] - \mathrm{E}_0\left[\sum_{t|b_t=-1}^{T} v_t\right] = \mathrm{E}_0[T](\mathrm{P}[b_t = 1]\mathrm{E}_0[v_t|b_t = 1]$$
$$-\mathrm{P}[b_t = -1]\mathrm{E}_0[v_t|b_t = -1])$$

- Sample information at T^*, where $v^+ = P[b_t = 1]E_0[v_t|b_t = 1]$

$$T^* = \arg\min_T\{|\theta_T| \geq \mathrm{E}_0[T]|2v^+ - \mathrm{E}_0[v_t]|\}$$

Volume/dollar imbalance bars

In practice

- Estimate $E_0[T]$ as an exponentially weighted moving average of T values from prior bars
- Estimate the second part as an exponentially weighted moving average of $b_t v_t$ values from prior bars

Tick runs bars

- Calculate the length of the current run

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

- $\mathrm{E}_0[\theta_T] = \mathrm{E}_0[T]\max\{\mathrm{P}[b_t = 1], 1 - \mathrm{P}[b_t = 1]\}$
- Sample information at T^*

$$T^* = \arg\min_T \{\theta_T \geq \mathrm{E}_0[T]\max\{\mathrm{P}[b_t = 1], 1 - \mathrm{P}[b_t = 1]\}\}$$

Tick runs bars

- In practice
  - Estimate $E_0[T]$ as an exponentially weighted moving average of T values from prior bars
  - Estimate $P[b_t = 1]$ as an exponentially weighted moving average of the proportion of buy ticks from prior bars
- Instead of measuring the length of the longest sequence (without offsetting), we count the number of ticks of each side without offsetting them
- In the context of forming bars, this turns out to be a more useful definition than measuring sequence lengths
  - Question: please compare tick runs bars with tick imbalance bars empirically.
- Try R

Volume/dollar runs bars

- Calculate volumes or dollars associated with a run

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

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

  $(1 - \mathrm{P}[b_t = 1])\mathrm{E}_0[v_t|b_t = -1]\}\}$

Volume/dollar runs bars

- In practice
  - Estimate $E_0[T]$ as an exponentially weighted moving average of T values from prior bars
  - Estimate $P[b_t = 1]$ as an exponentially weighted moving average of the proportion of buy ticks from prior bars
  - Estimate $E_0[v_t|b_t = 1]$ as an exponentially weighted moving average of the buy volumes from prior bars
  - Estimate $E_0[v_t|b_t = -1]$ as an exponentially weighted moving average of the sell volumes from prior bars
- Back to Course Scheduler