STAT430: Machine Learning for Financial Data

LarryHua.com/teaching

Spring 2019

Sampling features

- Motivations:
  - Reduce sample size: increasing training size does not necessarily lead to better results
  - Use samples with more informative features
- Sampling for reduction
  - Sequential sampling at a constant step size
  - Sampling randomly using a uniform distribution
- Event-based sampling
  - The CUSUM Filter

The CUSUM Filter

- Detect a shift in the mean value of a measured quantity away from a target value
- Consider iid observations $\{y_t\}_{t=1,\ldots,T}$ from a locally stationary process.
- Define the cumulative sums

$$S_t = \max\left\{0, S_{t-1} + y_t - \mathrm{E}_{t-1}\left[y_t\right]\right\}$$

- Sample a bar t if and only if $S_t \geq h$, at which point S_t is reset

$$S_t \geq h \Leftrightarrow \exists \tau \in [1, t] \left| \sum_{i=\tau}^{t} \left(y_i - \mathrm{E}_{i-1}\left[y_t\right]\right) \geq h \right.$$

- The filter is set up to identify a sequence of upside divergences from any reset level zero
- Skip some downward deviations

Symmetric CUSUM filter

- This concept of run-ups can be extended to include run-downs
- $S_t^+ = \max\left\{0, S_{t-1}^+ + y_t - \mathrm{E}_{t-1}\left[y_t\right]\right\},\ S_0^+ = 0$
- $S_t^- = \min\left\{0, S_{t-1}^- + y_t - \mathrm{E}_{t-1}\left[y_t\right]\right\},\ S_0^- = 0$
- $S_t = \max\left\{S_t^+, -S_t^-\right\}$
- Try R
  - Assuming $E_{t-1}\left[y_t\right] = y_{t-1}$

Feature matrix

```
unit_bar <- bar_unit(dat, unit=6e9)
i_CUSUM <- istar_CUSUM(unit_bar$C, h=1000)
iTmp <- c(0, i_CUSUM)

# fMat0 is used to store the feature matrix
fMat0 <- t(sapply(1:(length(i_CUSUM)),
                  function(i)
                  {
                    winTmp <- x[(iTmp[i]+1):(iTmp[i+1])]
                    C <- winTmp[length(winTmp)]
                    SD <- sd(winTmp)

                    # More features can be applied here

                    return(c(C,SD))
                  }
))
```

More hints for prediction

- Variable S_t could be based on some financial features such as structural break statistics, entropy, or market microstructure measurements.
- Once we have obtained this subset of event-driven bars, let the ML algorithm determine whether the occurrence of such events constitutes actionable intelligence.
- Back to Course Scheduler