Advances in Financial Machine Learning

Mithursan Krishnamoorthy

December 7, 2018

Contents

1 Part1:Data Analysis **1**
1.1 Goal: 1
1.2 Financial Data Structures 2
1.3 BARS 2
1.4 Standard Bars 2
1.5 Information Driven Bars 3
1.6 More Information Driven bars: 4
1.7 Dealing with Multi-Product Series
5
1.8 Sampling Features
6

2 Labeling **6**
2.1 Goal:
6
2.2 The Fixed Time Horizon Method 6
2.3 Computing Dynamic Thresholds 7
2.4 The Triple Barrier Method 7
2.5 Learning Side and Size 10
2.6 Meta Labeling 10

1 Part1:Data Analysis

1.1 Goal:

-Learn how to produce matrix X of financial features out of an unstructured dataset.
-Unsupervised learning algorithms can learn the patterns from the matrix X(eg. whether it contains hierarchical cluster)

1.2 Financial Data Structures

- Describes how to work with unstructured financial data and from that to derive a structured dataset for Ml algorithms
- 4 essential types of Financial Data

•**Fundamental Data** (Assets, Liabilities, Sales, Cost/Earnings, Macro Variables)
-Fundamental data is extremely regularized and low frequency. It is rather unlikely that there is much value left to be exploited. But may be useful in combination with other data types.

•**Market Data** (Price/Yield/Implied volatility, Volume, Dividend/coupons, Open interest.....)
- includes all trading activity that takes place in an exchange or trading venue
- most likely data provider would have to provide the information such as FIX messages.....
-unlike Fundamental Data, this data is abundant with over 10TB generated daily

•**Analytics** (Analyst recommendations, Credit ratings, Earnings expectations, News sentiment)
-Could think of analytics as Derivate Data, based on an original source which could be Fundamental, Market, Alternative, or collection of other Analytics

•**Alternative Data** (Satellite/CCTV images, Google searches, Twitter/chats, Metadata)

1.3 BARS

1.4 Standard Bars

1)Time Bars
2)Tick Bars
3)Volume Bars
4)Dollar Bars

-Purpose of these methods is to transform a series of observations that arrive at irregular frequency("inhomogeneous series") into a homogeneous series derived from regular sampling

•**Time Bars** - are obtained by sampling information at fixed intervals(eg, once every minutes)
-(ex, Open Price, Close Price, High Price, Low Price)
-Although popular should be avoided because Markets do not process info at constant time interval and time sampled series often exhibit poor statistical

properties(serial correlation, heteroscedasticity, non normality)

•**Tick bars**-idea behind it is that sample variables like (Open Price, Close Price, High Price, Low Price) will be extracted each time a pre-defined number of transactions take place(eg, 1000 tickets)
-"Price changes over a fixed number of transactions may have a Gaussian distribution. Price changes over a fixed time period may follow a stable Paretian distribution, whose variance is infinite."
-need to be aware of outliers
-Allows to synchronize sampling with a proxy of information arrival

•**Volume Bars**-Volume bars can sample every time a pre-defined amount of the securities units has been exchanged(ex, sample prices every time a futures contract exchanges 1000 units, regardless of the number of ticks involved
-(1973)Sample returns by volume achieved even better statistical properties than sampling by tick bars(i.e closer to an IID Gaussian distribution)

•**Dollar Bars**-formed by sampling an observation every time a pre-defined market value is exchanged.
"ex, suppose we wish to analyze a stock that has exhibited an appreciation of 100 percent over a certain period of time. Selling 1000 worth of that stock at the end of the period requires trading half the number of shares it took to buy 1000 worth of that stock at the beginning. The number of shares traded is a function of the actual value exchanged.

1.5 Information Driven Bars

-To sample more frequently when new information arrives
1)Tick Imbalance Bars
2)Volume/Dollar Imbalance Bars
3)TIBs, VIBs, DIBs(monitor order flow imbalance, as measure in terms of ticks, volumes, and dollar values exchanged)

•**Tick Imbalance Bars**
-We try to sample bars whenever tick imbalances exceed our expectation.
-Similar to STA457 project

- 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^*

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

-We try to determine the tick index, T, such that the accumulation of signed ticks exceeds a given threshold.
-Can't differentiate between order size of 10 and order 1(counts both as 1)

1.6 More Information Driven bars:

•Volume/Dollar Imbalance Bars
-The idea behind volume imbalance bars(VIB's) and dollar imbalance bars(DIBs) is we would like to sample bars when volume or dollar imbalances diverge from our expectations.
-TIBs, VIBs, DIBs monitor order flow imbalance as measures in terms of ticks, volumes, and dollar values exchanged.

•Tick Runs Bars
-Based on notions of tick rules and boundary condition discussed in TIBs, we can define a procedure to determine the index of the next sample, T.

-"Large traders will sweep the order books, use iceberg orders, or slick a parents order into multiple children, all of which leave a trace of runs in (bt) sequence"
-For that reason its better to monitor sequence of buys in the overall volume and take samples when that sequence diverges from our expectations.

•Volume/Dollar Runs Bars
-Volume Run Bar(VRBs), and Dollar Runs Bars(DRBs)
-We wish to sample bars whenever the volumes or dollars traded by one side exceed our expectation for a bar

1.7 Dealing with Multi-Product Series

-interested in modelling a time series of instruments, where the weights need to be dynamically adjusted over time.

-other time we may deal with products that pay irregular coupons/dividends
-We will have structural breaks if time series that was altered by
•The ETF Trick(add more)
-we wish to develop a strategy that trades a spread of futures

Nuisances
-spread is characterized by a vector of weights that changes over time.
-As a result, the spread itself may converge even if prices do not change.(when that happens, a model trading that series will be misled to believe that PnL has resulted from that weight induced convergence)
-Spreads require negative values, b/c they do not represent a price

-Ways to avoid these issues is to produce time series that reflects the value of 1 Dollar invested in a spread.
-Changes in the series will reflect changes in PnL, the series will be strictly position and implementation shortfall will be taken into account.
-This series will be used to model, generate signals, and trade as if it were an ETF

•PCA Weights

•Single Future Roll
-When dealing with a single futures contract, an equivalent and more direct approach is to form a time series of cumulative roll gaps, and detract that gaps series from the prices series

1.8 Sampling Features

-We learned how to produce a continuous, homogeneous and structured dataset from a collection of unstructured financial data
-Going to learn ways of sampling bars to produce a features matrix with relevant training examples.

•**Sampling for Reduction**
-Sampling features from dataset to reduce the amount of data used to fit the ML algorithm

•**Event-Based Sampling** - CUSUM Filter
-The CUSUM filter is a quality control method, designed to detect a shift in the mean value of a measure quantity away from a target value

2 Labeling

2.1 Goal:

-Describes ways to label financial data
-Supervised learning algorithms require that the rows of X are associated with an array of labels/values y, so that those labels/values can be predicted on unseen features samples
-create response variables for supervised learning

2.2 The Fixed Time Horizon Method

-Almost all ML papers label observations using the fixed time horizon method
-Features matrix X with I rows, each row (Xi) is the features sampled from some bars with index t=1,. . . ,T, where IT

$\{X_i\}_{i=1,\ldots,I}$. An observation X_i is assigned a label $y_i \in \{-1, 0, 1\}$,

$$y_i = \begin{cases} -1 & \text{if } r_{t_{i,0},t_{i,0}+h} < -\tau \\ 0 & \text{if } |r_{t_{i,0},t_{i,0}+h}| \leq \tau \\ 1 & \text{if } r_{t_{i,0},t_{i,0}+h} > \tau \end{cases}$$

where τ is a pre-defined constant threshold, $t_{i,0}$ is the index of the bar immediately after X_i takes place, $t_{i,0} + h$ is the index of the h-th bar after $t_{i,0}$, and $r_{t_{i,0},t_{i,0}+h}$ is the price return over a bar horizon h,

$$r_{t_{i,0},t_{i,0}+h} = \frac{p_{t_{i,0}+h}}{p_{t_{i,0}}} - 1$$

43

-For each Xi(row vector), we assign label yi
-Features = Independent variables/predictors;
-Label = response variable
-Limitations:
-h is fixed and time bars do not exhibit good statistical properties
-the threshold of tau is fixed regardless of the observed volatility
Alternatives:
1)Label per a varying threshold (sdt), estimated using a rolling exponentially weighted standard deviation of returns
2)Use volume/dollar bars, as their volatility's are much closer to constant(homoscedasticity)
Even these 2 improvements don't improve the flaw: path followed by prices
Stop Loss limits ???

2.3 Computing Dynamic Thresholds

-In practise we want to set profit taking and stop loss limits that are a function of the risks involved in a bet
-Run R code"Daily Volatility Estimates"

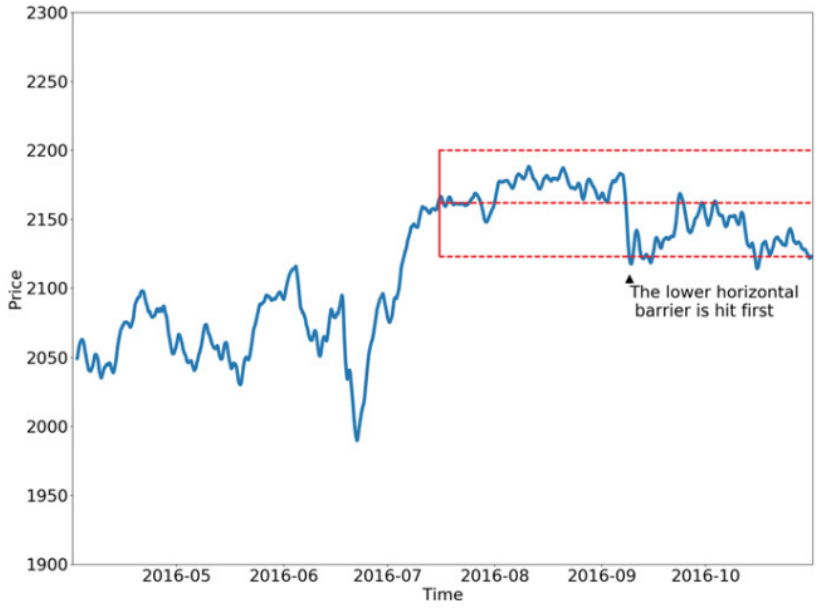
2.4 The Triple Barrier Method

-Labels an observation according to the first barrier touched out of three barriers
1)we set 2 horizontal barriers, and 1 vertical barrier
2)The horizontal barriers are defined by profit taking, and stop loss limits, which are a dynamic function of estimated volatility
3)the third barrier is defined in terms of the number of bars elapsed since the position was taken(an expiration limit)
4)If upper barrier is touched first, we label observation as 1
5)If the lower barrier is touched first, we label the observation as -1

6)If the vertical barrier is touched first, we have two choices: the sign of the return, or a 0(Author prefers sign of return as a matter of realizing a profit or loss within limits)
-Triple barrier method is path dependent(?)
-[pt, sl, t1]

- `close`: A pandas series of prices.
- `events`: A pandas dataframe, with columns,
  - `t1`: The timestamp of vertical barrier. When the value is `np.nan`, there will not be a vertical barrier.
  - `trgt`: The unit width of the horizontal barriers.
- `ptSl`: A list of two non-negative float values:
  - `ptSl[0]`: The factor that multiplies `trgt` to set the width of the upper barrier. If 0, there will not be an upper barrier.
  - `ptSl[1]`: The factor that multiplies `trgt` to set the width of the lower barrier. If 0, there will not be a lower barrier.
- `molecule`: A list with the subset of event indices that will be processed by a single thread. Its use will become clear later on in the chapter.

- Three useful configurations:
  - [1,1,1]: This is the standard setup, where we define three barrier exit conditions. We would like to realize a profit, but we have a maximum tolerance for losses and a holding period.
  - [0,1,1]: In this setup, we would like to exit after a number of bars, unless we are stopped-out.
  - [1,1,0]: Here we would like to take a profit as long as we are not stopped-out. This is somewhat unrealistic in that we are willing to hold the position for as long as it takes.
- Three less realistic configurations:
  - [0,0,1]: This is equivalent to the fixed-time horizon method. It may still be useful when applied to volume-, dollar-, or information-driven bars, and multiple forecasts are updated within the horizon.
  - [1,0,1]: A position is held until a profit is made or the maximum holding period is exceeded, without regard for the intermediate unrealized losses.
  - [1,0,0]: A position is held until a profit is made. It could mean being locked on a losing position for years.
- Two illogical configurations:
  - [0,1,0]: This is an aimless configuration, where we hold a position until we are stopped-out.
  - [0,0,0]: There are no barriers. The position is locked forever, and no label is generated.


2300
2250
2200
2150
2100
2050
2000
1950
1900
Price
2016-05
2016-06
2016-07
2016-08
2016-09
2016-10
Time
The lower horizontal
barrier is hit first

(a)


2500
2450
2400
2350
2300
2250
2200
Price
2017-02
2017-03
2017-04
2017-05
2017-06
2017-07
Time
The vertical barrier
is hit first

(b)

2.5 Learning Side and Size

-learn how to label examples so that an ML algorithm can learn both the side and the size of a bet

-interested in learning the side of a bet when we do not have an **underlying model** to set the sign of our position(long or short)
-under that circumstance, we cant differentiate between a profit taking barrier, and a stop loss barrier, since it requires knowledge of the side
-Learning size implies that either there are no horizontal barriers or that the horizontal barriers must be symmetric

2.6 Meta Labeling

-Learn what is the appropriate size of the bet

-We already have a model for setting the side of the bet(long, short)

-Now we need to learn the size of the bet, which includes the possibility of no bet at all(zero size)

-We often know whether we want to buy or sell a products, only remaining question is how much money we should risk in such a bet.

-Don't want Ml algorithm to learn the side, just want it to tell us what is appropriate size
- **Make some adjustments to previous code in order to handle meta labeling(look at code that expands getEvents to incorporate Meta-Labeling)
-Primary model: learning only the side of the best(eg, moving average crossing strategy)
-Quantamental investment research

How to use Meta-Labeling?
-Meta-Labeling is helpful when you want to achieve highest F1 scores.

-**F1 Scores**:measures the efficiency of a classifier as the harmonic average between precision and recall???

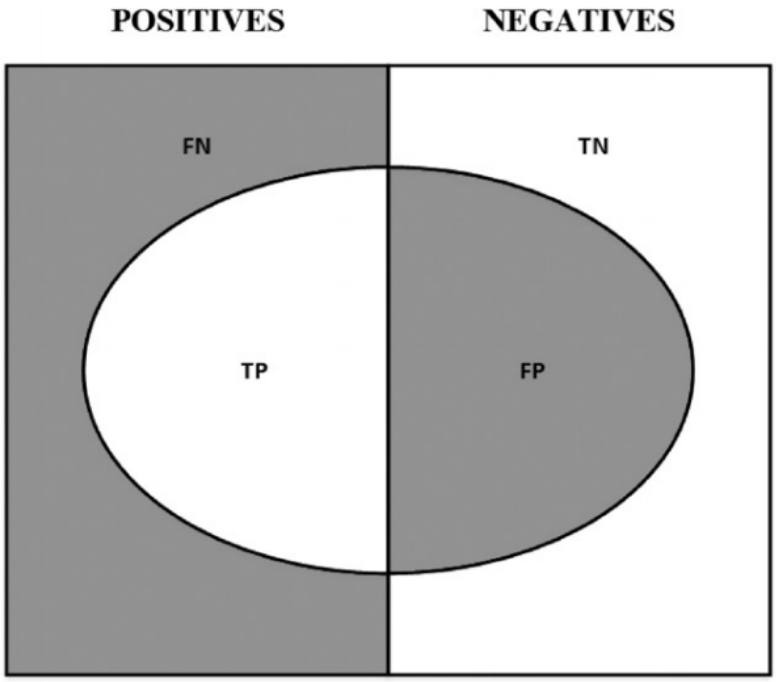


FIGURE 3.2 A visualization of the "confusion matrix"

-Binary Classifier predicts that some items exhibit the condition (ellipse), where the TP area contains the true positives and TN are contains the true negatives

-Leads to 2 kind of erros:False Positives(FP), and False Negatives(FN)

-Precision is the ratio between TP area and the area in the ellipse.

-Recall is the ratio between the TP area and the area in the left rectangle

-Accuracy is the sum of the TP and TN areas divided by the overall set of items(square)

-decreasing the FP are comes at a cost of increasing the FN area, because higher precision typically means fewer calls, hence lower recall.

-There is some combination of precision and recall that maximizes the overall efficiency of the classifier.

1) we want to build model that achieves high recall, even if the precision is not particularly high

2) we correct for the low precision by applying meta-labeling to the positives predicted by the primary model

-Meta-Labeling will increase your F! score by filtering out the false positives, where the majority of positives have already been identified by the primary model

Reasons why Meta-Labeling is powerful

1)Since ML algorithms are criticized as black boxes, Meta-Labeling allowed you to build an ML system on top of a white box(like a fundamental model founded on economic theory)

-ability to transform a fundamental model into an ML model makes meta-labeling useful

2)Effects of over fitting are limited when you apply meta-labeling, because ML will not decide the side of your bet, only the size

3)By decoupling the side prediction from the size prediction, meta-labeling enables sophisticated strategy structures(eg, features driving a rally may differ from the features driving a sell off, in that case you want to develop an Ml strategy exclusively for long positions, based on the buy recommendations of a primary model, and an ML strategy exclusively for short positions, based on the sell recommendations of an entirely different primary model)

4)Achieving high accuracy on small bets and low accuracy on large bets are disastrous.

-Makes sense to develop an ML algorithm solely focused on getting the sizing right(equally as important as identifying good opportunities)

•The Quantamental Way

-High demand among hedge funds for technologies that combine human expertise with quantitative methods

-Can add meta-labeling layer to any primary model, whether that is an ML algorithm, an econometric equation, technical trading rule, or a fundamental analysis.

ex, the meta-labeling ML algorithm could find that discretionary Portfolio Managers tend to make particularly good calls when there is a structural break, as they may be quicker to grasp a change in market regime, or another ex, it may be that PM's under stress, as evidenced by few hours of sleep, fatigue, change is weight tend to make inaccurate predictions.