

The background features abstract green geometric shapes. On the left, a solid green triangle points downwards. On the right, a complex arrangement of overlapping translucent green triangles in various shades of green and yellow-green creates a dynamic, layered effect.

AI WorkShop

金融データを扱った機械学習の演習

2021/07/16

土田晃司

目次

1. 金融データの基本形式
2. 金融で出力されるデータのバー(Bar)とは
 1. 標準バー
 2. 情報ドリブンバー
3. 特徴量サンプリング
4. 演習問題

1. 金融データの基本形式

金融データは、さまざまな形式で提供されています。データを加工する場合の目安として4つの基本的な金融データの形式を示します。

- ▶ ファンダメンタルデータ

金融政策、財政政策、マクロ経済指標

- ▶ 市場データ

価格、出来高、ポジション、指値注文・取消、インプライドボラティリティ

- ▶ アナリスティクス

アナリスト推奨、信用格付け、ニュースセンチメント（重要度）など

- ▶ オルタナティブデータ

検索ワード、Twitter、気象データ（災害）など

2. 金融で出力されるデータ方式バー（bar）

バーとはテーブル形式のデータのことで2つに区別される

▶ 標準バー

規則的なサンプリングに同期させて均一な系列に変換したデータ

種類：タイムバー、ティックバー、ボリュームバー、ドルバー

▶ 情報ドリブンバー

新しい情報が市場に到着した時に、高頻度でサンプリングしたデータ

種類：ティックインバランスバー、ボリューム/ドルインバランスバー、
ティックランバー、ボリューム/ドルインランバー

2. 1. 標準バー

▶ タイムバー

一定時間毎の平均の情報で、統計的には好ましくない？

閑散の時には過剰にサンプリング、頻繁の時には過少にサンプリング

例：時間あたりの4本値、出来高、テクニカル分析？



▶ ティックバー

取引数のことをで為替レートのことではない ガウス分布？

クローズ時や追証期限などの決済が集中するタイミングは注意が必要

例：指定した取引数に達する毎に、カウントした取引の加重平均レート、タイムスタンプ、バー取得時の中値、BID、ASKを抽出する



2. 1. 標準バー

▶ ボリュームバー

ティックバーよりもガウス分布に近い？

ティックバーでは拾えない、1取引で大きいアマウントの取引を拾える

例：指定した数量に達する毎に、加重平均レート、タイムスタンプ、取得時の中値、
BID、ASKを抽出する



▶ ドルバー

他の通貨の価値の変動を考慮したバー EUR/USDなどに使う？

例：指定した売買金額が取引される毎に、取引通貨の加重平均レート、タイムスタンプ、
取得時の取引通貨の中値、BID、ASK、円転レート抽出する

2. 2. 情報ドリブンバー

符号の累積値が動的に算出される閾値を超えたところでバーを生成する
売買価格が前回と比較してプラスだったら1、マイナスだったら-1として、
ある区間の不均衡の集計結果が期待値を超えたタイミングでバーを取得します。
期待値は不均衡の指数加重移動平均で計算し、ある区間は前の区間の指数加重移動平均で計算して求めます。

- ▶ ティックインバランスバー
期待値の対象が不均衡の件数となります
- ▶ ボリューム/ドルバランスバー
期待値の対象が取引数量になります
期待値の対象が売買数量になります

2. 2. 情報ドリブンバー

取引の連続性が動的に算出される期待値を超えたところでバーを生成する

▶ ティックランバー

売買に関係なく、取引頻度が一定時間内に期待値を超えたら、バーを取得する

▶ ボリューム・ドルランバー

売買に関係なく、取引数量が一定時間内に期待値を超えたら、バーを取得する

売買に関係なく、売買数量が一定時間内に期待値を超えたら、バーを取得する

3. 特徴量サンプリング

▶ データ量削減サンプリング

データ量を削除することで機械学習アルゴリズムに適合するため
等間隔サンプリングと一様サンプリングがある

▶ イベントベースサンプリング

構造変化、シグナル抽出、マイクロストラクチャー現象などのイベントが発生した場合にサンプリングするもので、それらは経済指標やボラティリティの急上昇、スプレッドの急拡大などが関連している可能性が高い

サンプリング方法例：CUSUMフィルタ

測定した値の平均値が目標値から離れていくかを検出するために設計された方法

4. 練習問題

- ▶ E-miniS&P 500 先物ティックデータに対して
- 1. ティックバー、ボリュームバー及びドルバーを作成せよ。
- 2. 週次で、ティックバー、ボリュームバー及びドルバーの数を数えよ。バー数を時系列でプロットせよ。週次バー数が最も安定しているのはどの方式か。
- 3. 3つのバー方式のリターンの系列相関を計算せよ。最も低い系列相関を持っている方式はどれか。
- 4. バー系列を月ごとのサブセットに分類する。バー方式ごとにサブセットのリターン分散を計算し、リターン分散の分散を計算する。分散が師匠となるのはどの方法か。
- 5. 3つのバー方式でリターンがのジャックベラ正規検定をせよ。最小検定量統計はどんな手法で得られるのだろうか。

4. 練習問題

- ▶ E-miniS&P 500 先物ティックデータに対して、ドルバーとドリンランバーを計算せよ、どちらのバーの系列相関が大きいのか。それななぜか。
- ▶ E-miniS&P 500 先物ドルバーに対して
 1. 移動平均回り5%幅のボリンジャーバンドを計算せよ。価格がバンドを外側へクロスする回数を数えよ。
 2. CUSUMフィルだを使用してバーをサンプリングする。ここで、 $\{y_t\}$ はリターン、 $h=0.05$ とする。サンプル数はいくつになるのか。
 3. 2つのサンプル系列の移動標準偏差を計算せよ。どちらの不均一性が小さいか。なぜこのような結果になるのか。

4. 練習問題

- ▶ 全問題のバーに対して
 1. CUSUMフィルだを使用してバーをサンプリングする。ここで、 $\{y_t\}$ は絶対リターン、 $h=0.05$ とする。サンプル数はいくつになるのか。
 2. サンプリングされたバーの移動標準偏差を計算せよ。
 3. どちらの不均一性が小さいか。なぜこのような結果になるのか。