3章冒頭

スコープの確認

3.1節、3.2節

前処理の手順、ポイント

データマイニングの分析目的

- 予測分析 — 既知のデータから予測モデルを作り、未知のデータの特徴を明らかにする
- 記述的分析 — データに対して、特徴や傾向間の関係性を抽出する

前処理

3章では予測的分析における前処理を扱う

1. データクレンジング（整理、標準化）【3.1節】

①目的変数データの設定

やることは4つ

- (1) データが反映している情報 — 分析目的に合致したデータを予測対象として選択する
- (2) データの入手可能性を確認 — 時間間隔が短いものや、詳細な情報は有償の場合が多い
- (3) データの範囲・計算方法を事前に確認する
  - 金融市場データの例
    - 株式指数：組み入れ銘柄の構成、計算等の重み
    - 個別銘柄：株式分割や負債などによる価値変動
  - 経済指標データの例
    - 景気動向指数：数値複数から合成されるので客観性は高いが、算出に時間がかかる
    - 景気ウォッチャー調査：調査対象調査なので主観的なゆらぎがあるが、速報性は高い
- (4) 予測期間の長さを決める — 決めるときに主に考慮する点
  - 取得できるデータの間隔
  - データの信頼度
  - 予測精度

②サンプリング間隔の決定

目的は、不等間隔時系列を等間隔時系列にすること — 例えば、ティックデータ等は等間隔ではない

やり方は主に2つ

- *ダウンサンプリング（データの間引き） — 十分に高頻度の場合はこっち
- *内挿（補間）

注意点 — もともと持っているデータの特性が失われないようにする
- よくあるのが、周期性
- グラフや基本統計量を見てから調整しよう

③欠損値と外れ値の処理

欠損値
- 時系列の場合、大抵は補間する
- ただし、理由が明らかな場合は補間しないこともある — 例えば、週末であるとき

外れ値
- 理論上ありえない場合には欠損値と同じ扱い
- 分析目的による

2. 特徴量抽出【3.2節】

①データの値を決めている要因について仮説を立てる

【Point】金融・経済の時系列データの値を決める要因

時系列データの値 = 自己相関 + 長期トレンド + 周期変動 + 外生変動 + ノイズ

- 自己相関 — 過去の自分自身の値が現時点の値に与える影響
- 長期トレンド — 長い期間に存在する緩慢した変動の傾向
- 周期変動 — ある決まった期間で繰り返し発生する変動の傾向
- 外生変動 — 予測対象とは別のメカニズムで生じた事象（外生的事象）が原因となって引き起こされる変動
- ノイズ — 特別な傾向や情報を含まないランダムな要因

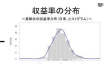
②可視化

最低限見ておくグラフ

- 時系列プロット

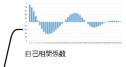
例

- 変動率のヒストグラム

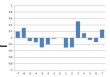


例

- コレログラム — 異なるラグについて自己相関係数もしくは交差相関係数を算出し、横軸にラグ、縦軸に相関係数をとったグラフ



自己相関係数



交差相関係数

③定常性の確認

- 定常性の定義 — この資料では省略
- 金融テキストマイニングが友好となるのは、非定常な場合であるため確認する。
- 確認方法 — 教科書では言及していないが

④見せかけの回帰の確認

- 単位根過程とは — 非定常過程であり、差分系列が定常であるか過程 — 例：ランダムウォーク
- 見せかけの回帰とは — 単位根過程である場合、関係のない単位根過程同士を回帰すると、回帰分析で有意な結果が出ること
- 単位根検定 — 本書の範囲外だが、外本本に分かりやすい解説あり。 — https://saitcooky.hatenablog.com/entry/2020/10/23/015651#%E5%8D%98%E4%BD%8D%E6%A0%B9%E6%A4%9C%E5%AE%9A

⑤変数変換

変換後のデータでも定常性がないか確認する
- 定常性があったら、自己相関による変動要因を説明した残差に含まれる外性要因をテキストデータで変換すると良い
- 他には長期トレンド要因、周期要因も

主な変数変換のやり方は以下の通り
- 差分変換 — 1階差分だけでなく、2階差分を目的変数とすることもある
- 対数変換 — 大きくブレていた変動が、データの持つ特徴を変えずに1つの変動に変換できる
- 対数差分変換 — 対数値の差分
  - 金融データでは、収益率の代わりに、対数リターン(対数値の差分)を用いることがある
- ベンチマーク評価 — 市場の平均株価からの超過リターンなど
- ロジット変換 — $Z = \log\left(\frac{Y}{1-Y}\right)$ — [0,1]の範囲のデータが、(−∞, ∞)の範囲に変換される
- 季節調整
  - 例えば、月別平均法
  - 例えば、パーソンズ法
- 移動平均