AI WorkShop

金融データを扱った機械学習の演習

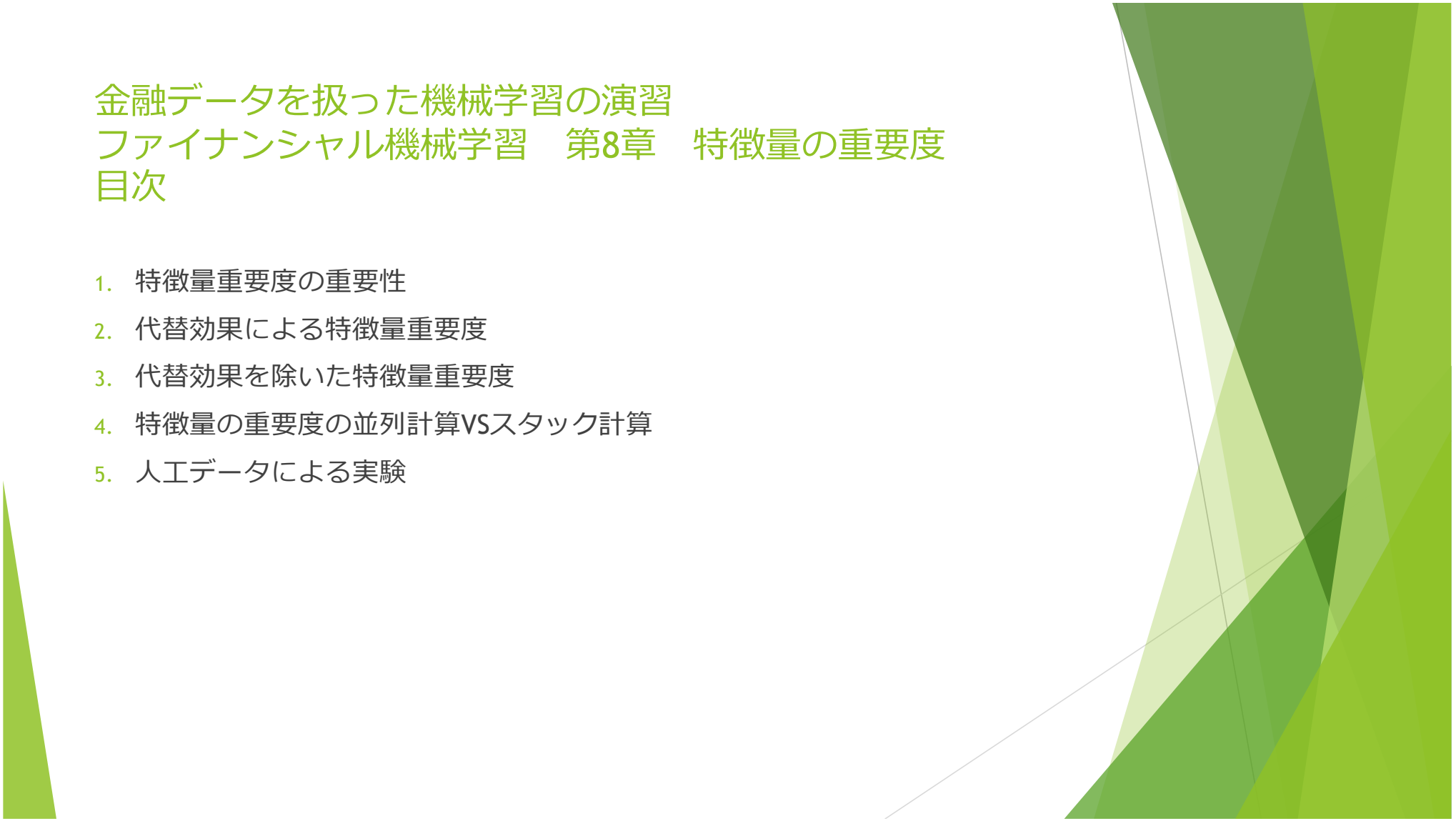
ファイナンシャル機械学習　第8章　特徴量の重要度

2022/02/18

土田晃司

金融データを扱った機械学習の演習
ファイナンシャル機械学習　第8章　特徴量の重要度
目次

1. 特徴量重要度の重要性
2. 代替効果による特徴量重要度
3. 代替効果を除いた特徴量重要度
4. 特徴量の重要度の並列計算VSスタック計算
5. 人工データによる実験



１．特徴量重要度の重要性

金融分野における機械学習で、データの一部を用いて機械学習のアルゴリズムのバックテストを何度も行うは誤り　＝＞　オーバフィットしてしまう

特徴量の重要度こそ時間をかけて確認する必要がある

・分類器の予測力を上げるシグナルを強める特徴量の追加
・ノイズにを追加するだけで特徴量を削除することができる
・どの情報が分類する上で必要な情報なのかがわかる

格言「バックテストはリサーチツールではない、特徴量の重要度こそがリサーチツールなのである」

2．代替効果による特徴量重要度

- 代替効果

ある特徴量の推定重要度が他の特徴量によって削除されるときに生じる効果

つまり、重要度が他の特徴量の影響により見つけづらくなる

対処する方法として、特徴量の主成分分析(次元の削除)を適用し、次に直交(２つの線や面が直角に「交わる」こと)な特徴量に対する重要度分析を行う

重要度を測る方法として

1. 平均不純度減少量（MDI）
2. 平均正解率減少量（MDA）

２．代替効果による特徴量重要度

- 平均不純度減少量（MDI）

ランダムフォレストなどのツリーベースの分類器特有な、重要度を測る

使用時の留意事項

1. マスキング効果：ある特徴量を無視して、他の特徴量を重視することがある
2. インサンプル：予測できなくても特徴量はある程度の重要度を持ってしまう
3. 特徴量は０から１の間の値になり、合計が１になる
4. ツリーベースの分類器以外は使用できない
5. ２つの同一の特徴量を半分で評価してしまう
6. いくつかの予測変数に偏る性質がある

留意点を踏まえたスニペットになっている

２．代替効果による特徴量重要度

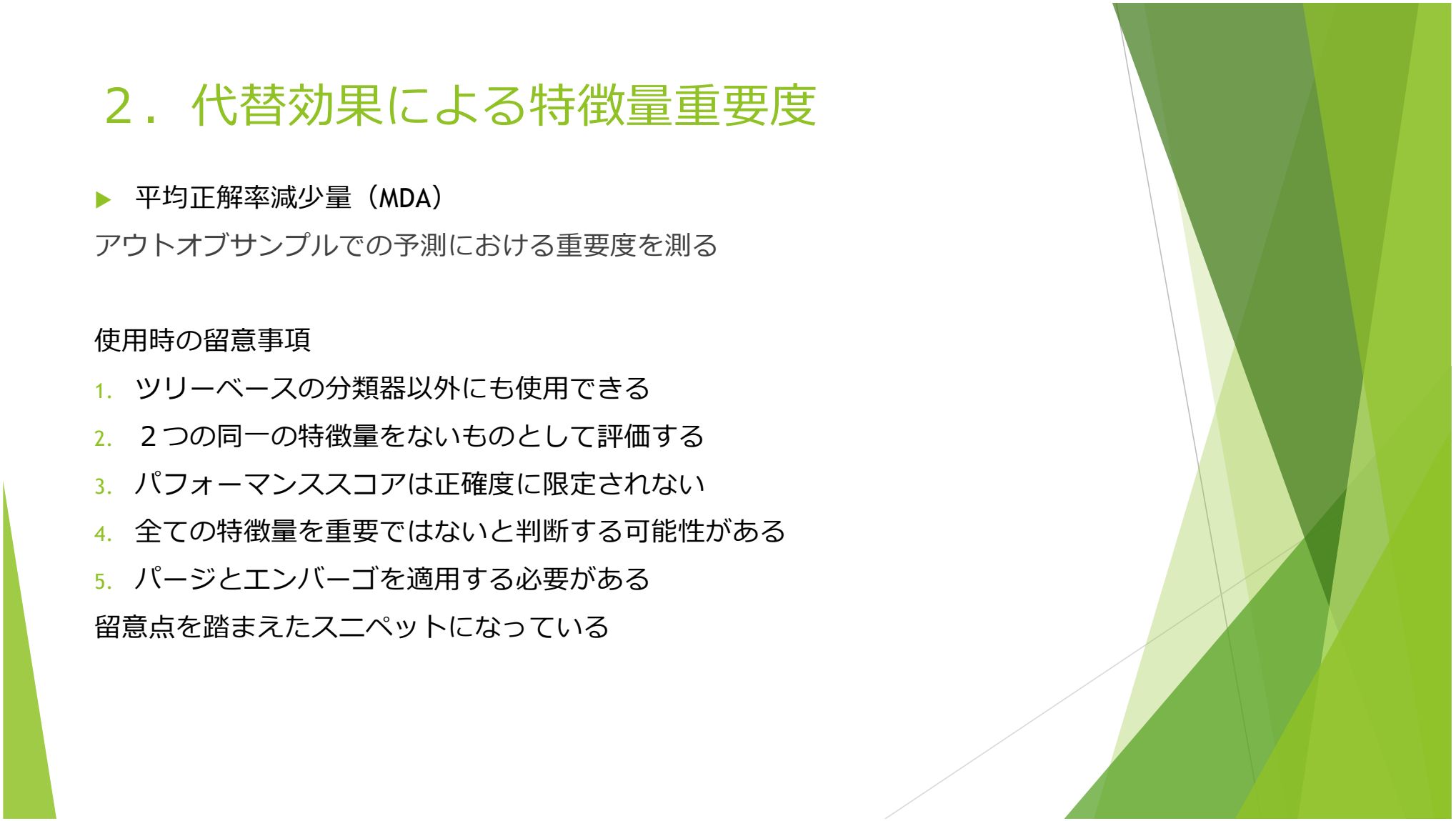
- 平均正解率減少量（MDA）

アウトオブサンプルでの予測における重要度を測る

使用時の留意事項

1. ツリーベースの分類器以外にも使用できる
2. ２つの同一の特徴量をないものとして評価する
3. パフォーマンススコアは正確度に限定されない
4. 全ての特徴量を重要ではないと判断する可能性がある
5. パージとエンバーゴを適用する必要がある

留意点を踏まえたスニペットになっている



３．代替効果を除いた特徴量重要度

代替効果で重要な特徴量が破棄される可能性への対処、データ構造を説明するために設計された特徴量に分析が行えるようになる。

単一特徴量重要度（SFI）

特徴量の直交化

３．代替効果を除いた特徴量重要度

- 単一特徴量重要度（SFI）

横断面（分析）で、アウトオブサンプルでの予測における重要度を測る。結合効果と階層的重要度が失われるという欠点があるので直交化をして適用する。

使用時の留意事項

1. ツリーベースの分類器以外にも使用できる
2. １度に１つの特徴量のみを扱うので代替効果はない
3. パフォーマンススコアは正確度に限定されない
4. 全ての特徴量を重要ではないと判断する可能性がある

留意点を踏まえたスニペットになっている

3．代替効果を除いた特徴量重要度

- 特徴量の直交化

代替効果の軽減、小さな値の固有値に関する特徴量を落として、次元削減ができる。データ構造を説明するために設計された特徴量の分析ができる。

主成分分析の結果とMDI,MDA,SFIの結果を比べて、特徴量の重要度が同じであれば、オーバーフィッティングしていない証明になるのでは。

４．特徴量重要度の並列計算VSスタック計算

- 並列計算：投資ユニバースの構成商品ごとに特徴量の重要度を計算し、平均を出す。代替効果が出るので、重要度が、分散してしまう可能性がある

- スタック計算：投資ユニバースを１つの投資商品と捉えて計算する。重要度が直接導きだされる。

スタック計算の方が重要度が直接導きだされるがオーバフィットを起こしにくいが、計算量が膨大でリソースが必要

5．人工データによる実験

人工データを作成して特徴量の重要度手法がどのように機能するかを確認する

以下の情報が含まれたデータセットを作成して、重要度手法の確認を行う

1. 有益な特徴量：ラベルの決定に用いられる特徴量
2. 冗長な特徴量：有益な特徴量をランダムに線形結合した特徴量（代替効果発生）
3. ノイズ：ラベルの決定に無関係な特徴量

演習

- 人工データの実験を行う
- Colaboratory

  https://colab.research.google.com/notebooks/intro.ipynb?hl=ja#
- Github

  https://github.com/ktsuchida11/AIWorkshop202202
- 参考になるコード

  https://github.com/boyboi86/AFML