



# 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7章 交差検証法

2022/01/21

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7章 交差検証法

### 目次

1. 交差検証法の目的
2. K-分割交差検証法
3. うまく機能しないK-分割交差検証法
4. K-分割交差検証法の改善 1 (ページ)
5. K-分割交差検証法の改善 2 (エンバーゴ)
6. 改善前後の違いを確認
7. 演習問題

# 1. 交差検証法の目的

機械学習の目的は、データの一般的な構造を学習し、将来の観測されていない特徴量を予測できるようにすることにある。

交差検証法では、1つのデータセットを、学習とテストのデータに分割して行うため、どのくらい学習できているかを確認することができる。

ファイナンスでは、交差検証を以下の状況で用いる。

- ▶ モデル開発
- ▶ バックテスト

モデル開発における交差検証法についての工夫を学ぶ

## 2. K-分割交差検証法

### ▶ K-分割交差検証法

データセットをK個の標本群に分割し、学習データとテストデータを入れ替えながら予測結果を複数回確認にして、各予測結果の平均を取って検証する方法

処理順序

1. データセットをK個の標本群に分割する
2. 機械学習アルゴリズムをi個除いた全ての標本群で学習する
3. 取り除いたi個の標本群でテストする

K=5分割の場合の処理



### 3. うまく機能しないK-分割交差検証法

ファイナンスではK-分割交差検証法がうまく機能しない

- ▶ 観測値が独立同分布な過程から抽出されると仮定できないこと
- ▶ モデルを開発する過程でテストデータセットが複数回使われ、複数のテストからの選択バイアスにつながってしまうこと

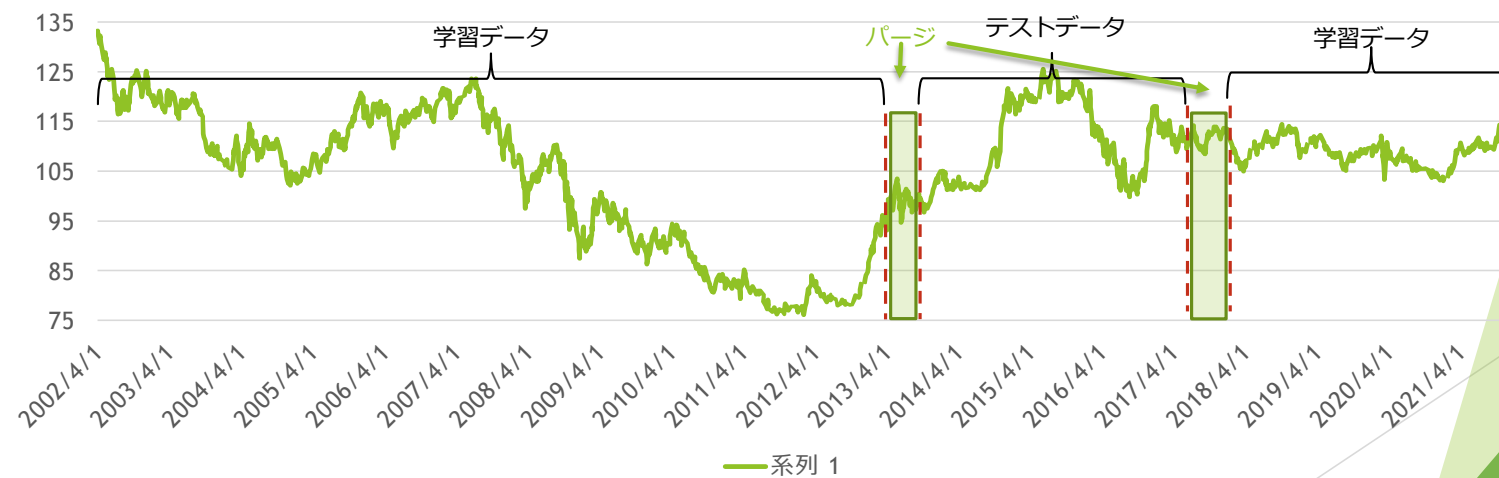
独立同分布：確率変数の列やその他の系が、それぞれの確率変数が他の確率変数と同じ確率分布を持ち、かつ、それぞれ互いに独立している場合をいう

情報のリークとは、学習データとテストデータに重複したデータを利用することで、特徴量の抽出がより良いものと誤ったものと極端になる可能性がある

情報のリーク（重複）をなくすための工夫をデータセットの分割時行うことで、解決する方法が提案されている

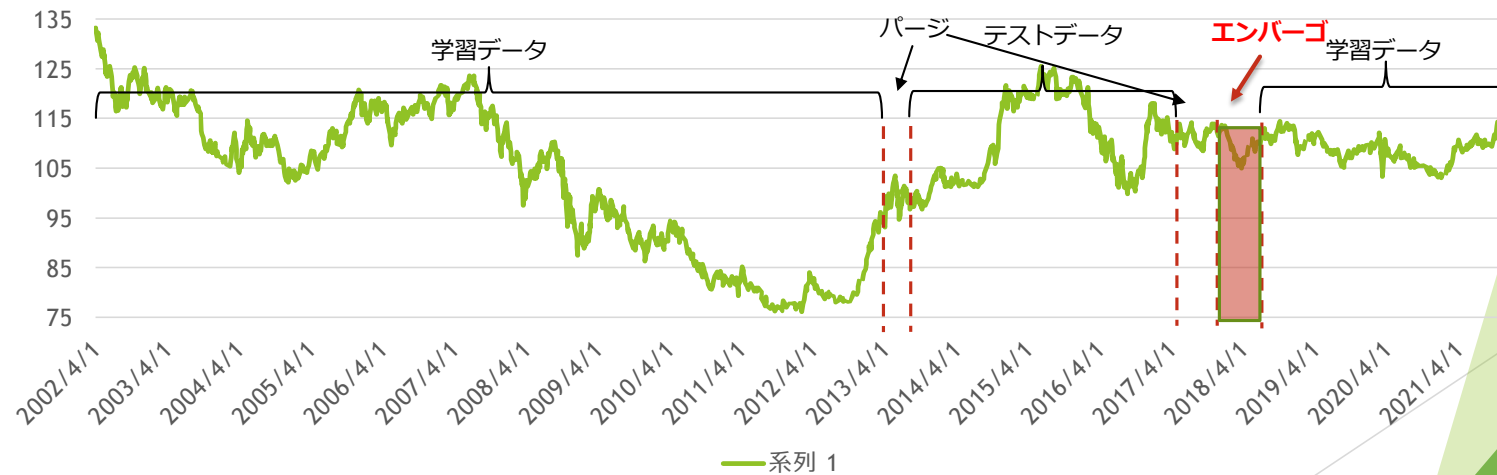
## 4. K-分割交差検証法の改善 1 (パージ)

パージングとは、学習データセットから、テストデータセットに含まれるラベルと時間が重複しているラベルを持つすべての観測データを除去すること



## 5. K-分割交差検証法の改善 2 (エンバゴ)

エンバゴとは、ファイナンシャルデータは系列データを含んでいることが多い  
ため、系列データの影響を強く受けないように学習データからテストデータに連  
続して続いているデータを取り除く



## 6. まとめ

- ▶ ファイナンスデータの交差検証法は、単純なk-分割交差検証法ではうまくいかない
  - ▶ 観測値が独立同分布な過程から抽出されるわけではないため
  - ▶ テストデータセットが複数回使われることでテストのようなデータしか予測できない選択バイアスになってしまう
- ▶ 演習では、通常のK-分割交差検証法での予測とパーキングとエンバーゴを実施したK-分割交差検証法の結果を比べます。

<https://github.com/boyboi86/AFML>