## 機械学習を 解釈する技術1

廣田雄亮

# はじめに

### 1章 機械学習の解釈性とは

機械学習を解釈性

### モデルの**ふるまい**を 分析者が**理解できる**状態

例

$$f(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

X<sub>1</sub>が1単位増加すると,予測値はβ<sub>1</sub>増加する. このことは全てのインスタンスで共通 → 特徴量と予測値の平均的な関係が明らか **解釈性が高い**,と言える 機械学習を解釈することの重要性(1/3)

データ分析のゴールは, 予測と情報抽出

その上でどのモデリングのアプローチを選ぶか

シンプルで解釈しやすいモデル例:線形モデル,決定木

•複雑なモデル

例:SVm, ランダムフォレスト

機械学習を解釈することの重要性(2/3)

ビジネスの現場においても

- 1. 予測精度 ← 絶対
- 2. モデルの解釈性

この2つが求められるが, 2つは**トレードオフ**にある



#### 機械学習を解釈することの重要性

モデルの**単純さ(解釈性)と<sup>測</sup>予測精度**には 対立関係がある その上で

⇒デどちらかを選択すると異分析のゴールを 正しく果たせない

•シンプルで解釈しやすいモデル

最初に予測精度を追求してからモデルの 後から予測の根拠を理解するべき 復年にモデレ

例:SVm, ランダムフォレスト

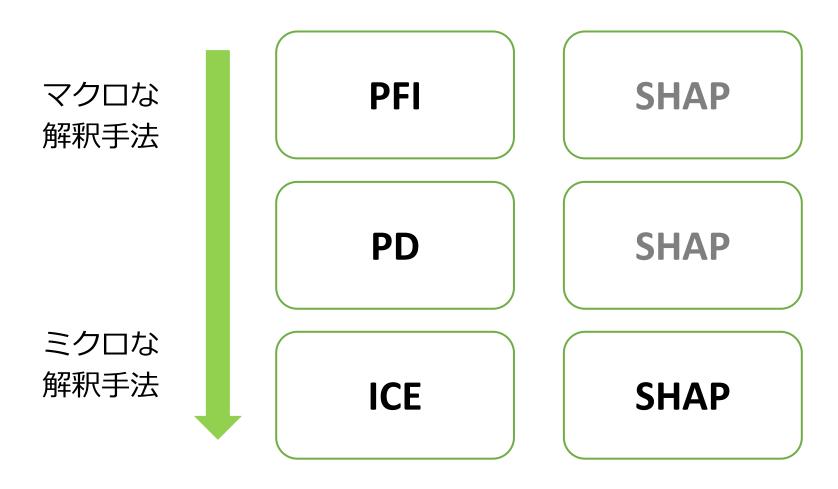
機械学習を解釈することの重要性(3/3)

また,

今後予測モデル自体の構築が容易になるほど, 予測モデルを正しく解釈し,適切に予測モデル を利用することの重要性が増す. 機械学習の解釈手法(1/3)

- •PFI: **P**ermutation **F**eature **I**mportance どの特徴量が重要か
- •PD: **P**artial **D**ependence 特徴量とモデルの予測値の平均的な関係
- •ICE: Individual Conditional Expectation 個別のインスタンス毎の特徴量と予測値の関係
- •SHAP: **Sh**apley **A**dditive ex**p**lanations モデルの出した予測値の理由

#### 機械学習の解釈手法(2/3)



4つの手法はあらゆる予測モデルに適用できる

#### 機械学習の解釈手法(3/3)

「弱い」使い方 比較的安全

「強い」使い方 注意が必要

#### モデルのデバッグ

事前知識と整合的か, 想定外の挙動がないか

→ 比較的安全

#### モデルの振る舞いを解釈

モデルは特徴量Aを重視している,大きくなると予測値が大きくなる

→ 一側面のみをとらえているだけ

#### 因果関係の探索

モデルの振る舞いを因果関係として解釈

→ 実験, 厳密な因果推論が必要