



AI総合演習

第4回: 機械学習モデルの設計と評価

瓜生真也 (デザイン型AI教育研究センター・助教)

講義内容

講義に関する資料（スライド、補足資料等）を  GitHub に置いておきます

<https://github.com/uribo/exeai> ダウンロード可能



1. ガイダンス
2. プログラミング入門
3. 機械学習の背景・数理
4. 機械学習モデルの設計と評価
5. 機械学習の手法
6. 機械学習モデルの解釈・説明性
7. 演習 1：プログラミング言語による機械学習モデルの実装
8. 深層学習の基礎
9. 実社会での応用：自然言語処理、推薦
10. 深層生成モデル
11. 演習 2：プログラミング言語による深層学習の実装
12. 課題解決型演習 1
13. 課題解決型演習 2
14. 課題解決型演習 3
15. 課題解決型演習の発表と振り返り

今日の目標

機械学習モデル構築の

一連の手続きを理解する

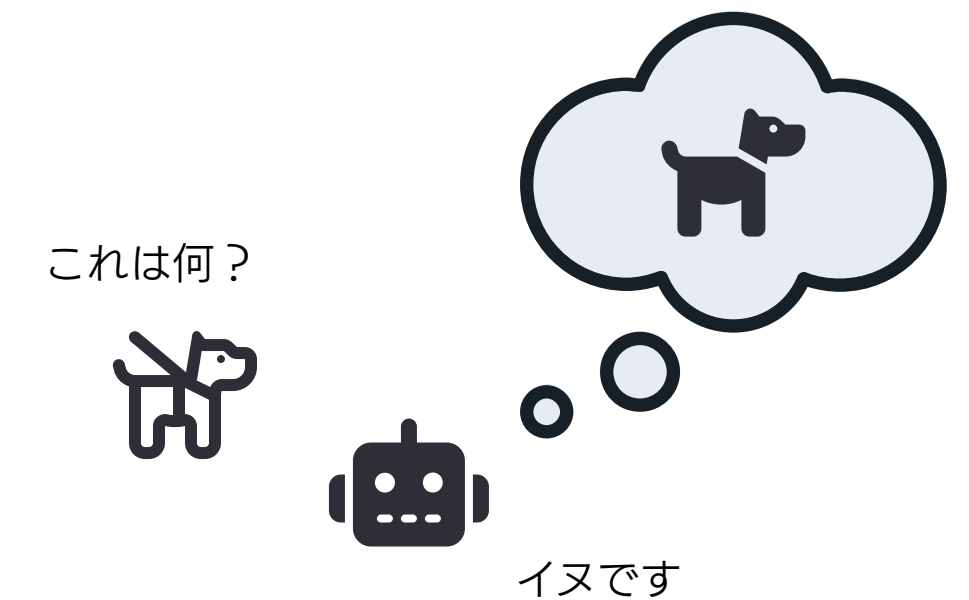
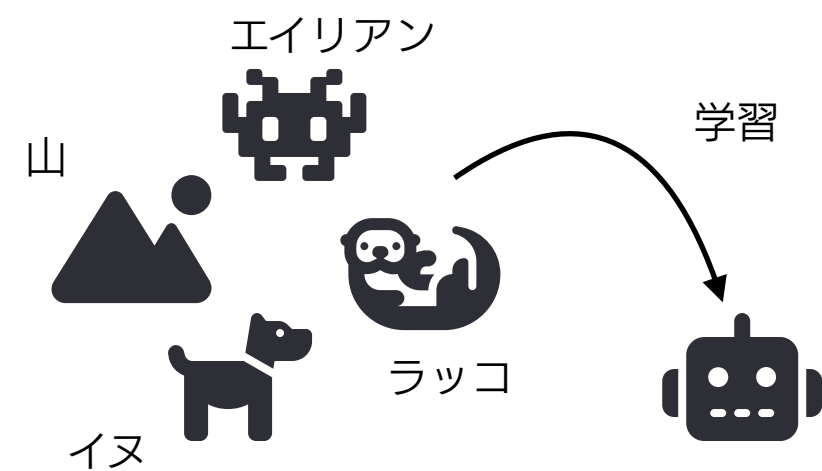
機械学習モデルの学習手法の違い

📍 (参照) 第一回の講義

目的や問題設定、条件に応じて異なる学習手法が存在する

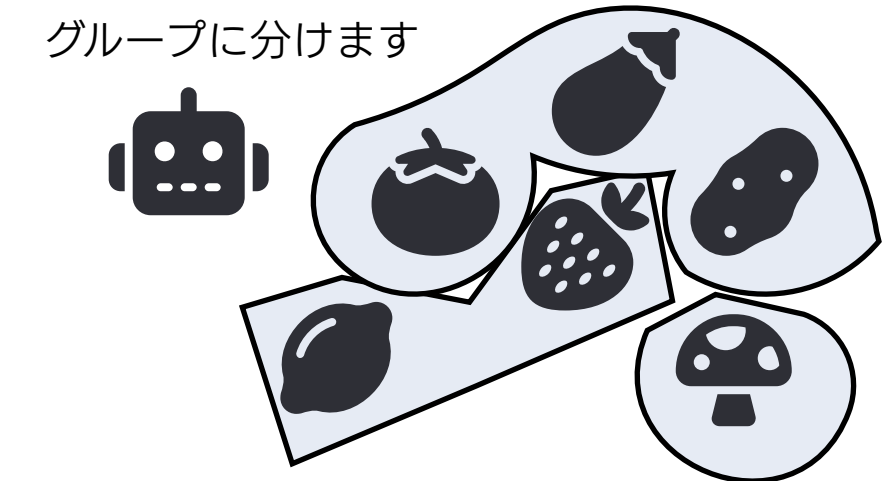
教師あり学習

問題と答えの組み合わせから傾向を学習、
新しいデータ（答えは不明）が与えられた時にデータの予測を行う
→ 回帰や分類問題など、特定の出力を予測するのに効果的



教師なし学習

答えのない状態でデータの特徴（構造やパターン）を学習、データの特徴を抽出する
→ クラスタリングや次元削減などデータの潜在的な構造を抽出する



📍 第五回で解説

強化学習

教師あり学習

教師あり学習の流れ

入力から出力 y を予測する関数 $y = f(x; \theta)$ を学習する

θ は入力に対する重み、パラメータ

🗄️ 訓練データ (学習データ) $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$

学習 訓練データを使ってモデルを学習 $f(x; \theta)$
→パラメータを調整

どうやって x から y を表現できる？

θ を最適な値とするには？

推論 学習モデルを使って、与えられたデータ (👁️テストデータ) から出力 y を予測する

回帰

例) 住宅の特徴量 (変数) から住宅価格を予測
→出力が連続値

分類

例) 画像の特徴量 (ピクセル情報) から対象物 (ラベル) を予測
→出力が離散値

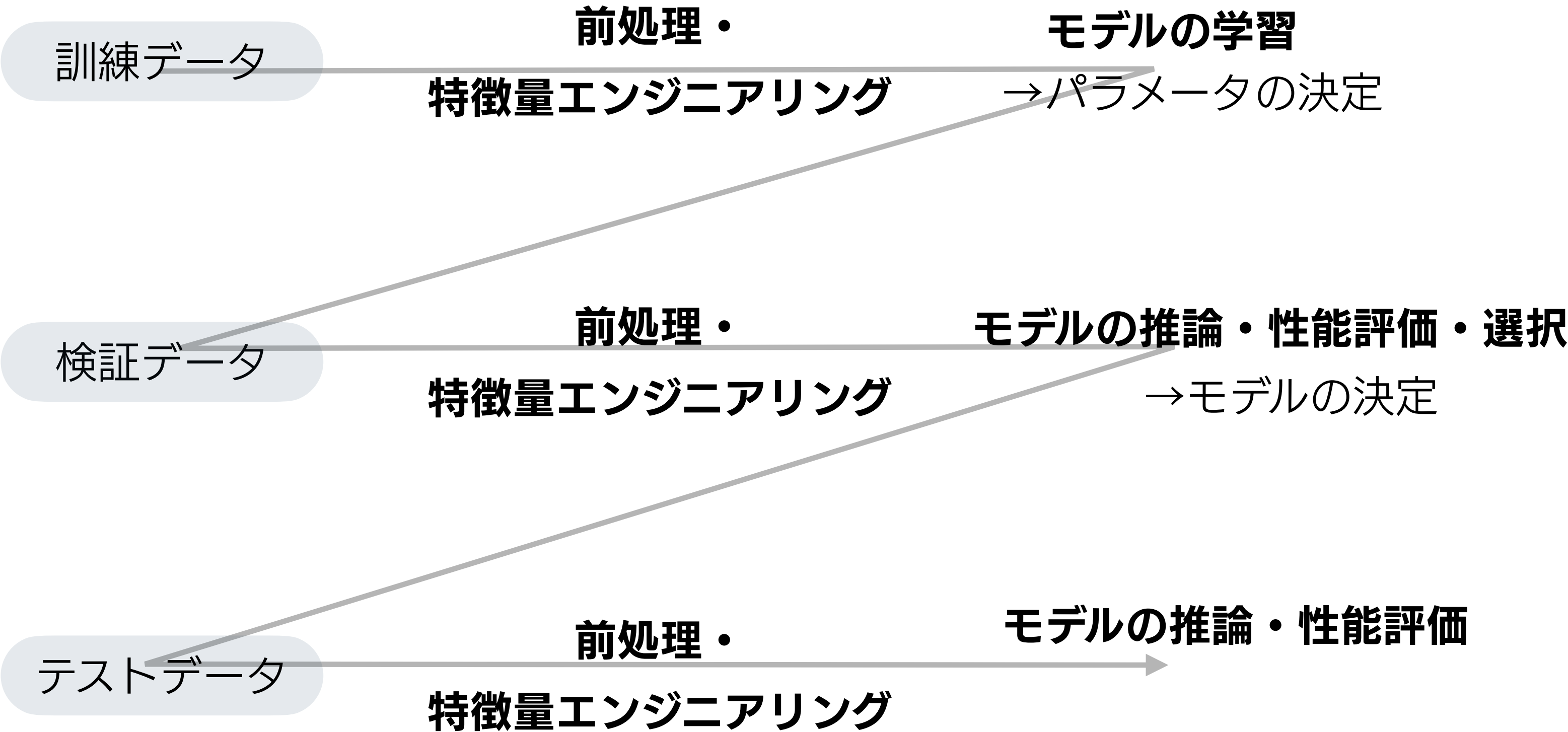


教師あり学習の流れ

取得済み
データ

探索的データ分析

データ分割



ペンギンデータの分類に挑戦

```
import seaborn as sns
penguins = sns.load_dataset("penguins")
```

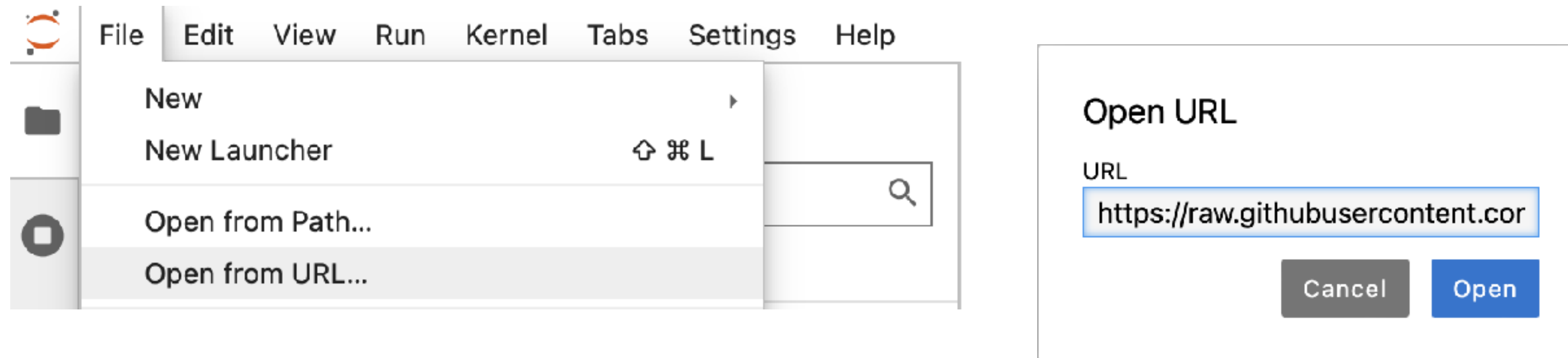

【課題】.ipynbを提出

提出期限: 来週の講義開始前まで

manabaのレポートとして提出してください

GitHubからanswer.ipynbをアップロードして記載

<https://raw.githubusercontent.com/uribo/exeai/main/week04/answer.ipynb>



注意: ファイル名は英数字のみにすること

日本語（漢字、片仮名、平仮名）、全角英数字、スペース、記号等は使わない

ファイルをダウンロードしても開けなくても問題ない（気にしない）

内容の確認、編集はJupyterHub上で行う

データ分割

```
X_train, X_test, y_train, y_test = train_test_split(  
    X, y, test_size=0.2, random_state=20230508  
)
```

汎化性能

訓練データに対する、未知のデータへの対応能力、予測精度

モデルの学習に用いるデータとは別に、汎化性能を調べるためのデータを用意する

→訓練データとテストデータ

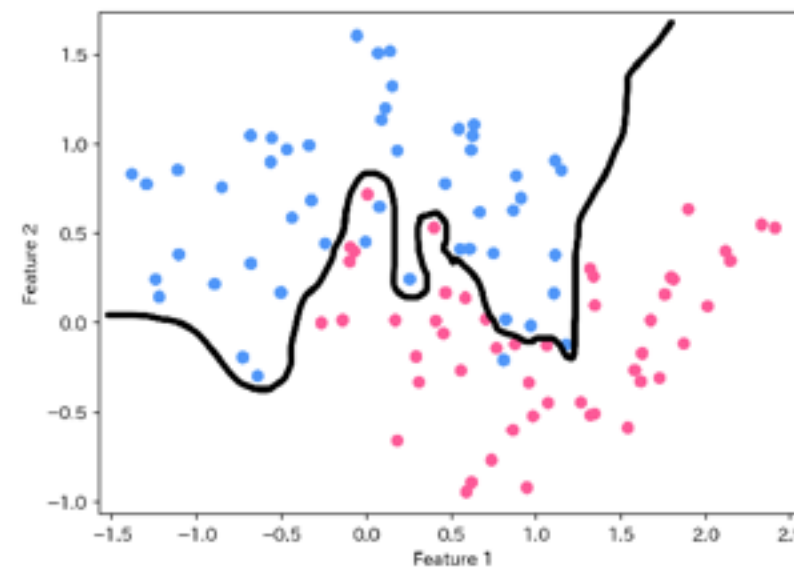
過学習、モデルの過剰適合

訓練データに過度に依存したモデルを構築したことにより、未知のデータへの予測精度が低下する

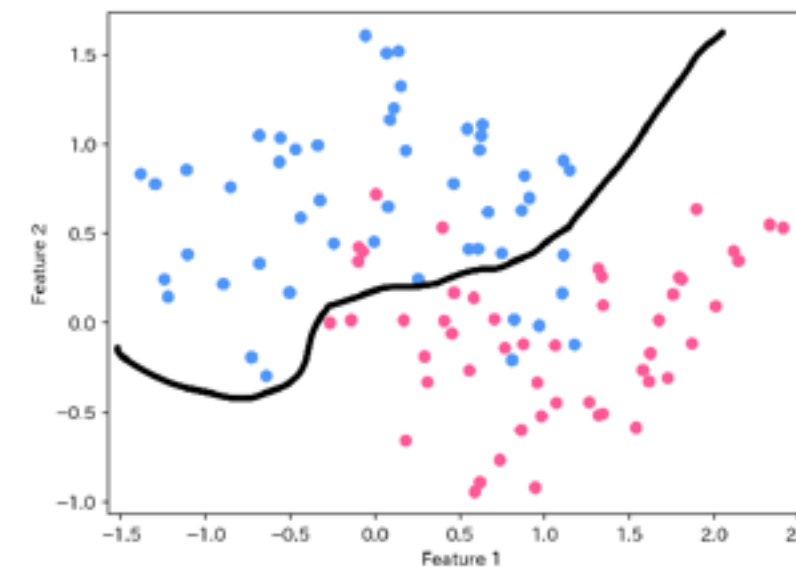
対策

- 交差検証法の採用
- 正則化
- データ増強
- モデルの簡略化

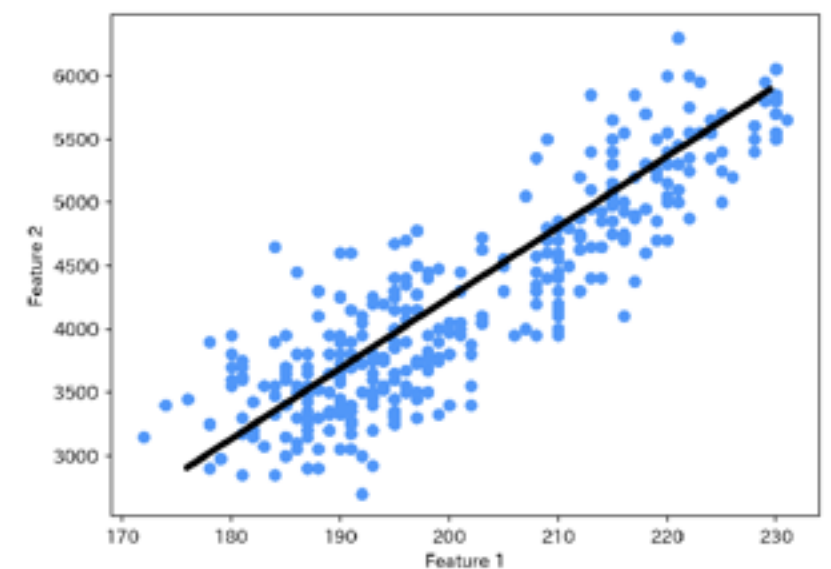
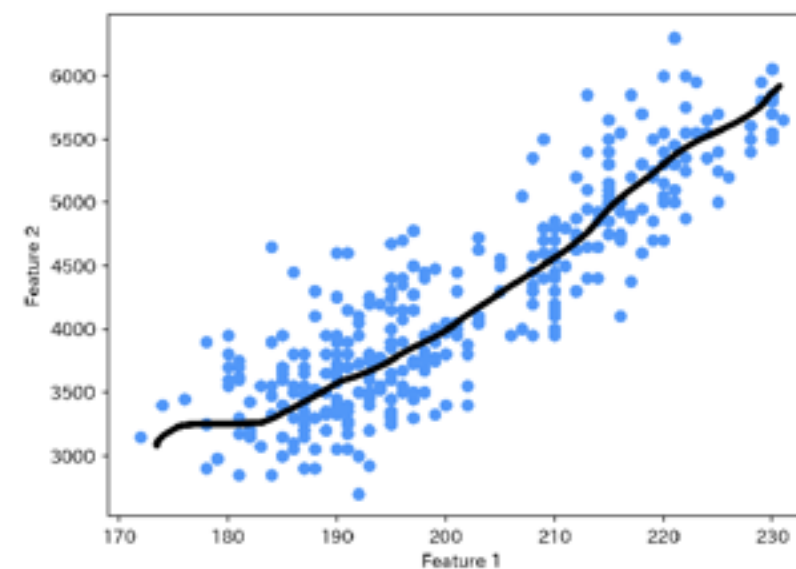
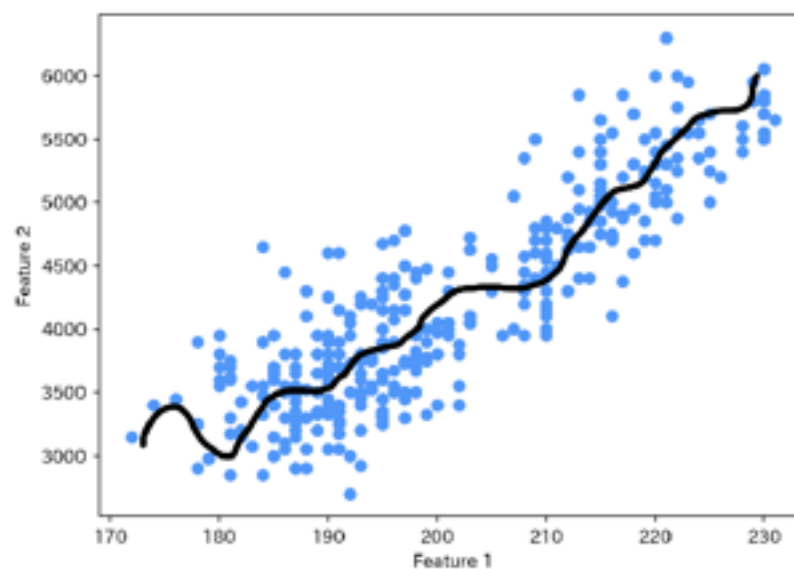
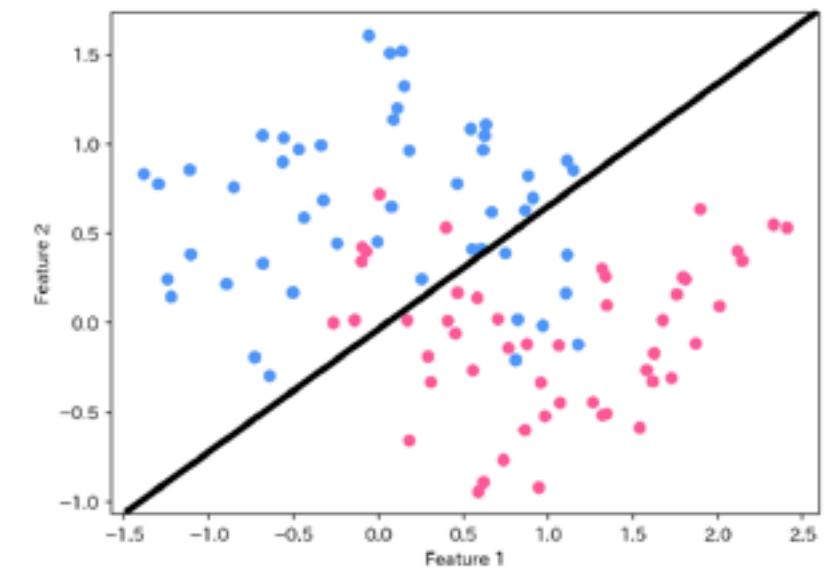
過学習



適度な学習



過小適合



ランダムに訓練データとテストデータに分割

ホールドアウト法

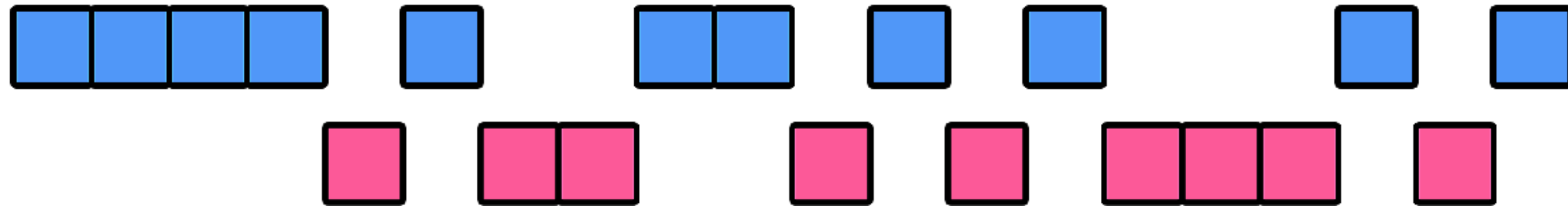


訓練データ



テストデータ

元データ



分割の方法によってはデータに偏りが生じ、過学習につながるおそれがある

例) 時系列データでのランダムな分割はNG (訓練データに未来のデータが含まれる)

交差検証法

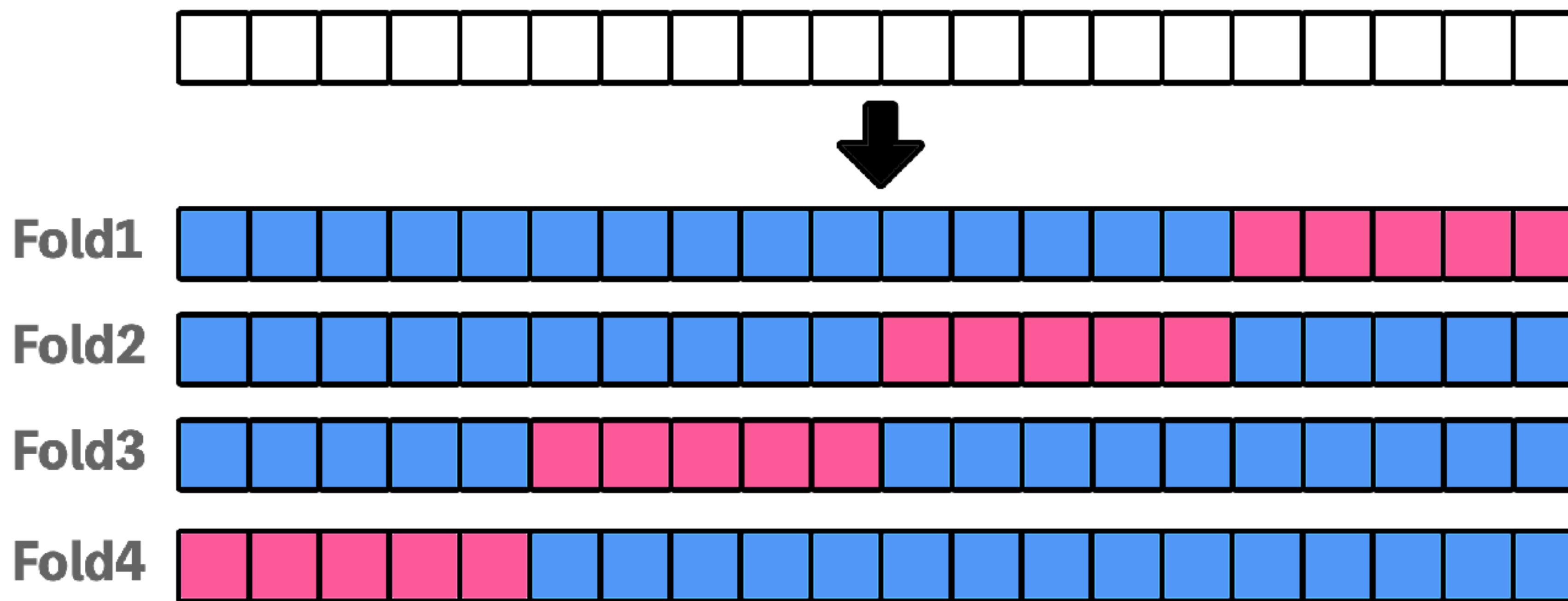
複数個の訓練データとテストデータの組み合わせを用意

過学習の影響を排除して、予測モデルの正確な精度が測定できる（分割時の偶然性による影響を軽減）

適切なハイパーパラメータの選択のためにも使われる

分割方法によっていくつかのバリエーションがある

k 分割交差検証

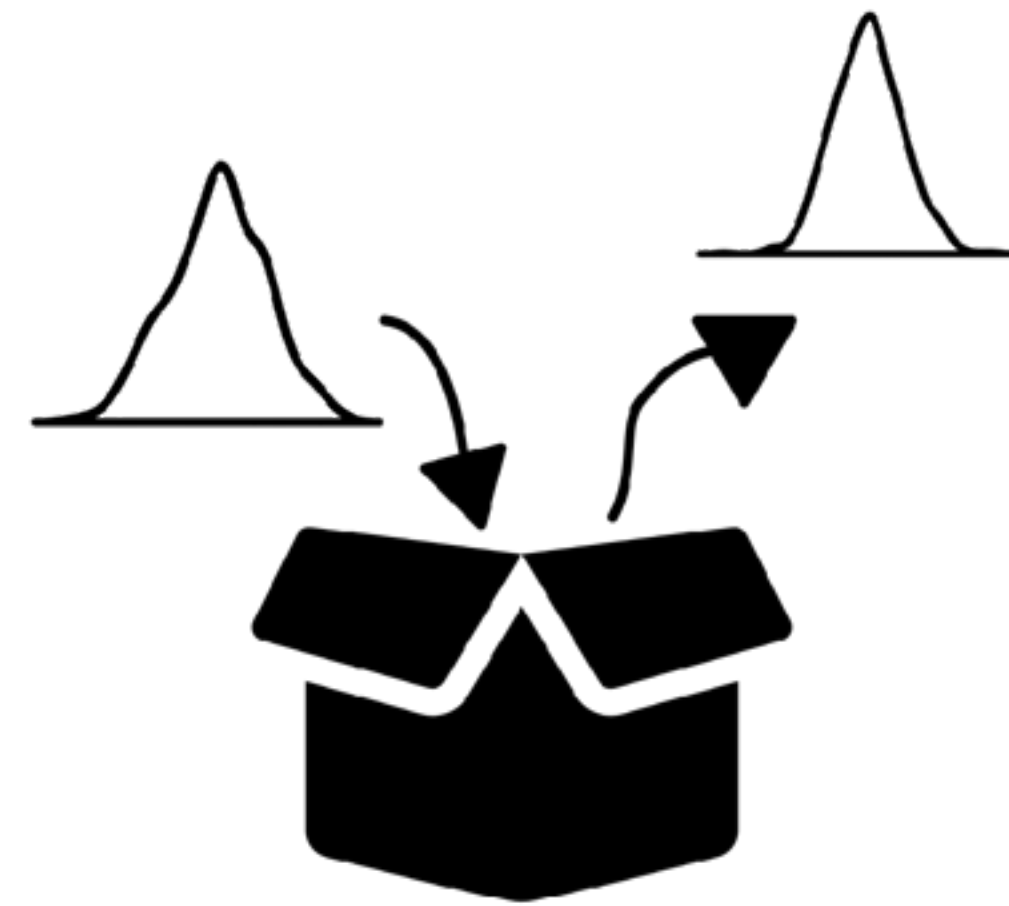
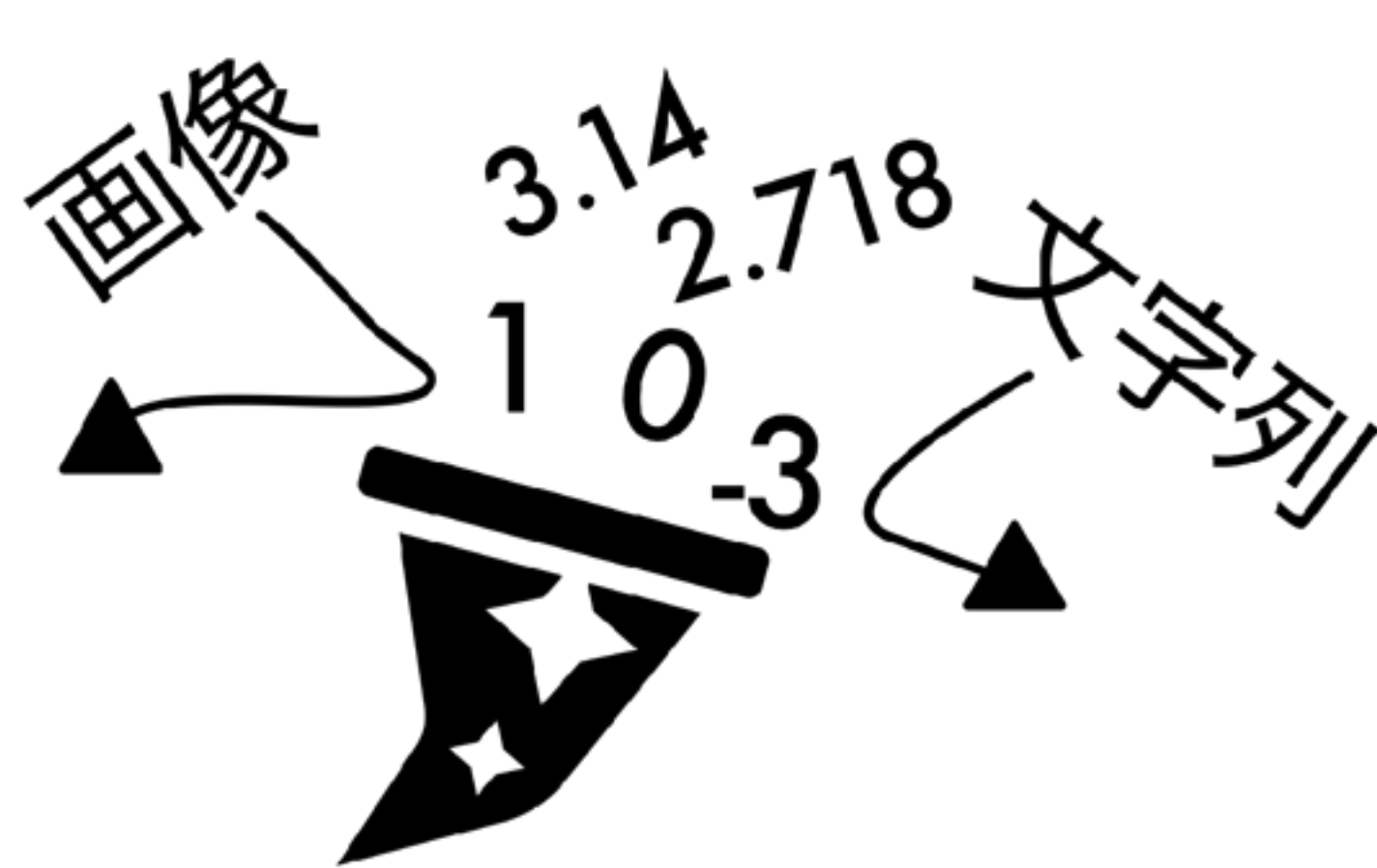


特徴量エンジニアリング

前処理・特徴量エンジニアリングの必要性

モデル上でのデータのふるまいを制御・変更するための手続き

モデル、アルゴリズムの多くは入力に仮定を置いたり、入力による制約を設けている



データ、モデル、目的に応じてさまざまな特徴量への処理が必要になる

→機械学習モデルの性能を大きく左右する

カテゴリ変数の数値化

ラベルエンコーディング

各カテゴリについて、対となる数値を割り当てる

```
from sklearn.preprocessing import LabelEncoder  
le = LabelEncoder()  
penguins["species"] = le.fit_transform(penguins["species"])
```

species	label
Adelie	0
Chinstrap	1
Gentoo	2

ダミー変数化

カテゴリの値を特徴量として扱う。

基準となるカテゴリの値（参照カテゴリ）の扱いにより

いくつかのバリエーションがある

island	island_Dream	island_Torgersen
Biscoe	0	1
Dream	0	0
Torgersen	1	0

```
import pandas as pd  
penguins_dummies = pd.get_dummies(penguins, columns=["island"], drop_first=True)
```

深層学習ではカテゴリ変数を低次元の連続値ベクトルに変換する「埋め込み」が使われる

前処理の必要性はモデルによって異なる

多くのモデルは入力データのスケールに敏感

例) 数字2桁の変数と7桁の変数がある場合、変数間の効果に差が生じる

→線形回帰、k-means、主成分分析などのモデルは変数間のスケールを揃える操作（スケーリング）が必要

→木ベースのモデル（決定木、ランダムフォレスト）は変数のスケールの影響を

受けないため、スケーリングは不要



```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaled_data = scaler.fit_transform(data)
```

欠損値や外れ値への対応はモデルや利用するライブラリによって対応が異なることが多い

例) ライブラリ側で自動的に欠損値を含むデータを削除

→k近傍法やサポートベクターマシンは外れ値の影響を受けやすい

数値データに潜む問題

例えば... スケールが大きくなる
歪んだ分布
外れ値を含む

変数間で複雑な関係をもつ
冗長な情報を含む

→適切な前処理・特徴量エンジニアリング、
適切なモデルの選択が求められる

ロジスティック回帰

```
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(X_train, y_train)
```

ロジスティック回帰

カテゴリカルな目的変数の予測に用いられる

目的変数の値… 0または1（二値変数）を予測する

入力変数とそれらの重みを組み合わせた線形関数を利用（線形回帰と同じ）

$$z = w_0 + w_1 * x_1 + w_2 * x_2 + \dots + w_n * x_n$$

その結果(ここではz)をシグモイド関数に入力として与える

→0から1の範囲からなる値を出力。データがあるカテゴリに属する確率として解釈できる

2つ以上のカテゴリが存在する場合…

各カテゴリに対して1つのロジスティック回帰モデルを用意し、
カテゴリに属する確率を求める。

候補となるカテゴリの中で最も確率の高いカテゴリを予測として採用する。

シグモイド関数とロジット関数

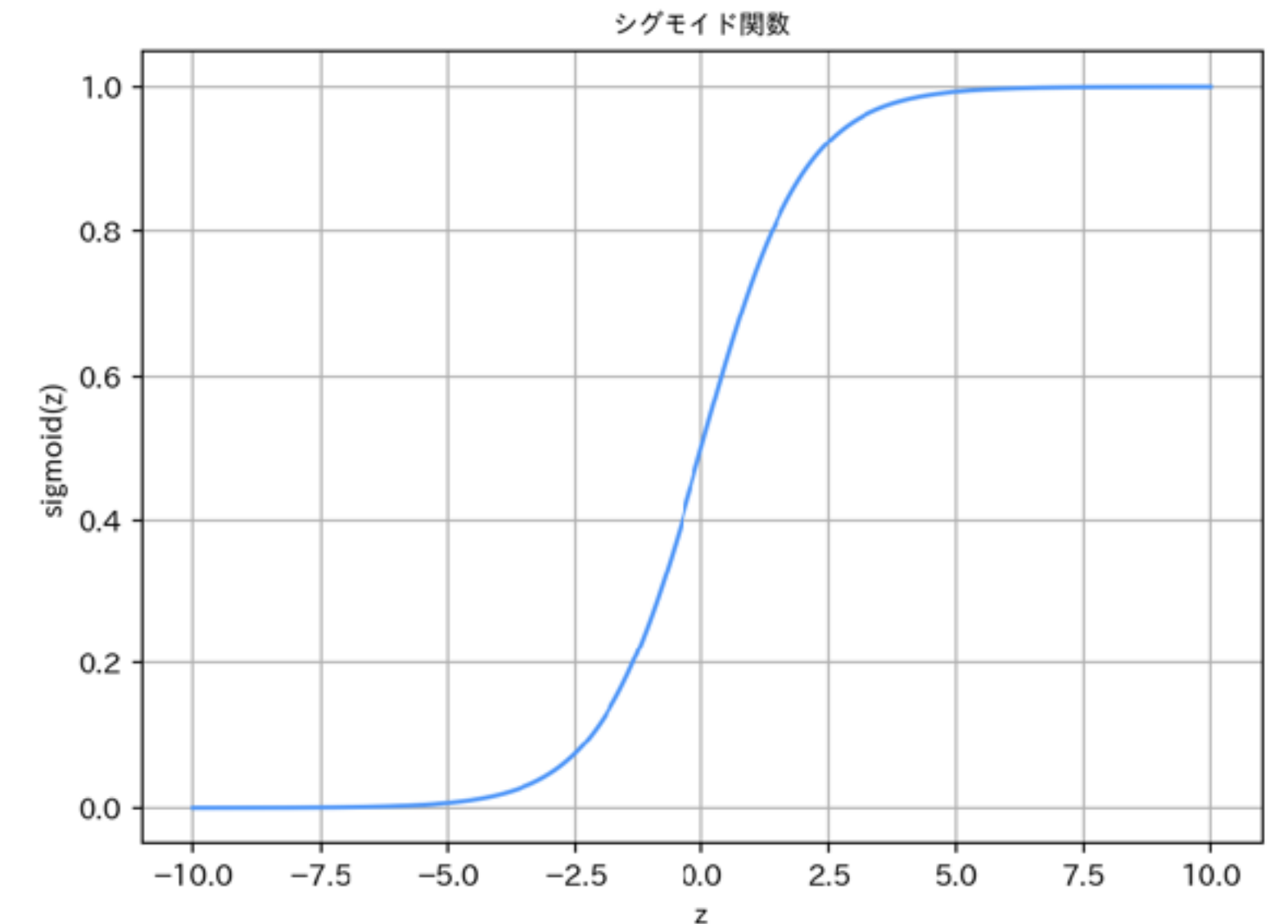
シグモイド関数

$$p = \frac{1}{1 + \exp^{-z}}$$

📍 (参照) 第三回の講義。自然対数の底をeまたはexpと表す

ロジット関数… シグモイド関数の逆関数

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$



オッズ比

事象の起こりやすさを2つの群で比較して示す

$$\frac{p}{1-p}$$

例) 男性が商品を購入する確率 (p_1) が0.8、女性が購入する確率(p_2)が0.2のとき、

オッズは次のように求められる

$$\frac{p_1}{1-p_1} = 4 \quad \frac{p_2}{1-p_2} = 0.25$$

比較する2群のオッズから比 (オッズ比) を求める。オッズ比が1より大きい場合、男性が女性に比べて商品を購入する確率が高いことを示す。1より小さい場合は男性よりも女性が商品を購入する確率が高いことを示す。

この場合、女性よりも男性が商品を購入する確率が16倍高いことを示す。

モデルの評価

分類問題におけるモデルの評価指標の例

真の値とモデルの予測結果を比較する

- 正解率(accuracy): モデルが正しく予測したデータの割合
- 適合率 (precision): 正と予測したデータのうち、実際に正である割合。
- 再現率(recall): 実際に正であるもののうち、正であると予測された割合。
- F1スコア(f1-score): 適合率と再現率の調和平均により得られた値。適合率と再現率のバランスを考慮した評価指標。この値が高いほど、適合率と再現率の両方が高いことを示す。

🔗 # 評価指標を個別に算出

```
accuracy_score(y_test, y_pred)
precision_score(y_test, y_pred)
recall_score(y_test, y_pred)
f1_score(y_test, y_pred)
```

🔗 # 評価指標をまとめて出力

```
classification_report(y_test, y_pred)
```

参考資料・URL

- 岡野原大輔『ディープラーニングを支える技術: 「正解」を導くメカニズム〈技術基礎〉』
(2022) 技術評論社. ISBN: 978-4-297-12560-8

瓜生居室: あり (電子版)、徳大図書館: あり、市立図書館: なし、県立図書館: なし

- 門脇大輔, 阪田隆司, 保坂桂佑, 平松雄司『Kaggleで勝つデータ分析の技術』 (2019)
技術評論社. ISBN: 978-4-297-10843-4

瓜生居室: あり (電子版)、徳大図書館: あり、市立図書館: なし、県立図書館: なし

- 八谷大岳『ゼロからつくるPython機械学習プログラミング入門』
(2020) 講談社. ISBN: 978-4-06-520612-6

瓜生居室: あり (電子版)、徳大図書館: なし、市立図書館: なし、県立図書館: なし

- Aurélien Géron (著), 下田倫大 (監訳), 長尾高弘 (訳)『scikit-learn、Keras、TensorFlowによる実践機械学習』 (2020) オライリー・ジャパン. ISBN: 978-4-87311-928-1

瓜生居室: あり (電子版)、徳大図書館: あり、市立図書館: なし、県立図書館: なし

- Alice Zheng, Amanda Casari (著), ホクソエム (訳)『機械学習のための特徴量エンジニアリング: その原理とPythonによる実践』 (2019) オライリー・ジャパン. ISBN: 978-4-87311-868-0

瓜生居室: あり (電子版)、徳大図書館: あり、市立図書館: なし、県立図書館: なし

https://speakerdeck.com/s_uryu/feature-engineering-recipes

