みんなのKaggle講座

Section2

Section2の概要

講座の内容

Section1. Kaggleの概要

Section2. 機械学習とKaggle

Section3. 精度向上のためのテクニック

Section4. Titanicの先へ

今回の内容

1. Section2の概要
2. 機械学習の概要
3. 機械学習のアルゴリズム
4. Pandasの基礎
5. Kaggleで機械学習を扱う
6. 演習

教材の紹介

- **Pythonの基礎:**

  python_basic

- **Section2の教材:**

  01_pandas_basic.ipynb

  02_titanic_random_forest.ipynb

  03_exercise.ipynb

演習の解答 Section1

- **House Prices - Advanced Regression Techniques**

  結果の提出にトライしよう！

  https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Data Explorer

957.39 kB

- data_description.txt
- sample_submission.csv
- test.csv
- train.csv

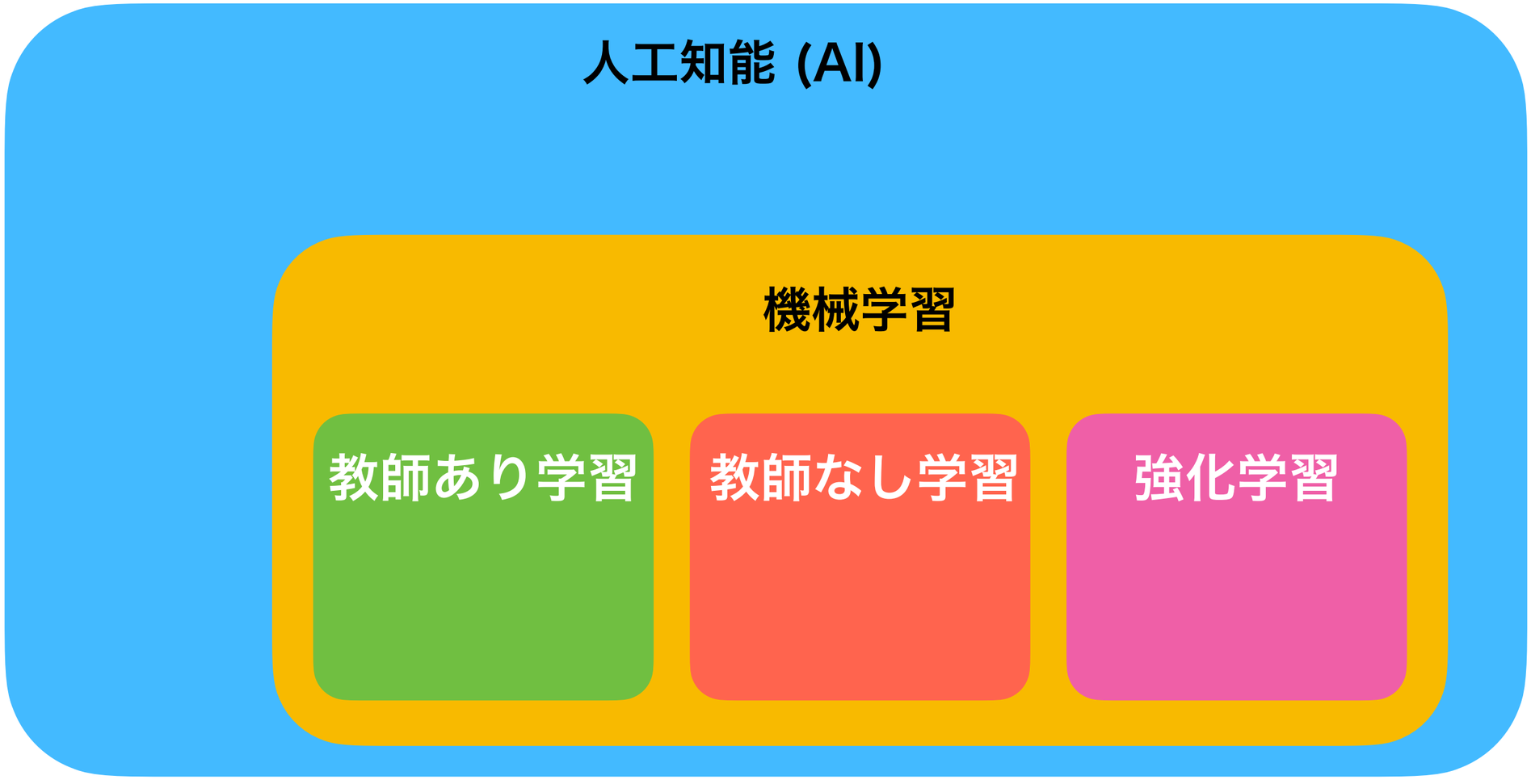
提出データの例

data_description.txt (13.37 kB)

```
 70      2-STORY 1945 & OLDER
 75      2-1/2 STORY ALL AGES
 80      SPLIT OR MULTI-LEVEL
 85      SPLIT FOYER
 90      DUPLEX - ALL STYLES AND AGES
120      1-STORY PUD (Planned Unit Development) - 1946 & NEWER
150      1-1/2 STORY PUD - ALL AGES
160      2-STORY PUD - 1946 & NEWER
180      PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190      2 FAMILY CONVERSION - ALL STYLES AND AGES
```

機械学習の概要

人工知能と機械学習



機械学習とは？

- **機械学習**は人工知能、あるいは統計学の一分野
- コンピュータプログラムが経験、学習を行う
- 「教師あり学習」、「教師なし学習」、「強化学習」に分類できる

機械学習のアルゴリズム

- 回帰
- k平均法
- サポートベクターマシン
- 決定木
- ニューラルネットワーク
- 強化学習
- アンサンブル学習
- etc…

機械学習の「モデル」

- 「モデル」は定量的なルールを数式などで表したもの
- モデルは多数の「学習するパラメータ」「ハイパーパラメータ」を、値や設定として持つ

	学習するパラメータ	ハイパーパラメータ
学習時	調整する	変更しない
モデル構築時	ランダムに設定することが多い	慎重に設定する

過学習と汎化能力

- **過学習**は、機械学習のモデルが特定のデータに過剰に適合してしまった状態
- 過学習に陥ると多様なデータに対応できる汎用性が失われる
- モデルの汎用性を発揮する能力は、**汎化能力**と呼ばれる
- 余談: 自然界、人間社会における過学習

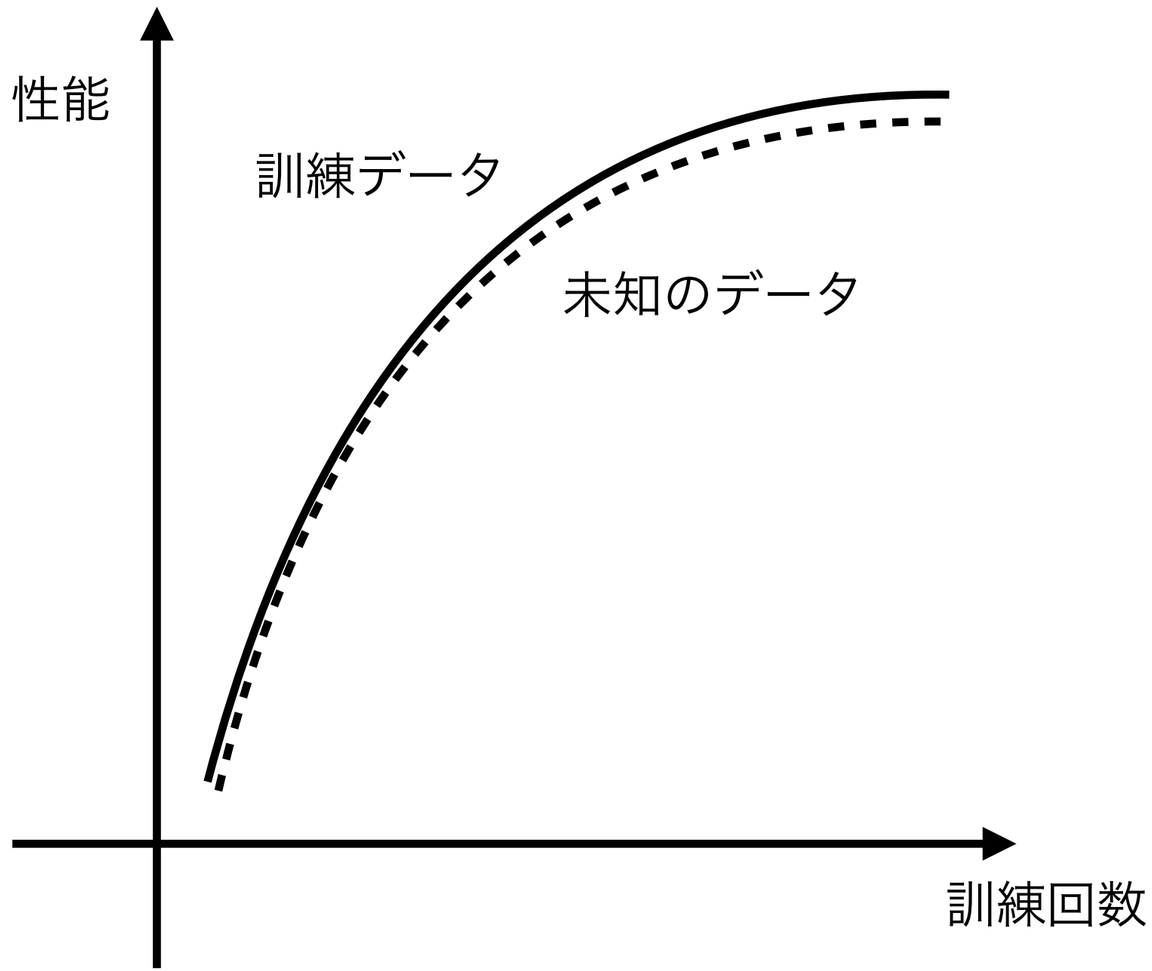
  → 大企業病、テストの一夜漬け、恐竜の絶滅、寿命の存在、etc…

汎化能力が高い例

- 汎化能力が高い場合は、訓練を重ねると、

  訓練データに対してだけではなく

  未知のデータに対しても性能を発揮する



未知のデータにも高い性能を発揮

汎化能力が低い例

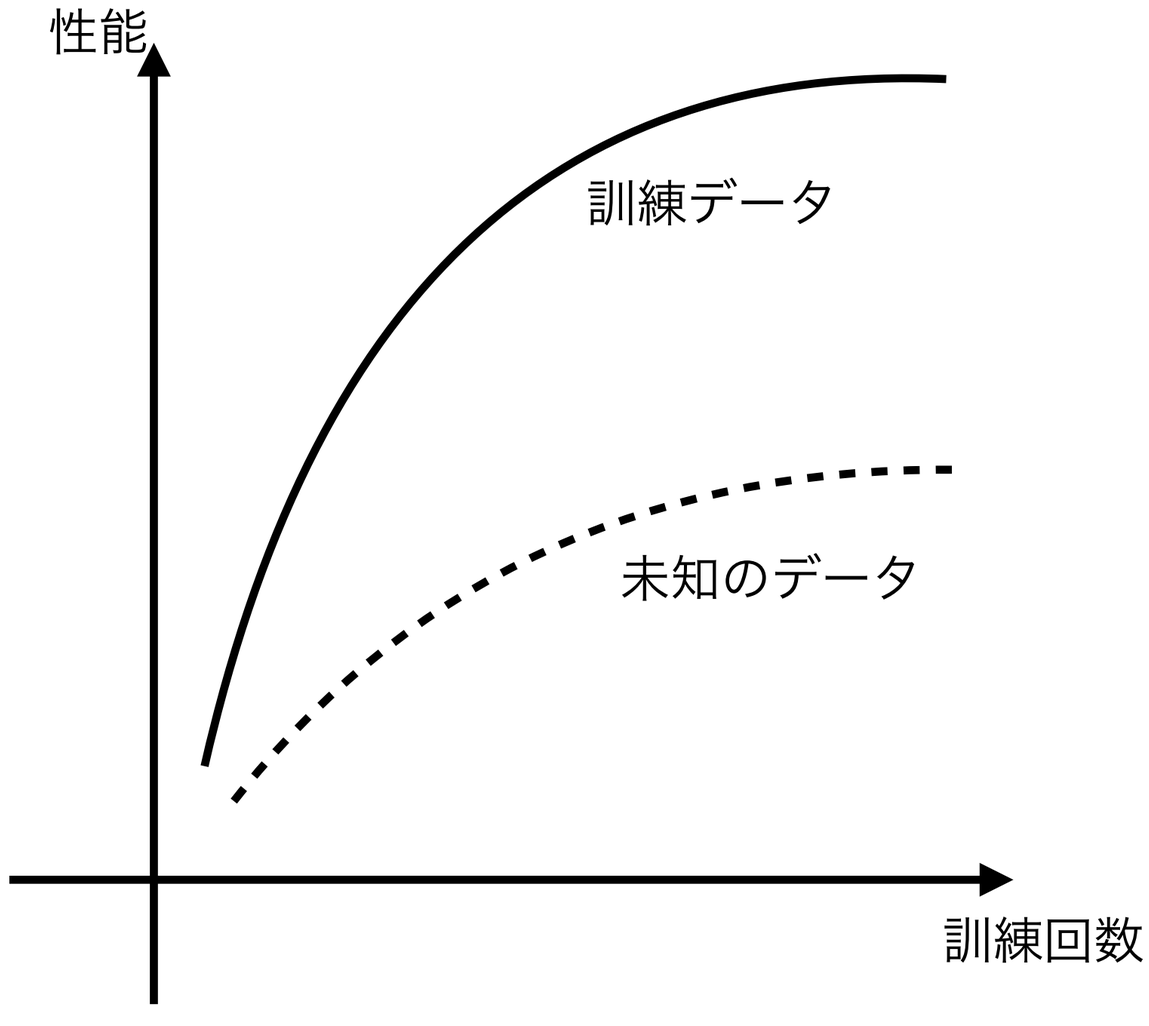
- 汎化能力が低い場合は、訓練を重ねると

  訓練データの性能は向上するが、

  未知のデータに対しての性能は向上しない

- このような過学習により、

  機械学習のモデルは汎用性を失ってしまう



訓練データへの過剰な適合（過学習）

機械学習のアルゴリズム

機械学習のアルゴリズム

- 回帰
- k平均法
- 多層パーセプトロン
- サポートベクターマシン
- 決定木
- etc…

回帰とは？

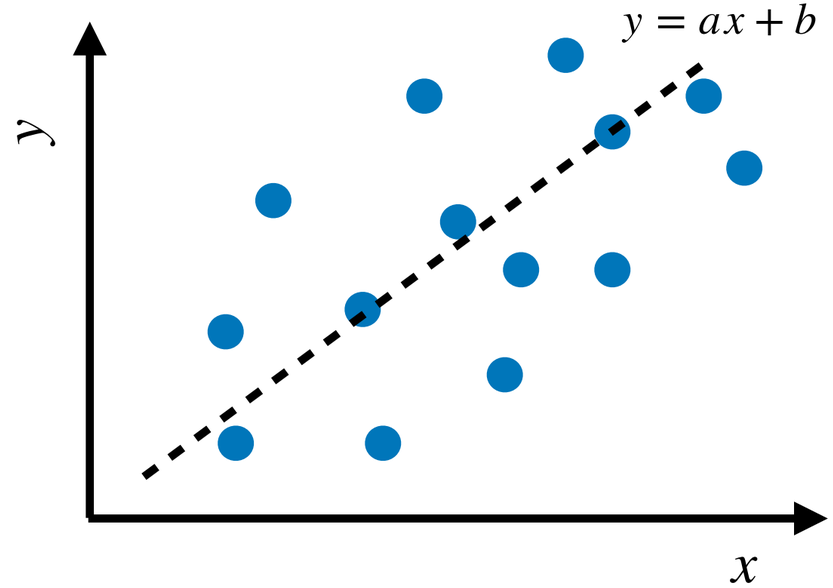
- 「回帰」は、関数をデータに当てはめることによって、

  ある変数yの変動を別の変数xの変動により説明/予測すること

- 「説明変数」は何かの原因となっている変数
- 「目的変数」はその原因を受けて発生した結果である変数
- 以下の「単回帰」は、最もシンプルな回帰

  $y = ax + b$

- 「重回帰」は説明変数が二変数以上になる回帰



k平均法とは？

「距離」に基づき、データをk個の

クラスタに分類する

1.各サンプルに、ランダムにk種類の

グループうちどれかを割り当てる

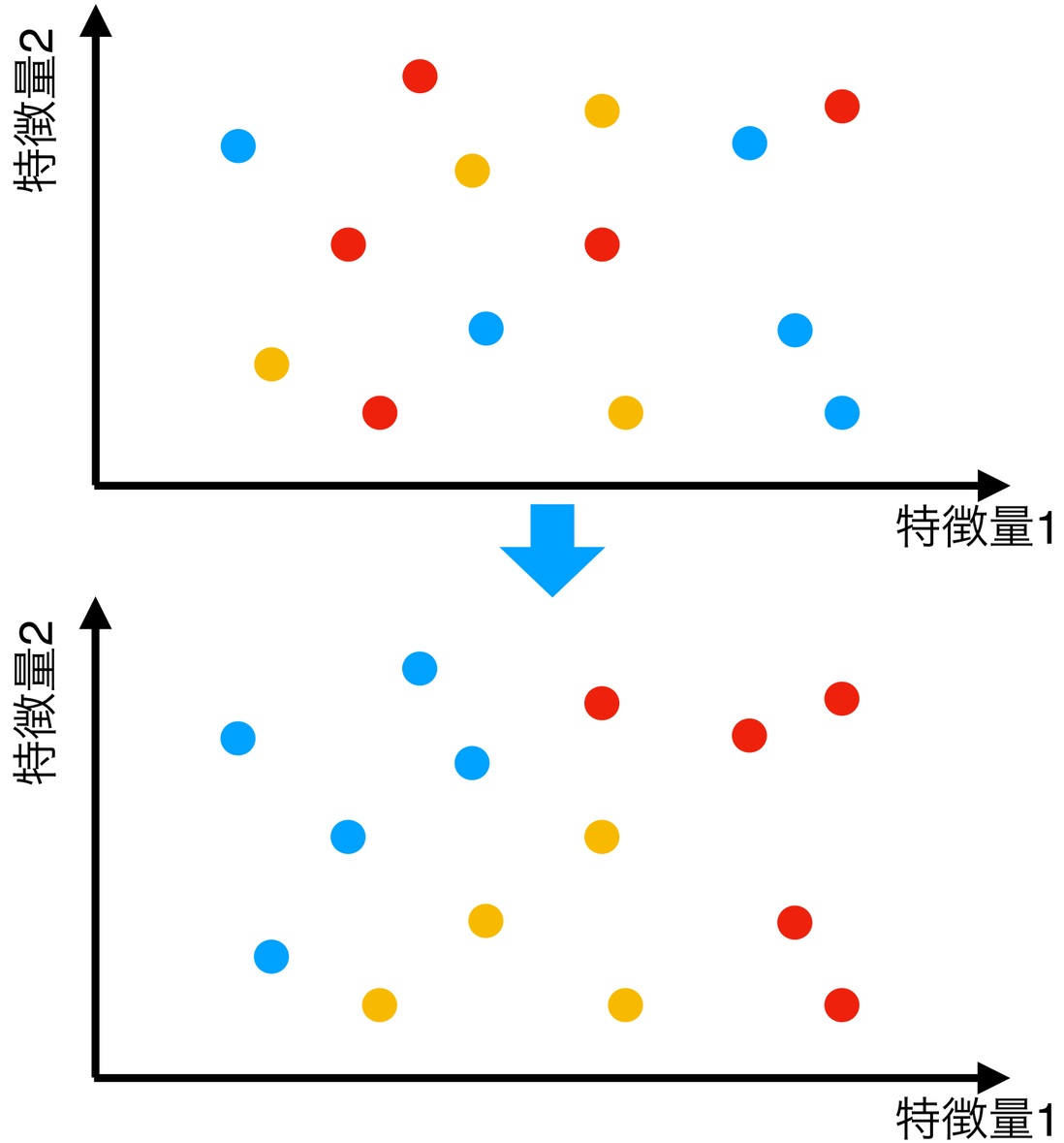
2.各グループの重心を計算する

3.各サンプルが属するグループを、

一番重心が近いグループに変更する

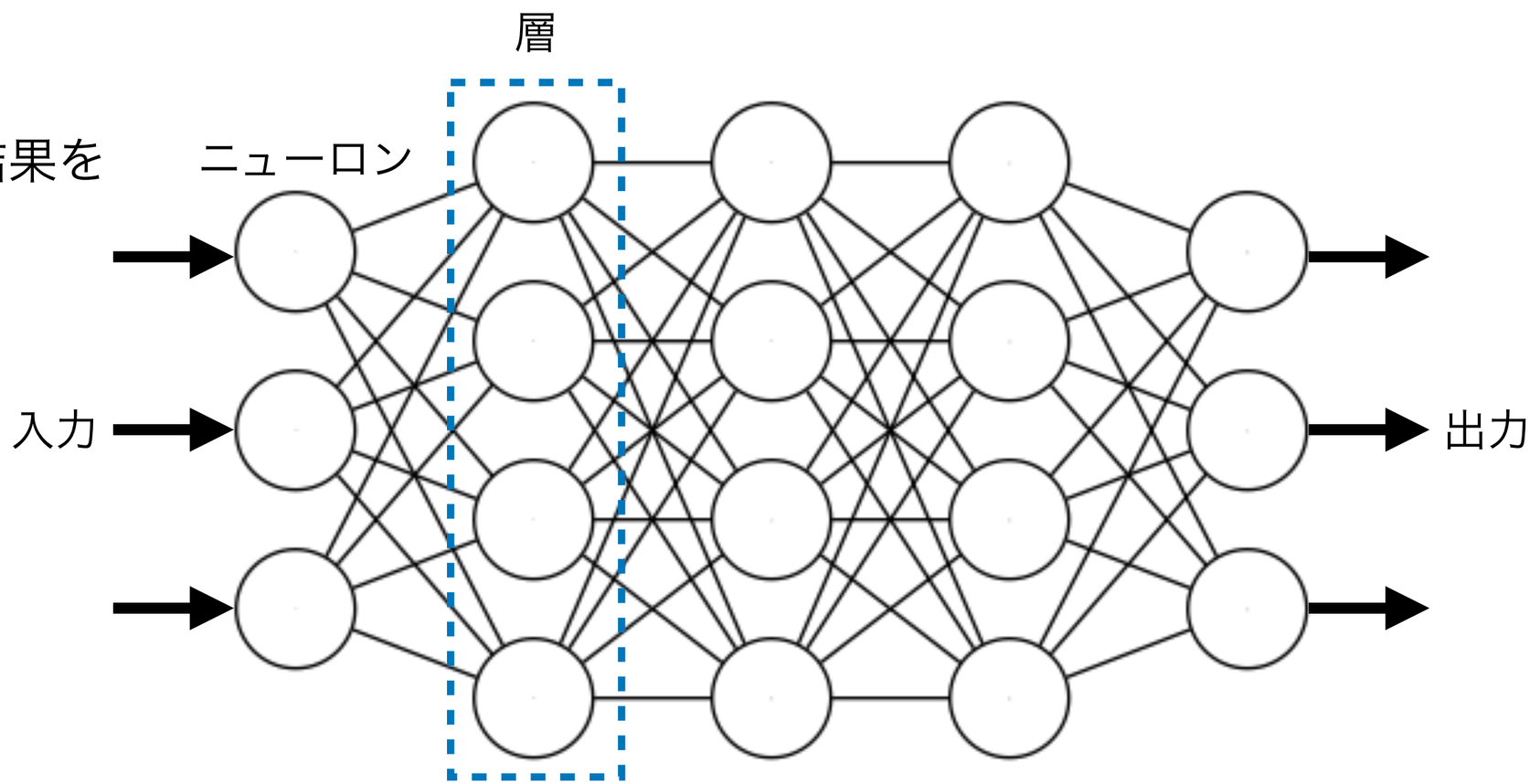
4.変化がなくなれば終了

変化がある場合は 2. に戻る



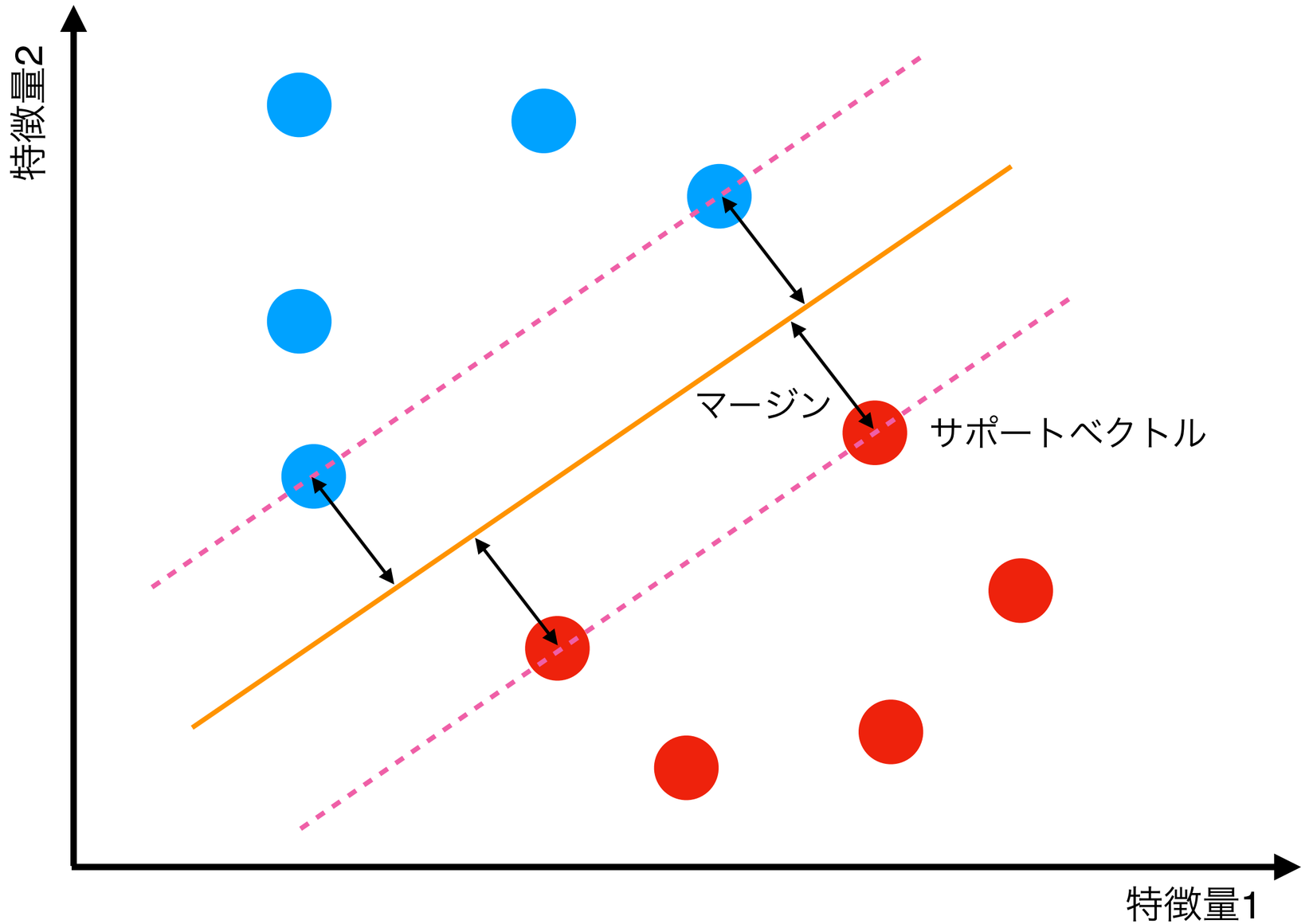
多層パーセプトロン

- 「多層パーセプトロン」（MLP）は、ニューロンを層状に並べたもの
- 数値を入力し、情報を伝播させ結果を出力する
- 出力は確率などの予測値として解釈可能で、ネットワークにより予測を行うことが可能
- ニューロンや層の数を増やすことで、高い表現力を発揮するようになる



サポートベクターマシン（SVM）とは？

- サポートベクターマシンでは、
  グループを明確に分ける
  境界により分類を行う
- 境界は「マージン最大化」
  により決定される
- 境界は「分類器」として機能し、
  データがどちらのグループに
  属するかを判別できる



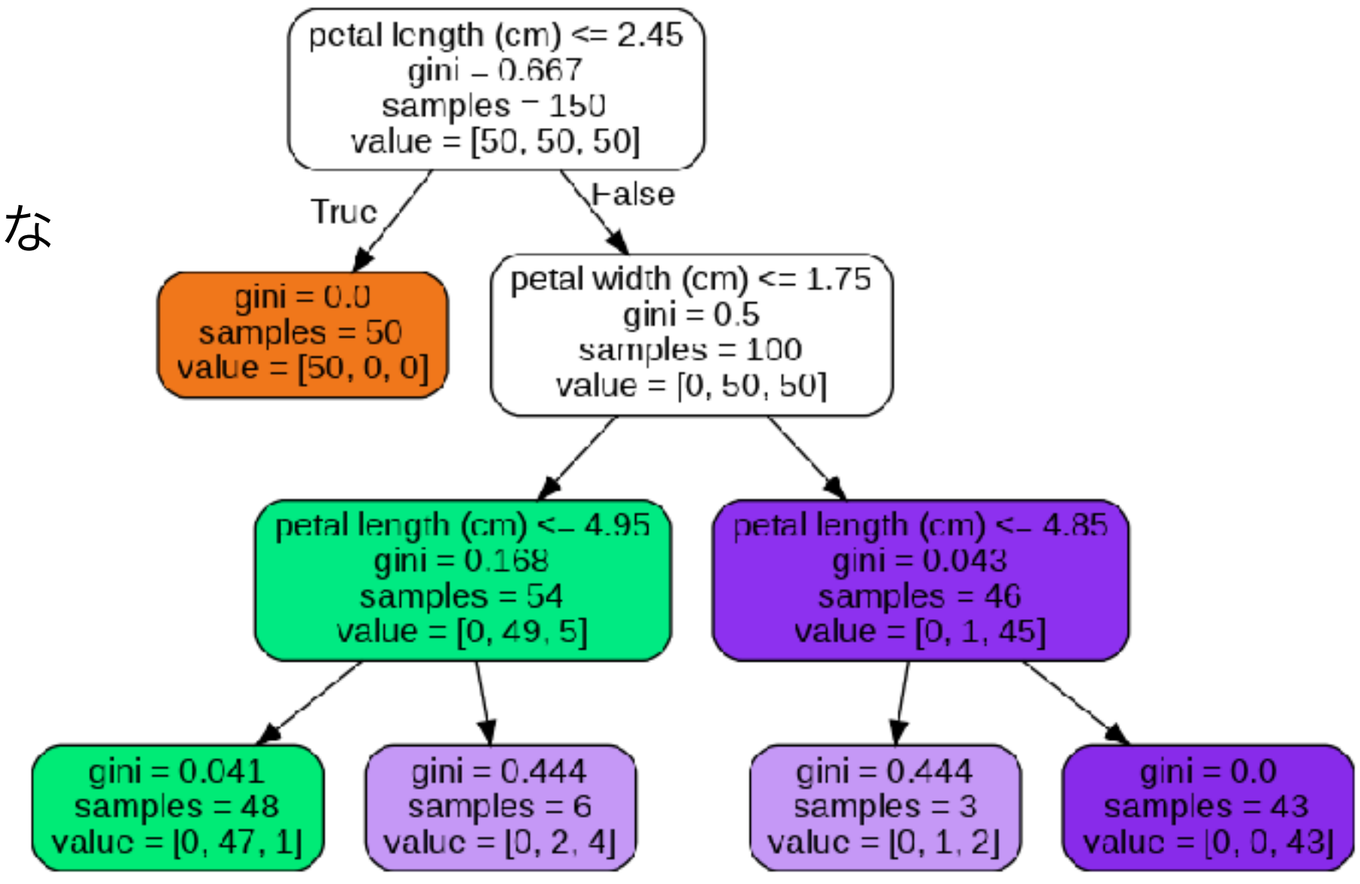
決定木とは？

- 決定木（decision tree）は、木の枝のような

  データ構造を用いて分類を行う

- 学習結果を視覚化が可能で、

  ルールを明確に表記できる



Pandasの基礎

Pandasの基礎

- 01_pandas_basic.ipynb

Kaggleで機械学習を扱う

使用するライブラリ: scikit-learn

- 世界中で広く使われているPythonの機械学習ライブラリ
- Google Colabではデフォルトでインストール済み
- 様々な機械学習アルゴリズムを含む
  - サポートベクターマシン
  - 回帰
  - K近傍法
  - 決定木
  - etc…

Kaggleで機械学習を扱う

- 02_titanic_random_forest.ipynb

演習

演習

- 03_exercise.ipynb

次回の内容

Section1. Kaggleの概要

Section2. 機械学習とKaggle

Section3. 精度向上のためのテクニック

Section4. Titanicの先へ