InterSystems
Creative data technology

開発者向けオンラインセミナー

機械学習101

堀田 稔
インターシステムズジャパン株式会社

2023年11月29日

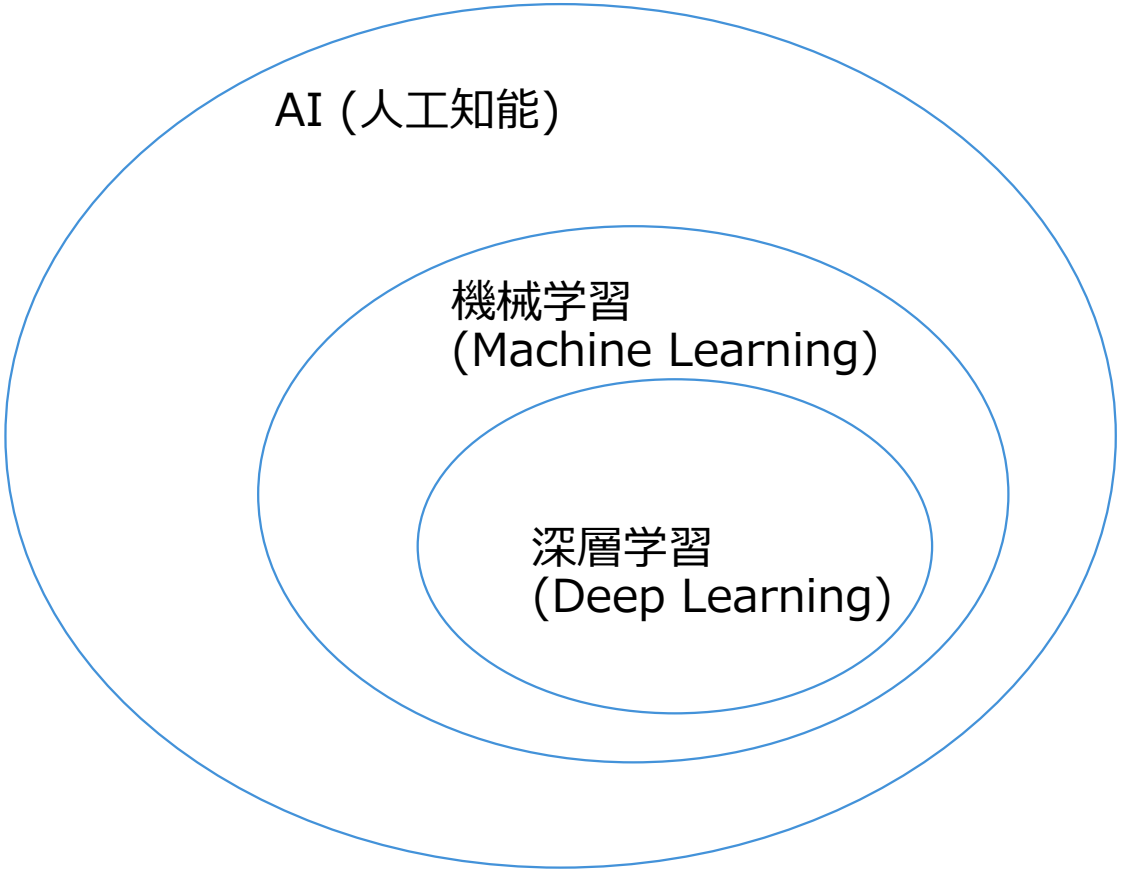
本セミナーの目標

- 機械学習の基本的な仕組みを理解する。
- 機械学習にまつわる用語を理解するための枠組みを紹介する。
- 学習のやり方や性能評価の基本を理解する。

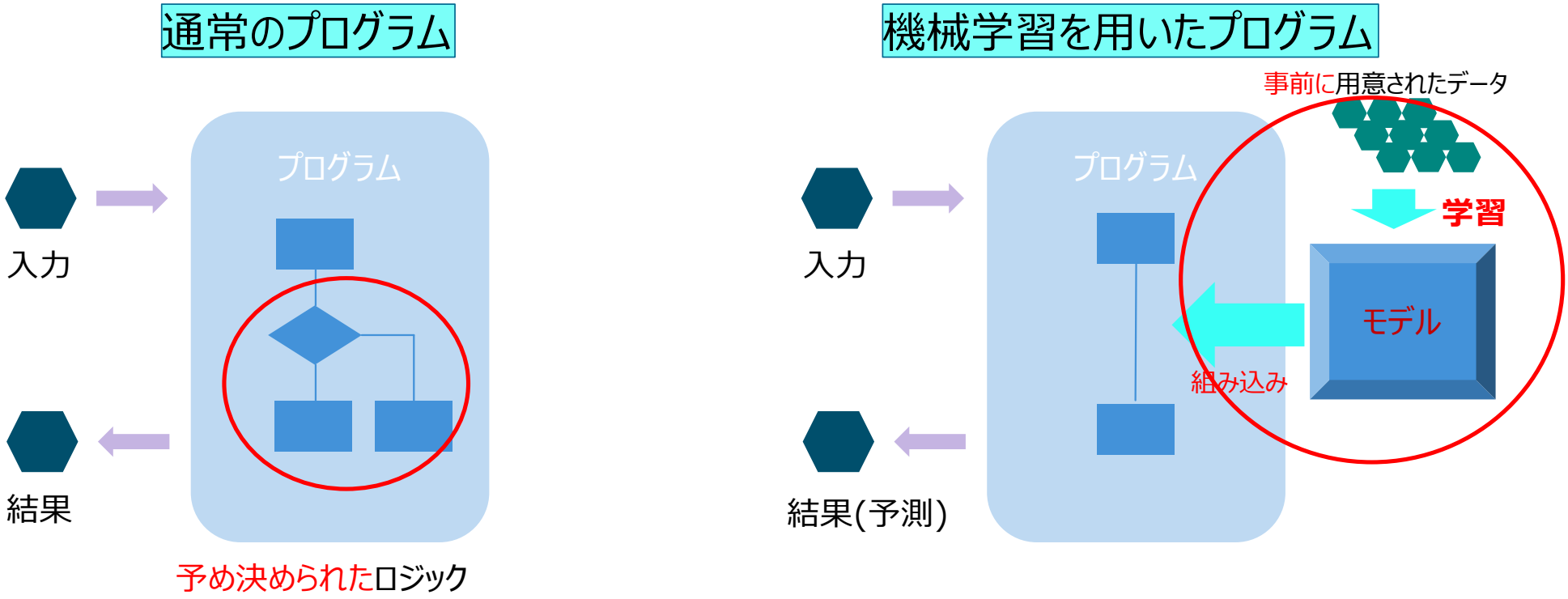
- 機械学習のライブラリを使う利用者にとっても、何が行われているのかを理解するのに必要な知識を得る。

- ※　説明の関係上数式が出てきますが、苦手な方は無理に理解する必要はありません。

AI – 機械学習 – 深層学習 の関係

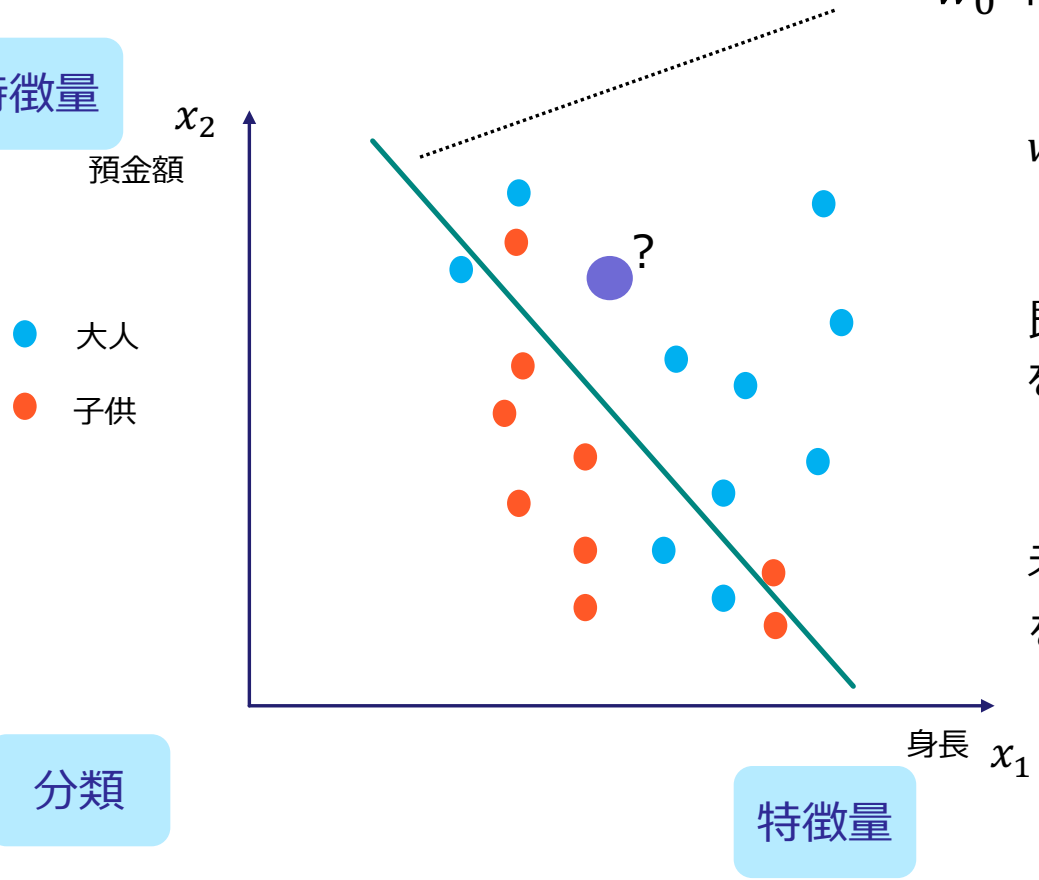


「通常」のプログラムと機械学習を用いたプログラムの違い



機械学習の基本的な考え方

特徴量



分類

特徴量

$$w_0 + w_1 x_1 + w_2 x_2 = 0$$

モデル

w_0　w_1　w_2 が直線を決定する

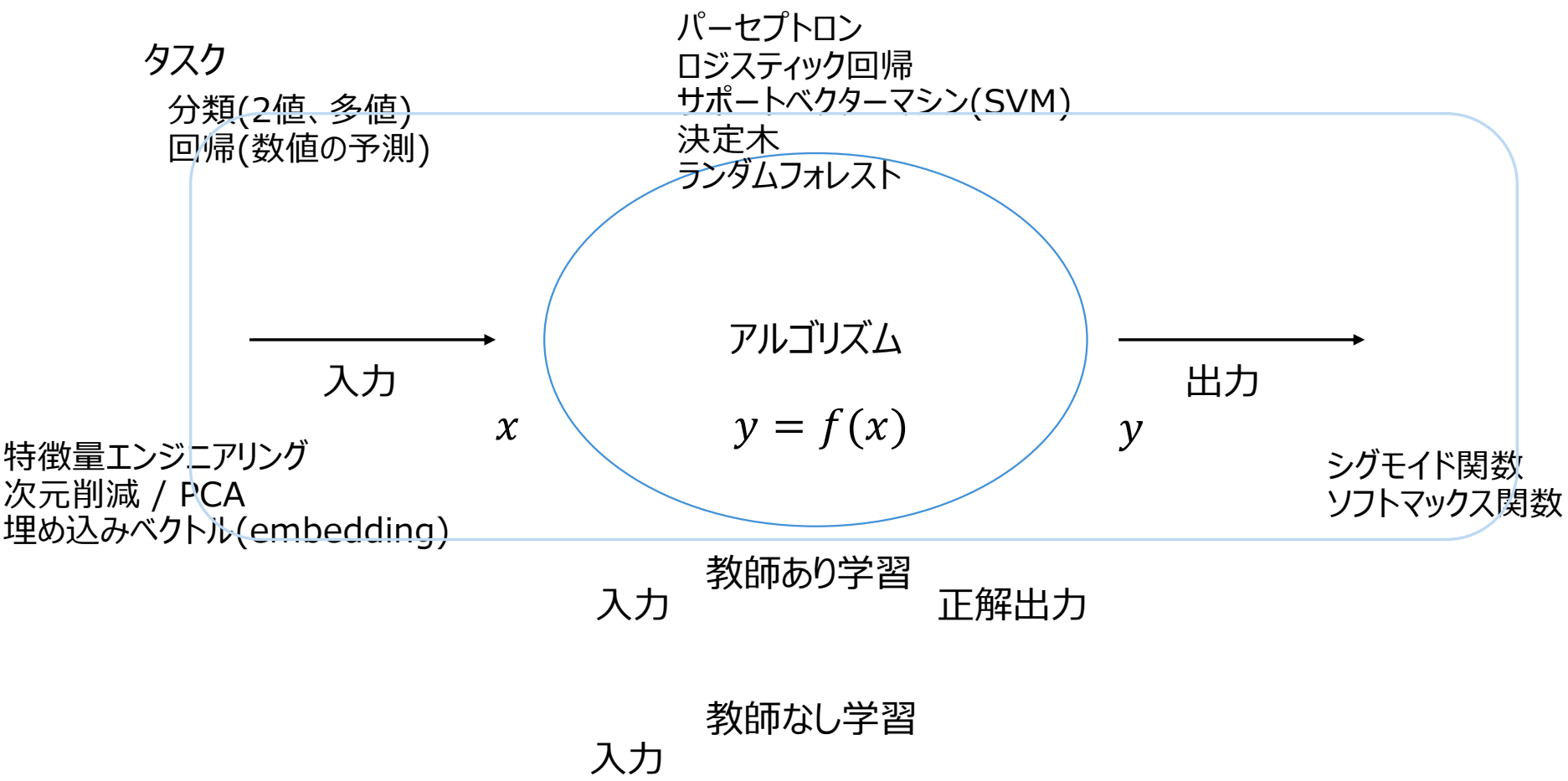
既知のデータから、最も誤りの少ない w_0, w_1, w_2 を求める

学習

未知のデータ x_1，x_2 対して $w_0 + w_1 x_1 + w_2 x_2$ を計算し、0より大きいか小さいかで大人か子供かを判断

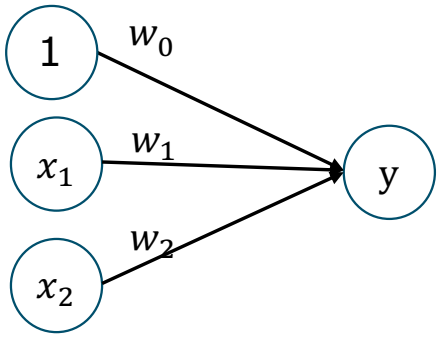
予測

機械学習の流れと用語

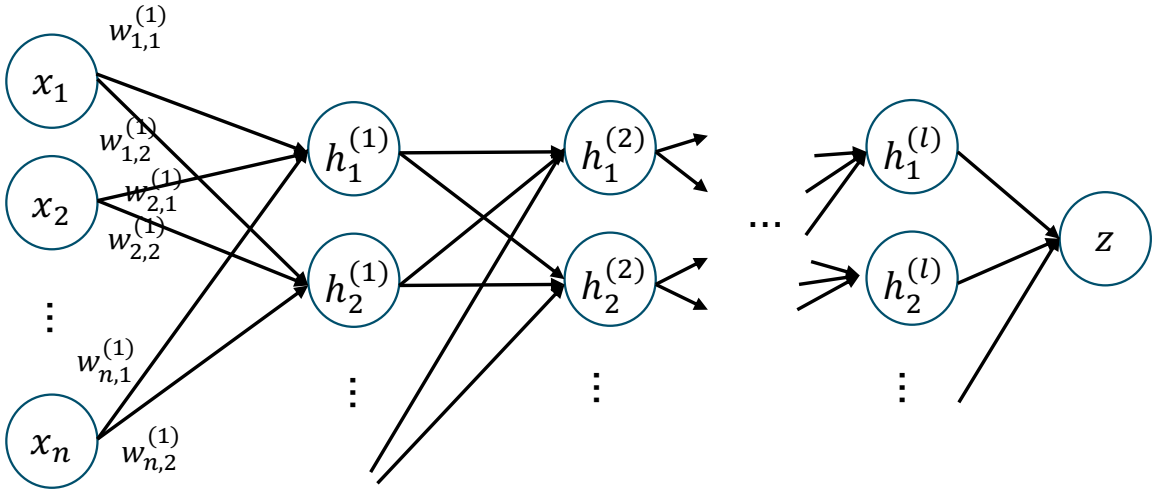

タスク
分類(2値、多値)
回帰(数値の予測)
パーセプトロン
ロジスティック回帰
サポートベクターマシン(SVM)
決定木
ランダムフォレスト
入力
x
アルゴリズム
$y = f(x)$
出力
y
特徴量エンジニアリング
次元削減 / PCA
埋め込みベクトル(embedding)
シグモイド関数
ソフトマックス関数
教師あり学習
入力
正解出力
教師なし学習
入力

深層学習

単純パーセプトロン



深層学習



- 多様な「関数」が表現できる
- パラメータが多く、学習に大量の計算資源が必要 → GPU の活用
- 層の重ね方や繋ぎ方にバリエーション
  - CNN, RNN,…

大規模言語モデルのキーワード： 埋め込みベクトル、Transformer

教師データの使い方

教師データ

学習

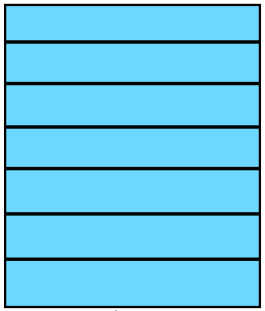
性能評価

モデル

性能OK！

学習データに対して最適なモデルになり、未知のデータに対する性能(汎化性能)が得られない（過学習)ことになりやすい。

教師データ



学習用データ

検証用データ

学習

性能評価

モデル

教師データを学習用データと検証用データに分けて、検証用データで性能を評価。

場合によっては、さらにテスト用データを用意することもある。

性能評価 (2値分類タスク)

代表的な指標

適合率(Precision) $= \dfrac{TP}{TP + FP}$

陽性と予測したデータの中の、真の陽性のデータの割合

高いほど予測の「ノイズ」が少ない

再現率(Recall) $= \dfrac{TP}{TP + FN}$

真の陽性のデータの中で、陽性と予測できたデータの割合

高いほど予測の「見落とし」が小さい

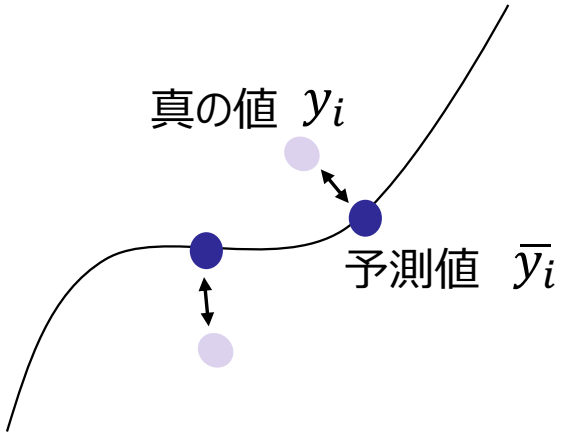
		予測値	
		Positive(陽性)	Negative(陰性)
真の値	Positive(陽性)	TP: True Positive (真陽性) 陽性を**正しく**陽性と予測	FN: False Negative (偽陰性) 陽性を**誤って**陰性と予測
	Negative(陰性)	FP: False Positive (偽陽性) 陰性を**誤って**陽性と予測	TN: True Negative (真陰性) 陰性を**正しく**陰性と予測

混同行列 (Confusion Matrix)

性能評価 (回帰タスク)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y_i})^2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \bar{y_i}|$$

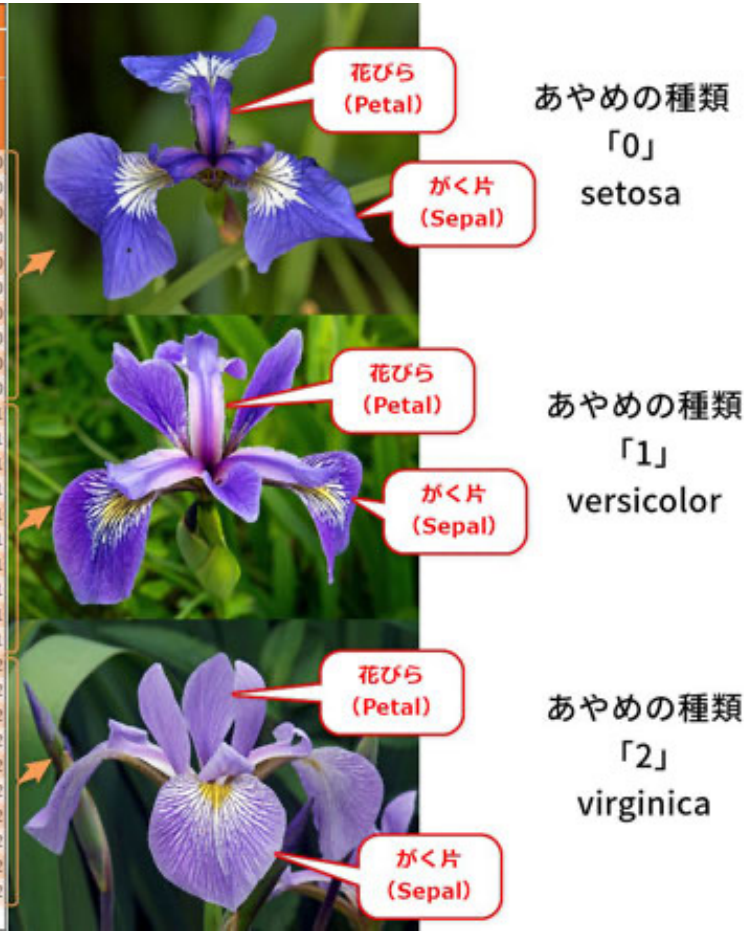


試してみよう：Irisの分類

Pythonで一般に使われる機械学習・データ分析ライブラリ scikit-learnを使って、Iris(アヤメ)の品種分類を試してみましょう

列番号→	1	2	3	4	目的変数↓
説明変数→	Sepal Length	Sepal Width	Petal Length	Petal Width	target
行番号↓	がく片の長さ（cm）	がく片の幅（cm）	花びらの長さ（cm）	花びらの幅（cm）	あやめの種類
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0
5	5.4	3.9	1.7	0.4	0
6	4.6	3.4	1.4	0.3	0
7	5.0	3.4	1.5	0.2	0
8	4.4	2.9	1.4	0.2	0
49	5.0	3.3	1.4	0.2	0
50	7.0	3.2	4.7	1.4	1
51	6.4	3.2	4.5	1.5	1
52	6.9	3.1	4.9	1.5	1
53	5.5	2.3	4.0	1.3	1
54	6.5	2.8	4.6	1.5	1
55	5.7	2.8	4.5	1.3	1
56	6.3	3.3	4.7	1.6	1
57	4.9	2.4	3.3	1.0	1
58	6.6	2.9	4.6	1.3	1
99	5.7	2.8	4.1	1.3	1
100	6.3	3.3	6.0	2.5	2
101	5.8	2.7	5.1	1.9	2
102	7.1	3.0	5.9	2.1	2
103	6.3	2.9	5.6	1.8	2
104	6.5	3.0	5.8	2.2	2
105	7.6	3.0	6.6	2.1	2
106	4.9	2.5	4.5	1.7	2
107	7.3	2.9	6.3	1.8	2
108	6.7	2.5	5.8	1.8	2
149	5.9	3.0	5.1	1.8	2

※合計データ数は150個。表の右端1個は目的変数（ラベル）。



https://atmarkit.itmedia.co.jp/ait/articles/2206/13/news032.html から引用