

# 機械学習

## 第1回 機械学習の概要

兵庫県立大学 社会情報科学部

川嶋宏彰

[kawashima@sis.u-hyogo.ac.jp](mailto:kawashima@sis.u-hyogo.ac.jp)

## 1. 機械学習の概要

- 機械学習の歴史
- どのような分野・講義が関係するのか？
- Google Teachable Machine を使ってみよう

## 2. 機械学習の簡単なモデルといくつかの重要な概念

- k近傍法 (k-NN) を例に

## 3. この講義の進め方

- 授業計画
- 成績評価
- 演習方法

# 分野間の関係

3

## ・ 人工知能 ⊃ 機械学習 ⊃ 深層学習

議論あり

人工知能

機械学習・パターン認識・データマイニング

深層学習  
(ディープラーニング)

画像，音声，自然言語などの  
各メディア技術

知識工学・記号推論

基礎  
知識

アルゴリズムとデータ構造

最適化

情報理論

信号処理

....

線形代数

微分積分

確率・統計

.....

# 人工知能とニューラルネット

4

- 人工知能は記号処理（探索・推論，知識）が主流であった
  - アプローチの対立：記号主義 vs コネクショニズム（ニューラルネット）
  - 現在は融合の流れ（知識 + ニューラルネット）
- 人工知能ブーム
  - 第一次（1950～60年代）：ダートマス会議 (1956)，推論・探索
  - 第二次（1980年代）：知識工学・エキスパートシステム
  - 第三次（2010年代～）：機械学習
- ニューラルネットブーム
  - 第一次（1960年代）：パーセプトロン
  - 第二次（1980後～90前）：バックプロパゲーション
  - 第三次（2010年代～）：深層学習（ディープラーニング）

コネクショニズム  
の復活

コネクショニズム  
の復活2

# 機械学習の歴史

5

- コンピュータの黎明期と機械学習の黎明期は重なる

- 1700～1900年代

- 最小二乗法の発見，統計学の発展

- 1940年代

- コンピュータの発展 (1936: チューリングマシンの提案)
- Hebb則: 脳のシナプス可塑性の計算モデル



A. Turing

(Wikipediaより)

- 1950年代: 機械学習の黎明期

- Alan M. Turing (1950) Computing Machinery and Intelligence, Mind 49:433-460. (学習する機械についても考察)
- Marvin L. Minsky (1951) 40ニューロン程度の学習機械製作
- Frank Rosenblatt (1958) パーセプトロン提案，実機作成

- 1986年 バックプロパゲーションの発表 → 多層化が可能に
  - 理論上，一定の条件で任意の連続な非線形関数を近似可能
  - しかし層が深いと学習困難 → 1990年代半ば頃からは冬の時代
- 1990～2000年代
  - カーネル法のブーム: Support Vector Machine による分類など
  - アンサンブル学習の発展: ブースティング, ランダムフォレストなど
  - 層の多いニューラルネットの学習方法の研究が進展
- 2010年代
  - 深層学習 (deep learning)


# Google Teachable Machine

7

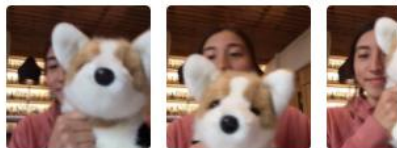
≡ Teachable Machine

<https://teachablemachine.withgoogle.com/>

## 新しいプロジェクト

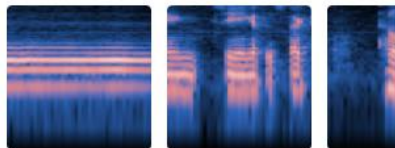
 ドライブから既存のプロジェクトを開きます。

 ファイルから既存のプロジェクトを開きます。



### 画像プロジェクト

ファイルやウェブカメラからの画像に基づいて学習させます。



### 音声プロジェクト

ファイルまたマイクからの1秒間の音声に基づいて学習させます。



### ポーズプロジェクト

ファイルやウェブカメラからの画像に基づいて学習させます。

# Google Teachable Machines

8

≡ Teachable Machine

Class 1  

画像サンプルを追加する:

 ウェブカメラ

 アップロード

Class 2  

画像サンプルを追加する:

 ウェブカメラ

 アップロード

⊞ クラスを追加

トレーニング

モデルをトレーニングする

詳細 

プレビュー

 モデルをエクスポートする

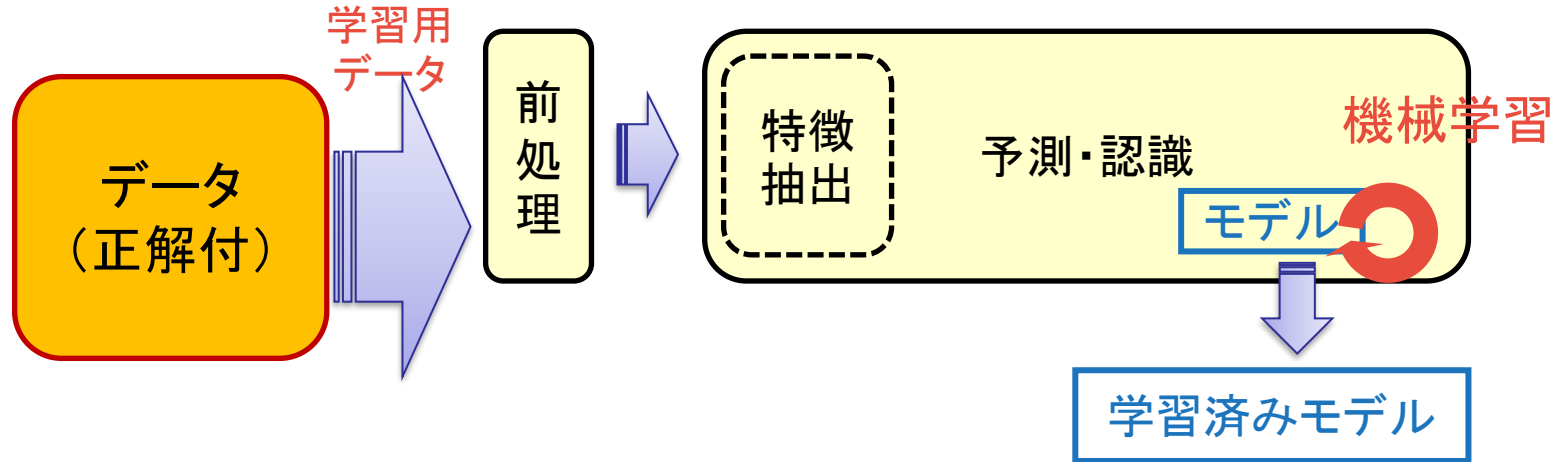
ここでプレビューするには、左にあるモデルをトレーニングしてください。



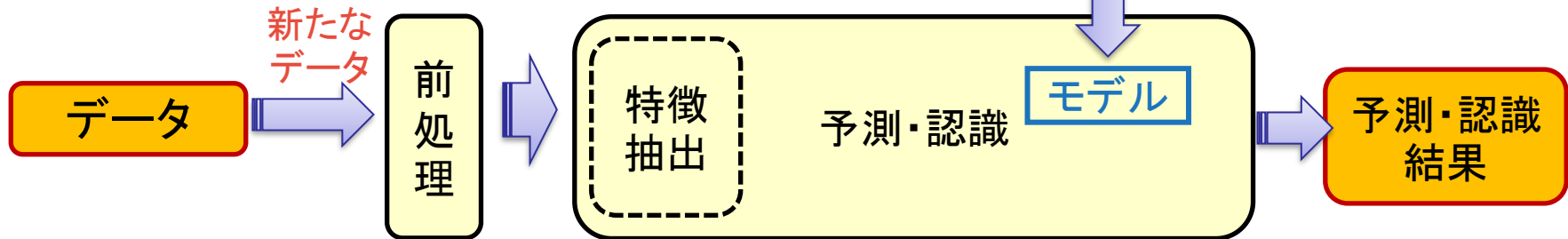
# 機械学習の流れ

9

## 学習・訓練時 (learning / training)



## 認識・推論時 (recognition / inference)




# 機械学習を学ぶためのオープンデータ

10

- 比較的小規模なデータでテストしはじめるのがよい

- UCI Machine Learning Repository
- Kaggle Datasets
- Scikit-learn や seaborn などの同梱データセット



今日はこのデータを使います















**UCI Machine Learning Repository**  
Center for Machine Learning and Intelligent Systems


Welcome to the UC Irvine Machine Learning Repository!


We currently maintain 585 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. For a general overview of the Repository, please visit our [About page](#). For information about citing data sets in publications, please read our [citation policy](#). If you wish to donate a data set, please [donation policy](#). For any other questions, feel free to [contact the Repository librarians](#).

Supported By:  In Collaboration With: 

Latest News:	Newest Data Sets:	Most Popular Data Sets (hits since 2007):
09-24-2018: Welcome to the new Repository admins Dheeru Dua and Eli Karra Taniskidou!	02-17-2021:  <a href="#">Hungarian Chickenpox Cases</a>	3891809:  <a href="#">Iris</a>
04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!	12-09-2020:  <a href="#">Myocardial infarction complications</a>	2102270:  <a href="#">Adult</a>
03-01-2010: Note from donor regarding Netflix data	10-14-2020:  <a href="#">Gait Classification</a>	1625539:  <a href="#">Wine</a>
10-16-2009: Two new data sets have been added.	10-03-2020:  <a href="#">Codon usage</a>	1475200:  <a href="#">Heart Disease</a>
09-14-2009: Several data sets have been added.	09-15-2020:  <a href="#">in-vehicle coupon recommendation</a>	1464326:  <a href="#">Wine Quality</a>
03-24-2008: New data sets have been added!	09-14-2020:  <a href="#">Dry Bean Dataset</a>	1459042:  <a href="#">Breast Cancer Wisconsin (Diagnostic)</a>
06-25-2007: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope		

Featured Data Set: [Forest Fires](#)

 Task: Regression  
Data Type: Multivariate  
# Attributes: 13  
# Instances: 517



Home  
Compete  
Data  
Code  
Communities  
Courses  
More

Search

Sign In Register

## Datasets

Explore, analyze, and share quality data. [Learn more](#) about data types, creating, and collaborating.

[+ New Dataset](#)


Search datasets

Filters

Datasets Tasks Computer Science Education Classification Computer Vision NLP Data Visualization


### Trending Datasets

See All



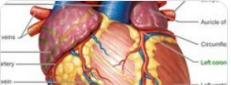
#### Netflix Movies and TV Shows

Shivam Bansal · Updated 2 months ago  
Usability **10.0** · 1 MB



#### Reddit Vaccine Myths

Gabriel Preda · Updated 19 hours ago  
Usability **10.0** · 222 KB



#### Heart Attack Analysis & Prediction Dataset

Rashik Rahman · Updated 11 days ago

# Palmerpenguins データセット

11

- 3種のペンギン

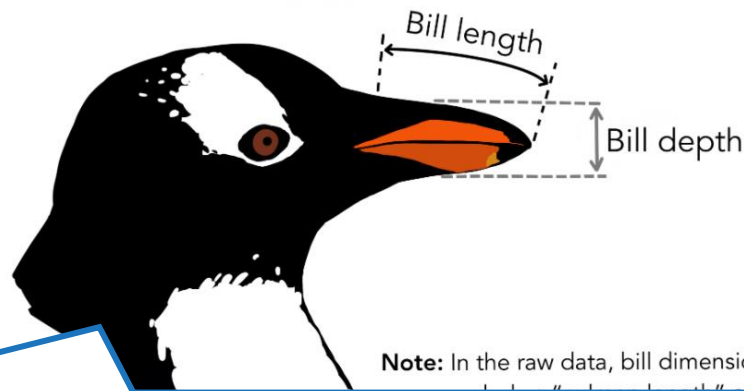
- アデリー (Adelie), ヒゲペンギン (Chinstrap), ジェンツー (Gentoo)

帽子のあごひものこと

CHINSTRAP!

GENTOO!

ADÉLIE!



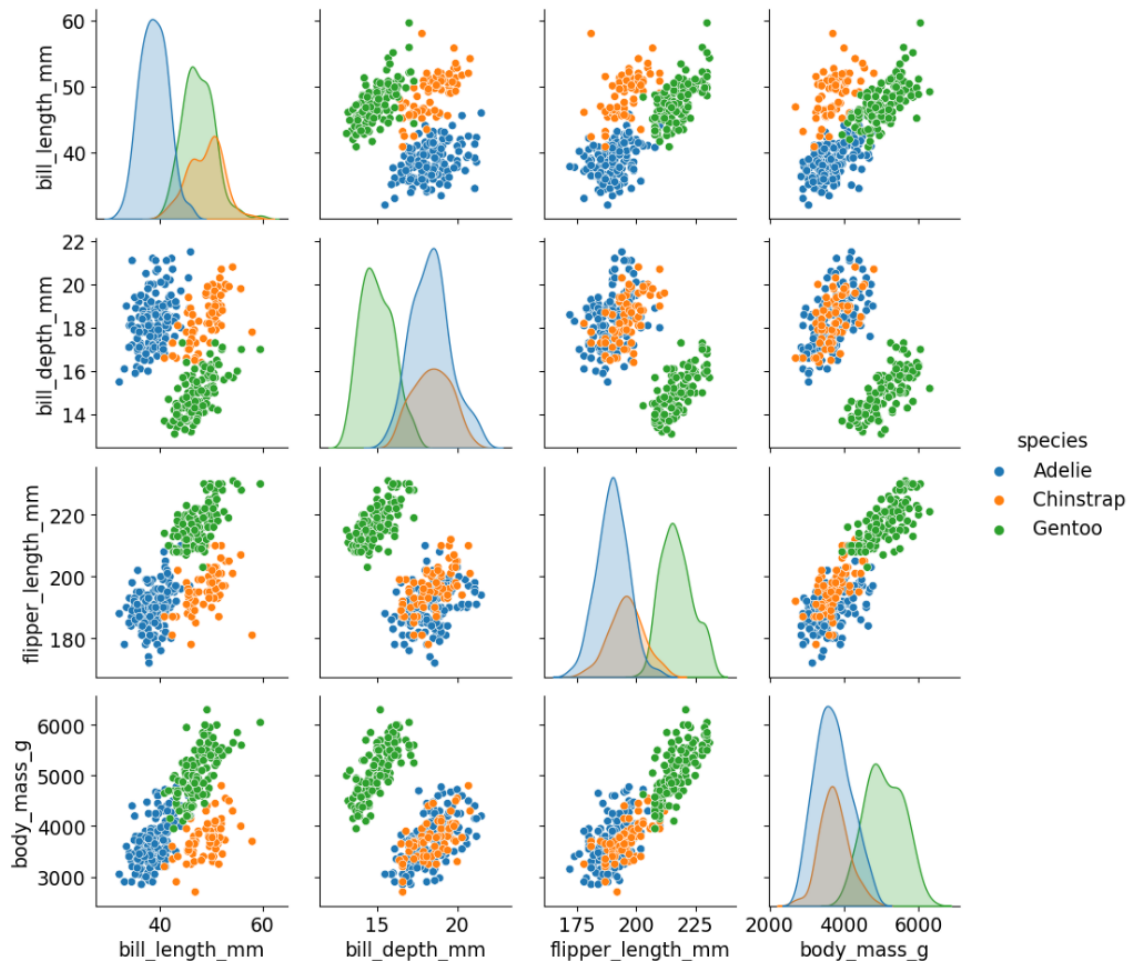
Artwork by @allison\_host

- 生息している南極の島の名前 (island)
- くちばしの長さ (bill\_length\_mm)
- くちばしの高さ (bill\_depth\_mm)
- フリッパー (翼) の長さ (flipper\_length\_mm)
- 体重 (body\_mass\_g), 性別 (sex) など

# まず散布図を見る

12

- 散布図行列 (pairplot)
  - くちばしの長さ (bill\_length\_mm)
  - くちばしの高さ (bill\_depth\_mm)
  - フリッパー (翼) (flipper\_length\_mm)
  - 体重 (body\_mass\_g)



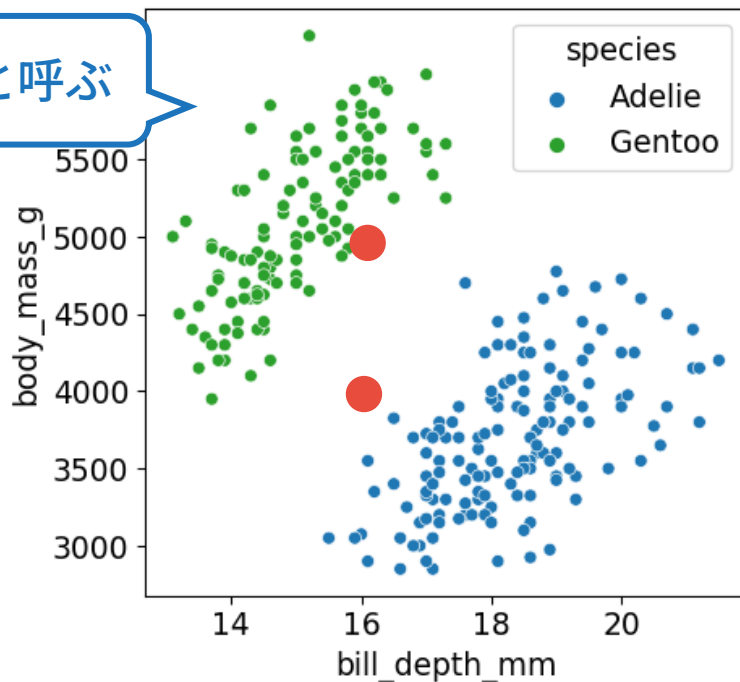
# ひとまず2種で考える

13

- クラス: AdelieとGentoo
- 特徴量: くちばしの高さ (bill\_depth\_mm), 体重 (body\_mass\_g)
- 分類タスク: この2特徴量だけでクラスを分類したい



特徴空間と呼ぶ



# k近傍法 (k-nearest neighbors または k-NN)<sup>14</sup>

- 判定したいデータ点:  $x = (x_1, x_2)$

$x_1$ : bill\_depth\_mm

$x_2$ : body\_mass\_g

- 学習データ:  $(x_1^{(i)}, x_2^{(i)})$  ( $i = 1, \dots, N$ )

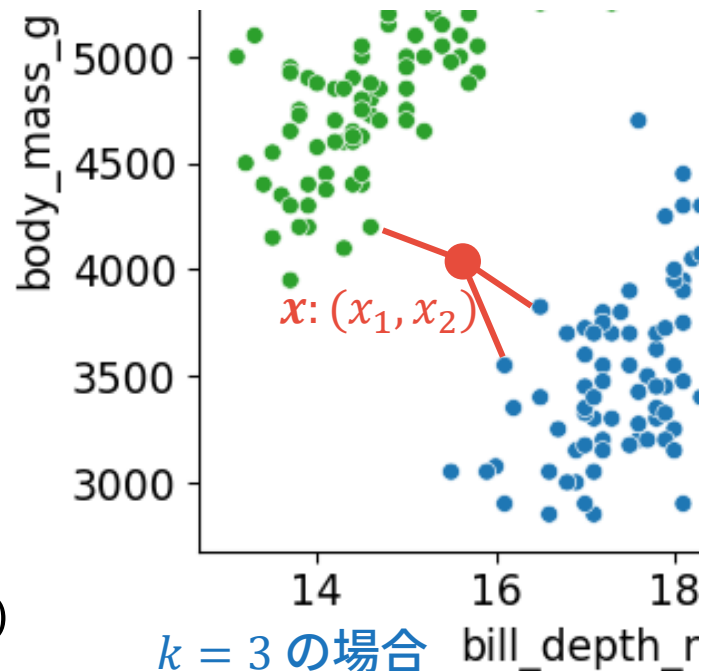
- $N$ : 学習データのサイズ (例:  $N = 273$ )

- 判定したいデータ点との距離

$$\sqrt{(x_1^{(i)} - x_1)^2 + (x_2^{(i)} - x_2)^2} \quad (i = 1, \dots, N)$$

- 判定方法

- 判定対象 $x$ と近いデータ点を k個見つけ, 各点がどのクラスに属するかを見て, 一番多いクラスを判定結果とする (距離で重みづける場合もあり)



# k近傍法

15

- 単純に実装したコードの例
  - Brute force

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ \vdots & \vdots \\ x_1^{(N)} & x_2^{(N)} \end{bmatrix}$$

$$x_{\text{test}} = [x_1 \quad x_2]$$

```
import numpy as np
import scipy
```

```
X = df2[features].values # 274個の2次元特徴量
y = df2['species'].values # 274個のクラスラベル
```

```
x_test = np.array([16, 4000]) # 判定したいデータ
k = 5 # 近い順に何個のデータまで見るか
```

```
dist2 = ((X - x_test) ** 2).sum(axis=1) # x_testとXの各点（各行）との距離の二乗
k_labels = y[np.argsort(dist2)][:k] # 距離の小さいk個の点のラベル
result = scipy.stats.mode(k_labels)[0][0] # 最頻ラベル
print(result)
```

(引き算: numpy の broadcast)

$$\begin{bmatrix} (x_1^{(1)} - x_1)^2 & (x_2^{(1)} - x_2)^2 \\ \vdots & \vdots \\ (x_1^{(N)} - x_1)^2 & (x_2^{(N)} - x_2)^2 \end{bmatrix}$$

この実装には大きな問題が2つ…

# k近傍法の計算量

16

- $N$  個のデータ点全てとの距離を，判定のたびに計算する！
  - 大規模な問題（ $N$  が大きいとき）には実用的ではない
- ここで「データ構造とアルゴリズム」が重要になる
  - 最近傍 (nearest neighbor) 探索の手法では「木構造」を使う
  - 例：k-d tree <https://ja.wikipedia.org/wiki/Kd木>
- Scikit-learnなどのライブラリでは木で実装

結論：実用に際しては機械学習ライブラリを使いましょう  
ただし，学ぶときには自分で実装してみるのもあり

```
from sklearn.neighbors import KNeighborsClassifier

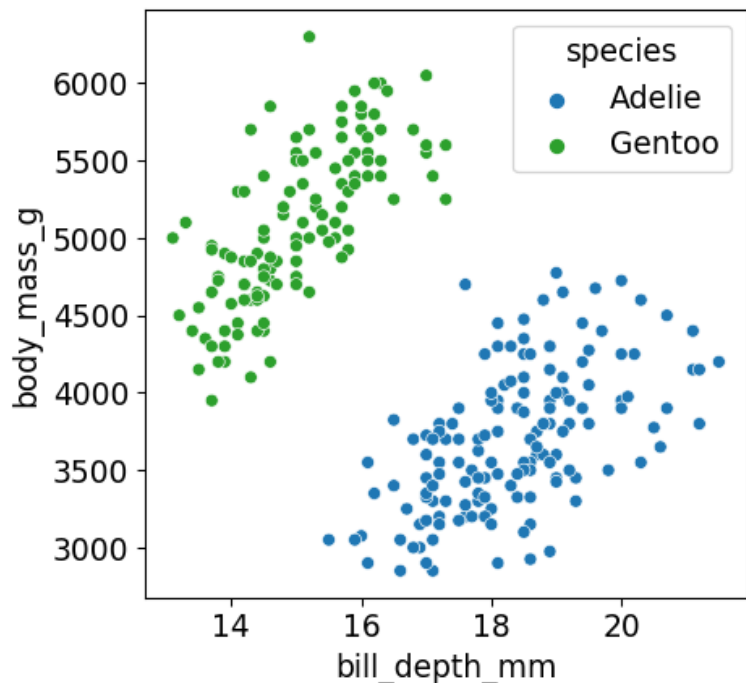
clf = KNeighborsClassifier(n_neighbors=k)
clf.fit(X, y) # 学習
y_pred = clf.predict([[16, 4000], [16, 5000]]) # 一度に複数判定
print(y_pred)
```



# 決定境界 (decision boundary)

17

- 特徴空間内の各点がどのクラスと判定されるか？
  - 決定境界 (decision boundary) : 各クラスに決定される領域の境界

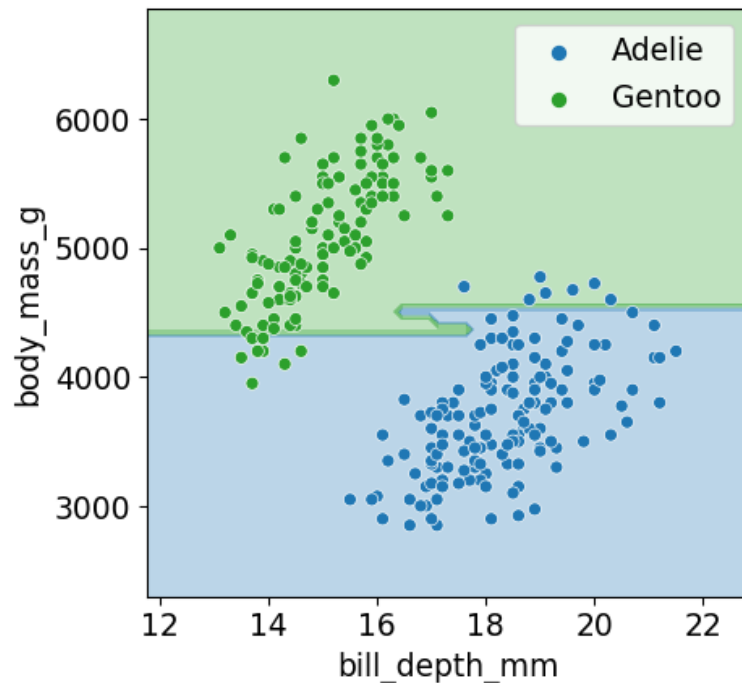
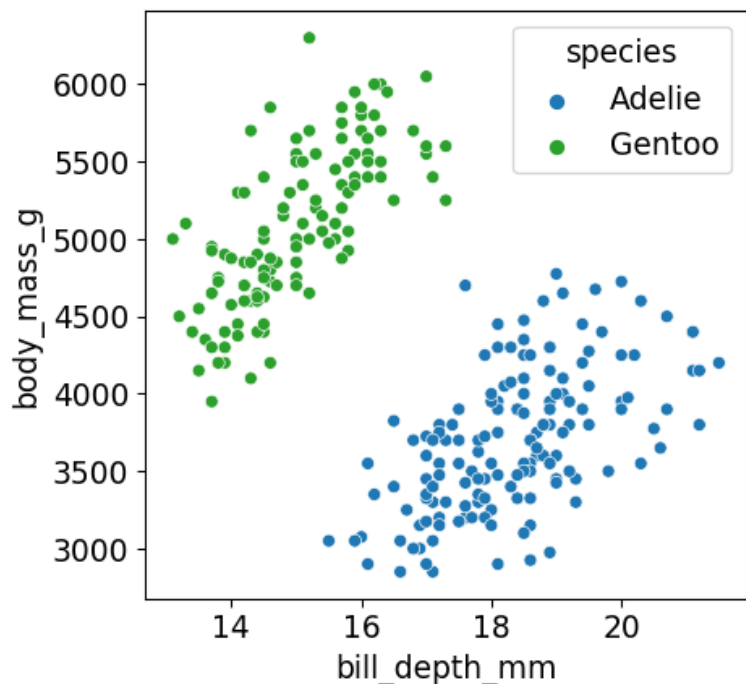


どのような決定境界  
ならば、うまく分類  
できるだろうか？  
(考えてみよう)

# 何もせずk近傍法を使ったときの決定境界

18

- 先ほどの実装における決定境界
  - なぜこのような決定境界になるのか？

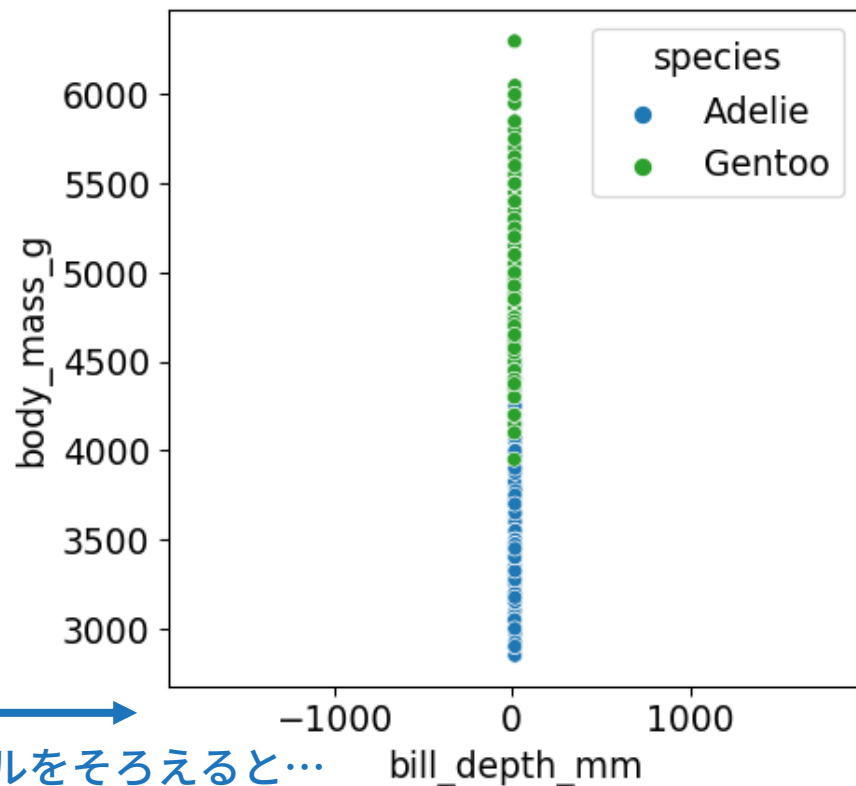
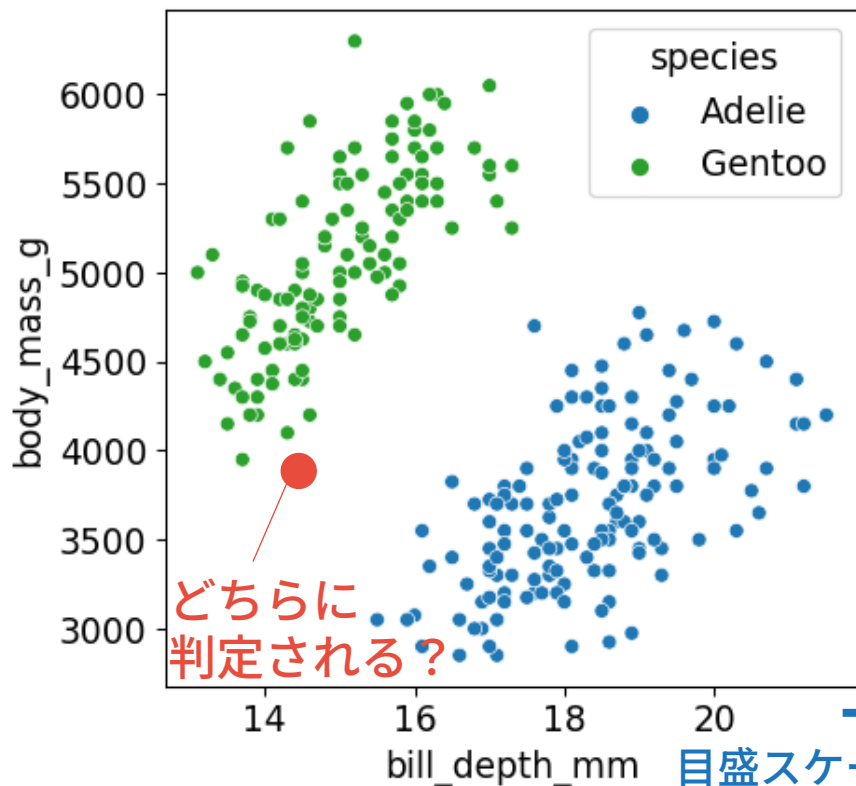


# k近傍法は距離を使う

19

$$\text{距離}^2 = \left(x_1^{(i)} - x_1\right)^2 + \left(x_2^{(i)} - x_2\right)^2$$

実はこんな空間で  
距離を見ている



目盛スケールをそろえると…

# スケーリングの前処理でデータの範囲をそろえる

20

- 各特徴量の平均0を分散1にする標準化 (standardization) が一般的

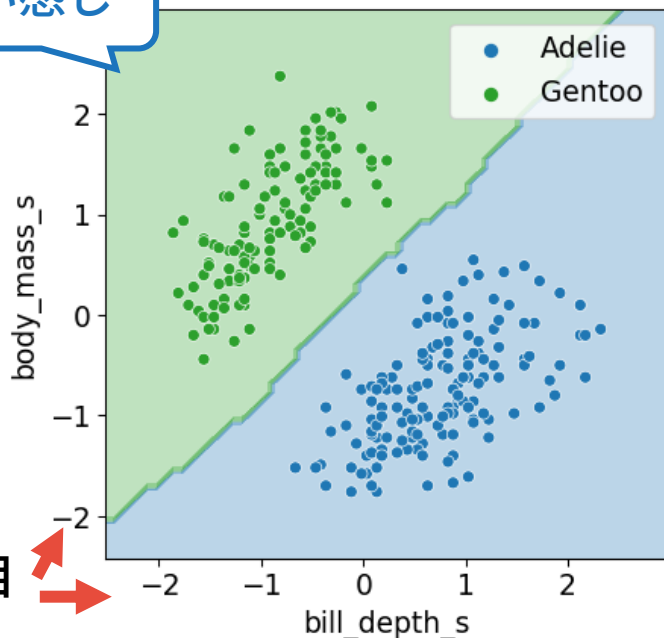
$$z_j^{(i)} = \frac{x_j^{(i)} - \bar{x}_j}{s_j} \quad (\text{データ } i \text{ の特徴量 } j \text{ を標準化})$$

$\bar{x}_j$ : 特徴量  $j$  の平均

$s_j$ : 特徴量  $j$  の標準偏差

いい感じ

標準化後のk-NNの決定境界



```
from sklearn.preprocessing import StandardScaler
```

```
scaler = StandardScaler()
```

```
Xs = scaler.fit_transform(X)
```

```
# StandardScaler を用いず以下のようにしてもよい
```

```
# Xs = (X - X.mean(axis=0))/X.std(axis=0)
```

```
clf_scaled = KNeighborsClassifier(n_neighbors=k)
```

```
clf_scaled.fit(Xs, y)
```

Scikit-learn にも  
用意されている

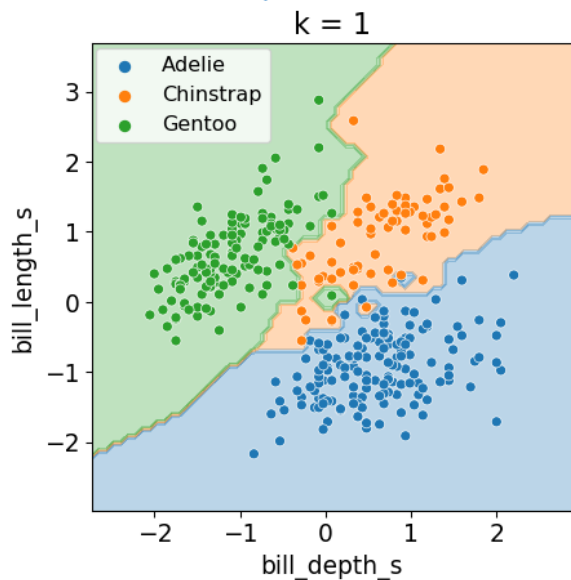
他にも (最大値 - 最小値) やIQRで割る方法など

# $k$ を変えてみる

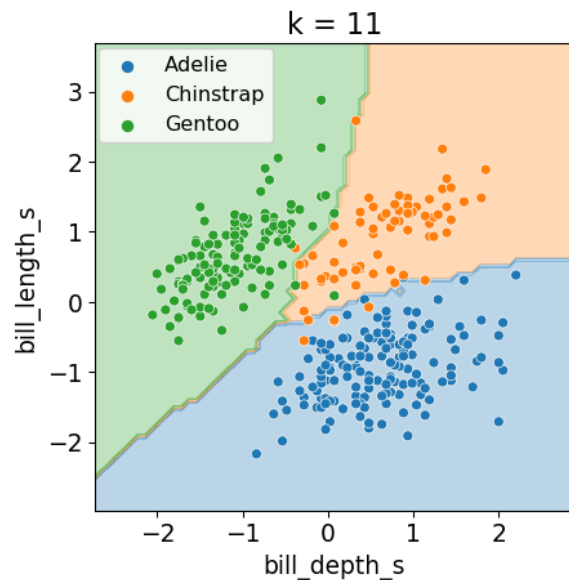
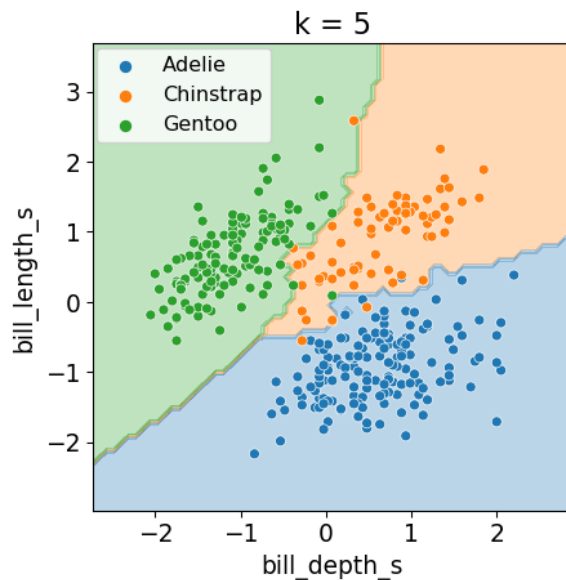
21

- 過学習 (過適合, overfitting): 学習データに適合しすぎることに
- 汎化 (generalization): 未知のデータに適用しやすくなることに
  - 機械学習の文脈での意味 (「汎化」は本来もっと広い意味)

overfitting!



汎化できている!



- k近傍法の復習
  - スケーリング（標準化）の重要性
  - kを変えたときの振る舞い: 汎化とオーバーフィッティング（過学習）
- 過学習しないように，汎化能力を高めるように学習したい！
  - 今後何度も出てきます
  - 対策方法：正則化，交差検証，データ拡張…
- 次回
  - 確率モデル：ガウス分布（密度関数）を仮定する機械学習

回	日付	トピック	
1	2025/04/07	機械学習の概要	2年後期「データマイニング」の復習
2	2025/04/14	確率モデルと機械学習	
3	2025/04/21	教師あり学習（回帰モデル）	
4	2025/04/28	教師あり学習（分類モデル）	
5	2025/05/12	アンサンブル学習	深層学習が流行る前の標準的手法（今でもよく使われている）
6	2025/05/19	演習1	
7	2025/05/26	ニューラルネットの基礎	基本的なニューラルネットとその演習
8	2025/06/02	ディープラーニング（深層学習）	
9	2025/06/09	演習2	
10	2025/06/16	畳み込みニューラルネット	
11	2025/06/23	演習3	いろいろなモデルや手法の紹介
12	2025/06/30	生成モデル	
13	2025/07/07	系列データを扱うモデル	
14	2025/07/14	強化学習（オンデマンド予定）	
15	2025/07/21	まとめと発展的话题（海の日が授業日！）	
16	2025/0?/?	評価（到達度の確認）	

- 成績評価の割合
  - レポート・小テスト: 40% (定期試験の勉強にもなるかも)
  - 定期試験: 60%
- レポートの提出方法
  - ユニパで提出
  - 演習回も一部レポート (Python の ipynb ファイルを配布)



- 教科書はありませんが以下を組み合わせてください

- スライド
- 録画ビデオ（あれば）
- コード
  - Colaboratory で公開するので各自実行できます
- テキスト「機械学習ことはじめ」
  - 授業の前半7回の一部を技術者用にまとめたもの
  - Python で機械学習を行うための基礎が身につく
  - 計6回分をウェブで配布予定
- 参考書
  - 次ページ以降

復習必須！

# 参考書 (シラバス記載+α)

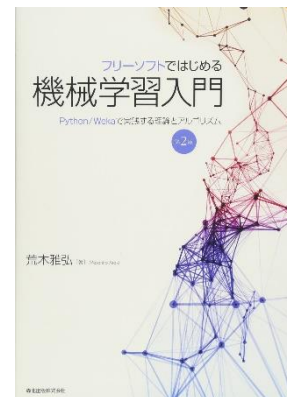
26

## • 参考書

- 荒木雅弘: フリーソフトではじめる機械学習入門 第2版, 森北出版 (2018)
  - アンサンブル学習や深層学習, 系列学習まで網羅
  - 数学的な説明+実装
- 八谷大岳: ゼロから作るPython機械学習プログラミング 入門, 講談社 (2020)
- 平井有三: はじめてのパターン認識, 森北出版 (2012)

## • 深層学習をはじめる

- 斎藤康毅: ゼロから作るDeep Learning, オライリー, (2016)



画像はamazon.co.jpより

- scikit-learn で広く学ぶ
  - A. Géron: scikit-learn, Keras, TensorFlowによる実践機械学習 第2版 (2020)
- 深く学ぶ
  - C. M. Bishop: パターン認識と機械学習 (上, 下), シュプリンガー (2012)
    - 特にベイズ推定, 統計モデルをじっくり学べる
    - いろいろな研究室で輪講されていた
  - T. Hastie他: The element of statistical learning (2<sup>nd</sup> ed.) , Springer (2009)
    - カラフルな図表. 英語版なら以下よりダウンロード可
    - <https://web.stanford.edu/~hastie/Papers/ESLII.pdf>



画像はamazon.co.jpより  
訳者は省略

- 用語（訳語）が教科書によって異なることがある
- classification: 分類，識別
  - クラスの決定（予測）まで含める
- discrimination: 識別，判別
  - クラス決定の手前までを指すことが多い
- 特に，classification は「分類」と「識別」のどちらもよく使われる（指定のテキストは「識別」）
- この授業では classification は「分類」とする

提出不要だが定期テストやレポート対策になることがある

今日のコード:

[https://colab.research.google.com/drive/1m\\_1tDb1HqLklwDWui72UV0r\\_xgF\\_NGzd#scrollTo=iebEQnXBPhzR](https://colab.research.google.com/drive/1m_1tDb1HqLklwDWui72UV0r_xgF_NGzd#scrollTo=iebEQnXBPhzR)

1. 3種のペンギンの散布図行列で気づいたことを述べよ
2.  $k$ 近傍法に関するコードを各自実行せよ
3.  $k$ を変えたときの影響について  $k = 1, 3, 5, 9, 15$  でプロットして考察せよ

# 予習（次回使います）

30

- 行列の計算

- 掛け算など
- 2変数の二次形式

$$(x \ y) \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} =$$

- 列ベクトル×行ベクトル（行列になります）

$$\begin{pmatrix} x \\ y \end{pmatrix} (x \ y) =$$

- ガウス分布（正規分布）

- 1次元ガウス分布（1変量正規分布のこと）

- 分散と共分散

- 2変量の共分散