機械学習エンジニアコース
Sprint

－ クラスタリング（k-means） －

DIVE INTO CODE

今回のモチベーション

目的はなにか

1. クラスタリングのアルゴリズムを知る
   スクラッチを通してk-meansを理解する

2. 線形代数の知識で導入されるアルゴリズム
   主成分分析を使う

3. クラスタ分析を使う

このスライドは?

ここでは、k-means法の基本的な知識を学びましょう

この課題の対象者

① scikit-learnのクラスタリングモデルを用いて、学習、推定するコードが書ける方
② 教師あり学習アルゴリズムを用いたことがある方

教師あり学習と教師なし学習

教師なし学習アルゴリズム（unsupervised learning algorithms）は、学習過程においてデータセットから獲得する**実測値の違い**に基づいて、機械学習アルゴリズムを**教師あり学習アルゴリズム**（supervised learning algorithms）から区別するため、利便的に用いられている名称である[1]。

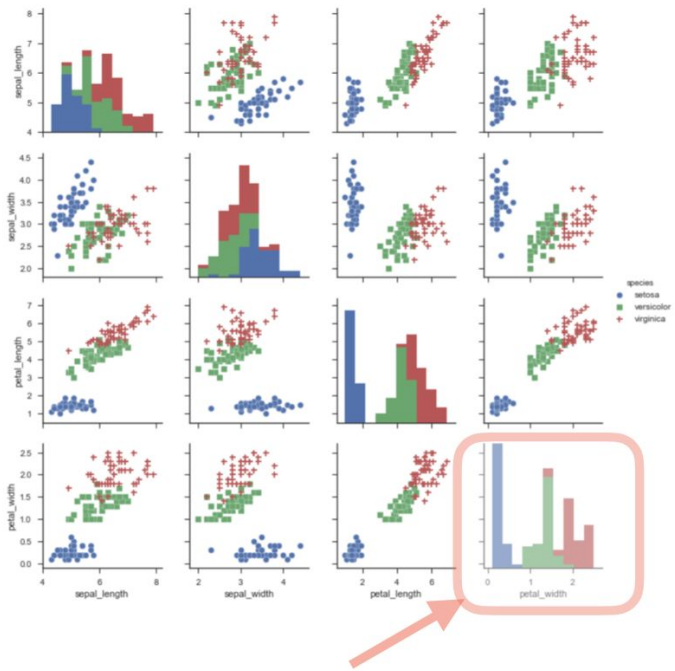
[1] Ian Goodfellow et al.'Deep Learning (Adaptive Computation and Machine Learning series)' 2016. Chapter 5.8参照。**ある実測値が説明変数なのか目的変数なのか**を判別するメタ的なテストが存在しないため、正式かつ厳格な定義はない（注; 著書のなかで判別できない具体的なケースをあげてはいないが「教師データ（正解ラベル）」とそうでないものを厳格に区別する基準がない、という意味と取れる）。大まかには、教師なし学習とは、人手で注釈をつける必要のない分布から情報を抽出する試み、と説明できる。

教師あり学習と教師なし学習

教師あり学習 vs 教師なし学習

教師あり学習は、(特徴量ベクトル $\boldsymbol{x}$ から求められる) 確率ベクトル $\mathbf{x}$ と、それに関連するベクトル $\mathbf{y}$ ($\mathbf{y}$の要素は例えば 0 or 1) の事例の組を学習し、**条件付き確率分布** $p(\mathbf{y}|\mathbf{x})$を推定することで、$\mathbf{x}$ をもとに $\mathbf{y}$ を予測できるように学習を行うアルゴリズムと考えることができる[2]。

別の仕方で説明すると、例えば irisデータセットの場合、教師あり学習とは、**アヤメが属する種を表す目的変数** $\mathbf{y}$ と、それぞれに関連づいた**アヤメの各部位の測定値からなる説明変数（特徴量）** $\boldsymbol{x}$を組で与えると、**目的変数を用いた評価**を受けながら学習し、推定時には**未知の説明変数**から**その種を予測する**ことを目的とする。



petal_widthの特徴量ベクトルから成るヒストグラム。
これをスケーリングし、全体の面積の総和が1になるとき、このベクトル**が確率ベクトル x** である。

[2] 機械学習アルゴリズムを最尤推定法で表現した場合。**条件付き確率分布 p(y|x)**とは、*x* と *y* について、*x* の値が特定の値であることを知ったときの *y* の確率分布を表す。

教師あり学習と教師なし学習

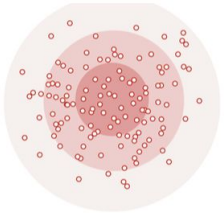
教師あり学習 vs 教師なし学習

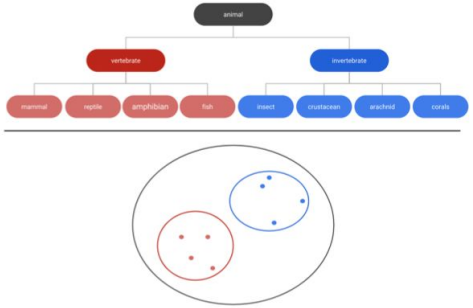
教師なし学習では、確率ベクトル $\mathbf{x}$ **のみ**から、その**確率分布** $p(\mathbf{x})$ や**主要な特性**を学習する。

明示的に分布を学習する事例としては密度推定(density estimation)[3]があり、その発展的タスクとして、暗に分布を学習するノイズ除去や画像生成などが事例としてあげられる。
さらに、分布よりも単純な**ある種の特性**をデータセットから学習するアルゴリズムとして、**クラスタリング**(Clustering)や、**主成分分析**(Principal Component Analysis)がある。これらは教師なし学習のなかでも古典的な部類に含まれる。

[3] データからそれがサンプリングされた（と仮定する）確率密度関数を推定する手法。

クラスタリング（k-means法）



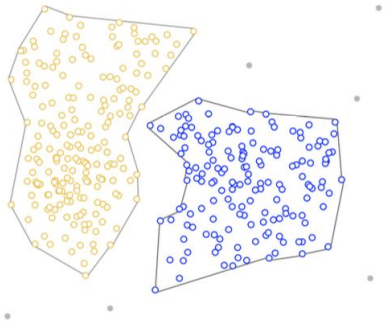


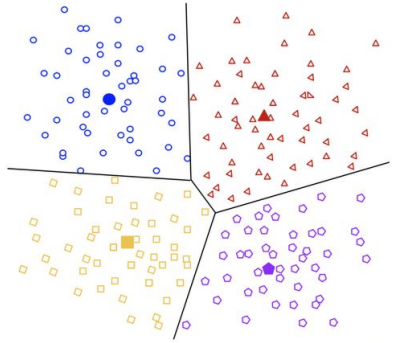
クラスタリングとは

クラスタリング (Clustering) とは、互いに近接するデータセットを集めて**クラスタ**[4]に分割するアルゴリズムのことである。
今回は、クラスタリング手法のひとつ[5]、k-meansを扱う。

代表的なクラスタリング手法を4つあげると、**重心ベース**(右下)、**密度ベース**(左下)、**分布ベース**(左上)、そして**階層的クラスタリング**(右上)である。

k-meansは重心ベースの手法に含まれる。





[4] クラスタとは、**特定の類似性のために集約されたデータ点の集合。**
[5] クラスタリングについて包括的に調査したサーベイ論文はこちら。 https://link.springer.com/article/10.1007/s40745-015-0040-1

http://tech.nitoyon.com/ja/blog/2013/11/07/k-means/

k-means法とはなにか

k-meansアルゴリズムは、*k*個（固定数）の重心を識別し、重心位置を最適化する（平均値を最小化する）ために反復的な計算を行い、すべてのデータ点を最も近接するクラスタに割り当てる仕組み。
クラスタ内の**誤差平方和**[6]を削減することにより、すべてのデータ点が各クラスタに割り当てられる。

[6] スライドp13参照。

k-means法

与えられた条件は何か

k-means法においては以下が仮定されている。

① 入力データは特徴量行列Xのみ（教師なし学習）
② ハイパーパラメータとして固定値 k を入力する

k-means法

Sprintの準備

k-meansの幾何学的説明

① データ分布からランダムサンプリングしたk個のデータ点をクラスタの重心とする（kはハイパーパラメータ）
② 各重心に対しすべてのデータ点とのユークリッド距離を計算する
③ 各重心との距離が最小となるデータ点郡を、その重心に帰属するクラスタとする
④ k個のクラスタ毎にデータの平均となる点を求め、新しい重心とする
⑤ ②へ戻る

k-means法

Sprintの準備

実装上の手順を確認する

① サンプル数のインデックスに対し、kクラス分のランダムな初期ラベルを割り当てる
② 各ラベル毎にデータ点をグルーピングし、クラスタを作成する
③ クラスタ毎にデータ点の平均値を求め、そのクラスタの重心とする
④ その重心から、すべてのサンプルのデータ点との距離を計算する
⑤ 各データ点から見て、距離が最小となる重心のクラスタにそのデータ点を割り当てる
⑥ ③～⑤を繰り返す
⑦ 収束条件（値が変化しない・定義した反復回数に達した等）を満たしたら、終了
⑧ 初期値を変更し①～⑦をn回繰り返し、SSEが最小のものを選ぶ

k-means法

SSEについて

クラスタ内**誤差平方和**（Sum of Squared Errors）
クラスタリングの性能評価関数。

関数

(データ点の座標)ー(重心座標)

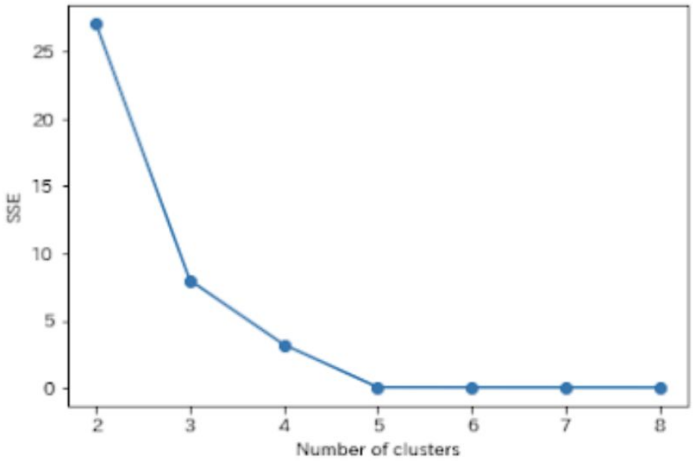
$$SSE = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|X_n - \mu_k\|^2$$

自分が属するクラスタならば 1
自分が属さないクラスタならば 0

エルボー法について

クラスタ数を決定する方法のひとつ。

右図のような縦軸をSSE、横軸をクラスタ数とするグラフを作り、ヒジ（エルボー）のように曲がっている箇所を見てクラスタ数を決定する。



k-means法

シルエット分析について

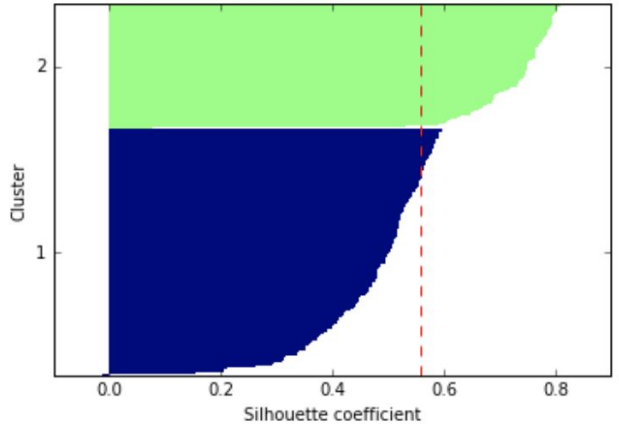
クラスタ数を決定する方法のひとつ。

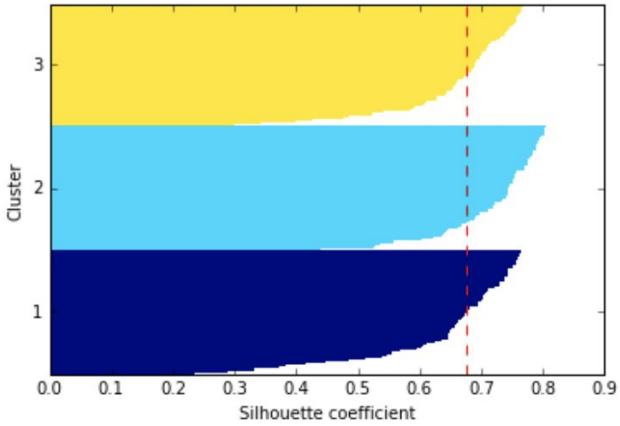
縦軸をサンプルのインデックス（クラスタごとにソート）、
横軸をシルエット係数としたシルエット図を作る。
点線はシルエット係数の平均を表す。

クラスタ数を変えた複数のシルエット図を作成し、以下のようなものを選ぶ。

- 厚さがだいたい等しい
- どのクラスタも点線を超えたサンプルがある程度ある。

例えば右の例だと、k=2（上）より、k=3（下）方が好ましい。





クラスタリング（k-means）
完