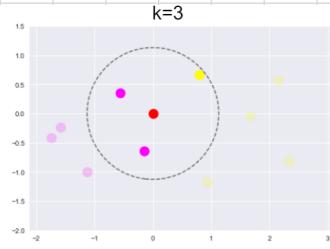
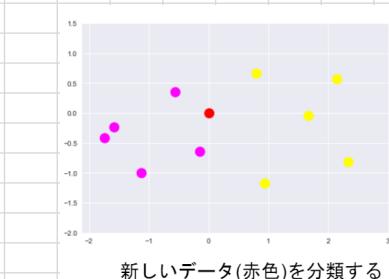


K近傍法(KNN)

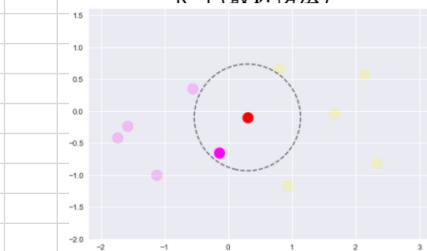
最近傍のデータをK個取ってきて、それらがもっと多く所属するクラスに識別



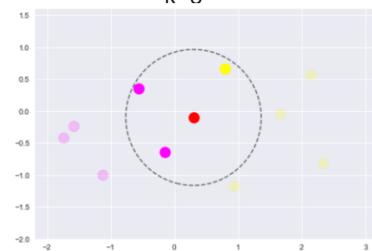
kを変化させると結果も変わる

k=1の場合: 最近傍法

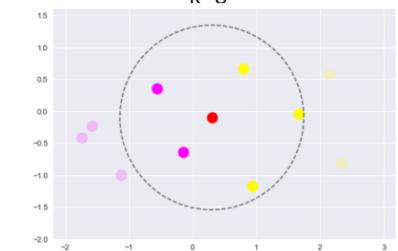
k=1 (最近傍法)



k=3

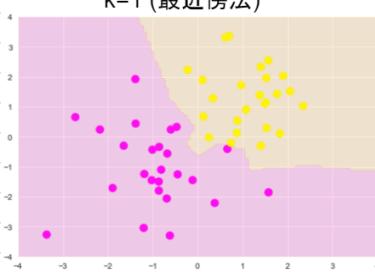


k=5

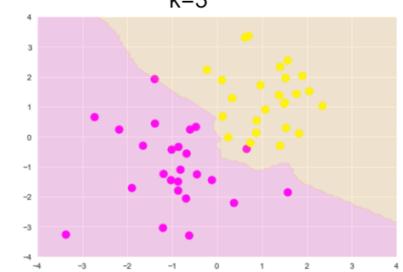


kを大きくすると決定境界は滑らかになる

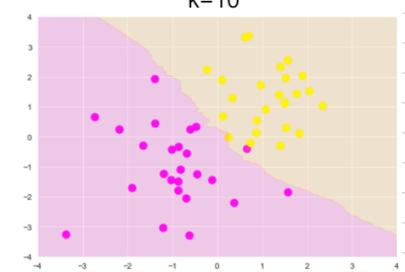
k=1 (最近傍法)



k=3



k=10



k近傍法 (kNN)

識別ルール

※rejectではなく、ランダムに割り当てても良い

$$\text{識別クラス} = \begin{cases} j & \{k_j\} = \max\{k_1, \dots, k_K\} \\ \text{reject} & \{k_i, \dots, k_j\} = \max\{k_1, \dots, k_K\} \end{cases}$$

①学習データ

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im}) \in \mathbb{R}^m$$

②所属するクラス

$$\Omega = \{C_1, C_2, \dots, C_K\} \quad \text{③} i \text{番目のデータが所属するクラス} \quad w_i \in \Omega$$

④入力にもっとも近いk個のデータ集合

$$k(\mathbf{x}) = \{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \dots, \mathbf{x}_{i_k}\}$$

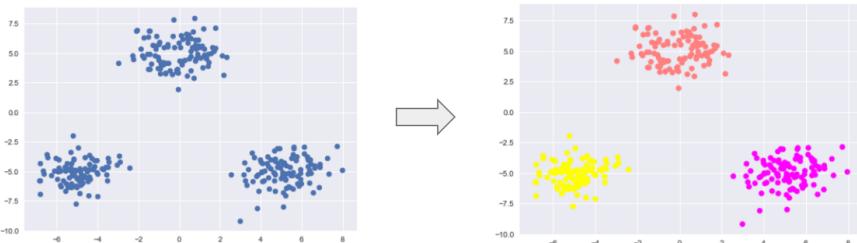
⑤④の中でクラスjに属するデータの数

$$k_j$$

K平均法(K-means)

教師なし学習
クラスタリング手法
与えられたデータを k 個のクラスタに分類する。

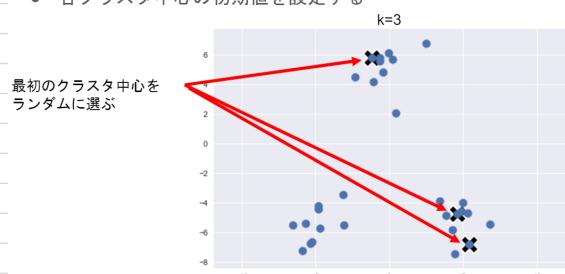
クラスタリング・・・特徴の似ているもの同士をグループ化



k平均法(k-means)のアルゴリズム

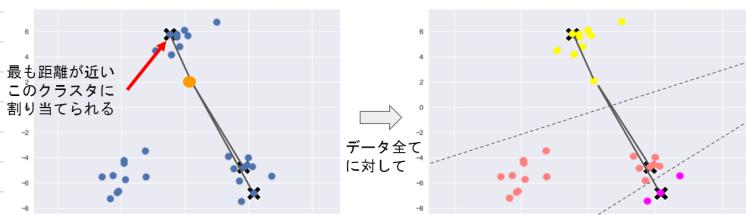
1) 各クラスタ中心の初期値を設定する

- 各クラスタ中心の初期値を設定する



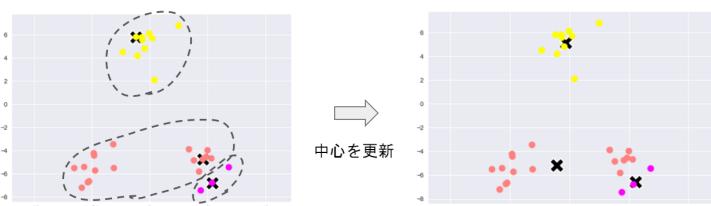
2) 各データ点に対して、各クラスタ中心との距離を計算し、最も距離が近いクラスタを割り当てる

- 各データ点に対して、各クラスタ中心との距離を計算し、最も距離が近いクラスタを割り当てる



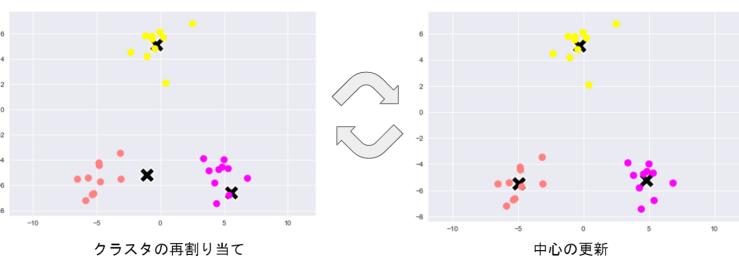
3) 各クラスタの平均ベクトル(中心)を計算する

- 各クラスタの平均ベクトル(中心)を計算する

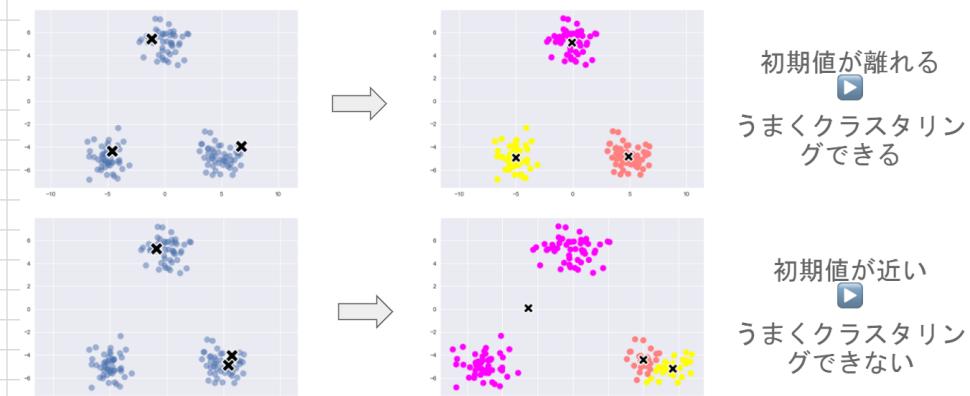


4) 収束するまで2, 3の処理を繰り返す

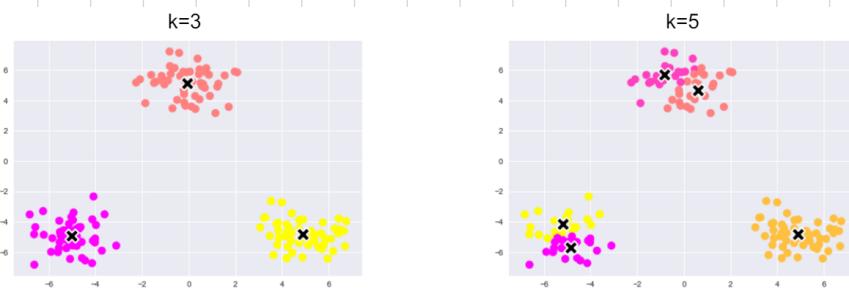
- クラスタの再割り当てと、中心の更新を繰り返す



中心の初期値を変えるとクラスタリング結果も変わりうる



kの値を変えるとクラスタリング結果も変わる



k-平均法 (k-means)

①学習データ

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im}) \in \mathbb{R}^m$$

②各クラスタの中心

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^N q_{ik} \mathbf{x}_i}{\sum_{i=1}^N q_{ik}}$$

③所属クラスタの識別

$$q_{ik} = \begin{cases} 1 & (\text{where } \mathbf{x}_i \in M(\boldsymbol{\mu}_k)) \\ 0 & (\text{else}) \end{cases} \quad M = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}$$

④各クラスタ内データと中心の距離の総和

$$J(q_{ik}, \boldsymbol{\mu}_k) = \sum_{i=1}^N \sum_{k=1}^K q_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

⑤中心を変化させたときの④の変化量

$$\frac{\partial J(q_{ik}, \boldsymbol{\mu}_k)}{\partial \boldsymbol{\mu}_k} = 2 \sum_{i=1}^N q_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k) = 0$$