情報管理

第８回：クラスタリング
教師なしクラスタリング(1)

今回の講義内容

第６回ではクラスタリングの方法として，「教師ありクラスタリング」を解説しました。

教師ありクラスタリングでは，学習データにクラスラベルが存在する状況でクラスタリングを行います。

今回は，学習データにクラスラベルが存在しない「教師ありクラスタリング」を説明します。

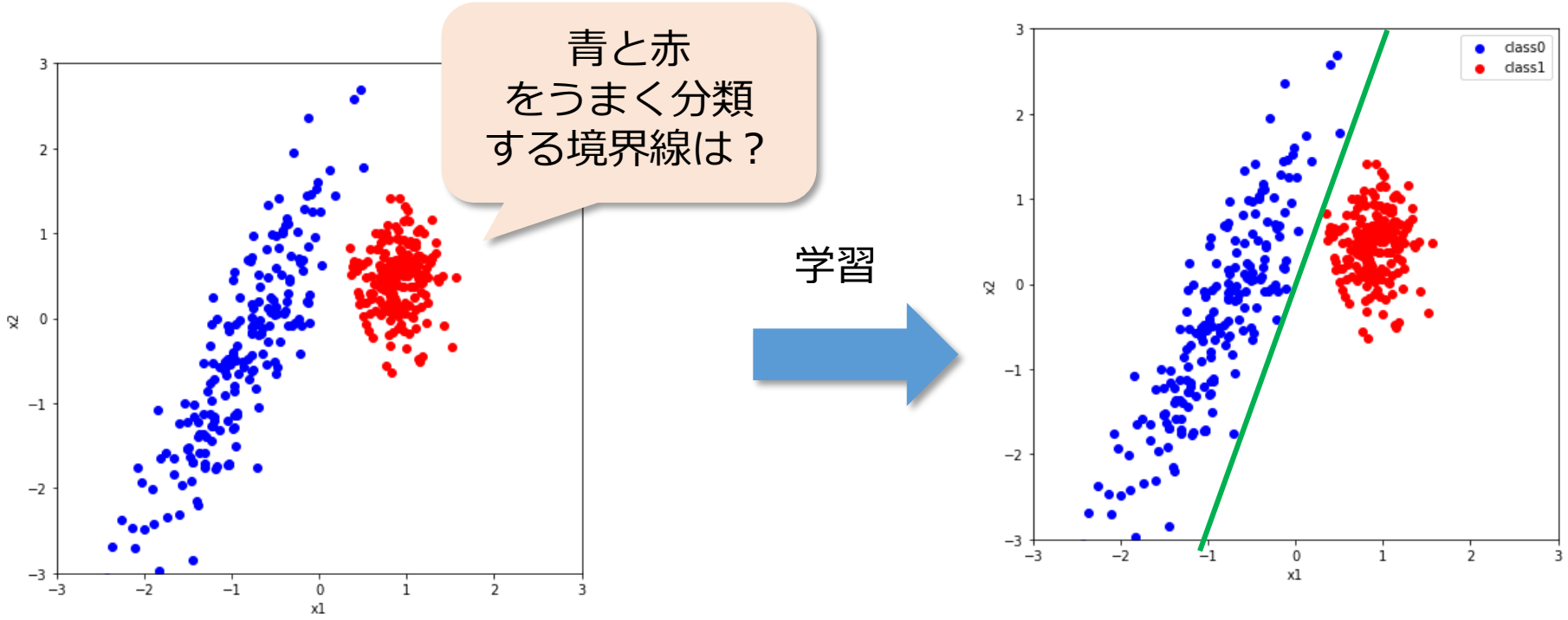
教師なしクラスタリングの方法

- K-means
- 混合正規分布

教師ありクラスタリング（おさらい）

教師とは？　→　学習データのクラスラベルのことです。

教師あり学習では，学習データをラベル通りにクラス分類する境界線を学習します。



分類の基準は？

大きく3種類あります。

- 識別関数を作成する。（第６回で説明した線形判別分析・ロジスティック回帰はこれ）
- 距離を測る。
- 類似度を測る。

} 今回使うのはこちら

距離や類似度にはたくさんの種類がありますが，ここでは

- ユークリッド距離
- 正規分布（多変量正規分布）

を紹介します。

ユークリッド距離

多次元ベクトル (D次元) で表現される，2個のサンプル

$$\boldsymbol{x} = [x_1, x_2, \ldots, x_D] \qquad y = [y_1, y_2, \ldots, y_D]$$

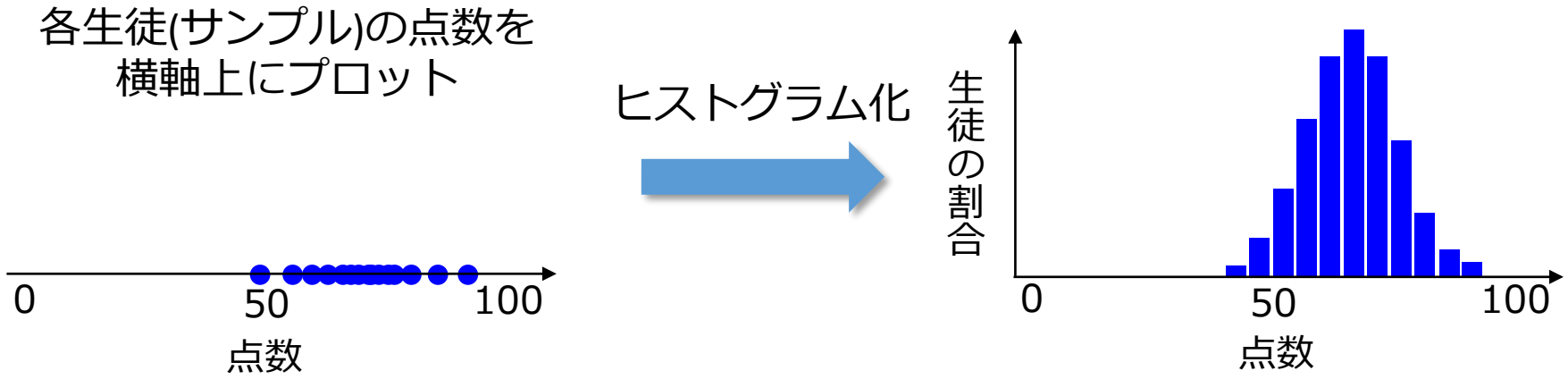
のユークリッド距離は以下のように定義されます。

$$distance(x, y) = \sqrt{\sum_{d=1}^{D} (x_d - y_d)^2}$$

距離が短いほど，そのサンプル同士は似ているということになります。

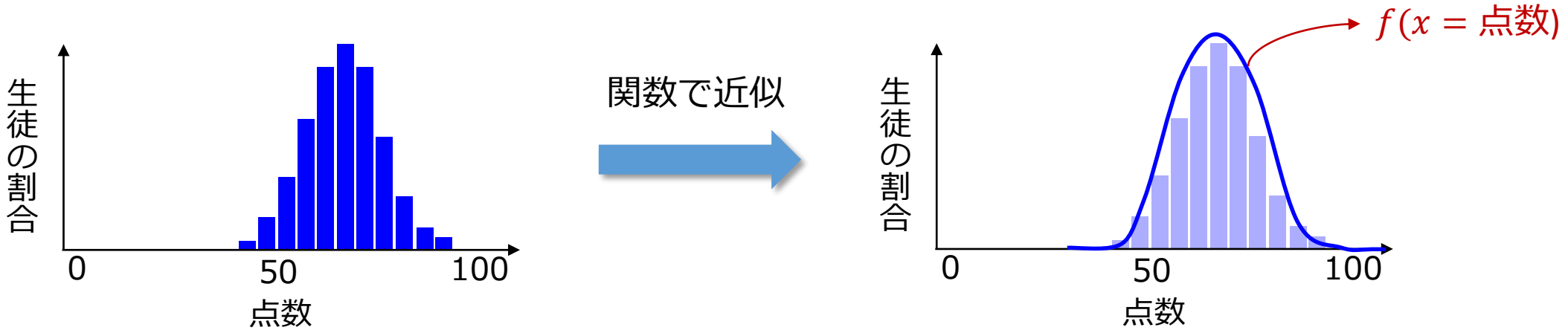
正規分布関数

あるクラスで，あるテストを実施したところ，以下のような分布になったとします。



正規分布関数

ヒストグラムを，関数で近似することを考えます。



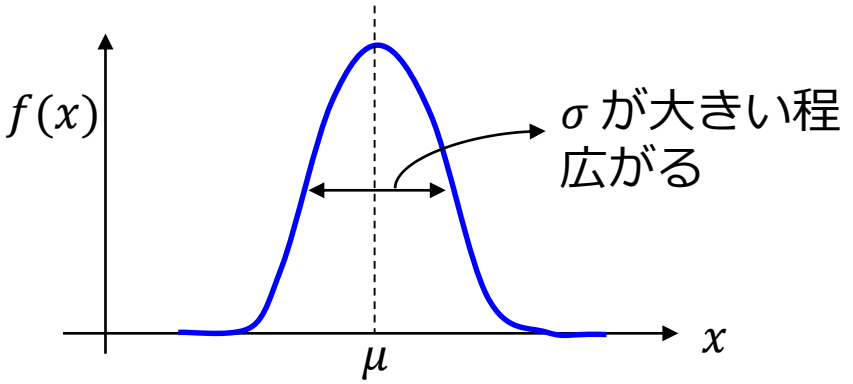
近似関数 f には様々なものが考えられますが，代表的な関数として，

正規分布関数があります。

正規分布関数

正規分布関数

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



μ: 平均値，σ: 標準偏差

正規分布関数は x が平均値 μ に近い程，高い値を取ります。

正規分布関数の値 $f(x)$ を類似度として捉えることが可能

また， σ が大きい程関数は横に広がります。

正規分布関数

前頁の正規分布関数は，スカラ値に対する定義です。

x が多次元のベクトル値の場合，多変量正規分布と呼び，以下で定義されます。

多変量正規分布

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^D|\Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

$\boldsymbol{x}$: 多次元ベクトルのデータ（次元数Dのベクトル）

$\boldsymbol{\mu}$: 平均値ベクトル (次元数 D のベクトル)

Σ: 分散共分散行列 ($D \times D$ の行列)

$|\Sigma|$: 行列式，　Σ^{-1}: 逆行列

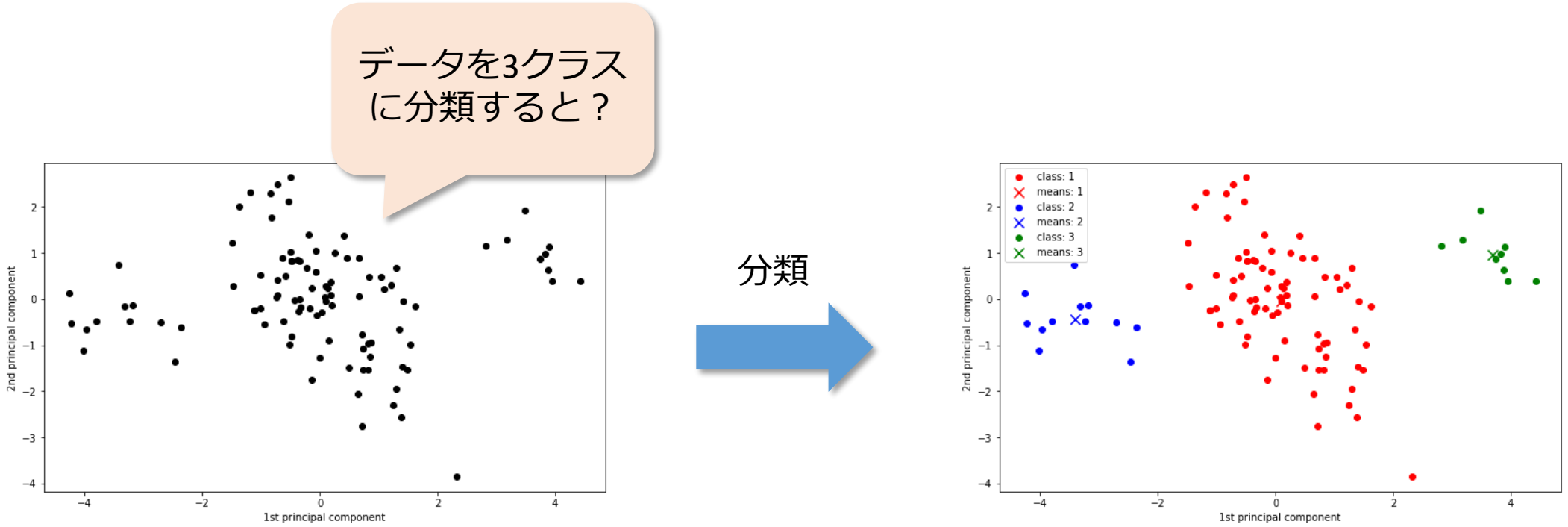
ユークリッド距離と正規分布関数の比較

08_01_similarity.ipynbを動かして，ユークリッド距離と正規分布関数を比較してみましょう。

教師なしクラスタリング手法１
K-means

教師なしクラスタリングとは

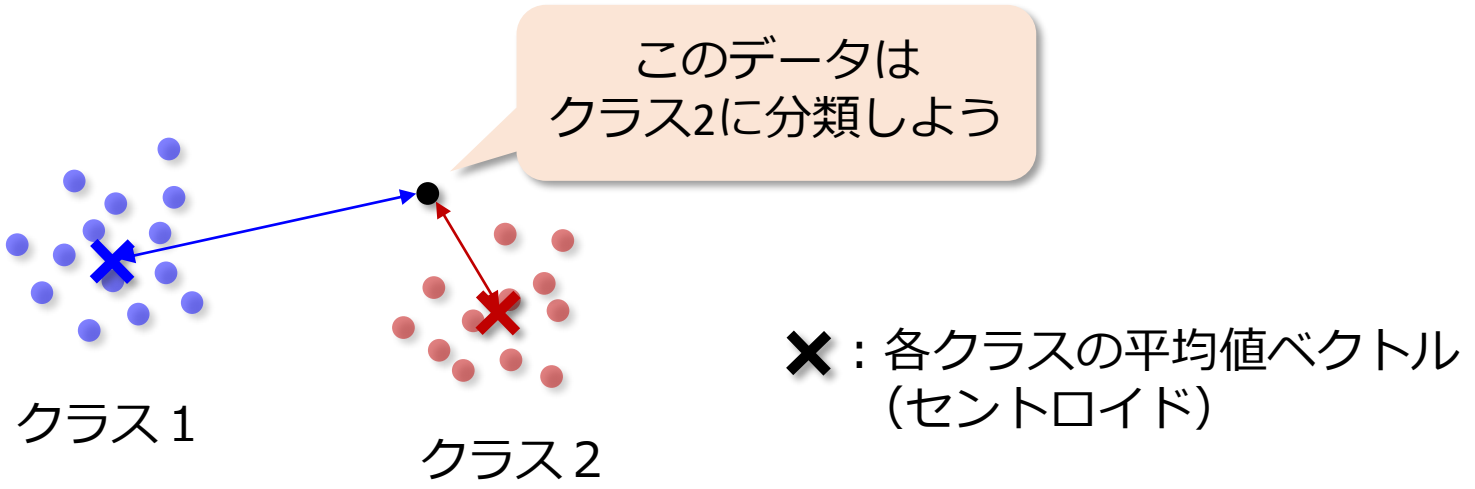
クラスラベルが無い場合のクラスタリングのこと。
この場合，データを何らかの基準で自動的に複数のクラスに分類します。



分類の基準として，先ほど説明した距離や類似度が登場します。

K-meansのアルゴリズム

各クラスの平均値ベクトル(アルゴリズム上ではセントロイドと呼ばれる)とのユークリッド距離をもとに，クラス分類を行います。



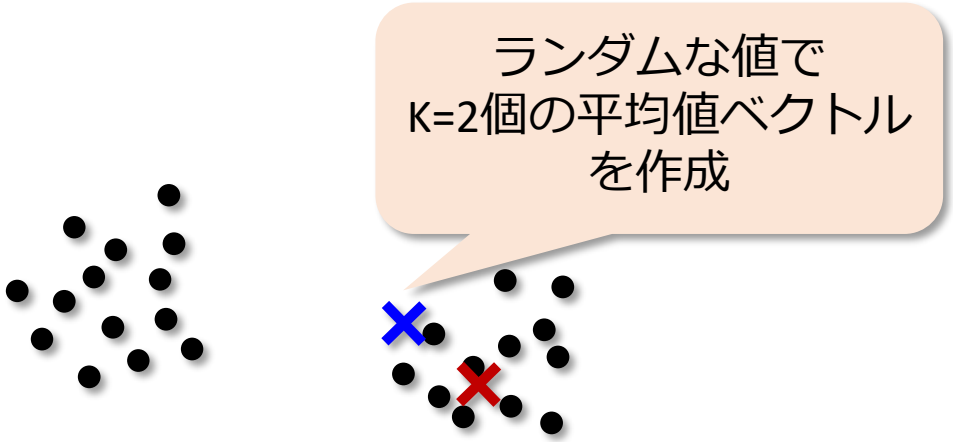
・・・クラスラベルが無い（どのデータがどのクラスか分からない）のに，どうして平均値ベクトルが計算できるのでしょうか？？

→ 正解は分からないので，平均値ベクトルを仮で作成し，徐々に修正することを考えます。

K-meansのアルゴリズム

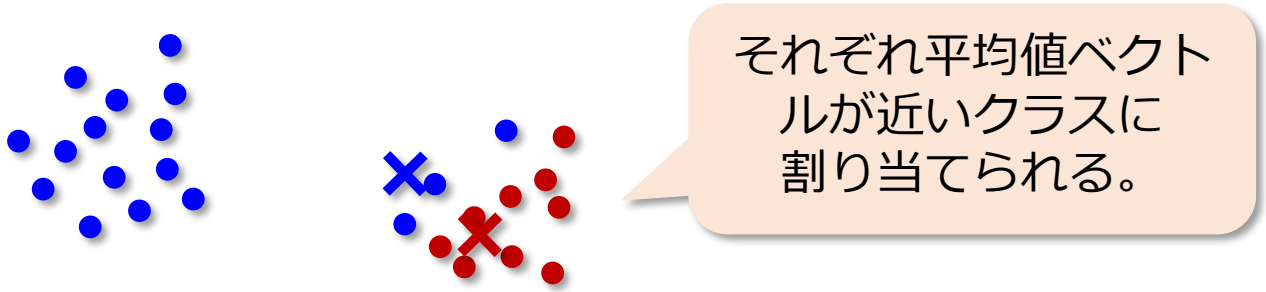
データをｋ個のクラスに分類する場合，

① 最初にｋ個の平均値ベクトルをランダムな値で作成します。



✖：各クラスの平均値ベクトル
（セントロイド）

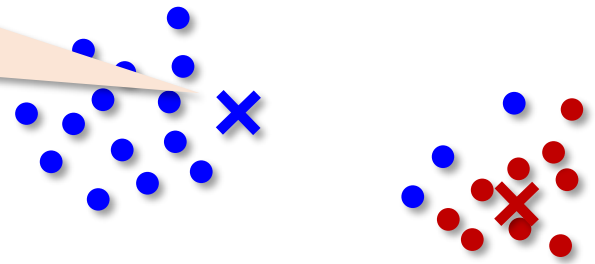
② 作成した平均値ベクトルと，各データとの距離を測り，分類を行います。



K-meansのアルゴリズム

③ 割り当てられた結果をもとに，平均値ベクトルを計算しなおします。

計算しなおした結果，
平均値ベクトルが移動

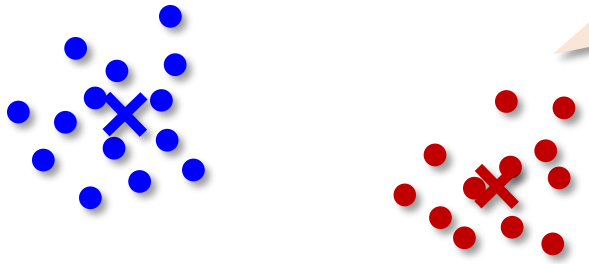


✖：各クラスの平均値ベクトル
（セントロイド）

ここではユークリッド距離の考え方が利用

前ページの②と③を，平均値ベクトルが移動しなくなるまで交互に繰り返します。

平均値ベクトルと分類が
互いに修正されながら
最適化されていく



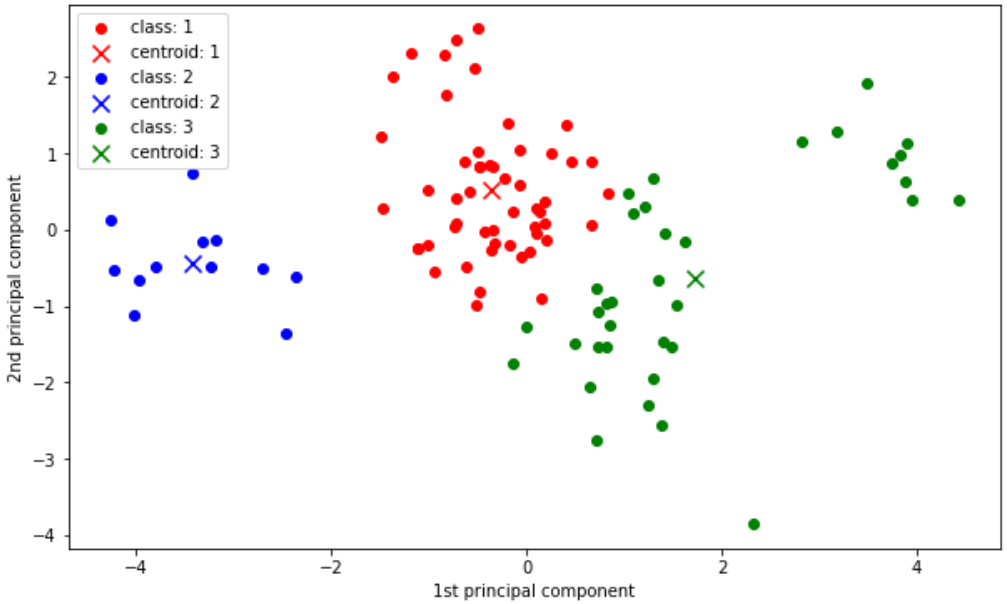
K-means を動かしてみよう

08_02_kmeans.ipynbを動かして，K-means による分類の様子を確認しましょう。

教師なしクラスタリング手法２
混合正規分布

K-means の課題点

08_02_kmeans.ipynbを K=3 として，3クラスに分類した場合，以下のようになります。



これは我々の直感とは異なる結果です。

K-means の課題点

なぜ K-means は直感と異なるクラスタリング結果をしたのでしょうか？

K-means は平均値ベクトルとのユークリッド距離しか考慮していないため，分布の広がり（＝分散共分散行列）を考慮したクラスタリングにならないためです。

では分布を考慮したクラスタリングはどのようにすれば良いでしょうか？

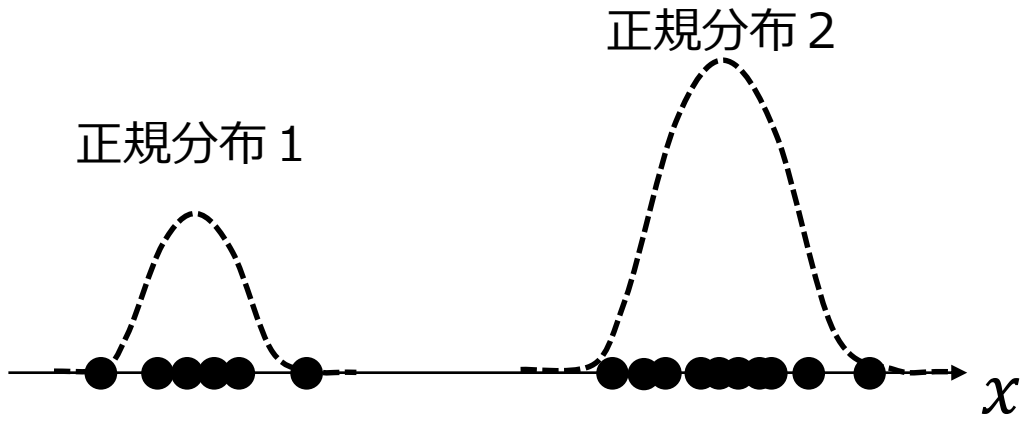
→ **混合正規分布**

混合正規分布

データ全体の分布を，複数の正規分布の足し合わせで表現する方法です。

混合正規分布には，以下の要素が含まれます。

- クラス毎(分布)の平均値ベクトル（これはK-meansにもあった）
- クラス毎の分散共分散行列
- クラス毎の重み
  下の図の場合クラス１よりクラス２の方が属しているデータ数が多いです。
  属しているデータ数に応じて，正規分布の高さに重みを加えます。
  （多くのデータはクラス２に属すのだから，大抵のデータはクラス２に属すべきだろう，という考え方。勢力の強さのようなイメージ。正式には「事前確率」と呼びます）



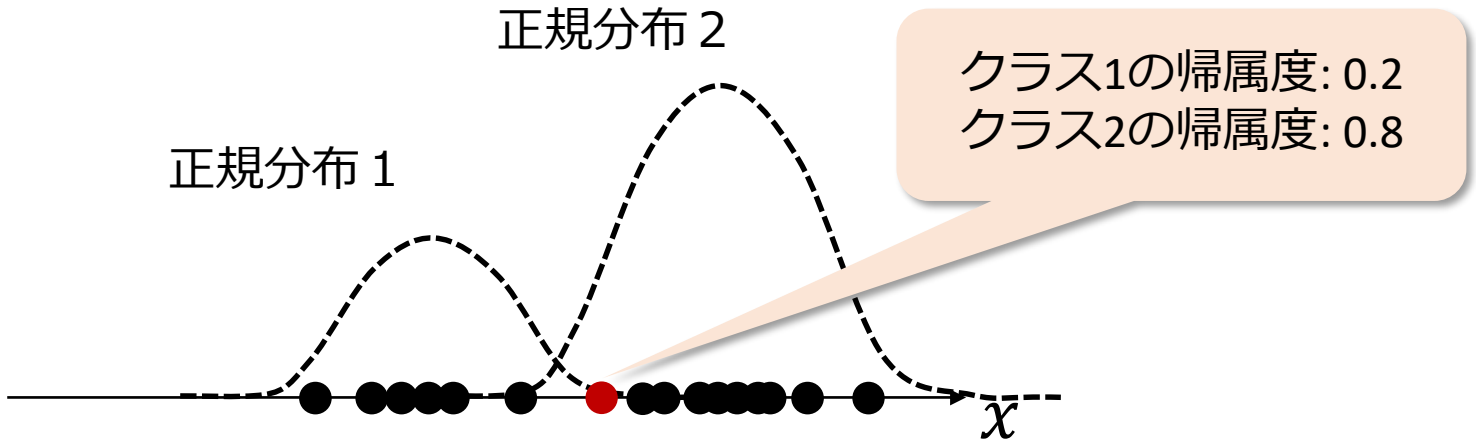
混合正規分布

K-meansでは，「各データはどれか一つのクラスだけに属する」という前提で分類を行っていましたが，混合正規分布では，

「クラス1の帰属度：0.4，クラス2の帰属度：0.6」というように，

各クラスに属する度合いを連続値として表現します。

（帰属度は全クラスで足すと 1 になります。）



混合正規分布のアルゴリズム

① K 個の平均値ベクトル μ_k ，分散共分散行列 Σ_k ，重み w_k の初期値を設定する。
（サンプルプログラムではK-meansの結果をもとに作成します。）

② 各データについて，各クラスに対する帰属度 r を求める。
n番目のデータx_n の k 番目のクラスへの帰属度 $r_{n,k}$ は以下の式で求められる。

$$s_{n,k} = w_k * f_k(x_n)$$ 式(2,1)

k番目の正規分布で計算した値 $f_k(x_n)$に，重み w_k をかけた

$f_k(x_n)$: k番目の正規分布パラメータμ_k, Σ_k で計算した正規分布関数の値

$s_{n,k}$: 重み付きの正規分布関数値

$$r_{n,k} = \frac{s_{n,k}}{\sum_{k=1}^{K} s_{n,k}}$$ 式(2,2)

$s_{n,k}$ に対して，クラス数 K で総和すると1になるように正規化

混合正規分布のアルゴリズム

③ 帰属度を用いて，各クラスの重みを計算する。

$$w_k = \frac{\sum_{n=1}^{N} r_{n,k}}{N}$$ 式(3,1)

各サンプルの k 番目のクラスへの帰属度を平均したもの

④ 帰属度を用いて，各クラスの平均値ベクトルを計算する。

$$\mu_k = \frac{\sum_{n=1}^{N} r_{n,k}\, x_n}{\sum_{n=1}^{N} r_{n,k}}$$ 式(4,1)

帰属度で重みをつけた平均値
（帰属度が全て1 の場合は
通常の平均値計算と同じ）

⑤ 帰属度を用いて，各クラスの分散共分散行列を計算する。

$$\Sigma_k = \frac{\sum_{n=1}^{N} r_{n,k}\,(x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} r_{n,k}}$$ 式(5,1)

帰属度で重みをつけた分散共分散行列
（帰属度が全て1 の場合は
通常の分散共分散行列と同じ）

②～⑤を，収束するまで繰り返す。

混合正規分布のアルゴリズム

K-means との大きな違い

- 平均値ベクトルだけでなく，分散共分散行列と分布の重みを考慮している。
- どれか一つのクラスにだけ分類するのではなく，帰属度という連続値で表現している。
  （逆に言うと，K-meansは帰属度が 0 か 1 の二値になっている。）

08_02_gmm.ipynb は python のライブラリを用いて，混合正規分布を実装しています。

きちんとしたソースコードの実装は，レポート課題です。

(混合正規分布は Gaussian Mixture Model; GMM とも呼ばれます。)

おわりに

今回は，クラスタリングの基本的な説明と，教師無しクラスタリングの方法について説明しました。

次回も教師無しクラスタリングの手法について解説する予定です。

レポート課題

第８回ファイル一式に含まれる，”report08.ipynb” をベースに，混合正規分布のアルゴリズムを実装してください。

p. 22，p.23を参考に，”report08.ipynb”の指示に従って，未実装の部分を補完してください。

ソースコードは穴埋め式なので，追記するだけで実装できますが，好みのスタイルに応じて，既に実装されている部分を消しても結構です。

レポート提出期限：6/28(水) AM10:30，ipynbファイルをhtmlファイルに変換して提出