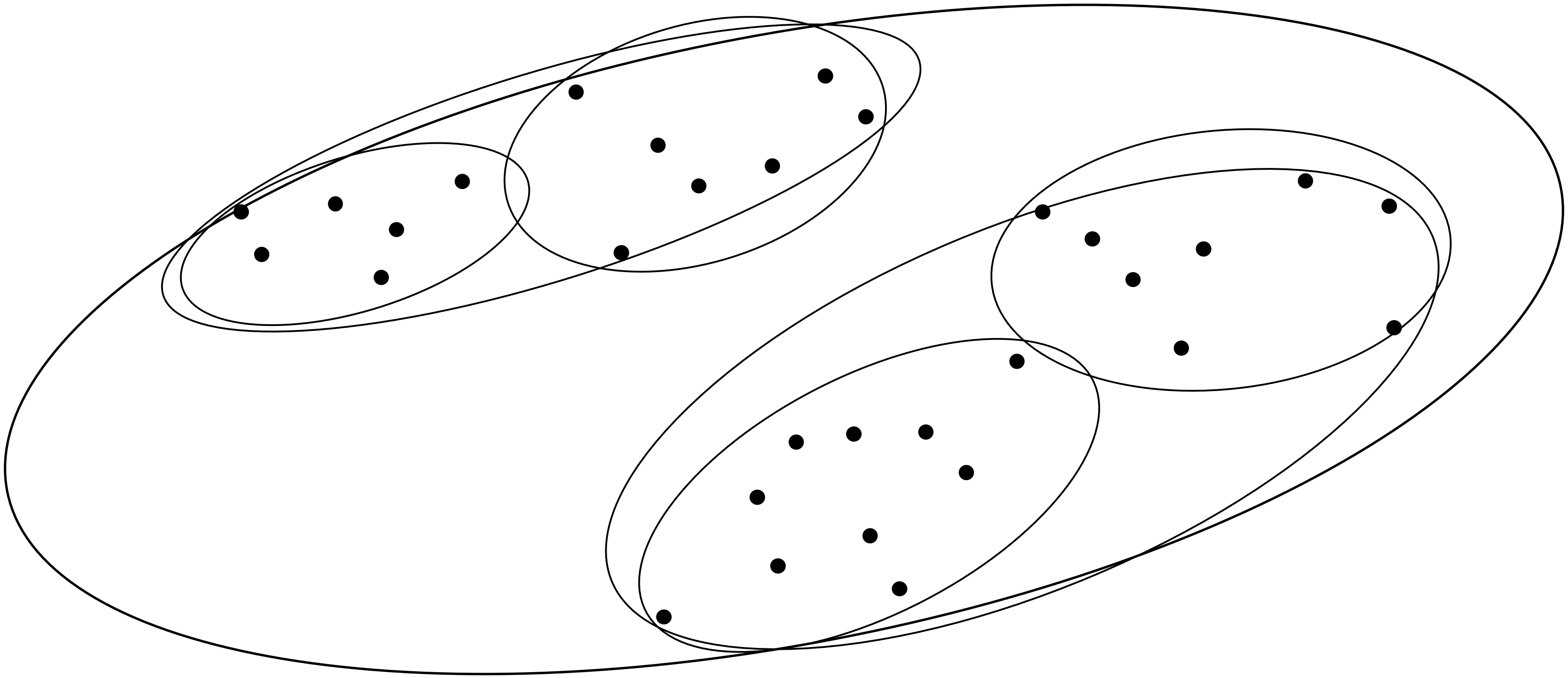
パターン認識
―クラスタリングとEMアルゴリズム―

担当:和田　俊和
部屋　A513
Email　twada@ieee.org

講義資料はhttp://wada1.sys.wakayama-u.ac.jp/PRA/

単純クラスタリング、k-meansクラスタリング、最大距離アルゴリズム、EMアルゴリズム

クラスタリング
―似たものをまとめる処理―

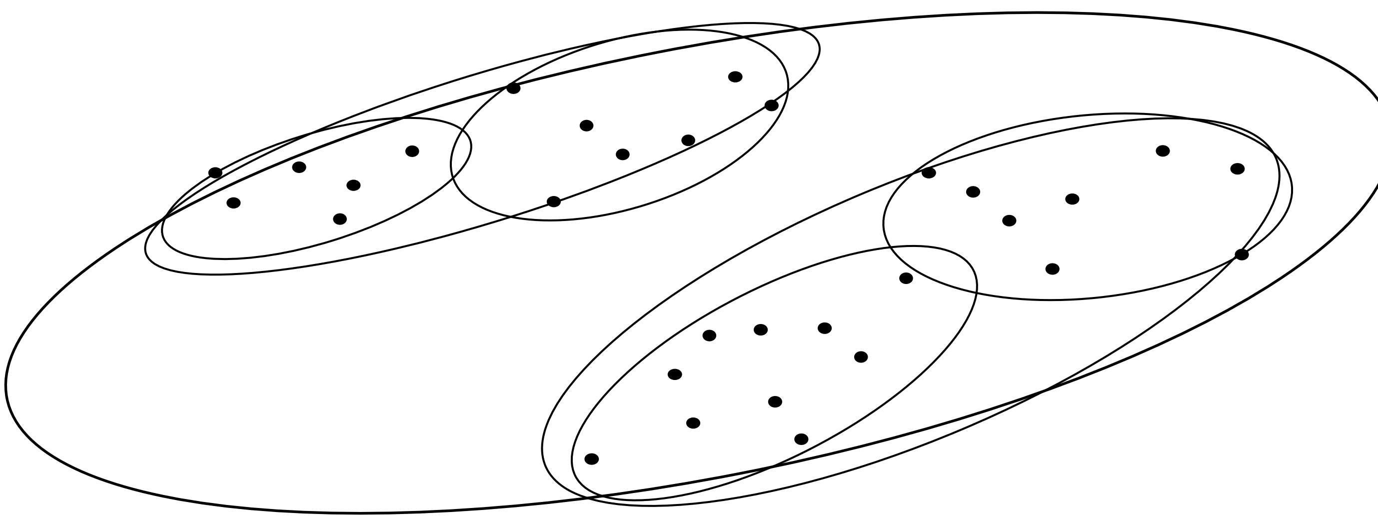


クラスタ（Cluster）=塊（かたまり）
Clustering = クラスタを作る処理

クラスタリング
＝教師なし学習

どのクラスに属するかが明示的に示されていないトレーニングデータに、データ間の類似性もしくは相違性に基づいてクラスラベルを付けていくこと。つまり、教師信号は与えられない。

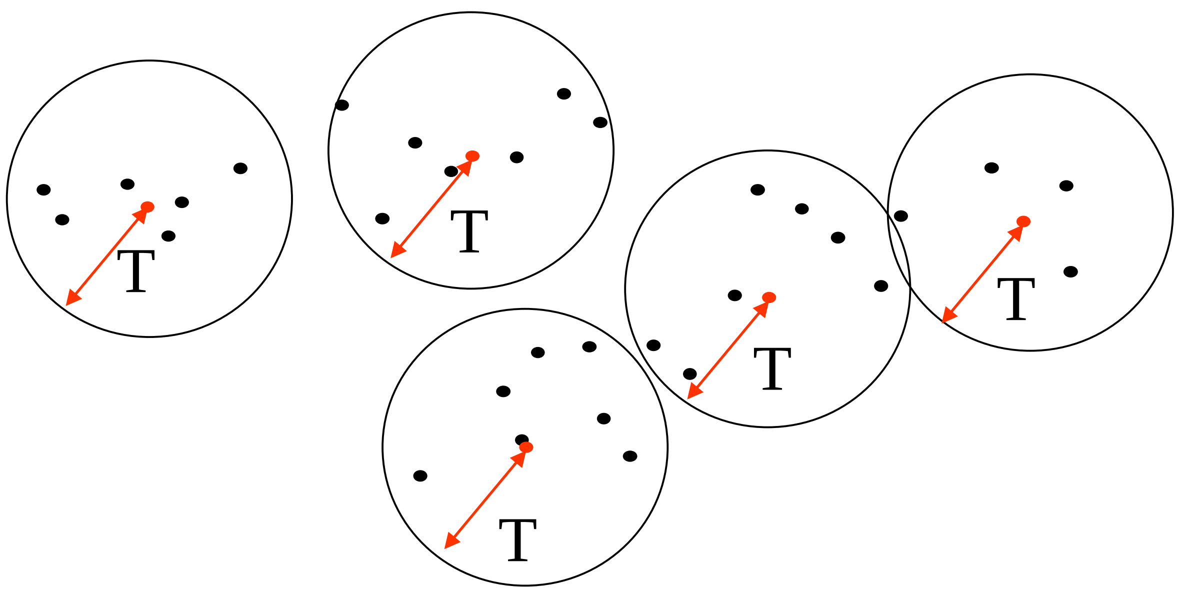
この問題を解くには、何らかの仮定を導入する必要がある。



単純クラスタリング

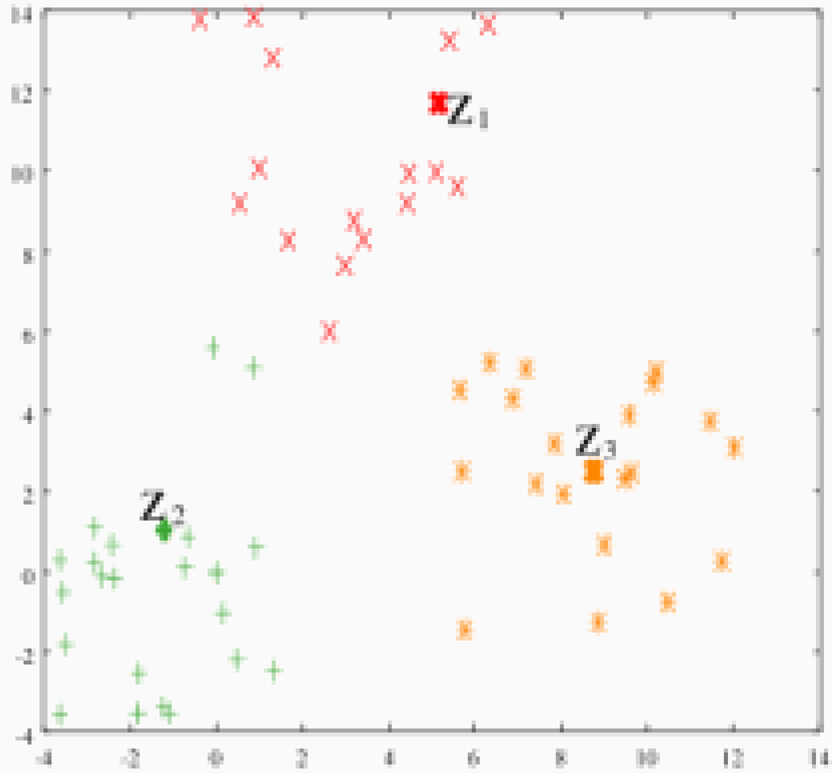
同一クラスタに属するパターン間の距離に関する制約を設ける

- 中心からの距離がT以内に存在するパターンを一つのクラスタとする。
- T以上離れている場合は、新しいクラスタ中心となる。



単純クラスタリング

同一クラスタに属するパターン間の距離に関する制約を設ける



最終的に３個のクラスタが生成された

単純クラスタリング

同一クラスタに属するパターン間の距離に関する制約を設ける

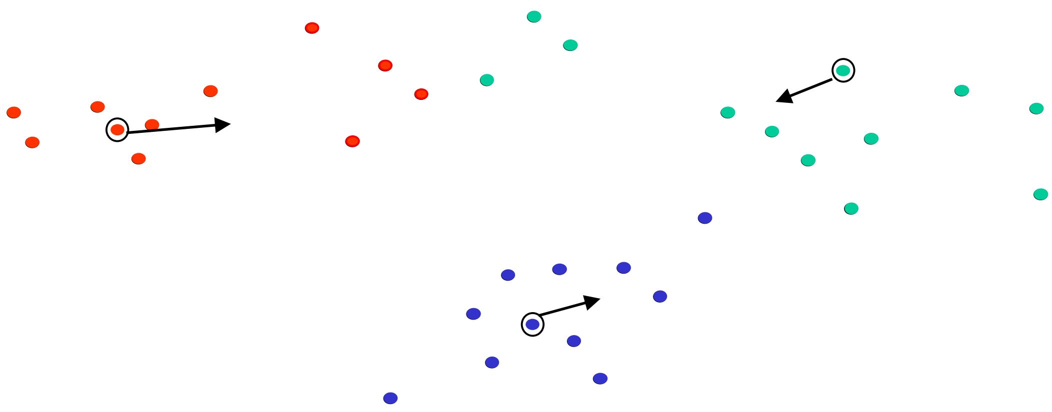
欠点:

- データを与える順序に依存した結果しか得られない
- 閾値Tを知る方法がない。

K-Meansクラスタリング

クラスタ数をあらかじめ決めておく

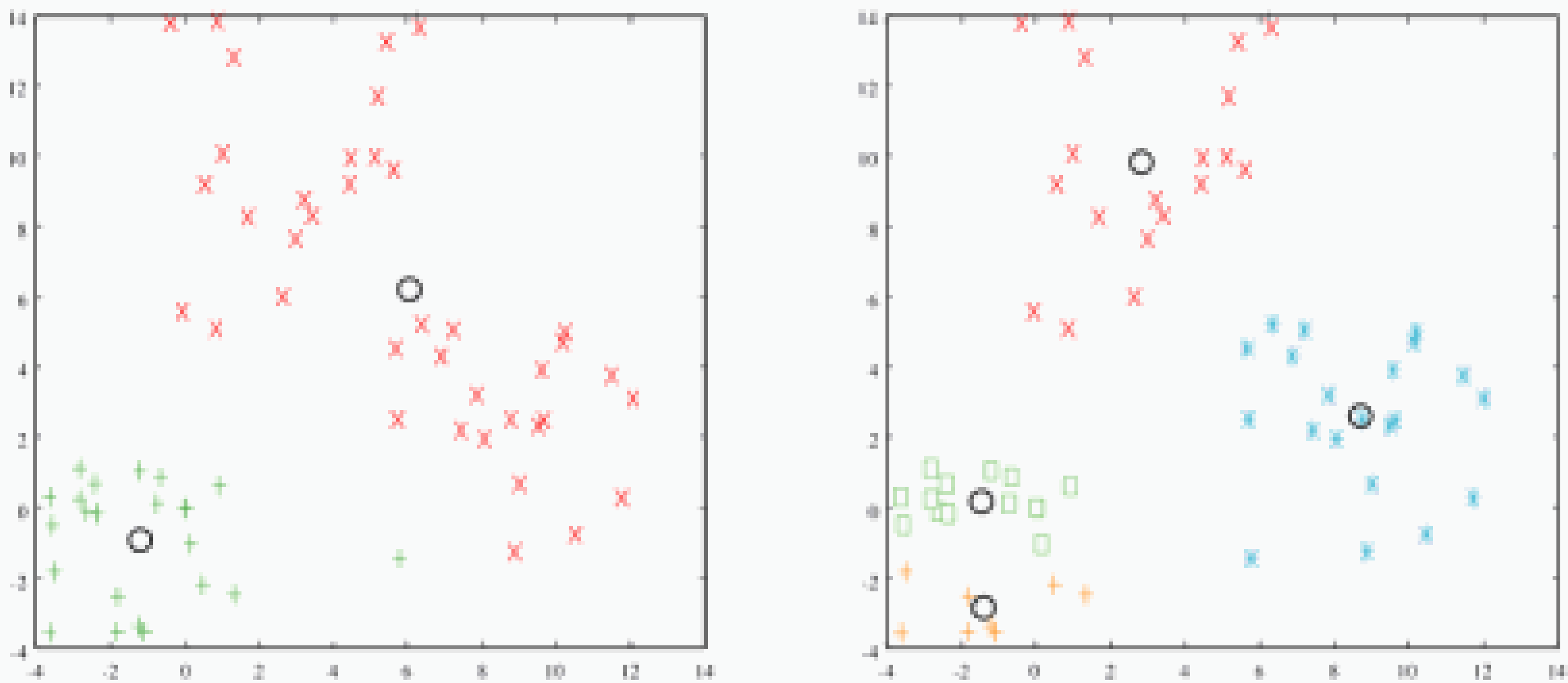
1. クラスタ中心をランダムに決めておき、
2. クラスタ中心からの距離を基にしてそのデータの帰属クラスタを決め
3. データの帰属性をもとにしてクラスタ中心を再計算する
4. クラスタ中心が移動していれば、2に戻る。



K-Meansクラスタリング

クラスタ数をあらかじめ決めておく

(クラスタの数の違いによるクラスタリング結果)



クラスタが２個の場合　　　クラスタが４個の場合

K-Meansクラスタリングデモ

K-Meansクラスタリング

クラスタ数をあらかじめ決めておく

欠点

- クラスタ数を既知としなければならない。
- 初期値に依存して結果が変わる。
- 計算が収束しない場合がある。

ISODATAアルゴリズム

K-means アルゴリズムに

- 同じクラスタに属するサンプルが閾値未満の場合、そのクラスタを作らない。
- クラスタ間距離が閾値未満の場合、それらのクラスタをまとめる
- クラスタ内の分散が大きくなりすぎるとクラスタを分割する

という条件を加えたもの。

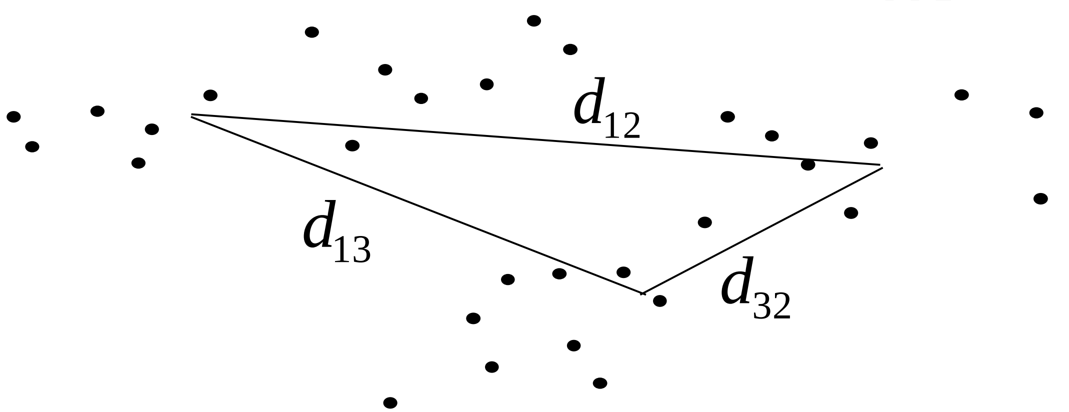
Tou, Julius T. and Rafael C. Gonzalez. 1974. Pattern Recognition Principles. Addison-Wesley Publishing Co.

最大距離アルゴリズム

最大クラスタ間距離を基準として、クラスタ間距離に関する制約を設ける

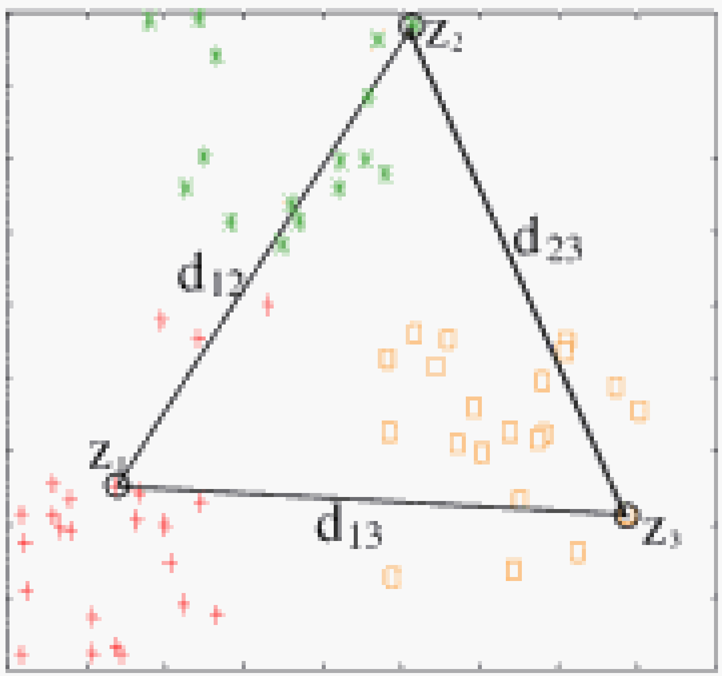
- 各クラスタ間の距離が最大クラスタ間距離のn/m以内になるようにクラスタリングを行う。

$$d_{12}\frac{n}{m} \leq d_{32}$$

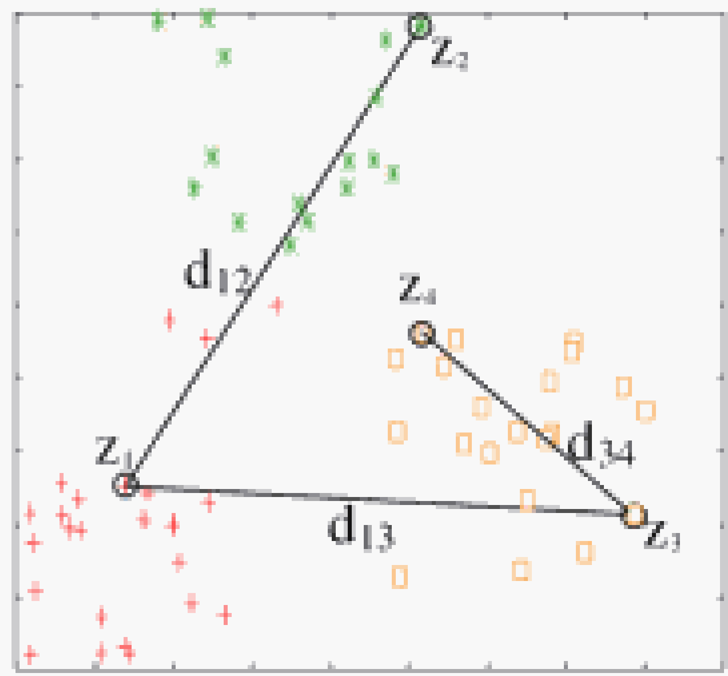


最大距離アルゴリズム

最大クラスタ間距離を基準として、クラスタ間距離に関する制約を設ける



$d_{12}\cdot\frac{n}{m} < d_{13}$ を満たすためZ_3がクラスタの中心に選ばれる



$d_{12}\cdot\frac{n}{m} < d_{34}$ を満たさないためZ_4はクラスタの中心にはならない

他のクラスタリング手法

- グラフを用いたクラスタリング（最小全域木を用いたクラスタリングなど）
- Fuzzy クラスタリング
- 階層的クラスタリング
- EMアルゴリズム
- その他

混合(確率密度)分布

サンプルが複数の分布の重み付き和に従うとき

$$p(\mathbf{x}|\omega_i) = \sum_{j=1}^{m} \xi_j p(\mathbf{x};\boldsymbol{\theta}_j)$$

サンプル$\boldsymbol{x}_k$からこのξ_jと$\boldsymbol{\theta}_j$を求めることができれば、
分布形状が決定できる。ちなみに、mは既知である。

各$\boldsymbol{x}_k$に関して、どのjの分布に従うかを決めることが
できれば、通常の最尤推定が適用できる。

不完全データ $X = \{\mathbf{x}_1, \quad \mathbf{x}_2, \quad \cdots, \quad \mathbf{x}_n\}$

$$\Downarrow$$

完全データ $Y = \{(\mathbf{x}_1, J_1), \quad (\mathbf{x}_2, J_2), \quad \cdots, \quad (\mathbf{x}_n, J_n)\}, \quad 1 \le J_i \le m$

EM アルゴリズムの概要

1. E (Expectation) ステップ : 次で定義される完全データの対数尤度の条件付き期待値 を計算する.（ここでは、$\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\cdots\boldsymbol{\theta}_m,\xi_1,\cdots,\xi_m)$ と見なす。）

$$Q(\boldsymbol{\theta}\,|\,X;\boldsymbol{\theta}^{(t)})=E_{\boldsymbol{\theta}(t)}\left[\log p(\mathbf{y};\boldsymbol{\theta})\,|\,X\right]$$

$$=\int_{\mathbf{y}=\mathbf{Y}(\mathbf{X})} p(\mathbf{y}\,|\,X;\boldsymbol{\theta}^{(t)})\log p(\mathbf{y};\boldsymbol{\theta})d\mathbf{y}$$

具体的な計算方法は後に述べる。

2. M (Maximization) ステップ $Q(\boldsymbol{\theta}\,|\,X;\boldsymbol{\theta}^{(t)})$ を $\boldsymbol{\theta}$ について 最大化したものを $\boldsymbol{\theta}^{(t+1)}$ とおき、t＝t＋１として１に戻る。

E step の詳細

1. 分布モデル

$$p(\mathbf{x};\boldsymbol{\theta})=\sum_{j=1}^{m}\xi_j\, p(\mathbf{x};\boldsymbol{\theta}_j)$$

X がJ番目の要素分布に従う確率を

$$\boxed{q^{(t)}(J\,|\,\mathbf{x})=\frac{\xi_J\, p(\mathbf{x};\boldsymbol{\theta}_J^{(t)})}{\sum_{k=1}^{m}\xi_k\, p(\mathbf{x};\boldsymbol{\theta}_k^{(t)})}}$$

とすると

これによって重み付けをした尤度の和として、

$$E_{\boldsymbol{\theta}(t)}\left[\log p(\mathbf{y};\boldsymbol{\theta})\,|\,X\right]=\sum_{i=1}^{n}\sum_{k=1}^{m}q^{(t)}(k\,|\,\mathbf{x}_i)\log\left\{\xi_k\, p(\mathbf{x}_i;\boldsymbol{\theta}_k)\right\}$$

が得られる。この式を最大化するξ_kとθ_kを求める。

M step の詳細

問題: $E_{\boldsymbol{\theta}(t)}\left[\log p(\mathbf{y};\boldsymbol{\theta}) \mid X\right]$ を $\sum_{k=1}^{m} \xi_k = 1$ という条件の下で、

最大化する $\boldsymbol{\theta}$ を求める。

$$E_{\boldsymbol{\theta}(t)}\left[\log p(\mathbf{y};\boldsymbol{\theta}) \mid X\right] = \sum_{i=1}^{n}\sum_{k=1}^{m} q^{(t)}(k \mid \mathbf{x}_i) \log\{\xi_k p(\mathbf{x}_i;\boldsymbol{\theta}_k)\}$$

に Lagrange の未定係数項を加えて式の変形をしていくと、結果的に、次式が得られる。

$$\boxed{\xi_J^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} q^{(t)}(J \mid \mathbf{x}_i)}$$

$\boldsymbol{\Theta}_k$ に関してはこの式から求める。 ⇒

$$\sum_{i=1}^{n} q^{(t)}(J \mid \mathbf{x}_i) \frac{\partial \log p(\mathbf{x}_i;\boldsymbol{\theta}_J^{(t)})}{\partial \boldsymbol{\theta}_J} = 0$$

M step の詳細：混合正規分布の場合

$$\xi_J^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} q^{(t)}(J \mid \mathbf{x}_i)$$

$$\boxed{\boldsymbol{\mu}_J^{(t+1)} = \frac{1}{n\xi_J^{(t+1)}}\sum_{i=1}^{n} q^{(t)}(J \mid \mathbf{x}_i)\mathbf{x}_i}$$

$$\boxed{\Sigma_J^{(t+1)} = \frac{1}{n\xi_J^{(t+1)}}\sum_{i=1}^{n} q^{(t)}(J \mid \mathbf{x}_i)(\mathbf{x}_i - \boldsymbol{\mu}_J^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_J^{(t+1)})^T}$$

混合正規分布のあてはめ

EM¥MixtureEMj.html