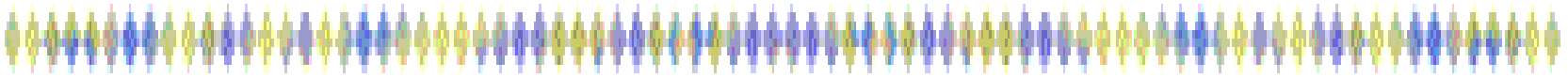
EMアルゴリズム

―クラスタリングへの適用と最近の発展―



3. 混合分布とEMアルゴリズム

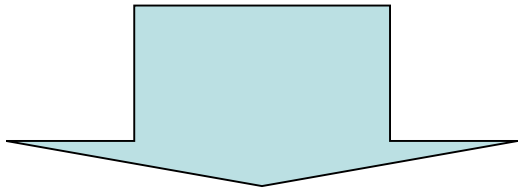
発表日 : 2007/9/5

発表者 : **阿部 竜之介**

最尤推定

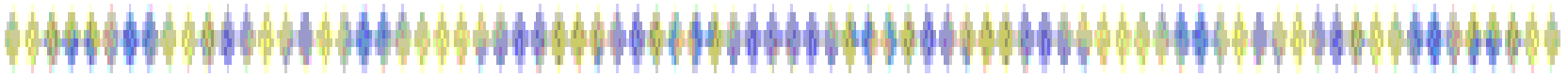
<u>最尤推定とは</u>

■得られたデータ x に対して、θというパラメータを持つ確率分布のモデル　$p(x;\theta)$ を当てはめ、その値が一番大きくなるようなθを求める



EMアルゴリズムを用いて求める

混合分布



混合分布とは

■ K 個の確率モデル $f_1(x;\xi_1),\ldots,f_K(x;\xi_K)$ があったとき、重み付きの和を

$$p(x;\theta)=\sum_{j=1}^{K}\omega_j f_j(x;\xi_j)$$

とし（θ は全てのパラメータ $\omega_1,\ldots,\omega_K,\xi_1,\ldots,\xi_K$ をまとめて表したもの）、重み ω_j が確率分布であるとすると、$p(x;\theta)$ も確率分布となり、こうして得られる分布 p を混合分布という

EMアルゴリズム　1/3

EMアルゴリズムとは

- EMアルゴリズムは不完全データ x の分布に関する最適化問題を、完全データ y の分布に関する最適化問題の繰り返し演算に帰着させる方法
- EステップとMステップの2つのステップからなる

1. Expectationステップ：以下の Q を計算する

$$Q(\theta \mid \theta^{(t)}) = E\left[\log p(y;\theta) \mid x,\theta^{(t)}\right]$$
$$= \int_{Y(x)} p(y \mid x;\theta^{(t)}) \log p(y;\theta)\, dy$$

2. Maximizationステップ： Q を最大にするような θ を求め $\theta^{(t+1)}$ とおく

$Y(x)$：y全体の集合
$\theta^{(t)}$ ：t 回目のθ

EMアルゴリズム　2/3

■実際に混合分布に適用してみると、以下のような式になる

$$Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{K} q_{ij}^{(t)} \log\{\omega_j f_j(x_i;\xi_j)\}$$

ここで、

$$q_{ij}^{(t)} = \frac{\omega_j^{(t)} f_j(x_i;\xi_j^{(t)})}{\sum_{j'}^{K} \omega_{j'}^{(t)} f_{j'}(x_i;\xi_{j'}^{(t)})}$$

とおいた

■ Q を最大にする θ を見つけるには Q の極値を調べればよい

EMアルゴリズム　3/3

■結果として、ω_j 、ξ_j については

$$\omega_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} q_{ij}^{(t)} \quad , \quad \frac{\partial}{\partial \xi_j}\sum_{i=1}^{n} q_{ij}^{(t)} \log f_j(x_i;\xi_j) = 0$$

という（ξ_j は上式右の方程式を解き）解が得られる

■ $f_j(x;\xi_j)$ が正規分布ならば、その平均 μ_j と分散 σ_j^2 は以下のような重み付き平均と重み付き分散になる

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} q_{ij}^{(t)} x_i}{\sum_{i=1}^{n} q_{ij}^{(t)}} \quad , \quad \sigma_j^{2(t+1)} = \frac{\sum_{i=1}^{n} q_{ij}^{(t)} (x_i - \mu_j^{(t+1)})^2}{\sum_{i=1}^{n} q_{ij}^{(t)}}$$

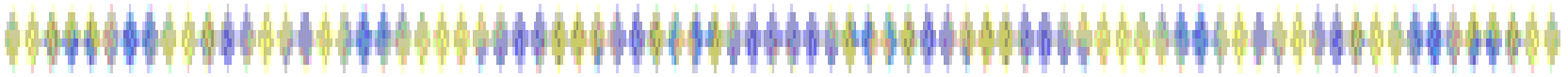
ファジィクラスタリングとの関係

■正規混合分布の重みを $\omega_j = 1/K$ 、分散を$\sigma_j^2 = \sigma^2$ として固定すると、EMアルゴリズムは式(5)(クリスプK-means法の拡張として考えられた$q_{ij}^{(t)}$ の式)をメンバーシップ関数とするファジィK-means法と全く同じ形をしていることが分かる

◆提案されてきた全てのファジィK-means法が混合分布のEMアルゴリズムで説明できるわけではない

◆分布の仮定が大きく崩れたときや次元が大きいときなどにはクリスプまたはファジィK-means法の方がよい結果を与えることがある

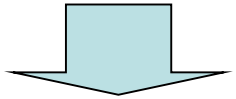
外れ値の扱い



外れ値とは

■実データを用いたクラスタリングでどのクラスタにも属さないと思われる値のこと

◆外れ値によりクラスタリングの結果に大きな影響を与えることがあるため、なんらかの方法で取り除く必要がある



●コーシー分布のような外れ値を許容するような分布を用いる

●「外れ値」の要素分布というのを作り、それを一つ加えた混合分布でモデル化する