自然言語処理のための機械学習入門

3. クラスタリング

・・・

3.4 混合正規分布によるクラスタリング
3.5 EM アルゴリズム
3.6 クラスタリングにおける問題点や注意点

新納浩幸

混合正規分布

$\boldsymbol{x}$: d 次元のデータ

クラスタ k のデータは正規分布から生成される

$$N(\boldsymbol{m}_k, \sigma^2)$$

$$\begin{aligned} P(\boldsymbol{x}) &= \sum_{i=1}^{K} P(c_i) P(\boldsymbol{x}|c_i) \\ &= \sum_{i=1}^{K} P(c_i) \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{m}_i|^2}{2\sigma^2}\right) \end{aligned}$$

混合正規分布によるクラスタリング

$\boldsymbol{x}$ を $\arg\max_c P(c|\boldsymbol{x})$ に属させればよい

パラメータは・・・

観測データから
これらを推定する

$P(c_1), P(c_2), \cdots, P(c_K)$

$\boldsymbol{m}_1, \boldsymbol{m}_2, \cdots, \boldsymbol{m}_K$

σ^2

最尤法 → EMアルゴリズム

最尤法

$P(\boldsymbol{x};\boldsymbol{\theta})$　：パラメトリックな生成モデル

$D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$　：観測データ

$\prod_{i=1}^{N} P(\boldsymbol{x}_i;\boldsymbol{\theta})$　を最大にする θ を求める

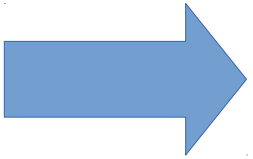
尤度

$\sum_{i=1}^{N} \log P(\boldsymbol{x}_i;\boldsymbol{\theta})$　を最大にする θ を求める

対数尤度

EM 法

混合正規分布の対数尤度を最大化するパラメータを直接求めるのは困難



EM 法

隠れ変数 c を 1 つ導入する，y = (x, c)

x の分布の最適化問題と y の分布の最適化問題を交互に解くことで、パラメータを求める方法

EM 法

y=(x,c) ： 完全データ、 q(y)： y の分布

$\boldsymbol{\theta}^{(t)}$: t 回目の繰り返しで得られたパラメータ

E-step

$$Q = E\left(\log q(y|\boldsymbol{x}, \boldsymbol{\theta}^{(t)})\right)$$

M-step

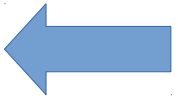
Q を最大にする θ を求める ⟶ $\boldsymbol{\theta}^{(t+1)}$

E-step と M-step を収束するまで繰り返す

混合正規分布の EM 法

隠れ変数 c を x のクラスに設定

パラメータが分かれば、クラスは分かる
クラスが分かれば、パラメータも分かる



この関係が本質的

ここから混合正規分布のパラメータを求めるのは、
かなり大変、、、省略、、、私の書籍参照

混合正規分布の推定パラメータ

$$\alpha_c^{(t)} = \hat{P}(c)$$

$$f_c^{(t)}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi\hat{\sigma^2})^d}} \exp\left(-\frac{|\boldsymbol{x} - \boldsymbol{m}_c|^2}{2\hat{\sigma^2}}\right)$$

$$g_{ic}^{(t)} = \frac{\alpha_c^{(t)} f_c^{(t)}(\boldsymbol{x}_i)}{\sum_{k=1}^{K} \alpha_k^{(t)} f_k^{(t)}(\boldsymbol{x}_i)}$$

混合正規分布の推定パラメータ

$$\alpha_c = \frac{1}{N}\sum_{j=1}^{N} g_{ic}^{(t)}$$

$$\boldsymbol{m}_c^{(t+1)} = \frac{\sum_{i=1}^{N} g_{ic}^{(t)} \boldsymbol{x}_i}{\sum_{i=1}^{N} g_{ic}^{(t)}}$$

$$\sigma^{2(t+1)} = \frac{1}{dN}\sum_{i=1}^{N}\sum_{c=1}^{K} g_{ic}^{(t)} \sum_{j=1}^{d}(x_j^i - m_{cj}^{(t+1)})^2$$

クラスタリングの問題点・注意点

・クラスタ数

MDL の利用など

・初期値

K-means や EM アルゴリズム の結果は初期値に依存

・計算時間、アンダーフロー

・評価