# クラスタリング (凸クラスタリングの導出)

橋本 玄基

2016/05/20

## 1 概要

クラスタリングは教師なし学習の一つであり，データ群を外的要因無く，アルゴリズムに従って分類する手法である．近年流行している深層学習は，教師付き学習であり．ラベルの貼られた大量のデータを学習して未来のデータを予測するために用いられる．対して教師なし学習は，正しいラベルが不明であるデータに用いられ，妥当であろうラベル分けを行いデータ解析の補助のために用いられる．例として，顧客のグループ化による広告メールの差別化等があげられる．橋本，清野らによるプログラミング教育のためのログデータ解析についてもログデータはラベルの無いデータであるため教師なし学習であるクラスタリングを用いて解析を行っている．

## 2 種類

### 2.1 階層クラスタリング

階層クラスタリングは性質が近いデータをまとめていき，小さな分類を重ね大きな分類を作成する手法である．最短距離法，ウォード法などがあげられる．どの時点の階層を見るかで任意のクラスタに分類でき，樹形図等からデータの関係を非常に細かく見ることができるが，計算量が膨大かつモデルが複雑であるため，大量データの解析には向いていない．

### 2.2 非階層クラスタリング

非階層クラスタリングは分布や境界値などを更新することで，分類をする手法である．k-means 法，混合正規分布モデル，凸クラスタリング，ノンパラメトリックベイズモデルなどがあげられる．

## 3 k-means 法 (k 平均法)

### 3.1 概要

k-means 法は，予めクラスタ数を決定した上で，クラスタリングを行う．2 つの手順，クラスタ毎の中心の更新と最も近いクラスタ中心に所属クラスを更新を 2 つの手順を繰り返してクラスタを確定する．k-means 法はアルゴリズムが単純であり，高速である．

### 3.2 アルゴリズム

クラスタ数が $k$, それぞれのクラスタに属するデータ集合を $C_1, C_2, \cdots, C_k$, クラスタ中心を $\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_k$ とする．このとき観測データ集合 $X = \{\boldsymbol{x_i} \mid i = 1, 2, \cdots, n\}$ であるとするとアルゴリズムは以下のようになる．

1. $X$ の中からデータをランダムに $k$ 個選びそれぞれを $\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_k$ とする．
2. 全データ $\boldsymbol{x}_i (i = 1, 2, \cdots, n)$ を最も近いクラスタ中心に対応するクラスタ集合に所属させる．
3. 各クラスタ中心を式 1 の通りに更新する．

$$\boldsymbol{c}_i = \frac{1}{|C_i|} \sum_{\boldsymbol{x} \in C_i}^{j} \boldsymbol{x} \tag{1}$$

4. クラスタ集合に変化がなくなるまで手順 (2)，(3) を繰り返す．

### 3.3 k-means++

3.2 節より分かる通り，k-means 法は初期値にランダムな値が用いられるため，収束する結果も必ず良くなるとは限らない．また，同じ結果になった際も初期値によては繰り返し回数がとても多くなってしまう．この問題をいくらか改善するために k-means++ 法が提案された．アルゴリズムは以下のとおりである．

1. クラスタ中心 $\boldsymbol{c}_1$ を $X$ の中からランダムに 1 つ選ぶ．
2. $D(\boldsymbol{x})(\boldsymbol{x} \in X)m$ を求める．$D(\boldsymbol{x})$ は $\boldsymbol{x}$ と決定された最も近いクラスタ中心の最短距離を表す．
3. 式 2 の確率分布に従って次のクラスタ中心 $\boldsymbol{c}_i$ を決定する．

$$\frac{D(\boldsymbol{x})^2}{\sum_{\boldsymbol{x} \in X} D(\boldsymbol{x})^2} \tag{2}$$

4. k-means 法の手順 (2) から (4) を行う．

### 3.4 k-means 法の弱点

k-means 法はクラスタ数が既知でなければ使用できない．また大きな弱点として，クラスタ毎のデータ数はほぼ等しくなるため，分布数が大きく異なるクラスを正しく分類することはできない．

## 4 混合正規分布モデル (GMM)

### 4.1 概要

混合正規分布モデルは, あるデータ群はいくつかの正規分布によって発生したと考え, それぞれの正規分布のパラメータ (平均と分散) を求める数理モデルである. 各正規分布をクラスタとみなし，クラスタリングの手段として用いられることになる. 混合正規分布モデルによるクラスタリングの特徴として,1 つのクラスタに属するのてはなく, どのクラスタであるかの確率が出力されるソフトなクラスタリングである点があげられる.

### 4.2 アルゴリズム

クラスタ数が $c$ のとき，クラスタを $\omega_i$，事前確率を $\pi_i$，正規分布の平均を $\boldsymbol{\mu}_i$，分散（2 次元以上の場合は分散共分散行列）を $\boldsymbol{\Sigma}_i$ とおく $(i = 1, 2, \cdots, c)$．また，$\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)^t$，$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_c, \pi_1, \cdots, \pi_c)^t$ と定義する．そして，$n$ 個のデータを $\boldsymbol{x}_k (k = 1, 2, \cdots, n)$ とおいた時，アルゴリズムは以下のようになる．なお，$P(x)$ は確率関数，$p(x)$ は確率密度関数である．

1. $\pi_i, \boldsymbol{\theta}_i$ の初期値を与える．
2. 各 $\boldsymbol{x}_k$ に対して $P(\omega_i|\boldsymbol{x}_k; \boldsymbol{\theta})$ を求める．

$$P(\omega_i|\boldsymbol{x}_k; \boldsymbol{\theta}) = \frac{\pi_i \cdot p(\boldsymbol{x}_k|\omega_i; \boldsymbol{\theta}_i)}{\sum_{j=1}^{c} \pi_j \cdot p(\boldsymbol{x}_k|\omega_j; \boldsymbol{\theta}_j)} \tag{3}$$

3. 次式によって $\hat{\pi}_i$,$\hat{\boldsymbol{\mu}}_i$,$\hat{\boldsymbol{\Sigma}}_i$ を求めた後，$\pi_i = \hat{\pi}_i$,$\boldsymbol{\mu}_i = \hat{\boldsymbol{\mu}}_i$,$\boldsymbol{\Sigma}_i = \hat{\boldsymbol{\Sigma}}_i$ としてパラメータを更新する．

$$\hat{\pi}_i = \frac{1}{n} \sum_{k=1}^{n} P(\omega_i|\boldsymbol{x}_k; \boldsymbol{\theta}) \tag{4}$$

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{k=1}^{n} P(\omega_i|\boldsymbol{x}_k; \boldsymbol{\theta}) \boldsymbol{x}_k}{\sum_{k=1}^{n} P(\omega_i|\boldsymbol{x}_k; \boldsymbol{\theta})} \tag{5}$$

$$\hat{\boldsymbol{\Sigma}}_i = \frac{\sum_{k=1}^{n} P(\omega_i|\boldsymbol{x}_k; \boldsymbol{\theta})(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_i)^t}{\sum_{k=1}^{n} P(\omega_i|\boldsymbol{x}_k; \boldsymbol{\theta})} \tag{6}$$

4. 対数尤度 $\log p(\boldsymbol{x}; \boldsymbol{\theta})$ を求め増分が定めた閾値以下なら終了し，そうでなければ (2) に戻る．

### 4.3 混合正規分布モデルの弱点

k-means 法と同様に，クラスタ数が既知である必要がある．しかし，情報量基準を用いてモデル選択を行うことが可能である．しかしそのためには各クラス多数で計算を繰り返すため，計算量がとても多くなる．

## 5 凸クラスタリング

### 5.1 概要

前述の 2 つの手法は，共にクラスタ数が既知である必要があるとともに，結果が初期値に依存するため，必ず大域的最適解が求まる保証が無い．これを解決するために考えられたのが凸クラスタリングである．凸クラスタリングは混合正規分布モデルに正規分布の分散共分散行列は一定であり，クラスタ間で共通であることと，各クラスタのプロトタイプ (クラスタの中心) はパターン集合の中から選ぶことの 2 つの条件を加えたものである．

## 5.2 導出

$d$ 次元空間上に $n$ 個のデータ $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ があり $c$ 個のクラスタ $\omega_1, \omega_2, \cdots, \omega_c$ に分類されるとする．凸クラスタリングでは $c$ は未知である．このデータ集合に対する尤度 $p(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n; \boldsymbol{\theta})$ は各々の尤度を取り，以下のとおりである．

$$p(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n; \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k; \boldsymbol{\theta}) \tag{7}$$

$$p(\mathbf{x}_k; \boldsymbol{\theta}) = \sum_{i=1}^{c} \pi_i \cdot p(\mathbf{x}_k | \omega_i; \boldsymbol{\theta}_i) \tag{8}$$

$$\sum_{i=1}^{c} \pi_i = 1 \tag{9}$$

$\pi_i$ は $\omega_i$ の混合比 (事前確率)，$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_c, \pi_1, \pi_2, \cdots, \pi_c)^t$，$\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)^t$ である．

条件の 1 つめから凸クラスタリングでは分散共分散行列は既知であるとされ推定は行われない．そこで，$\boldsymbol{\Sigma_i}$ として以下の対角行列を考える．$\boldsymbol{I_d}$ は $d$ 次元 の単位行列とする．

$$\boldsymbol{\Sigma_i} = \sigma^2 \boldsymbol{I_d} \tag{10}$$

すると，パラメータ $\boldsymbol{\theta}_i$ は $\boldsymbol{\mu}_i$ のみとなるため式 7，8 から $p(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n; \boldsymbol{\theta})$ は

$$p(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n; \boldsymbol{\theta}) = \prod_{k=1}^{n} \left( \sum_{i=1}^{c} \pi_i \cdot p(\mathbf{x}_k | \omega_i; \boldsymbol{\mu}_i) \right) \tag{11}$$

と書け，対数尤度 $\log p(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n; \boldsymbol{\theta})$ は以下のようになる．

$$\log p(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n; \boldsymbol{\theta}) = \sum_{k=1}^{n} \log \left( \sum_{i=1}^{c} \pi_i \cdot p(\mathbf{x}_k | \omega_i; \boldsymbol{\mu}_i) \right) \tag{12}$$

条件の 2 つめから，プロトタイプはデータ集合の中から選ばれる．混合正規分布におけるプロトタイプは $\boldsymbol{\mu}_i$ なので

$$\boldsymbol{\mu}_i \in \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \tag{13}$$

である．混合正規分布モデルにそって式 12 を最大にする $c$ と $c$ 個の$\pi_i$と $\boldsymbol{\mu}_i$ を式 13 の条件で求める．$c$ の値を固定し，それぞれに対し，パラメータの推定を行う方法が単純であるが，計算量が膨大になるため，次の方法を用いる．

最初に $c = n$ とおき，各データをプロトタイプとし，パラメータ推定をすると不要なクラスタの混合比 $\pi_i \simeq 0$ となると予想できるため，0 でない $\pi_i$ を持つクラスタの数が最終的な $c$ の値とする．このとき

$$\boldsymbol{\mu}_i = \mathbf{x}_i (i = 1, 2, \cdots, n) \tag{14}$$

となり，式 9 は

$$\sum_{i=1}^{n} \pi_i = 1 \tag{15}$$

と書き換えられる．正規分布 $p(\mathbf{x}_k|\omega_i;\boldsymbol{\mu}_i)$ は式 14 より $p(\mathbf{x}_k|\omega_i;\mathbf{x}_i)$ と表される．これを $f_{ik}$ と定義する．

$$\begin{aligned} f_{ik} &\stackrel{\text{def}}{=} p(\mathbf{x}_k|\omega_i;\mathbf{x}_i) \\ &= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_k-\mathbf{x}_i)^t\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_k-\mathbf{x}_i)\right] \\ &= \frac{1}{(2\pi_i\sigma^2)^{d/2}} \exp\left[-\frac{1}{2\sigma^2}||\mathbf{x}_k-\mathbf{x}_i||^2\right] \end{aligned} \tag{16}$$

式 16 は，以下から求められる．

$$\begin{aligned} |\boldsymbol{\Sigma}_i|^{1/2} &= \{(\sigma^2)^d\}^{1/2} \\ &= (\sigma^2)^{d/2} \end{aligned} \tag{17}$$

$$\begin{aligned} &(\mathbf{x}_k-\mathbf{x}_i)^t\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_k-\mathbf{x}_i) \\ &= \begin{bmatrix}\mathrm{x}_{k1}-\mathrm{x}_{i1} & \cdots & \mathrm{x}_{kd}-\mathrm{x}_{id}\end{bmatrix} \begin{bmatrix}\sigma^{-2} & & \\ & \ddots & \\ & & \sigma^{-2}\end{bmatrix} \begin{bmatrix}\mathrm{x}_{k1}-\mathrm{x}_{i1} \\ \vdots \\ \mathrm{x}_{kd}-\mathrm{x}_{id}\end{bmatrix} \\ &= \frac{1}{\sigma^2}\sum_{j=1}^{d} \mathrm{x}_{kj}-\mathrm{x}_{ij} \\ &= \frac{1}{\sigma^2}\sqrt{\sum_{j=1}^{d} \mathrm{x}_{kj}-\mathrm{x}_{ij}}^2 \\ &= \frac{1}{\sigma^2}||\mathbf{x}_k-\mathbf{x}_i||^2 \end{aligned} \tag{18}$$

$f_{ik}$ はデータが与えられた時点で定数のみの式となるため $f_{ik}$ 自体を定数とみなせる．以上のことから式 12 は下式の通りになる．

$$\begin{aligned} \log p(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n;\boldsymbol{\theta}) &= \sum_{k=1}^{n}\log\left(\sum_{i=1}^{c}\pi_i\cdot p(\mathbf{x}_k|\omega_i;\boldsymbol{\mu}_i)\right) \\ &= \sum_{k=1}^{n}\log\left(\sum_{i=1}^{c}\pi_i\cdot p(\mathbf{x}_k|\omega_i;\mathbf{x}_i)\right) \\ &= \sum_{k=1}^{n}\log\left(\sum_{i=1}^{c}\pi_i\cdot f_{ik}\right) \end{aligned} \tag{19}$$

凸クラスタリングは混合正規分布モデルの特別な場合なのでアルゴリズム自体は 4.2 と同じであるが，今までの条件から $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ は定数となっているため，更新するパラメータは $\pi_i$ のみである．よってアルゴリズムのうち用いる式は式 4，3 のみであり，式 5，6 は必要ない．そのため凸クラスタリングの更新式は 4 に 3 を代入

し，以下のとおりとなる．

$$\begin{aligned}
\pi_i &\leftarrow \frac{1}{n}\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k) \\
&= \frac{1}{n}\sum_{k=1}^{n} \frac{\pi_i \cdot p(\mathbf{x}_k|\omega_i;\boldsymbol{\theta}_i)}{\sum_{j=1}^{c} \pi_j \cdot p(\mathbf{x}_k|\omega_j;\boldsymbol{\theta}_j)} \\
&= \frac{1}{n}\sum_{k=1}^{n} \frac{\pi_i \cdot p(\mathbf{x}_k|\omega_i;\mathbf{x}_i)}{\sum_{j=1}^{c} \pi_j \cdot p(\mathbf{x}_k|\omega_j;\mathbf{x}_j)} \\
&= \frac{1}{n}\sum_{k=1}^{n} \frac{\pi_i \cdot f_{ik}}{\sum_{j=1}^{c} \pi_j \cdot f_j k}
\end{aligned} \tag{20}$$

式 9 のもとで式 19 を最大化する問題は (証明の難易度が高いため省略するが)，凸計画問題であるため，凸クラスタリングは一回の志向で大域的最適解を求めることができる．

## 5.3 アルゴリズム

アルゴリズムは以下のようになる．記号の定義は前節と同様である．

1. 分散 $\sigma^2$ を設定する．
2. 与えられたデータ $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ を用いて以下の式を求める

$$\begin{aligned}
f_{ik} &\stackrel{\text{def}}{=} p(\mathbf{x}_k|\omega_i;\mathbf{x}_i) \\
&= \frac{1}{(2\pi_i\sigma^2)^{d/2}} \exp\left[-\frac{1}{2\sigma^2}||\mathbf{x}_k - \mathbf{x}_i||^2\right]
\end{aligned}$$

3. 事前確率である混合比 $\pi_1, \pi_2, \cdots, \pi_n$ の初期値を設定する．
4. $\pi_i$ を以下の式で更新する．

$$\pi_i \leftarrow \frac{1}{n}\sum_{k=1}^{n} \frac{\pi_i \cdot f_{ik}}{\sum_{j=1}^{n} \pi_i \cdot f_{ik}} \tag{21}$$

5. 以下の対数尤度の増分が閾値以下なら終了し，そうでなければ 3 を繰り返す．

$$\log p(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n;\boldsymbol{\theta}) = \sum_{k=1}^{n} \log\left(\sum_{i=1}^{n} \pi_i \cdot f_{ik}\right) \tag{22}$$

## 5.4 凸クラスタリングの弱点

凸クラスタリングは，k-means, 混合正規分布モデルと違い，凸計画化問題であるため，必ず大域的最適解が求まる利点があるが，分散 $\sigma^2$ がハイパーパラメータであり，かつすべてのクラスタで等しくなる．また，式の導出から超球形の分布となる．そしてその分散も適切にしなければ効果的なクラスタリングを行うことができない．また，凸クラスタリングの 2 つ目の条件は，データが十分に大きく，クラスタが密であることを前提にしているため，十分なデータ数が必要である点がある．