クラスタリング

階層的クラスタリング, k 平均法, ソフトクラスタリング

中田和秀

東京科学大学 工学院 経営工学系

機械学習入門

https://www.nakatalab.iee.e.titech.ac.jp/text/nakata.html

概要

特定のタスクに対する正解は無いものの、データ自体は大量にあるというケースは多い。その場合には教師なし学習を行うことになる。ここでは、その一つであるクラスタリングについて説明をする。クラスタリングによって、データ点を幾つかのグループにまとめることができる。

目次：

記号の使い方：

- $A := B$ は、B で A を定義する、B を A に代入することを意味する
- $[n]$ は n までのインデックスの集合を表し $[n] := \{1, 2, \cdots, n\}$

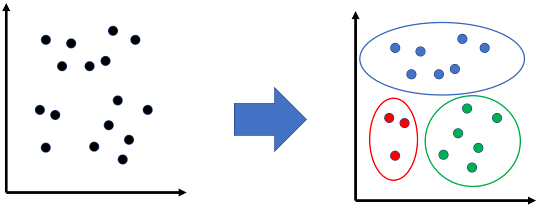
クラスタリング

教師なし学習：
与えられたデータに対し、データに潜むパターン・構造・知見を見出す。

クラスタリング

データを幾つかのクラスタにグループ分けする手法

例： 顧客層の発見、行動パターンの類型化、企業の分類、故障原因の抽出など



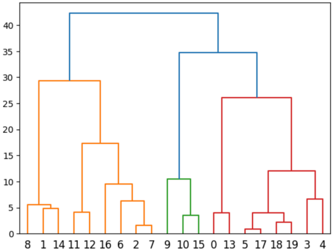
以下で説明する手法

- 階層的クラスタリング
- k 平均法
- ソフトクラスタリング（混合正規分布）

階層的クラスタリング

バラバラの状態から「距離」の近い順に融合して次第に大きなクラスタを作る手法
樹形図 (dendrogram) で表現する。

樹形図



学習アルゴリズム

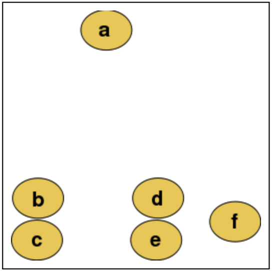
ステップ０　データ点一つが一つのクラスタとする。

ステップ１　クラスタ間の距離を計算

ステップ２　最も距離が小さなクラスタを合併する

ステップ３　クラスタが二つ以上あればステップ１に戻る

学習アルゴリズムのイメージ



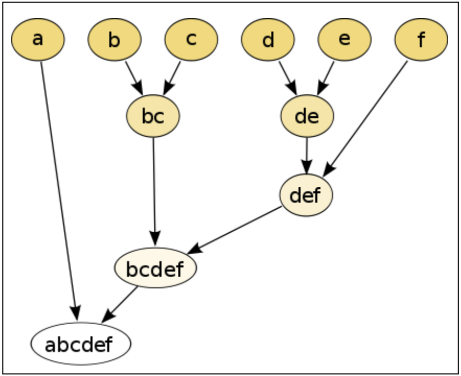


Figure: https://en.wikipedia.org/wiki/Hierarchical_clustering

クラスタ間の距離

- 単連結法：

  $$D(C_i, C_j) := \min_{\boldsymbol{x} \in C_i,\ \boldsymbol{y} \in C_j} d(\boldsymbol{x}, \boldsymbol{y})$$

  大きなクラスタができやすい

- 完全連結法：

  $$D(C_i, C_i) := \max_{\boldsymbol{x} \in C_i,\ \boldsymbol{y} \in C_j} d(\boldsymbol{x}, \boldsymbol{y})$$

  同じようなサイズのクラスタができやすい

- 群平均法：

  $$D(C_i, C_j) := \frac{1}{|C_i||C_j|} \sum_{\boldsymbol{x} \in C_i,\ \boldsymbol{y} \in C_j} d(\boldsymbol{x}, \boldsymbol{y})$$

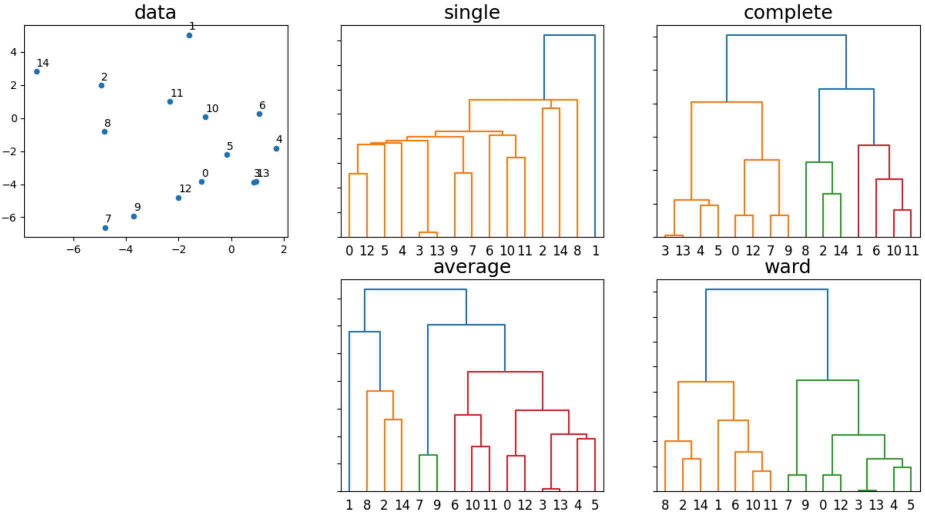
  単連結法と完全連結法の間の性質

- Ward 法：　$\boldsymbol{\mu}$ は平均ベクトル

  $$D(C_i, C_j) := \sum_{\boldsymbol{x} \in C_i \cup C_j} \|\boldsymbol{x} - \boldsymbol{\mu}_{ij}\|^2 - \sum_{\boldsymbol{x} \in C_i} \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2 - \sum_{\boldsymbol{x} \in C_j} \|\boldsymbol{x} - \boldsymbol{\mu}_j\|^2$$

  クラスタ内変動の増分で距離を定義しており、良さそうな結果になりやすい

数値例


data
single
complete
average
ward
0 12 5 4 3 13 9 7 6 10 11 2 14 8 1
3 13 4 5 0 12 7 9 8 2 14 1 6 10 11
1 8 2 14 7 9 6 10 11 0 12 3 13 4 5
8 2 14 1 6 10 11 7 9 0 12 3 13 4 5

Ward 法の効率的な計算方法

Ward 法

$$D(C_i, C_j) := \sum_{\boldsymbol{x} \in C_i \cup C_j} \|\boldsymbol{x} - \boldsymbol{\mu}_{ij}\|^2 - \sum_{\boldsymbol{x} \in C_i} \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2 - \sum_{\boldsymbol{x} \in C_j} \|\boldsymbol{x} - \boldsymbol{\mu}_j\|^2$$

- クラスタ C_i の平均ベクトルは $\boldsymbol{\mu}_i := \frac{1}{|C_i|} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{x}$
- クラスタ C_i と C_j を統合してクラスタ C_{ij} を作る
- クラスタ C_{ij} の平均ベクトルを $\boldsymbol{\mu}_{ij}$ とする

次のように簡単に計算が可能

$$\boldsymbol{\mu}_{ij} = \frac{|C_i|}{|C_i| + |C_j|} \boldsymbol{\mu}_i + \frac{|C_j|}{|C_i| + |C_j|} \boldsymbol{\mu}_j$$

$$D(C_i, C_j) = \frac{|C_i|\ |C_j|}{|C_i| + |C_j|} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2$$

式変形は次スライド

式変形 1

$$
\begin{aligned}
\boldsymbol{\mu}_{ij} :=& \frac{1}{|C_{ij}|} \sum_{\boldsymbol{x} \in C_{ij}} \boldsymbol{x} \\
=& \frac{1}{|C_i \cup C_j|} \sum_{\boldsymbol{x} \in C_i \cup C_j} \boldsymbol{x} \\
=& \frac{1}{|C_i| + |C_j|} \left(\sum_{\boldsymbol{x} \in C_i} \boldsymbol{x} + \sum_{\boldsymbol{x} \in C_j} \boldsymbol{x} \right) \\
=& \frac{|C_i|}{|C_i| + |C_j|} \frac{1}{|C_i|} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{x} + \frac{|C_j|}{|C_i| + |C_j|} \frac{1}{|C_j|} \sum_{\boldsymbol{x} \in C_j} \boldsymbol{x} \\
=& \frac{|C_i|}{|C_i| + |C_j|} \boldsymbol{\mu}_i + \frac{|C_j|}{|C_i| + |C_j|} \boldsymbol{\mu}_j
\end{aligned}
$$

式変形 2

$$\begin{aligned}
\sum_{\boldsymbol{x}\in C} d(\boldsymbol{x},\boldsymbol{\mu})^2 &= \sum_{\boldsymbol{x}\in C} (\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu}) \\
&= \sum_{\boldsymbol{x}\in C} \boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{\mu}^T\boldsymbol{x} + \boldsymbol{\mu}^T\boldsymbol{\mu} \\
&= \sum_{\boldsymbol{x}\in C} \boldsymbol{x}^T\boldsymbol{x} \;-\; 2\boldsymbol{\mu}^T \sum_{\boldsymbol{x}\in C} \boldsymbol{x} \;+\; |C|\boldsymbol{\mu}^T\boldsymbol{\mu} \\
&= \sum_{\boldsymbol{x}\in C} \boldsymbol{x}^T\boldsymbol{x} \;-\; 2\boldsymbol{\mu}^T|C|\boldsymbol{\mu} \;+\; |C|\boldsymbol{\mu}^T\boldsymbol{\mu} \\
&= \sum_{\boldsymbol{x}\in C} \boldsymbol{x}^T\boldsymbol{x} \;-\; |C|\boldsymbol{\mu}^T\boldsymbol{\mu}
\end{aligned}$$

よって、

$$\sum_{\boldsymbol{x}\in C_i} \|\boldsymbol{x}-\boldsymbol{\mu}_i\|^2 = \sum_{\boldsymbol{x}\in C_i} \boldsymbol{x}^T\boldsymbol{x} - |C_i|\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i$$

$$\sum_{\boldsymbol{x}\in C_j} \|\boldsymbol{x}-\boldsymbol{\mu}_j\|^2 = \sum_{\boldsymbol{x}\in C_j} \boldsymbol{x}^T\boldsymbol{x} - |C_j|\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j$$

$$\sum_{\boldsymbol{x}\in C_{ij}} \|\boldsymbol{x}-\boldsymbol{\mu}_{ij}\|^2 = \sum_{\boldsymbol{x}\in C_{ij}} \boldsymbol{x}^T\boldsymbol{x} - |C_{ij}|\boldsymbol{\mu}_{ij}^T\boldsymbol{\mu}_{ij}$$

式変形 3

$$
\begin{aligned}
D(C_i, C_j) &:= \sum_{\boldsymbol{x} \in C_{ij}} \|\boldsymbol{x} - \boldsymbol{\mu}_{ij}\|^2 - \sum_{\boldsymbol{x} \in C_i} \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2 - \sum_{\boldsymbol{x} \in C_j} \|\boldsymbol{x} - \boldsymbol{\mu}_j\|^2 \\
&= \left(\sum_{\boldsymbol{x} \in C_{ij}} \boldsymbol{x}^T \boldsymbol{x} - |C_{ij}| \boldsymbol{\mu}_{ij}^T \boldsymbol{\mu}_{ij} \right) - \left(\sum_{\boldsymbol{x} \in C_i} \boldsymbol{x}^T \boldsymbol{x} - |C_i| \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i \right) \\
&\quad - \left(\sum_{\boldsymbol{x} \in C_j} \boldsymbol{x}^T \boldsymbol{x} - |C_j| \boldsymbol{\mu}_j^T \boldsymbol{\mu}_j \right) \\
&= -|C_{ij}| \boldsymbol{\mu}_{ij}^T \boldsymbol{\mu}_{ij} + |C_i| \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i + |C_j| \boldsymbol{\mu}_j^T \boldsymbol{\mu}_j \\
&= -|C_{ij}| \left(\frac{|C_i| \boldsymbol{\mu}_i + |C_j| \boldsymbol{\mu}_j}{|C_i| + |C_j|} \right)^T \left(\frac{|C_i| \boldsymbol{\mu}_i + |C_j| \boldsymbol{\mu}_j}{|C_i| + |C_j|} \right) + |C_i| \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i + |C_j| \boldsymbol{\mu}_j^T \boldsymbol{\mu}_j \\
&= \frac{|C_i|\ |C_j|}{|C_i| + |C_j|} \left(\boldsymbol{\mu}_i^T \boldsymbol{\mu}_i - 2 \boldsymbol{\mu}_i^T \boldsymbol{\mu}_j + \boldsymbol{\mu}_j^T \boldsymbol{\mu}_j \right) \\
&= \frac{|C_i|\ |C_j|}{|C_i| + |C_j|} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2
\end{aligned}
$$

階層的クラスタリングの特徴

- 樹形図で視覚的に捉えることができる。
- 学習後、任意のクラスタ数での分割結果を知ることができる。
- 各特徴量のスケールをあわせておく必要がある。特に情報が無ければ、各特徴量の平均が 0、分散が 1 になるように線形変換を行うことが多い。
- 最初に全データ間の距離の計算が必要 (データ数の 2 乗)。データ数が多くなると計算に時間がかかる（また、可視化も無意味になる）。
- ここではバラバラの状態からクラスタを作っていく「凝縮的クラスタリング」を説明したが、一つのクラスタにまとまった状態から分解していく「分割的クラスタリング」もある。

k 平均法

データ：$\{\boldsymbol{x}_d\}_{d\in[D]} \quad \boldsymbol{x}_d \in \mathbb{R}^n$

目的：K 個のクラスタに分類したい（K は予め決めておいた数）

アイデア：

データ点が所属しているクラスタの代表点までの「距離」の和を最小化

- クラスタ k の代表点：$\boldsymbol{\mu}_k \in \mathbb{R}^n$
- クラスタ k が属するデータ点の集合：$C_k \subset \{\boldsymbol{x}_d\}_{d\in[D]}$

学習

変数は $\boldsymbol{\mu}_k$ と $C_k \quad (k \in [K])$

$$\begin{aligned} \min \quad & \sum_{k\in[K]} \sum_{\boldsymbol{x}\in C_k} d(\boldsymbol{x}, \boldsymbol{\mu}_k) \\ \text{s.t.} \quad & \bigcup_{k\in[K]} C_k = \{\boldsymbol{x}_d\}_{d\in[D]}, \\ & C_i \cap C_j = \emptyset \qquad\qquad (i \neq j,\ i, j \in [K]). \end{aligned}$$

学習アルゴリズム

最適解を見つけることは難しい。

$\rightarrow$ 近似最適解を計算する。

学習アルゴリズム

交互最適化

- ステップ 0　ランダムに $\boldsymbol{\mu}_k$ を定める
- ステップ 1　$\boldsymbol{\mu}_k \in \mathbb{R}^n$ を固定して、各データ点が属するクラスタを最適化
- ステップ 2　クラスタを固定して、$\boldsymbol{\mu}_k \in \mathbb{R}^n$ を最適化
- ステップ 3　目的関数値が変化しなければ終了。そうでなければステップ 1 に戻る

- 「距離」として $d(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} - \boldsymbol{y}\|^2$ を利用することが多い [1]
- 以下では、この距離におけるアルゴリズムを説明する

[1] 距離の公理は満たしていないが、2 点間の遠さを表す指標にはなる

ステップ１の説明

学習

変数は $\boldsymbol{\mu}_k$ と $C_k \quad (k \in [K])$

$$
\begin{aligned}
\min \quad & \sum_{k \in [K]} \sum_{\boldsymbol{x} \in C_k} d(\boldsymbol{x}, \boldsymbol{\mu}_k) \\
\text{s.t.} \quad & \bigcup_{k \in [K]} C_k = \{\boldsymbol{x}_d\}_{d \in [D]}, \\
& C_i \cap C_j = \emptyset \qquad\qquad (i \neq j,\ i, j \in [K]).
\end{aligned}
$$

$\boldsymbol{\mu}_k$ は固定された状態で、各データ点が属するクラスタを最適化する。

- データ点 $\boldsymbol{x}_d$ は一番近い代表点 $\boldsymbol{\mu}_k$ のクラスタに所属するのが最適。
- $k^* := \operatorname*{argmin}_{k \in [K]} d(\boldsymbol{x}_d, \boldsymbol{\mu}_k)$ としたとき、$\boldsymbol{x}_d \in C_{k^*}$

  ※ 一番近い代表点が２つ以上あるときはランダムに選ぶ

ステップ２の説明

学習

変数： $\boldsymbol{\mu}_k$ と C_k $(k \in [K])$

$$\begin{aligned} \min \quad & \sum_{k \in [K]} \sum_{\boldsymbol{x} \in C_k} d(\boldsymbol{x}, \boldsymbol{\mu}_k) \\ \text{s.t.} \quad & \bigcup_{k \in [K]} C_k = \{\boldsymbol{x}_d\}_{d \in [D]}, \\ & C_i \cap C_j = \emptyset \qquad (i \neq j,\ i, j \in [K]). \end{aligned}$$

クラスタは固定された状態で、$\boldsymbol{\mu}_k$ を最適化

クラスタごとに独立して最適化が可能 $\quad \displaystyle \min_{\boldsymbol{\mu}_k} \sum_{\boldsymbol{x} \in C_k} d(\boldsymbol{x}, \boldsymbol{\mu}_k)$

$d(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} - \boldsymbol{y}\|^2$ のときの最適解は平均ベクトル（重心）

$$\boldsymbol{\mu}_k^* := \frac{1}{|C_k|} \sum_{\boldsymbol{x} \in C_k} \boldsymbol{x}$$

ステップ 2 の説明

$$
\begin{aligned}
f(\boldsymbol{\mu}_k) := \sum_{\boldsymbol{x} \in C_k} \|\boldsymbol{x} - \boldsymbol{\mu}_k\|^2 &= \sum_{\boldsymbol{x} \in C_k} (\boldsymbol{x} - \boldsymbol{\mu}_k)^T (\boldsymbol{x} - \boldsymbol{\mu}_k) \\
&= \sum_{\boldsymbol{x} \in C_k} (\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k - 2\boldsymbol{x}^T \boldsymbol{\mu}_k + \boldsymbol{x}^T \boldsymbol{x}) \\
&= |C_k| \boldsymbol{\mu}_k^T \boldsymbol{\mu}_k - 2 \left(\sum_{\boldsymbol{x} \in C_k} \boldsymbol{x} \right)^T \boldsymbol{\mu}_k + \sum_{\boldsymbol{x} \in C_k} \boldsymbol{x}^T \boldsymbol{x}
\end{aligned}
$$

2 次関数の微分を考える。

$$
\nabla f(\boldsymbol{\mu}_k) = 2|C_k| \boldsymbol{\mu}_k - 2 \sum_{\boldsymbol{x} \in C_k} \boldsymbol{x}, \qquad \nabla^2 f(\boldsymbol{\mu}_k) = 2|C_k| \boldsymbol{I}
$$

$\nabla^2 f(\boldsymbol{\mu}_k)$ は半正定値なので、$f(\boldsymbol{\mu}_k)$ は凸関数。

最適解の必要十分条件 $\nabla f(\boldsymbol{\mu}_k^*) = \mathbf{0}$ を解くと、

$$
\boldsymbol{\mu}_k^* := \frac{1}{|C_k|} \sum_{\boldsymbol{x} \in C_k} \boldsymbol{x}
$$

有限回の反復での終了

k 平均法は有限回の反復で終了する。

- ステップ１で各データ点の所属クラスタが決まれば、ステップ２で $\boldsymbol{\mu}_k$ が決まり、ステップ３での目的関数値も一つに定まる
- データ点の所属クラスタのパターンは有限なので、ステップ３での目的関数値も有限個の異なる値しかとれない。
- ステップ１とステップ２では目的関数値は増えることはない。
- ステップ３時点での目的関数が変わらなければ終了。
減ったとしても、いずれ目的関数が変わらなくなり終了。

最適化アルゴリズムの実行例

赤い X：代表点、色：クラスタ。　左上が正解で、中上から先が k 平均法の各反復。

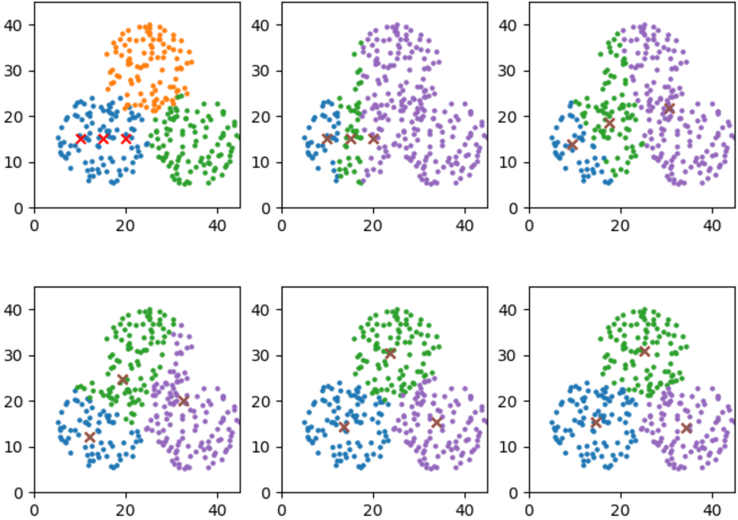


Figure: http://taustation.com/k-means-clustering/

最尤推定

データ点は K 個の多変量正規分布 $N(\boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})$ のどれかから生成された点であると考える。

データ点 $\boldsymbol{x}_d$ がクラスタ k から生成される確率：$q_{dk} \in [0, 1]$

尤度関数

$$l(\mathcal{D}|\boldsymbol{U}, \boldsymbol{Q}) := \prod_{d\in[D]} \sum_{k\in[K]} q_{dk} \left\{ \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{x}_d - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_d - \boldsymbol{\mu}_k) \right\} \right\}$$

$$\boldsymbol{U} = \{\boldsymbol{\mu}_k\}_{k\in[K]},\ \boldsymbol{Q} \in \mathbb{R}^{D\times K}$$

尤度の最大化を考える。データ点 $\boldsymbol{x}_d$ 毎に q_{dk} は最大化できる。

$$\begin{aligned} \max_{q_{d1},\cdots,q_{dK}} \quad & \sum_{k\in[K]} c_{dk} q_{dk} \\ \text{s.t} \quad & \sum_{k\in[K]} q_{dk} = 1, \\ & 0 \le q_{dk} \le 1 \qquad (k \in [K]). \end{aligned}$$

ただし、$c_{dk} := \dfrac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\dfrac{1}{2\sigma^2} (\boldsymbol{x}_d - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_d - \boldsymbol{\mu}_k) \right\}$

最尤推定（続き）

この問題の最適値は $\max_{k \in [K]} c_{dk}$ である。

【証明】　前ページにある最適化問題の最適値を z^* とする。

- 最適解を q^*_{dk} $(k \in K)$ とする。

$$
\begin{aligned}
z^* := \sum_{k \in [K]} c_{dk} q^*_{dk} &\leq \sum_{k \in [K]} \left(\max_{k' \in [K]} c_{dk'} \right) q^*_{dk} \\
&= \max_{k' \in [K]} c_{dk'} \sum_{k \in [K]} q^*_{dk} = \max_{k' \in [K]} c_{dk'}
\end{aligned}
$$

- $\widetilde{q_{dk}} := \begin{cases} 1 & (k = \operatorname*{argmax}_{k' \in [K]} c_{dk'}) \\ 0 & (\text{o.w.}) \end{cases}$ とすると、実行可能解である。

  そのときの目的関数値は $\sum_{k \in [K]} c_{dk} \widetilde{q_{dk}} = \max_{k' \in [K]} c_{dk'}$ となる。

  よって、$z^* \geq \max_{k' \in [K]} c_{dk'}$ と分かる。

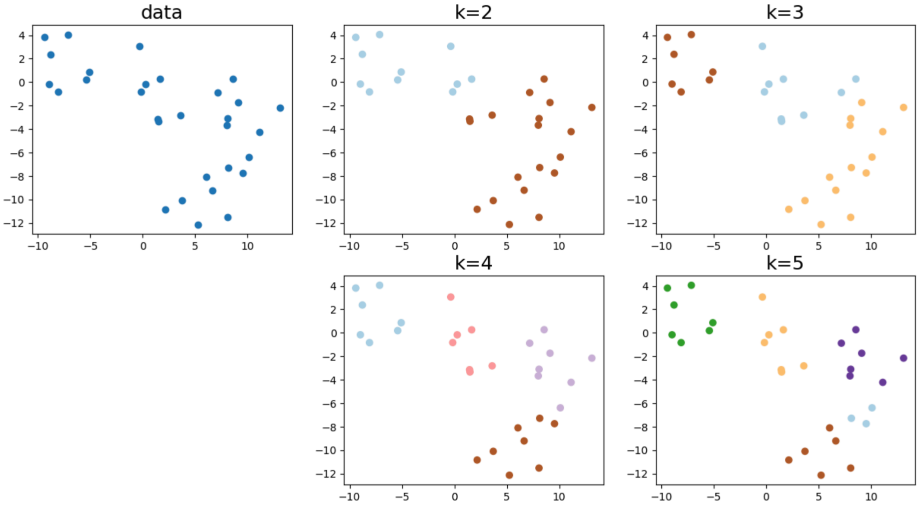
以上より、$z^* = \max_{k \in [K]} c_{dk}$ である。

最尤推定（続き）

$$
\begin{aligned}
& \max_{\boldsymbol{U},\boldsymbol{Q}} \quad l(\mathcal{D}|\boldsymbol{U},\boldsymbol{Q}) \\
\iff & \max_{\boldsymbol{U}} \quad \prod_{d\in[D]} \max_{k\in[K]} c_{dk} \\
\iff & \max_{\boldsymbol{U}} \quad \prod_{d\in[D]} \max_{k\in[K]} \left\{ \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{x}_d - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_d - \boldsymbol{\mu}_k) \right\} \right\} \\
\iff & \max_{\boldsymbol{U}} \quad \sum_{d\in[D]} \max_{k\in[K]} \left\{ -\frac{1}{2\sigma^2} (\boldsymbol{x}_d - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_d - \boldsymbol{\mu}_k) \right\} \\
\iff & \min_{\boldsymbol{U}} \quad \sum_{d\in[D]} \min_{k\in[K]} (\boldsymbol{x}_d - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_d - \boldsymbol{\mu}_k) \\
\iff & \min_{\boldsymbol{U}} \quad \sum_{d\in[D]} \min_{k\in[K]} d(\boldsymbol{x}_d, \boldsymbol{\mu}_k)
\end{aligned}
$$

ユークリッド距離の２乗を用いた k 平均法は、この分布における最尤推定

計算例



当然だが、k の指定によってクラスタリングは異なる

k 平均法の特徴

特徴

- 学習結果は初期点に依存するため、幾つか初期点から学習を行い，結果の良いもの（目的関数値の低いもの）を採用する。
- 予めクラスタ数 K を決めておく必要がある。
- 各特徴量のスケールをあわせておく必要がある。特に情報が無ければ、各特徴量の平均が 0、分散が 1 になるように線形変換を行うことが多い。
- 距離としてユークリッド距離の 2 乗を用いない場合は、代表点 $\boldsymbol{\mu}_k$ の計算が難しくなる。
  - $d(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} - \boldsymbol{y}\|_1 = \sum_{i=1}^{n} |x_i - y_i|$ $\rightarrow$ 各次元でメディアンの計算
  - $d(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} - \boldsymbol{y}\|$ $\rightarrow$ Fermat-Weber 問題
  - 一般の距離： 代表点をデータ点に限定する $\rightarrow$ K-medoid 法

計算時間の比較例

データ数	k 平均法 ($k = 10$)	階層的クラスタリング (Ward 法)
1.0×10^3	0.05s	0.08s
3.0×10^3	0.11s	0.68s
1.0×10^4	1.06s	10.9s
3.0×10^4	1.50s	68s
1.0×10^5	2.62s	メモリ不足
3.0×10^5	9.85s	
1.0×10^6	27.7s	
3.0×10^7	59.7s	
1.0×10^8	259s	

ソフトクラスタリング

混合モデル

- K 個の分布の中から一つが選ばれる。
  k 番目の分布が選ばれる確率：π_k $\qquad \boldsymbol{\pi} \geq 0, \quad \boldsymbol{e}^T \boldsymbol{\pi} = 1$
- 選ばれた分布でデータが生成される。
  $p(\boldsymbol{x}) := p_k(\boldsymbol{x}; \boldsymbol{\theta}_k)$ $\qquad \boldsymbol{\theta}_k$ は分布のパラメタ $\qquad (k \in [K])$

確率密度： $p(\boldsymbol{x}) := \sum_{k \in [K]} \pi_k \, p_k(\boldsymbol{x}; \boldsymbol{\theta}_k)$

例：混合正規分布の場合

$$p_k(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) := \frac{1}{(2\pi)^{n/2}} \frac{1}{\det(\boldsymbol{\Sigma}_k)^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) \right\}$$

ソフトクラスタリング

$\boldsymbol{x}_d$ が k 番目のクラスタ（分布）に属する確率

$$p(C_k | \boldsymbol{x}_d) = \frac{p(C_k) \, p(\boldsymbol{x}_d | C_k)}{\sum_{k' \in [K]} p(C_{k'}) \, p(\boldsymbol{x}_d | C_{k'})} = \frac{\pi_k \, p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)}{\sum_{k' \in [K]} \pi_{k'} \, p_{k'}(\boldsymbol{x}_d; \boldsymbol{\theta}_{k'})}$$

パラメタの推定

独立にサンプリングされたデータ：　$\mathcal{D} := \{\boldsymbol{x}_d\}_{d\in[D]}$

$$p(\mathcal{D}) = \prod_{d\in[D]} p(\boldsymbol{x}_d) = \prod_{d\in[D]} \sum_{k\in[K]} \pi_k \, p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)$$

を最大にするパラメタ $\boldsymbol{\pi} \in \mathbb{R}^K, \boldsymbol{\theta}_k \ (k \in [K])$ を見つける

対数尤度の最大化

変数：$\boldsymbol{\pi} \in \mathbb{R}^K, \quad \boldsymbol{\theta}_k \quad (k \in [K])$

$$\begin{aligned} \max \quad & \sum_{d\in[D]} \log \sum_{k\in[K]} \pi_k \, p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k) \\ \text{s.t.} \quad & \boldsymbol{e}^T \boldsymbol{\pi} = 1, \ \boldsymbol{\pi} \geq \mathbf{0}. \end{aligned}$$

この最適化問題の解法を幾つかの方向から考える。

最適性条件

最適解であるための必要条件　（証明は 2 ページ後）

$$r_{dk} = \frac{\pi_k\, p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)}{\sum_{k' \in [K]} \pi_{k'}\, p_{k'}(\boldsymbol{x}_d; \boldsymbol{\theta}_{k'})} \qquad (d \in [D],\ k \in [K]) \qquad (1)$$

$$\pi_k = \frac{1}{D} \sum_{d \in [D]} r_{dk} \qquad (k \in [K]) \qquad (2)$$

$$\sum_{d \in [D]} r_{dk} \nabla_{\boldsymbol{\theta}_k} \log p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k) = \mathbf{0} \qquad (k \in [K]) \qquad (3)$$

一般にこの方程式を解くことは容易ではない

$\rightarrow$ 一部の変数を固定した方程式を解く

(1)　$\boldsymbol{\pi}, \boldsymbol{\theta}_k$ を固定して、r_{dk} を計算

(2), (3)　r_{dk} を固定して、$\boldsymbol{\pi}, \boldsymbol{\theta}_k$ を計算

最適化アルゴリズム

混合分布のパラメタ推定

ステップ０ 初期点 $\boldsymbol{\pi}$, $\boldsymbol{\theta}_k$ $(k \in [K])$ を決める。

ステップ１ $$r_{dk} := \frac{\pi_k\, p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)}{\sum_{k' \in [K]} \pi_{k'}\, p_{k'}(\boldsymbol{x}_d; \boldsymbol{\theta}_{k'})} \qquad (d \in [D],\ k \in [K])$$

ステップ２ $$\pi_k := \frac{1}{D} \sum_{d \in [D]} r_{dk} \qquad (k \in [K])$$

ステップ３ 次の方程式を $\boldsymbol{\theta}_k$ について解く [a]

$$\sum_{d \in [D]} r_{dk} \nabla_{\boldsymbol{\theta}_k} \log p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k) = \mathbf{0} \qquad (k \in [K])$$

ステップ４ 終了条件を満たしていなければ、ステップ１に戻る

[a] 解けるかどうかは p_k の形による。例えば正規分布の場合は簡単に計算できる。

最適性条件の証明1

以下では、$\boldsymbol{\theta} := \{\boldsymbol{\theta}_k\}_{k\in[K]}$ とする。

ひとまず、不等式条件 $\boldsymbol{\pi} \geq \mathbf{0}$ は無視して、ラグランジュの未定乗数法を考える。

$$L(\boldsymbol{\pi}, \boldsymbol{\theta}, \lambda) := \sum_{d\in[D]} \log \sum_{k\in[K]} \pi_k \, p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k) \; + \; \lambda(1 - \boldsymbol{e}^T \boldsymbol{\pi})$$

$$\nabla_\lambda L(\boldsymbol{\pi}, \boldsymbol{\theta}, \lambda) = 1 - \boldsymbol{e}^T \boldsymbol{\pi} = 0$$

より、$\sum_{k\in[K]} \pi_k = 1$ だと分かる。
先に、次のように記号を定義する [2]。

$$r_{dk} := \frac{\pi_k \, p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)}{\sum_{k'\in[K]} \pi_{k'} \, p_{k'}(\boldsymbol{x}_d; \boldsymbol{\theta}_{k'})} \qquad (d \in [D], \; k \in [K])$$

[2] $\boldsymbol{x}_d$ が k 番目の分布に属する確率 $P(\boldsymbol{C}_k | \boldsymbol{x}_d)$

最適性条件の証明２

$$\nabla_{\pi_k} L(\boldsymbol{\pi}, \boldsymbol{\theta}, \lambda) = \sum_{d \in [D]} \frac{1}{\sum_{k' \in [K]} \pi_{k'} \, p_{k'}(\boldsymbol{x}_d; \boldsymbol{\theta}_{k'})} p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k) - \lambda$$
$$= \sum_{d \in [D]} \frac{1}{\pi_k} r_{dk} - \lambda = 0$$

より、$\pi_k \lambda = \sum_{d \in [D]} r_{dk}$ であると分かる。両辺で $k \in [K]$ に対して和を取ると、

$$\sum_{k \in [K]} \pi_k \lambda = \sum_{k \in [K]} \sum_{d \in [D]} r_{dk}$$
$$\lambda = \sum_{d \in [D]} \sum_{k \in [K]} r_{dk} \qquad (\because \sum_{k \in [K]} \pi_k = 1)$$
$$= \sum_{d \in [D]} 1 = D \qquad (\because \sum_{k \in [K]} r_{dk} = 1)$$

最適性条件の証明３

よって、$\pi_k = \frac{1}{D} \sum_{d \in [D]} r_{dk}$ という関係が得られる。
このとき、$\pi_k \geq 0$ であるから、非負条件は自動的に満たされる。

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}_k} L(\boldsymbol{\pi}, \boldsymbol{\theta}, \lambda) &= \sum_{d \in [D]} \frac{\pi_k}{\sum_{k' \in [K]} \pi_{k'} \, p_{k'}(\boldsymbol{x}_d; \boldsymbol{\theta}_{k'})} \nabla_{\boldsymbol{\theta}_k} p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k) \\
&= \sum_{d \in [D]} \frac{r_{dk}}{p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)} \nabla_{\boldsymbol{\theta}_k} p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k) \\
&= \sum_{d \in [D]} r_{dk} \frac{\nabla_{\boldsymbol{\theta}_k} p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)}{p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)} \\
&= \sum_{d \in [D]} r_{dk} \nabla_{\boldsymbol{\theta}_k} \log p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k) = \mathbf{0}
\end{aligned}
$$

以上より、最適性条件を導くことができた [3]。

[3] 逆に、得られた３つの条件から最初に述べた必要十分条件を導くこともできる

混合正規分布の場合

正規分布の場合

$$p_k(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) := \frac{1}{(2\pi)^{n/2}} \frac{1}{\det(\boldsymbol{\Sigma}_k)^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)\right\}$$

方程式 $\sum_{d\in[D]} r_{dk} \nabla_{\boldsymbol{\theta}_k} \log p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k) = \mathbf{0}$ の解は、次のようになる。

$$\boldsymbol{\theta}_k := \{\boldsymbol{\mu}_k,\ \boldsymbol{\Sigma}_k\}$$

$$\boldsymbol{\mu}_k := \frac{\sum_{d\in[D]} r_{dk} \boldsymbol{x}_d}{\sum_{d\in[D]} r_{dk}} \qquad (k \in [K])$$

$$\boldsymbol{\Sigma}_k := \frac{\sum_{d\in[D]} r_{dk} (\boldsymbol{x}_d - \boldsymbol{\mu}_k)(\boldsymbol{x}_d - \boldsymbol{\mu}_k)^T}{\sum_{d\in[D]} r_{dk}} \qquad (k \in [K])$$

重み付き平均と重み付き分散共分散行列

証明は２ページ後

混合正規分布の場合

混合正規分布のパラメタ推定

ステップ0　初期点 $\boldsymbol{\pi}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \ (k \in [K])$ を決める。

ステップ1　$$r_{dk} := \frac{\pi_k\, p_k(\boldsymbol{x}_d; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k' \in [K]} \pi_{k'}\, p_{k'}(\boldsymbol{x}_d; \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})} \qquad (d \in [D],\ k \in [K])$$

ステップ2　$$\pi_k := \frac{1}{D} \sum_{d \in [D]} r_{dk} \qquad (k \in [K])$$

ステップ3　$$\boldsymbol{\mu}_k := \frac{\sum_{d \in [D]} r_{dk} \boldsymbol{x}_d}{\sum_{d \in [D]} r_{dk}} \qquad (k \in [K])$$

ステップ4　$$\boldsymbol{\Sigma}_k := \frac{\sum_{d \in [D]} r_{dk} (\boldsymbol{x}_d - \boldsymbol{\mu}_k)(\boldsymbol{x}_d - \boldsymbol{\mu}_k)^T}{\sum_{d \in [D]} r_{dk}} \qquad (k \in [K])$$

ステップ5　終了条件を満たしていなければ、ステップ１に戻る

$$p_k(\boldsymbol{x}_d; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) := \frac{1}{(2\pi)^{n/2}} \frac{1}{\det(\boldsymbol{\Sigma}_k)^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_d - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_d - \boldsymbol{\mu}_k) \right\}$$

証明 1

$$\log p_k(\boldsymbol{x}_d; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log\det \boldsymbol{\Sigma}_k - \frac{1}{2}(\boldsymbol{x}_d - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_d - \boldsymbol{\mu}_k)$$

であるので、

$$\sum_{d\in[D]} r_{dk} \nabla_{\boldsymbol{\mu}_k} \log p_k(\boldsymbol{x}_d; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \sum_{d\in[D]} r_{dk} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_d - \boldsymbol{\mu}_k) = \mathbf{0}$$

よって、

$$\sum_{d\in[D]} r_{dk}(\boldsymbol{x}_d - \boldsymbol{\mu}_k) = \mathbf{0}$$

$$\sum_{d\in[D]} r_{dk}\boldsymbol{x}_d = \sum_{d\in[D]} r_{dk}\boldsymbol{\mu}_k$$

$$\boldsymbol{\mu}_k = \frac{\sum_{d\in[D]} r_{dk}\boldsymbol{x}_d}{\sum_{d\in[D]} r_{dk}}$$

証明２

$$\sum_{d\in[D]} r_{dk} \nabla_{\boldsymbol{\Sigma}_k} \log p_k(\boldsymbol{x}_d; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$
$$= \sum_{d\in[D]} \frac{r_{dk}}{2} \left(-\boldsymbol{\Sigma}_k^{-1} + \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_d - \boldsymbol{\mu}_k)(\boldsymbol{x}_d - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}\right) = \boldsymbol{O}$$

$$\sum_{d\in[D]} r_{dk} \boldsymbol{\Sigma}_k = \sum_{d\in[D]} r_{dk}(\boldsymbol{x}_d - \boldsymbol{\mu}_k)(\boldsymbol{x}_d - \boldsymbol{\mu}_k)^T$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{d\in[D]} r_{dk}(\boldsymbol{x}_d - \boldsymbol{\mu}_k)(\boldsymbol{x}_d - \boldsymbol{\mu}_k)^T}{\sum_{d\in[D]} r_{dk}}$$

別の見方

必要条件の変数の一部を固定する方法では、反復が進んだときによいものになっているのか分からない。

同じ解法の別の解釈として、次の２つがある。

- 補助関数法
- 有限離散分布に対する EM アルゴリズム　$\rightarrow$ 「生成モデル」で説明

以下で説明を行う。

補助関数法の導入

$$\max \sum_{d\in[D]} \log \sum_{k\in[K]} \pi_k \, p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)$$

$f_0(x) := \log x,\ \ f_k(\boldsymbol{\theta}) := \pi_k \, p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)$ とする。

- $-f_0(x)$ は凸関数（最大化なのでマイナスをつける）
- $f_k(\boldsymbol{\theta}) \geq 0 \quad (\forall \boldsymbol{\theta})$

補助関数法の話を適用すると、

- $\displaystyle\sum_{k\in[K]} r_{dk} = 1,\ r_{dk} \geq 0$ を満たす r_{dk} に対して、

  $$\log \sum_{k\in[K]} \pi_k \, p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k) \geq \sum_{k\in[K]} r_{dk} \log \frac{\pi_k \, p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)}{r_{dk}}$$

- $r_{dk} := \dfrac{\pi_k \, p(\boldsymbol{x}_d; \boldsymbol{\theta}_k)}{\sum_{k'\in[K]} \pi_{k'} \, p(\boldsymbol{x}_d; \boldsymbol{\theta}_{k'})}$ のとき等号が成り立つ。

これを $d \in [D]$ に関する和として適用すると、補助関数法が適用できる。

補助関数法の内部で解く最適化問題

補助関数法のステップ 2

変数：$\boldsymbol{\pi}$, $\boldsymbol{\theta}_k$ $(k \in [K])$

$$\begin{aligned} \max \quad & \sum_{d\in[D]} \sum_{k\in[K]} r_{dk} \log \frac{\pi_k\, p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)}{r_{dk}} \\ \text{s.t.} \quad & \boldsymbol{e}^T \boldsymbol{\pi} = 1,\ \boldsymbol{\pi} \geq \boldsymbol{0}. \end{aligned}$$

$$\begin{aligned} & \sum_{d\in[D]} \sum_{k\in[K]} r_{dk} \log \frac{\pi_k\, p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)}{r_{dk}} \\ &= \sum_{d\in[D]} \sum_{k\in[K]} r_{dk} (\log \pi_k + \log p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k) - \log r_{dk}) \\ &= \sum_{k\in[K]} \left(\sum_{d\in[D]} r_{dk} \right) \log \pi_k + \sum_{k\in[K]} \left(\sum_{d\in[D]} r_{dk} \log p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k) \right) + \text{定数} \end{aligned}$$

$1 + K$ 個の最適化問題に分割できる。

ステップ 2 での最適化

$$\text{問題 } A: \quad \max_{\boldsymbol{\pi}} \sum_{k\in[K]} \left(\sum_{d\in[D]} r_{dk} \right) \log \pi_k \quad \text{s.t} \quad \boldsymbol{e}^T \boldsymbol{\pi} = 1,\ \boldsymbol{\pi} \geq \mathbf{0}.$$

$$\underset{(k\in[K])}{\text{問題 } B_k}: \quad \max_{\boldsymbol{\theta}_k} \sum_{d\in[D]} r_{dk} \log p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)$$

問題 A： 最適解は $\pi_k^* := \frac{1}{D} \sum_{d\in[D]} r_{dk} \quad (k \in [K])$ 証明は次ページ

問題 B_k： 重み付き最尤推定

分布 $p_k(\boldsymbol{x}; \boldsymbol{\theta}_k)$ で、$\{\boldsymbol{x}_d\}_{d\in[D]}$ のサンプルが得られたときの最尤推定は以下のもの

$$\max_{\boldsymbol{\theta}_k} \prod_{d\in[D]} p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k) \quad \Longleftrightarrow \quad \max_{\boldsymbol{\theta}_k} \sum_{d\in[D]} \log p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)$$

問題 A の最適解の証明

不等式条件 $\boldsymbol{\pi} \geq \mathbf{0}$ は無視して、ラグランジュの未定乗数法を適用する。

$$L(\boldsymbol{\pi}, \lambda) := \sum_{k \in [K]} \left(\sum_{d \in [D]} r_{dk} \right) \log \pi_k + \lambda (1 - \boldsymbol{e}^T \boldsymbol{\pi})$$

$$\nabla_{\pi_k} L(\boldsymbol{\pi}, \lambda) = \left(\sum_{d \in [D]} r_{dk} \right) \frac{1}{\pi_k} - \lambda = 0 \qquad (k \in [K])$$

$$\nabla_{\lambda} L(\boldsymbol{\pi}, \lambda) = 1 - \boldsymbol{e}^T \boldsymbol{\pi} = 0$$

上式より、

$$\lambda \pi_k = \sum_{d \in [D]} r_{dk}$$

$$\sum_{k \in [K]} \lambda \pi_k = \sum_{k \in [K]} \sum_{d \in [D]} r_{dk}$$

$$\lambda = D$$

よって、$\pi_k = \frac{1}{D} \sum_{d \in [D]} r_{dk}$ となり、$\boldsymbol{\pi} \geq \mathbf{0}$ も満たすので最適解。

補助関数法

混合分布に対する補助関数法

ステップ０ 初期点 $\boldsymbol{\pi}, \boldsymbol{\theta}_k \ (k \in [K])$ を決める。

ステップ１ $r_{dk} := \dfrac{\pi_k \, p(\boldsymbol{x}_d; \boldsymbol{\theta}_k)}{\sum_{k' \in [K]} \pi_{k'} \, p(\boldsymbol{x}_d; \boldsymbol{\theta}_{k'})} \qquad d \in [D], \ k \in [K]$

ステップ２

$$\pi_k := \frac{1}{D} \sum_{d \in [D]} r_{dk} \quad (k \in [K])$$

$k \in [K]$ に対して、重み付き最尤推定を行う。

$$\max_{\boldsymbol{\theta}_k} \sum_{d \in [D]} r_{dk} \log p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)$$

ステップ３ 終了条件を満たしていなければ、ステップ１に戻る

補助関数法と変数固定法の関係

重み付き最尤推定問題： $\max_{\boldsymbol{\theta}_k} \sum_{d\in[D]} r_{dk} \log p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k)$

最適解の必要条件は

$$\nabla_{\boldsymbol{\theta}_k} \sum_{d\in[D]} r_{dk} \log p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k) = \mathbf{0}$$

$$\sum_{d\in[D]} r_{dk} \nabla_{\boldsymbol{\theta}_k} \log p_k(\boldsymbol{x}_d; \boldsymbol{\theta}_k) = \mathbf{0}$$

この方程式を解くものと思えば、p.29 の解法と同じものになる。

k 平均法と混合ガウス分布の関係 1

k 平均法で解く最適化問題　(p.20)

変数： $\boldsymbol{\pi}_d \in \mathbb{R}^K \ (d \in [D]), \quad \boldsymbol{\mu}_k \in \mathbb{R}^n \ (k \in [K])$

$$\begin{aligned} \max \quad & \textstyle\sum_{d \in [D]} \log \sum_{k \in [K]} [\boldsymbol{\pi}_d]_k \, p_k(\boldsymbol{x}_d; \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I}) \\ \text{s.t.} \quad & \boldsymbol{e}^T \boldsymbol{\pi}_d = 1, \ \boldsymbol{\pi}_d \geq \boldsymbol{0} \ \ (d \in [D]). \end{aligned}$$

混合ガウス分布での最適化問題と比較して、

- 分散共分散行列を $\sigma^2 \boldsymbol{I}$ で固定
- データ点ごとにカテゴリ分布 $\boldsymbol{\pi}$ を持つ

という違いがある。

混合正規分布　(p.27)

変数： $\boldsymbol{\pi} \in \mathbb{R}^K, \quad \boldsymbol{\mu}_k \in \mathbb{R}^n, \ \boldsymbol{\Sigma}_k \in \mathbb{R}^{n \times n} \ \ (k \in [K])$

$$\begin{aligned} \max \quad & \textstyle\sum_{d \in [D]} \log \sum_{k \in [K]} \pi_k \, p_k(\boldsymbol{x}_d; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ \text{s.t.} \quad & \boldsymbol{e}^T \boldsymbol{\pi} = 1, \ \boldsymbol{\pi} \geq \boldsymbol{0}. \end{aligned}$$

k 平均法と混合ガウス分布の関係 2

別の関係性を考える。

混合ガウス分布において、正規分布のばらつきが極端に小さい状況を考える。
つまり、

$$\boldsymbol{\Sigma}_k = \lim_{\sigma^2 \to 0} \sigma^2 \boldsymbol{I}$$

とする。このとき、

$$\begin{aligned} r_{dk} &:= \lim_{\sigma^2 \to 0} \frac{\pi_k p(\boldsymbol{x}_d; \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})}{\sum_{k' \in [K]} \pi_{k'} p(\boldsymbol{x}_d; \boldsymbol{\mu}_{k'} \sigma^2 \boldsymbol{I})} \\ &= \begin{cases} 1 & (k = \operatorname*{argmax}_{k' \in [K]} p(\boldsymbol{x}_d; \boldsymbol{\mu}_{k'}, \boldsymbol{I})) \\ 0 & (\text{o.w.}) \end{cases} \\ &= \begin{cases} 1 & (k = \operatorname*{argmin}_{k' \in [K]} \|\boldsymbol{x}_d - \boldsymbol{\mu}_{k'}\|) \\ 0 & (\text{o.w.}) \end{cases} \end{aligned}$$

すると、p.34 のアルゴリズムは k 平均法のアルゴリズムと一致する。

ソフトクラスタリングの特徴

特徴

- 重み付きの最尤推定ができる分布ならば適用できる。
- k 平均法よりも柔軟にクラスタリングができる。
  一方、解釈しづらくなる。
- 各クラスタ（分布）への所属確率をベクトルにして特徴量にすることも可能