卒業研究中間報告書

クラスタ数推定に用いる最適な情報量基準の探求

指導教員

藤田 一寿

津山工業高等専門学校

情報工学科

萩原 涼介

平成 29 年 8 月 18 日

目次

1. はじめに

クラスタリングとはデータを教師なし学習により任意の数のクラスタに分ける手法である．クラスタリングはデータ解析，データマイニング，パターン認識など様々な分野で用いられる．多くのクラスタリング手法では，予めクラスタ数を指定しクラスタリングを行う．しかし，データに対し最適なクラスタ数を指定しなければ，最適なクラスタリング結果を得ることはできない．その為，クラスタ数を推定することは重要な課題となっている．

既存のクラスタ数推定手法の多くは，情報量規準に基づきクラスタ数の推定を行っている．情報量規準とは簡単に言えば確率分布とデータの分布の当てはまり具合を表す．その情報量基準は多くの研究者により様々なものが提案されている．しかし，どの情報量規準がどのようなデータに対し有効かは分かっていない．そこで本研究では，クラスタ数推定に用いる情報量規準として最適なものを数値実験を通し明らかにする．本研究では，クラスタ数推定の手法として X-means を用いる．

2. 先行研究

2.1 クラスタリング

クラスタリングとはデータを教師なし学習により任意の数のクラスタに分ける手法である．多くのクラスタリング手法においては，データの類似度をユークリッド距離やマンハッタン距離などの距離尺度によって定義し，それによってクラスタを抽出する．クラスタリングはデータ解析，データマイニング，パターン認識など様々な分野で用いられる．

2.2 K-means

K-means[1] は，多次元空間上のデータ点集合について，各データが属するクラスタを同定する最もよく使われるクラスタリング手法の一つである．K-means は，以下の 2 つの手順を繰り返すことでクラスタリングを行う．

(1) 各データ点とデータ点の距離を求め，各データ点を最も近いセントロイドのクラスタに割り当てる．
(2) クラスタに所属するデータの平均を新たなセントロイドとする．

Algorithm 1 K-means Algorithm

Require: N: データ数, K: クラスタ数, λ: 収束条件, $\boldsymbol{X} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$

K 個のセントロイドとしてランダムに配置する．

3. 手法

3.1 Kullback-Leibler 情報量

偶然を伴う現象は，ある確率分布に従う確率変数の実現値であると考えることができる．この確率分布を近似するモデル（以後「モデル」）はデータを生成する真の確率分布にどの程度近いかによって評価することができる．また，データにモデルを当てはめることは，データから真の確率分布を推定しているものとみなすことができる．このようにモデルと真の分布が共に確率分布であると見なし，モデルの評価や推定を行う．

真の分布とモデルの近さを図る客観的な規準として Kullback-Leibler 情報量 (以後「K-L 情報量」）がある．連続型の確率分布のとき，$g(x)$ を真の確率密度関数，$f(x)$ をモデルが定める確率密度関数とすると，モデル

に関する真の分布の K-L 情報量は $\log\{g(X)/f(X)\}$ の期待値を取り (1) 式で表される．

$$
\begin{aligned}
I(g \mid f) &= E_X\left(\log\left\{\frac{g(X)}{f(X)}\right\}\right) \\
&= \int_{-\infty}^{\infty} \log\left\{\frac{g(x)}{f(x)}\right\} g(x)dx
\end{aligned}
\tag{1}
$$

ただし，log は自然対数で，注記がない限り一貫してこの意味で用いる．

このように，真の分布がわかっている場合には K-L 情報量によってモデルの良し悪しを比較できた．しかし，通常，真の分布は未知で，真の分布から得られたデータだけが与えられていることがい．したがって，データから K-L 情報量を推定する必要がある．$g(x)$，$f(x)$ をそれぞれ真の分布とモデルに対応する密度関数とすると，(1) 式より，

$$
\begin{aligned}
I(g \mid f) &= \int_{-\infty}^{\infty} \left\{\log \frac{g(x)}{f(x)}\right\} g(x)dx \\
&= -\int_{-\infty}^{\infty} \{\log f(x)\} g(x)dx - \left(-\int_{-\infty}^{\infty} \{\log g(x)\} g(x)dx\right)
\end{aligned}
$$

となるが，右辺の第 2 項は定数であり，したがって右辺第 1 項が大きいほど K-L 情報量 $I(g \mid f)$ は小さくなることがわかる．右辺第 1 項の $\int_{-\infty}^{\infty}\{\log f(x)\}g(x)dx$ は，確率密度関数 $\log f(x)$ の期待値 $E(\log f(x))$ であり，平均対数尤度と呼ばれている．ここで，

$$
\sum_{i=1}^{n} \log f(x_i)
$$

を対数尤度と呼ぶことにすると，n 個の独立な観測値 $\{x_1, x_2, \cdots, x_i\}$ が得られると，この平均対数尤度は，対数尤度の n 分の 1

$$
\frac{1}{n}\sum_{i=1}^{n} \log f(x_i)
$$

で近似される．したがって，符号に注意すると，対数尤度が大きいほど，そのモデルは真の分布に近いと考えられる．このようにして，対数尤度を K-L 情報量の推定値を考えることにすると異なったタイプのモデルの良し悪しも比較できるのである．

ところで，確率変数 $(X_1, X_2, \cdots, X_n)$ の同時密度関数が $f(x_1, x_2, \cdots, x_n \mid \theta)$ で与えられているものとする．θ は確率密度関数を規定するパラメータである．この時，観測値 $(x_1, x_2, \cdots, x_n)$ は与えられたものとして固定し，f を θ の関数と考える時，この関数を**尤度**と呼び，$L(\theta)$ で表す．すなわち，

$$
L(\theta) = f(x_1, x_2, \cdots, x_n \mid \theta)
$$

である．特に，確率変数が独立な場合には $(X_1, X_2, \cdots, X_n)$ の確率密度関数は，各 $X_i (i = 1, \cdots, n)$ の確率密度関数の積に等しいことから，

$$
L(\theta) = f(x_1 \mid \theta) f(x_2 \mid \theta) \cdots f(x_n \mid \theta)
$$

となる．この両辺の対数をとると，すでに求められた対数尤度関数

$$
l(\theta) = \sum_{i=1}^{n} \log f(x_i \mid \theta)
$$

が導かれる．

ここでは，平均対数尤度の推定量から対数尤度を直接導入した．しかし，モデルが確率分布の形で与えられている場合には，まず観測値の同時分布から尤度を定義し，その対数として対数尤度を求めるほうが都合が良い．$(X_1, X_2, \cdots, X_n)$ が独立でない場合にも，尤度の対数として対数尤度

$$l(\theta) = \log f(x_1, \cdots, x_n \mid \theta)$$

が定義できる．

3.2 最尤法

ここまで，データに基づいて K-L 情報量の大小を比較するためには対数尤度を比較すれば良いことを示した．あらかじめ与えられたいくつかのモデルがある場合には，対数尤度が最大となるモデルを選択することによって，近似的には真の分布にいちばん近いモデルが得られることになる．したがって，モデルがいくつかの調整できるパラメータを保つ場合には，対数尤度を最大とするようにパラメータの値を選ぶことによって良いモデルが得られることがわかる．この推定を最大尤度法，略して**最尤法**と呼ばれている．また，最尤法で導かれた推定量は**最尤推定量**と呼ばれ，この最尤推定量によって定められるモデルが最尤モデルである．最尤モデルの対数尤度を**最大対数尤度**という．

3.3 情報量規準

情報量規準とは，最尤推定によって当てはめられたモデルが複数個あるときに，その中の一つを選択する規準である．ここでは，最も有名な情報量規準のひとつである AIC(Akaike Information Criterion; 赤池情報量規準)[2] を例に取り理論の解説を行う．

まず，複数個のモデルがあるときに，各モデルの良し悪しを評価する規準として，最尤モデルの平均対数尤度のデータ $\boldsymbol{x}$ に関する期待値（**期待平均対数尤度**）を導入する．データ $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ を確率変数 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$ の実現値とする．X_i は独立に同じ真の分布 $g(\cdot) = f(\cdot \mid \boldsymbol{\theta}^*)$ に従うものとする．ここで，f は K 個のパラメータで規定される自由パラメータ数 K のモデル

$$\mathrm{MODEL}(K) : f(\cdot \mid \boldsymbol{\theta}),\ \boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_K)$$

である．$\boldsymbol{\theta}^* = (\theta_i^*, \theta_2^*, \cdots, \theta_K^*)$ を真のパラメータという．このモデルのパラメータ空間を

$$\Theta_k = \{\boldsymbol{\theta} \in \Theta_K \mid \theta_{k+1} = \theta_{k+2} = \cdots = \theta_K = 0\}$$

に制約して得られる制約モデルを MODEL(k) と書くことにする．$k \leq l$ なら，MODEL(k) は MODEL(l) を制約したモデルになる．この場合，

$$\text{MODEL}(k)\text{の自由パラメータ数} = k$$

が成り立つ．$\Theta_1 \subset \Theta_2 \subset \cdots \subset \Theta_K$ であるから，$\boldsymbol{\theta}^* \in \Theta_k$ を満足する最小の k が存在する．この k をモデルの真の自由パラメータ数といい，k^* で表す．

データ $\boldsymbol{x}$ が与えられたときの，MODEL(k) のパラメータの最尤推定量 $\hat{\boldsymbol{\theta}}_k$ は対数尤度

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i \mid \boldsymbol{\theta})$$

を，$\boldsymbol{\theta} \in \Theta_k$ の範囲で最大化することによって得られる．f は確率密度関数である．このとき，最大対数尤度は

$$l(\hat{\boldsymbol{\theta}}_k) = \max_{\theta \in \Theta_k} l(\boldsymbol{\theta})$$

で与えられる．分布 $f(\cdot \mid \boldsymbol{\theta})$ の平均対数尤度は，Z を X_i と同じ分布に従い $\boldsymbol{X}$ とは独立な確率変数として，

$$E_Z\{\log f(Z \mid \boldsymbol{\theta})\} = \int f(z \mid \boldsymbol{\theta}^*)dz$$

で与えられる．対数尤度との対応を考えて，平均対数尤度の n 倍

$$l^*(\boldsymbol{\theta}) = nE_Z\{\log f(Z \mid \boldsymbol{\theta})\}$$

を定義しておく．$l^*(\boldsymbol{\theta})$ が大きいほど，分布 $f(\cdot \mid \boldsymbol{\theta})$ の真の分布 $f(\cdot \mid \boldsymbol{\theta}^*)$ に対する近似がいいことになる．

最尤モデルの良さは $l^*(\hat{\boldsymbol{\theta}}_k)$ で評価できる．これは確率変数 $\boldsymbol{X}$ の実現値 $\boldsymbol{x}$ に依存する確率変数であるので，その期待値を取り，

$$l_n{}^*(k) = E_{\boldsymbol{X}}\{l^*(\hat{\boldsymbol{\theta}}_k)\} = \int l^*(\hat{\boldsymbol{\theta}}_k)\prod_{i=1}^{n} g(x_i)d\,\boldsymbol{x}$$

でモデルを評価することにする．$l_n{}^*(k)$ をモデルの**期待平均対数尤度**という．期待平均対数尤度は，個々の実現地に依存しない，データの真の分布およびモデルだけで決まる量である．この値が大きいほど良いモデルということになる．

3.4 X-means

先に紹介した K-means は，クラスタ数を事前に指定する必要が生じる．しかし，実際にデータのクラスタリングを行う際，クラスタ数が事前に与えられることは少ない．

X-means[3] は，データ分布が混合等方 Gauss 分布から生成されたと想定してクラスタ数推定及びクラスタリングを行う手法である．K-means の逐次繰り返しと，BIC[4] (Bayesian Information Criterion; ベイズ情報量規準) による分割停止規準を用いることで，クラスタ数を推定しクラスタリングを実行する．

具体的には以下の手順で行われる．

(1) クラスタ数 k を初期化する (通常は $k = 2$) .
(2) K-means を実行する．
(3) 次の処理を $j = 1$ から $j = k$ まで繰り返す．
    (a) クラスタ j の BIC_j を計算する．
    (b) クラスタ j に所属するデータに対し，クラスタ数 2 として K-means を行う．
    (c) クラスタ数 2 としてクラスタリングした結果に対し $\mathrm{BIC'}_j$ を計算する．
    (d) BIC_j と $\mathrm{BIC'}_j$ を比較し，$\mathrm{BIC'}_j$ が大きければクラスタ数 k に 1 を足す．
(4) 前の処理で k が増加した場合は処理 2 へ戻る．そうでない場合は終了する．

X-means で用いる BIC は次のように求められる．d 次元のデータ $\boldsymbol{D} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_d)$ を K 個のクラスタに分割することを考える．モデル M_j の評価に用いる BIC は以下で与えられる．

$$\mathrm{BIC}(M_j) = \hat{l}_j(D) - \frac{p_j}{2}\ln R \tag{2}$$

p_j はモデル M_j のパラメータ数であり，R は M_j のデータ数，$\hat{l}_j(D)$ は p 変量 Gauss 分布の対数尤度関数である．

等方 Gauss 分布を考えると分散 σ^2 は (3) 式により表される．

$$\hat{\sigma}^2 = \frac{1}{R-K}\sum_i \left(\boldsymbol{x}_i - \boldsymbol{\mu}_{(i)}\right)^2 \tag{3}$$

すると，確率は次で表される．

$$\hat{P}(x_i) = \frac{R_{(i)}}{R} \frac{1}{\sqrt{2\pi}\hat{\sigma}^d} \exp\left(-\frac{1}{2\hat{\sigma}^2}||\boldsymbol{x}_i - \boldsymbol{\mu}_{(i)}||^2\right) \tag{4}$$

ここで $\boldsymbol{\mu}_i$ は d 次元の平均ベクトルである．したがって対数尤度関数は

$$\begin{aligned} l(D) &= \log \prod_i P(x_i) \\ &= \sum_i \left(\log \frac{1}{\sqrt{2\pi}\sigma^d} - \frac{1}{2\sigma^2}||\boldsymbol{x}_i - \boldsymbol{\mu}_{(i)}||^2 + \log \frac{R_{(i)}}{R}\right) \end{aligned} \tag{5}$$

となる．ここでクラスタ $n(1 < n < K)$ のデータ D_n に着目する．クラスタ n のデータ数を R_n と表記すると，(5) 式は以下で表される．

$$\begin{aligned} \hat{l}(D_n) = &-\frac{R_n}{2}\log(2\pi) - \frac{R_n \cdot d}{2}\log(\hat{\sigma}^2) - \frac{R_n - K}{2} \\ &+ R_n \log R_n - R_n \log R \end{aligned} \tag{6}$$

4. クラスタリング実験

4.1 実験環境

実験には Python3.5 を用い，機械学習のライブラリとして TensorFlow を用いてアルゴリズムを実装した．

4.2 K-means によるクラスタリング

図 1 のデータをクラスタ数 5 としてクラスタリングした結果，図 2 のような結果になった．

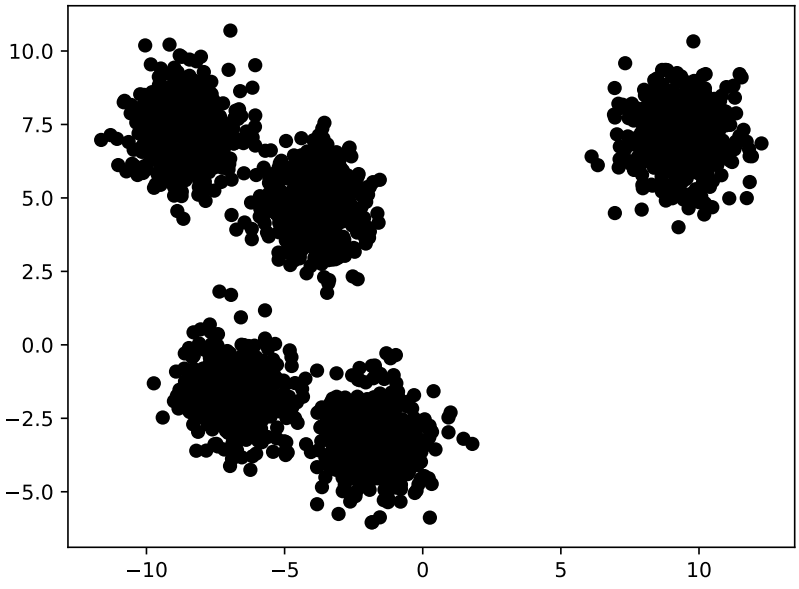

図 1　クラスタリング前のデータ

4.3 X-means によるクラスタリング

K-means と同様に，クラスタ数が 5 つのデータを生成し，X-means により分割した結果，図 3 のようになった．図より，クラスタ数を 5 つとして適切にクラスタリングを行っていることがわかる．

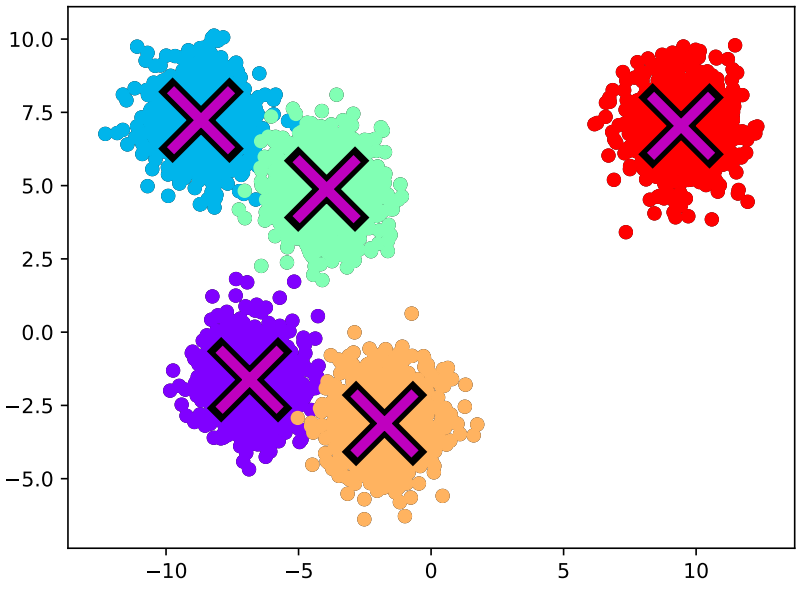

図 2　K-means によるクラスタリング結果

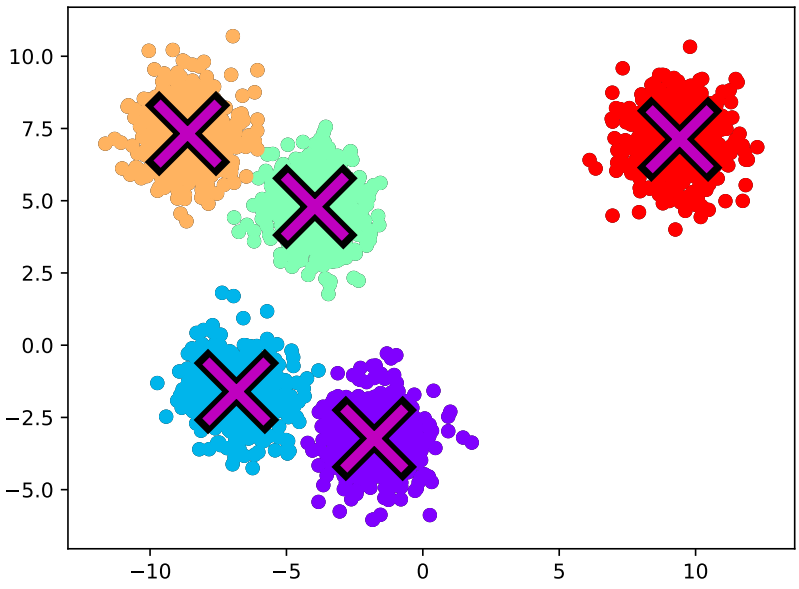

図 3　X-means によるクラスタリング結果

5. おわりに

本実験では，機械学習の基礎学習及び K-means, X-means の実装を行った．先に示した実験結果より，X-means は適切にクラスタ数を推定し，クラスタリングを行うことが確認された．現在は分割停止規準として BIC を用いているが，今後様々な情報量規準を用いて実験を行い，最適な情報量規準を明らかにしたい．クラスタリング精度について検証を行っていきたい．

参考文献

1) James MacQueen et al.: Some methods for classification and analysis of multivariate observations, Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, No. 14, pp. 281–297 (1967).
2) Akaike, H.: Information theory and an extension of the maximum likelihood principle, Proceedings of the 2nd International Symposium on Information Theory, pp. 267-281 (1973).
3) Dan Pelleg, Andrew W Moore, et al.: Xmeans: Extending K-means with Efficient Estimation of the Number of Clusters., ICML, Vol. 1, pp. 727–734 (2000).
4) Gideon Schwarz et al.: Estimating the dimension of a model, The annals of statistics, Vol. 6, No.2, pp. 461–464 (1978).