情報工学実験Ⅴ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

出席番号： 32 番　　報告者： 萩原 涼介　　指導教員： 藤田 一寿　　提出日： 2017 年 7 月 11 日

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．

2. 実験の手法

2.1 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クラスタリング手法の一種である．具体的には，以下の 2 つの手順を繰り返すことで具体的にクラスタリングを行う．

1）各データに割り当てられているクラスタのセントロイドを求める
2）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．

2.2 x-means

x − means[2] 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の決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[3](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，クラスタ数を決定しクラスタリングを実行する．

具体的にはは以下の手順で行われる．

1）クラスタ数を小さなにして k-means を実行
2）各クラスタにおける BIC を算出する
3）それぞれのクラスタのセントロイドを 2 つに分割し，k-means を再度実行
4）分割したそれぞれのクラスタにおける BIC を算出
5）分割前と後の BIC を比較し，BIC が大きくなっていれば採用する
6）2 から 5 を繰り返し，変化がなくなればクラスタリングが完了する

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．

$$\mathrm{BIC}(M_j) = \hat{l}_j(D) - \frac{p_j}{2}\ln R \tag{1}$$

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 (2) 式により表される．

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

すると，確率は次で表される．

$$\hat{P}(x_i) = \frac{R_{(i)}}{R}\frac{1}{\sqrt{2\pi}\hat{\sigma}^d}\exp\left(-\frac{1}{2\hat{\sigma}^2}||\boldsymbol{x}_i - \boldsymbol{\mu}_{(i)}||^2\right) \tag{3}$$

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

$$\begin{aligned} l(D) &= \log\prod_i P(x_i) \\ &= \sum_i\left(\log\frac{1}{\sqrt{2\pi}\sigma^d} - \frac{1}{2\sigma^2}||\boldsymbol{x}_i - \boldsymbol{\mu}_{(i)}||^2 + \log\frac{R_{(i)}}{R}\right) \end{aligned} \tag{4}$$

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，(4) 式

は以下で表される．

$$\hat{l}(D_n) = -\frac{R_n}{2}\log(2\pi) - \frac{R_n \cdot d}{2}\log(\hat{\sigma}^2) - \frac{R_n - K}{2} + R_n \log R_n - R_n \log R \quad (5)$$

2.3 実験環境

実験には Python3.5 を用い，TensorFlow1.2.1 と呼ばれるオープンソースのライブラリを用いてアルゴリズムを実装した．

3. 結果

3.1 k-means によるクラスタリング

図 1 のデータをクラスタ数 3 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．

なお，クラスタリングの打ち切り条件は，セントロイドの差が 1.0×10^{-10} 以下のときとした．

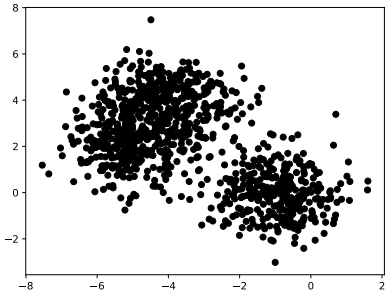

図 1　クラスタリング前のデータ

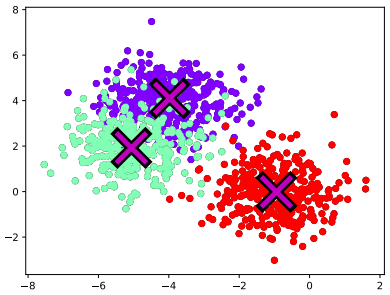

図 2　クラスタリング後のデータ

4. おわりに

5. 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) 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).

3) Gideon Schwarz et al.: Estimating the dimension of a model, The annals of statistics, Vol. 6, No.2, pp. 461–464 (1978).