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 22 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．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. 実験の経過

4 月から 6 月は機械学習と確率統計の基礎学習および Python の勉強を行った．6 月，7 月にクラスタリング手法である K-means と X-means の実装を行った．

3. 実験の手法

3.1 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．

3.2 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[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．

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

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

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

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．

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

3.3 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. 実験結果

4.1 K-means によるクラスタリング

図 1 のデータをクラスタ数 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．

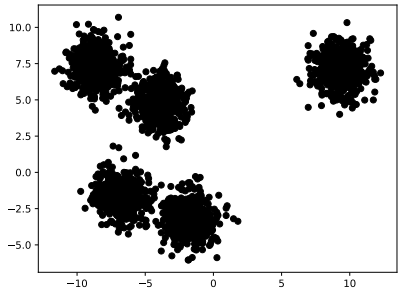

図 1　クラスタリング前のデータ

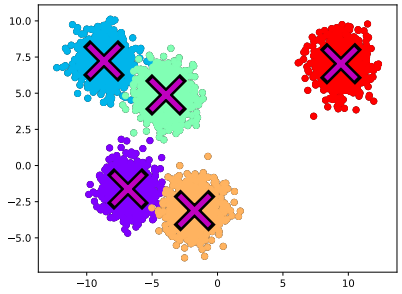

図 2　K-means によるクラスタリング結果

4.2 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．

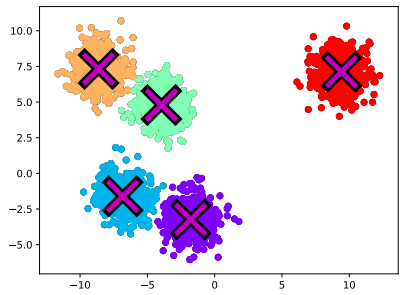

図 3　X-means によるクラスタリング結果

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．

6. 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).