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 8 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. 実験の経過と手法

4 月から 6 月は機械学習および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 k-means や x-means の実装を行った．

k-means 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．

x-means[1] は，k-means の逐次繰り返しと，BIC(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，クラスタ数を自動的に決定することができるクラスタリング手法である．

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

$\hat{l}(D)$ はデータに関する対数尤度関数を最大化したものである．$p/2$ はパラメータ空間の次元数であり，R はパラメータ数

$$\hat{\sigma}^2 = \frac{1}{R-K}\sum_i \left(x_i - \mu_{(i)}\right)^2 \tag{2}$$

$$\hat{P}(x_i) = \frac{R_{(i)}}{R}\frac{1}{\sqrt{2\pi}\hat{\sigma}^M}\exp\left(-\frac{1}{2\hat{\sigma}^2}||x_i - \mu_{(i)}^2\right) \tag{3}$$

x-means におけるクラスタ数推定は以下のように行われる．

1）クラスタ数を小さな値 (2 や 3) 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

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実装した．

3. 結果

3.1 k-means によるクラスタリング

4. おわりに

参考文献

[1] Dan Pelleg, Andrew W Moore, et al. “X-means: Extending K-means with Efficient Estimation of the Number of Clusters.” In: ICML. Vol. 1. 2000, pp. 727–734.