カーネル k-means 法におけるガウスカーネルのパラメータ推定

関谷　祐理　(指導教員：吉田裕亮)

1　はじめに

正解ラベルの付いていない観測データを, その散布に従っていくつかのクラスタに分類する手法をクラスタリングと呼ぶ. 正解ラベル, すなわち教師信号がないことから「教師なしの学習」,「自己組織化」などとも呼ばれる. クラスタリング手法は, 線形手法と非線形手法の 2 つに大きく類別され, 非線形手法ではカーネル法を用いたカーネル k-means などが有名である.

一般に, クラスタリングを行う際にはクラスタ数 K は既知として K 個のクラスタに分割する. しかしながら, 現実的な問題では K が未知である場合が多いため, 線形手法においては最適なクラスタ数を推定するには, AIC などの情報量基準を用いて評価される. また, カーネル法を用いた非線形手法においては, カーネル関数を用いて標本データを写像するため, カーネルの設定によりクラスタの構成が大きく変わる傾向にある. そのためカーネル k-means では, クラスタ数 K の推定とカーネル関数のパラメータ調整が問題とされている.

本研究では, カーネル k-means で非線形クラスタリングを行う際に, カーネルパラメータとクラスタ数 K を適切に選択できる評価基準のひとつを提案する. これは確率分布に基づくものではない評価基準である.

2　カーネル k-means 法

カーネル k-means 法とは, 線形分離不可な非線形データに対応したクラスタリング手法である. カーネル関数を用いて標本を高次元の特徴空間に写像し, 特徴空間上で k-means 法を実行する.

サンプル点集合 $X = \{x_1, x_2, \cdots, x_n\}$ が与えられ, その特徴ベクトルを $\phi(x_1), \phi(x_2), \cdots, \phi(x_n)$, 分類するクラスタを $C_1, C_2, \cdots, C_c$, クラスタ重心を $\mu_1, \mu_2, \cdots, \mu_c$ とする. ($|C_k|$ は k 番目のクラスタ要素数)

[2.1]　適当に c 個のクラスタ重心 $\mu_k (k = 1, 2, \cdots, c)$ を決める.

[2.2]　特徴空間での, 各データとクラスタ重心との距離をカーネル関数を用いて計算する.

$$\begin{aligned}\left\|\phi(x_i) - \mu_k\right\|^2 &= \left\|\phi(x_i) - \frac{1}{|C_k|}\sum_{x_j \in C_k}\phi(x_j)\right\|^2 \\ &= k(x_i, x_i) - \frac{2}{|C_k|}\sum_{x_j \in C_k} k(x_i, x_j) \\ &\quad + \frac{1}{|C_k|^2}\sum_{x_j \in C_k}\sum_{x_l \in C_k} k(x_j, x_l).\end{aligned}$$

[2.3]　[2.2] に基づいて C_k と μ_k を以下のように更新する.

$$C_k = \left\{x_i \mid \mu_k = \arg\min_{\mu_j} \|\phi(x_i) - \mu_j\|^2\right\},$$

$$\mu_k = \frac{1}{|C_k|}\sum_{x_j \in C_k}\phi(x_j).$$

[2.4]　ステップ [2.3] を収束するまで繰り返す.

なお本研究で扱うカーネルは, ガウスカーネル

$$k(x_i, x_j) = \exp(-\beta \|x_i - x_j\|^2).$$

である.

3　クラスタ評価基準

線形クラスタリングの評価基準として, 次の変動行列がある.

サンプル点集合 $X = \{x_1, x_2, \cdots, x_n\}$ が与えられたとき, 分類するクラスタを $C_1, C_2, \cdots, C_c$, クラスタ重心を $\mu_1, \mu_2, \cdots, \mu_c$ とする.

全変動及びクラス内, クラス間もバラツキは $n \times n$ 行列で表され, 以下の式のような関係がある.

$$T = W + B$$

$$T = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T \qquad \text{全変動行列}$$

$$W = \sum_{k=1}^{c}\sum_{x_i \in C_k}(x_i - \mu_k)(x_i - \mu_k)^T \qquad \text{クラスタ内行列}$$

$$B = \sum_{k=1}^{c}|C_k|(\mu_k - \bar{x})(\mu_k - \bar{x})^T \qquad \text{クラスタ間行列}$$

これらの関係を利用したクラスタ評価基準に, トレース評価基準式 $\min \mathrm{tr}W$ がある. しかしながら, これは

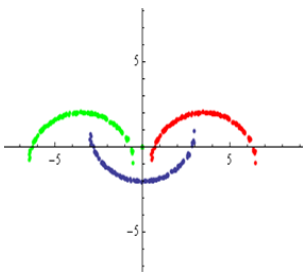
$$\mathrm{tr}W = \sum_{k=1}^{K} W_k = \sum_{k=1}\frac{1}{|C_k|}\sum_{j \in C_k}\sum_{i<j}\|x_i - x_j\|^2$$

となるので, クラスタ内のデータ間の平方和を最小にすることと同等になるため, 確定的な方法ではないといえる.

4　提案手法と結果

本研究では, 特徴空間上での変動行列をクラスタ評価基準に用いる.

N=450, 正解クラスタ数 K=3 の非線形データに対し, クラスタ数 k=3 でカーネル k-means を行う. 実行後, 以下の式から T, W を計算する. (μ は全サンプルの重心)



$$T = \sum_{i=1}^{n}\|\phi(x_i) - \mu\|^2$$

$$W = \sum_{k=1}^{c}\sum_{x_i \in C_k}\|\phi(x_i) - \mu_k\|^2$$

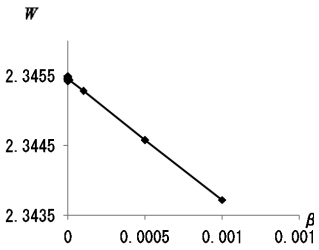


図 1: β と W

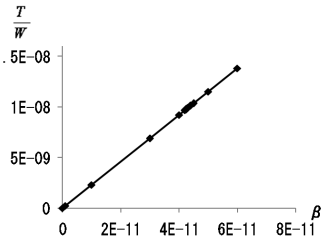


図 2: β と T/W

図 1, 図 2 より β の減少に伴い, W の値は減少, T/W の値は増加する傾向にあることが分かった. W, T/W の 2 つの値を組み合わせてパラメータの推定を試みる.

4.1 カーネルパラメータ β の推定

提案手法

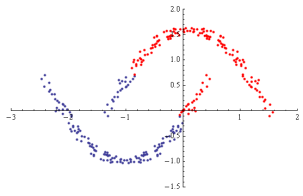
[A1] “正解クラスタ数 K=2 のデータ”と “正解クラスタ K=2 のデータ + 一様乱数によるノイズ”の 2 種類の標本データを用意する.

[A2] クラスタ数を正解のクラスタ数に固定し, 非線形クラスタリングを行う.

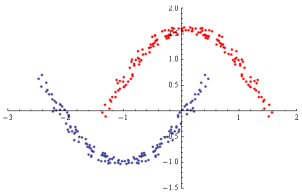
[A3] β と $W+T/W$ の値をグラフにプロットし, これらの関係を調べる.

結果

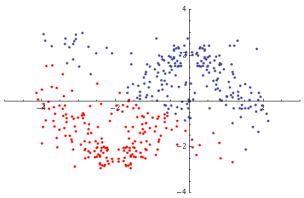
以下は手順 [A1]〜[A3] の中でクラスタリングを行った結果と, そのときの β の値である.



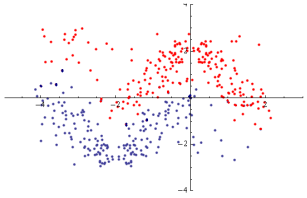
β=0.004 の結果



β=0.0036 の結果

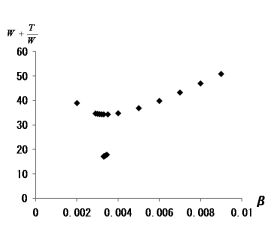


$\beta=9.1\times10^{-9}$ の結果

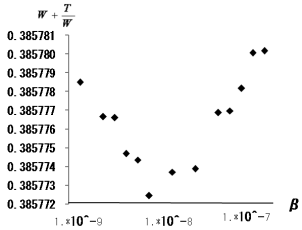


$\beta=8\times10^{-9}$ の結果

一様乱数によるノイズの有無に関わらず, 正しくクラスタリングが行われたときに T/W は最小値をとることが確認できた. そのため, 適切なクラスタリングを行うには, T/W を最小値にとるような β を選択すればよいと考えられる.



K=2 のデータ群



K=2 のデータ群と
一様乱数によるノイズ

4.2 クラスタ数 k の推定

提案手法

[B1] 正解クラスタ数 K=2〜5 の非線形データと楕円型データを用意する.

[B2] 各データに対し, クラスタ数 k=2〜6 の非線形クラスタリングを実行する.

[B3] クラスタ数 k の評価基式を $W+\frac{a}{k}\frac{T}{W}$ と仮定し, $k=K$ のときに評価式が最小になるような a の範囲を推定する.

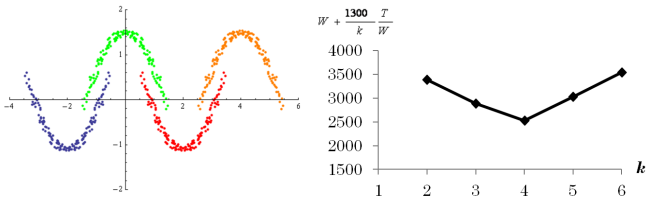
[B4] [B1] にない標本データを用意し, クラスタリングを行う. k=K のとき, [B3] で得られた評価式が最小値をとるか確認する.

結果

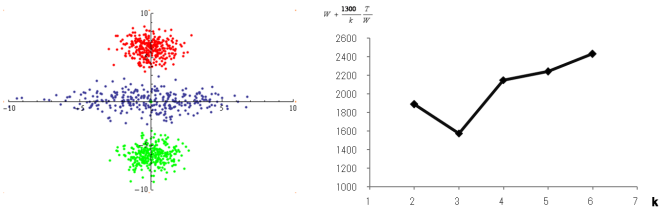
手順 [B1]〜 [B4] を行った結果 a=1300 と推定できたため, 評価式は $W+\frac{1300}{k}\frac{T}{W}$ と考えられる.

以下は, 求めた評価式を用いて非線形クラスタリングを行った結果である. k=K のとき, 評価式が最小値になるのを確認できた.

例 1:非線形データに対するクラスタ数 k の推定



例 2:線形データに対するクラスタ数 k の推定



5 まとめと今後の課題

本研究では, データ間距離を用いたクラスタ評価基準を利用し, カーネル k-means におけるクラスタ数 k とカーネルパラメータ β の推定方法を提案することができた.

しかしながら, カーネル k-means の欠点である初期依存や局所解, カーネル関数の選択の問題を十分考慮しなかったため, 本研究結果はそれらの問題を依然含んでいる可能性がある. 今後は, 初期状態に依存しないカーネル DP-means(非線形手法) などを用いたクラスタ評価基準を考案していきたい.

参考文献

1. 赤池昭太郎, カーネル多変量解析〜非線形データの新しい展開〜, 岩波書店, 東京, 2009.