2.5.1　カーネル密度推定法

８月１日

田中洸一

内容

- 問題提起
- 密度推定量の導出
- カーネル密度推定法
- ガウスカーネルによる確率密度モデル
- カーネル密度推定法の利点と欠点

問題提起

- ノンパラメトリック手法の1つとしてカーネル推定法を提示

　ヒストグラムアプローチを踏まえて密度推定量を導出

↓

カーネル密度推定を求め、考察

密度推定量の導出（1）

- すでに観測値の集合が得られているあるD次元のユークリッド空間の中から、未知の確率密度p(x)を推定したい。

※ちなみに、ユークリッド空間とは、n次元における二点
$\mathrm{a} = (a_1, a_2, \ldots, a_n)\ \mathrm{b} = (b_1, b_2, \ldots, b_n)$ に対してユークリッド距離、

$$d(\mathrm{a}, \mathrm{b}) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

が定められている距離空間のこと

密度推定量の導出（2）

ヒストグラム法を踏まえ、xを含む領域の確率

$$P = \int_{\Re} p(\mathrm{x}) d\mathrm{x} \qquad (2.242)$$

※RはxをふくむAる小さな領域

ここでデータ集合をNとすると、R内の点の総数Kは、
二項分布に従う

$$\mathrm{Bin}(K|N,P) = \frac{N!}{K!(N-K)!} P^K (1-P)^{N-K} \qquad (2.243)$$

密度推定量の導出（3）

二項分布の平均と分散（式は（２．１１）（２．１２））より、

データ点が領域内にある平均割合が

$$\mathrm{E}\left[K/N\right] = P$$

平均の周りの分散は

$$\mathrm{var}\left[K/N\right] = P(1-P)/N$$

よって、大きなNについて、この分布は平均の周囲で鋭く尖ったものとなり、以下のように示せる

$$K \simeq NP \qquad (2.244)$$

密度推定量の導出（4）

Rが確率密度p（x）がこの領域内でほぼ一定と
みなせるほど十分に小さいと仮定すれば、

$$P \simeq p(\mathrm{x})\mathrm{V} \qquad (2.245)$$

※Vは領域Rの体積

（2．244）と（2．245）より、

$$p(\mathrm{x}) = \frac{K}{N\mathrm{V}} \qquad (2.246)$$

これが密度推定量となる

密度推定量

（2．246）密度推定量について、

1．Kを固定し、データからVの値を推定

↓

K近傍法（後の章でやる）

2．Vを固定して、データからKを推定

↓

<u>カーネル推定法</u>（今からやる）

カーネル密度推定法（1）

以下を設定、定義

- 確率密度を求めたいデータ点をx
- xを中心とする小さな超立方体を領域R
- 領域内にある点の数Kを数えるカーネル関数

$$k(\mathrm{u}) = \begin{cases} 1, & |u_i| \le 1/2, \quad i = 1,\ldots,D\text{のとき} \\ 0, & \text{それ以外} \end{cases} \qquad (2.247)$$

(2.247) より $k\left((\mathrm{x}-\mathrm{x}_n)/h\right)$ はxを中心とする一辺がhの立方体の内部に、データ点x_nがあれば1に、そうでなければ0

カーネル密度推定法（2）

よって、立方体内部の総点数

$$K = \sum_{n=1}^{N} k\left(\frac{\mathrm{x} - \mathrm{x}_n}{h}\right) \qquad (2.248)$$

これを（2．246）に代入して、xでの推定密度が得られる

$$p(x) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k\left(\frac{\mathrm{x} - \mathrm{x}_n}{h}\right) \qquad (2.249)$$

※D次元超立方体V = h^D

ガウスカーネルによる確率密度モデル（1）

（2．249）でカーネル密度推定法を導いたが、ヒストグラム法で生じた問題（密度の区間が縁で不連続）が生じてしまう。

↓

より滑らかなカーネル関数を選び、より滑らかな密度モデルが必要

↓　解決策

カーネル関数をガウスカーネルに

ガウスカーネルによる確率密度モデル（2）

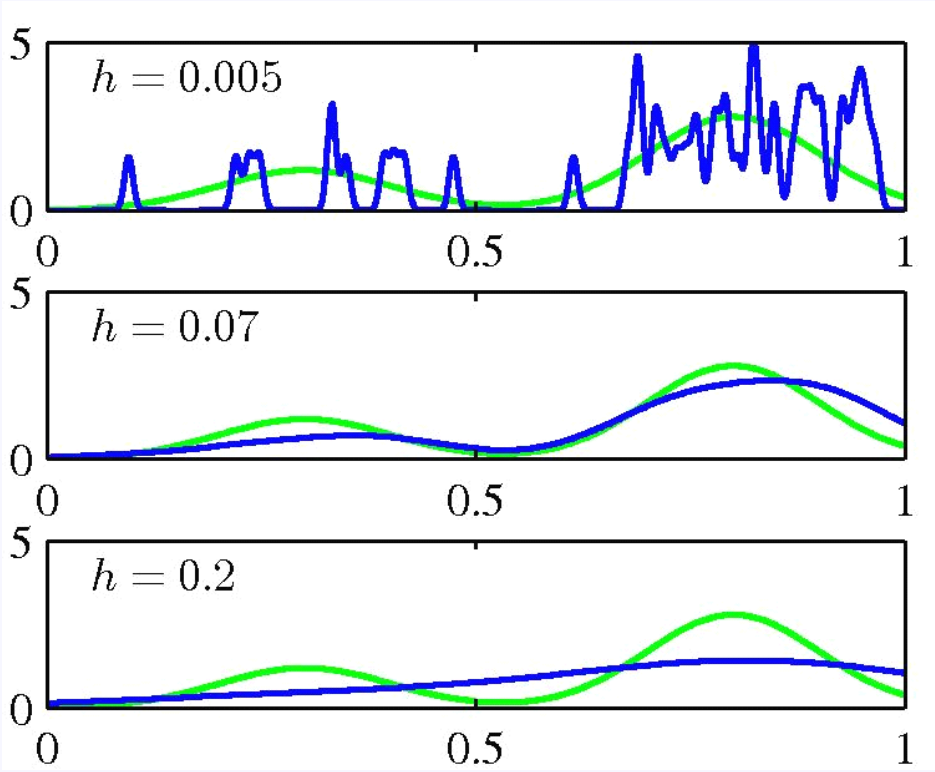
確率密度モデル

$$p(x) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{\left(2\pi h^2\right)^{D/2}}\exp\left\{-\frac{\|\mathrm{x}-\mathrm{x}_n\|^2}{2h^2}\right\} \quad (2.250)$$

※hはガウス分布の標準偏差

パラメータhが平滑化パラメータの役割を果たす（次のスライドにて例を提示）

ガウスカーネルによる
確率密度モデル(3)



←*h*が小さいとノイズが多い

←*h*が適切ならＯＫ！

←*h*が大きいと分布の
特徴がかき消される

hの最適化はモデル複雑度の問題

カーネル密度推定法の利点と欠点

利点

↓

「訓練」段階では、訓練集合を保存するだけ！
（計算必要なし）

欠点

↓

密度評価の計算コストが、
データ集合の大きさに比例！