カーネル密度推定法

氏 名： 西本　洋紀

2019 年 5 月 27 日

1　目的

ある分布 $p(\mathbf{x})$ に従って、N 個の観測値が得られたとする。この時、N 個の観測値から、分布 $p(\mathbf{x})$ を推定したい。

2　方法

$\mathbf{x}$ を含むある小さな領域 $\mathcal{R}$ を考える。この領域に割り当てられた確率 P は

$$P = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x} \tag{1}$$

と書ける。
領域 $\mathcal{R}$ のデータ点の総数 K は、二項分布に従う。

$$Bin(K|N,P) = \binom{N}{K} p^K (1-p)^{N-K} \tag{2}$$

よって、$\frac{K}{N}$ の分布の平均と分散は

$$\mathbf{E}[K/N] = P \tag{3}$$

$$var[K/N] = \frac{P(1-P)}{N} \tag{4}$$

となる事がわかる。
$N \to \infty$ のとき、$\frac{K}{N}$ の分布は平均周りで尖った形になり、

$$K \simeq NP \tag{5}$$

となる。
また、$\mathcal{R}$ が小さく、確率密度 $p(\mathbf{x})$ がこの領域内で一定とみなせるとしたら、

$$P \simeq p(\mathbf{x}) V \tag{6}$$

となる。ただし、V は、$\mathcal{R}$ の体積である。

V を固定し、K をデータから推定できたとすると、

$$p(\mathbf{x}) \simeq \frac{K}{NV} \tag{7}$$

により、$p(\mathbf{x})$ を推定できる。あとは K をデータから推定をする部分を考えれば良い。

確率密度を求めたい点を $\mathbf{x}$、この点を中心とする超立方体を $\mathcal{R}$ とする。また、観測値の集合を $\{\mathbf{x_n}\}$ とする。さらに、カーネル関数

$$k(\mathbf{u}) = \begin{cases} 1 & (\|u_i\| \leq \frac{1}{2}) \\ 0 & (o.w.) \end{cases} \tag{8}$$

を定義する。この方法では、$k(\frac{\mathbf{x}-\mathbf{x_n}}{h})$ は $\mathbf{x}$ を中心とする一辺が h の立方体 $\mathcal{R}$ の内部にデータ点があれば 1 に、そうでなければ 0 になる。

立方体 $\mathcal{R}$ 内のデータ点の総数は

$$K = \sum_{n=1}^{N} k(\frac{\mathbf{x}-\mathbf{x_n}}{h}) \tag{9}$$

と書け、$\mathbf{x}$ での推定密度は

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h^D} k(\frac{\mathbf{x}-\mathbf{x_n}}{h}) \tag{10}$$

となる。上式の確率密度のクラスは、カーネル密度推定法や **Parzen 推定法**と呼ばれる。

このとき、立方体の縁で、人為的な不連続が生じてしまう。より滑らかなカーネル関数を使えば、より滑らかな密度が得られる。カーネル関数 $k(\mathbf{u})$ には、以下の条件を満たす任意の関数を選ぶことができる。

$$k(\mathbf{u}) \geq 0 \tag{11}$$

$$\int k(\mathbf{u}) d\mathbf{u} = 1 \tag{12}$$

一般によく選ばれるカーネル関数はガウスカーネルで、次の確率密度モデルになる。

$$p(\mathbf{x}) = \frac{1}{(2\pi h^2)^{D/2}} exp\{-\frac{(\|\mathbf{x}-\mathbf{x_n}\|)^{\mathbf{2}}}{2h^2}\} \tag{13}$$

ただし、h はガウス分布の標準偏差を表す。パラメータ h は平滑化パラメータの役割を果たし、h が小さすぎると、結果はノイズの多い密度モデルになるが、大きすぎると、データの局所的な性質が捉えにくくなる。最良の密度は、中間的な h の値で得られる。