第12章 カーネル密度推定法

12.3 パーゼン窓法とカーネル密度推定法
12.4 尤度公差確認法

07t4072f　岡崎　駿

概要

領域Rの体積Vを固定

領域Rに含まれる訓練標本の数kを
訓練標本から求める

パーゼン窓法(1)

- 前提条件
  - 領域R としてある点 x を中心とする一辺の長さが b の超立方体を使用
  - パターン空間の次元: d
  - 領域 R に含まれる訓練標本の数k
  - 領域 R の体積: V

パーゼン窓法（2）

- 領域R の体積

$$V = b^d$$

- 領域R に含まれる訓練標本の数k

$$k = \sum_{i=1}^{n} W\left(\frac{\mathbf{x} - \mathbf{x}_i}{b}\right)$$

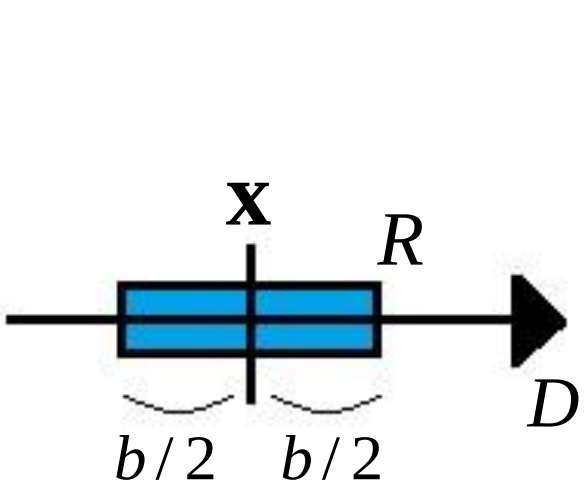
$$W(\mathbf{x}) = \begin{cases} 1 & \max_{i=1,\ldots,d} \left|x^{(i)}\right| \le \frac{1}{2} \\ 0 & \text{それ以外} \end{cases}$$

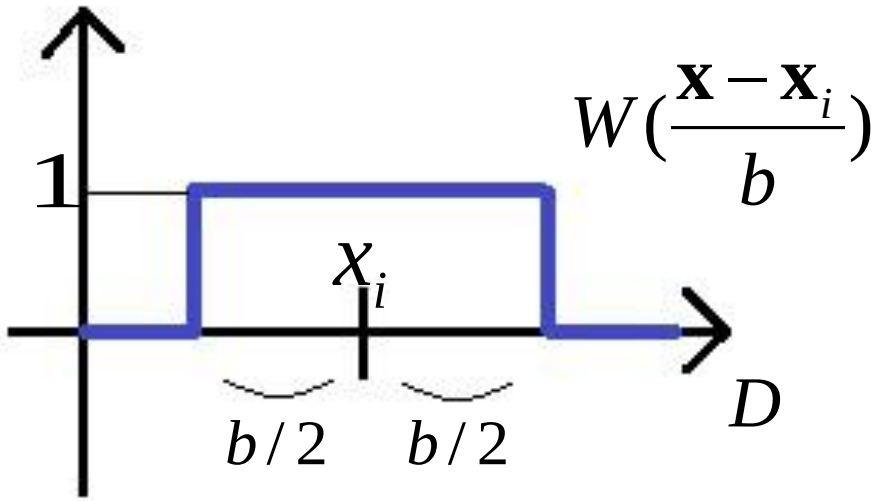
$$\mathbf{x} = (x^{(1)}, \ldots, x^{(d)})^T$$

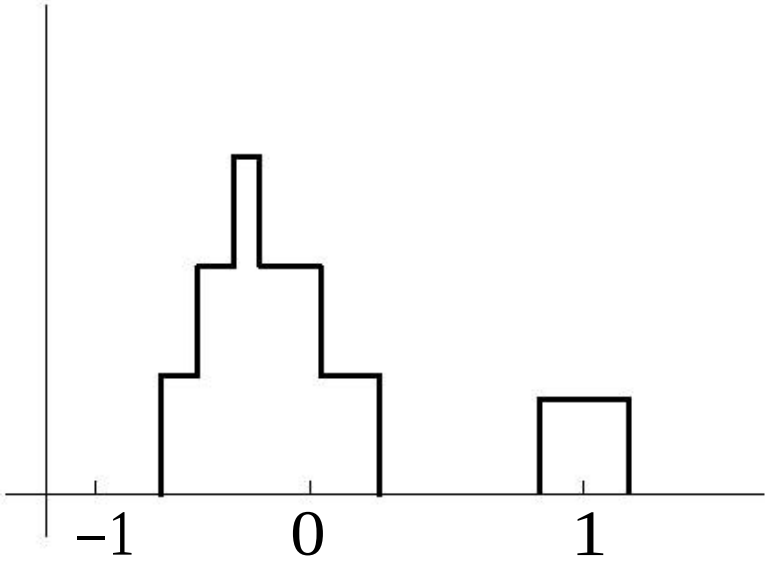
- 推定量

$$\hat{p}_{Parzen}(\mathbf{x}) = \frac{1}{nb^d} \sum_{i=1}^{n} W\left(\frac{\mathbf{x} - \mathbf{x}_i}{b}\right)$$

パーゼン窓法(3)







パーゼン窓法(4)

- 長所
  - 各ビンの幅が訓練標本から適応的に決定
- 短所
  - 推定結果が領域のつなぎ目で不連続になる

カーネル密度推定法で解決可能！

カーネル密度推定法（1）

- 特徴
  - パーゼン窓関数の代わりにカーネル関数を用いる
- 推定量

$$\hat{p}_{KDE}(\mathbf{x}) = \frac{1}{nb^d}\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{b}\right) \qquad K(\mathbf{x}) \geq 0 \ for\ any\ \mathbf{x} \in D, \int_D K(\mathbf{x})d\mathbf{x} = 1$$

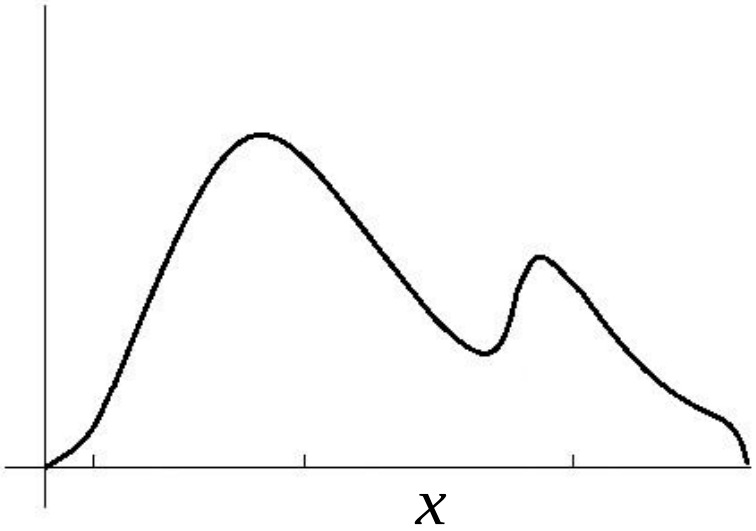
- カーネル関数
  - よく使われるのは「ガウスカーネル」

$$K(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2}\right)$$

  - バンド幅（領域の長さ）はガウス関数の標準偏差に対応

カーネル密度推定法(2)

- 結果



滑らかな確率密度関数推定量が得られる

だが、バンド幅を適切選ばないと良い近似にならない

尤度交差確認法（1）

- 目的
  - 最適なバンド幅のモデルを決定する
- アルゴリズムの概要
  - 訓練標本を「推定用」と「確認用」に分割
  - 推定用：確率密度関数を推定
  - 確認用：その尤度を評価
  - バンド幅の値の候補をいくつか用意し、最適なものを選択する

尤度交差確認法（2）

単純に「推定用」と「確認用」に分割しただけでは、標本の分割方法にモデル選択が依存する

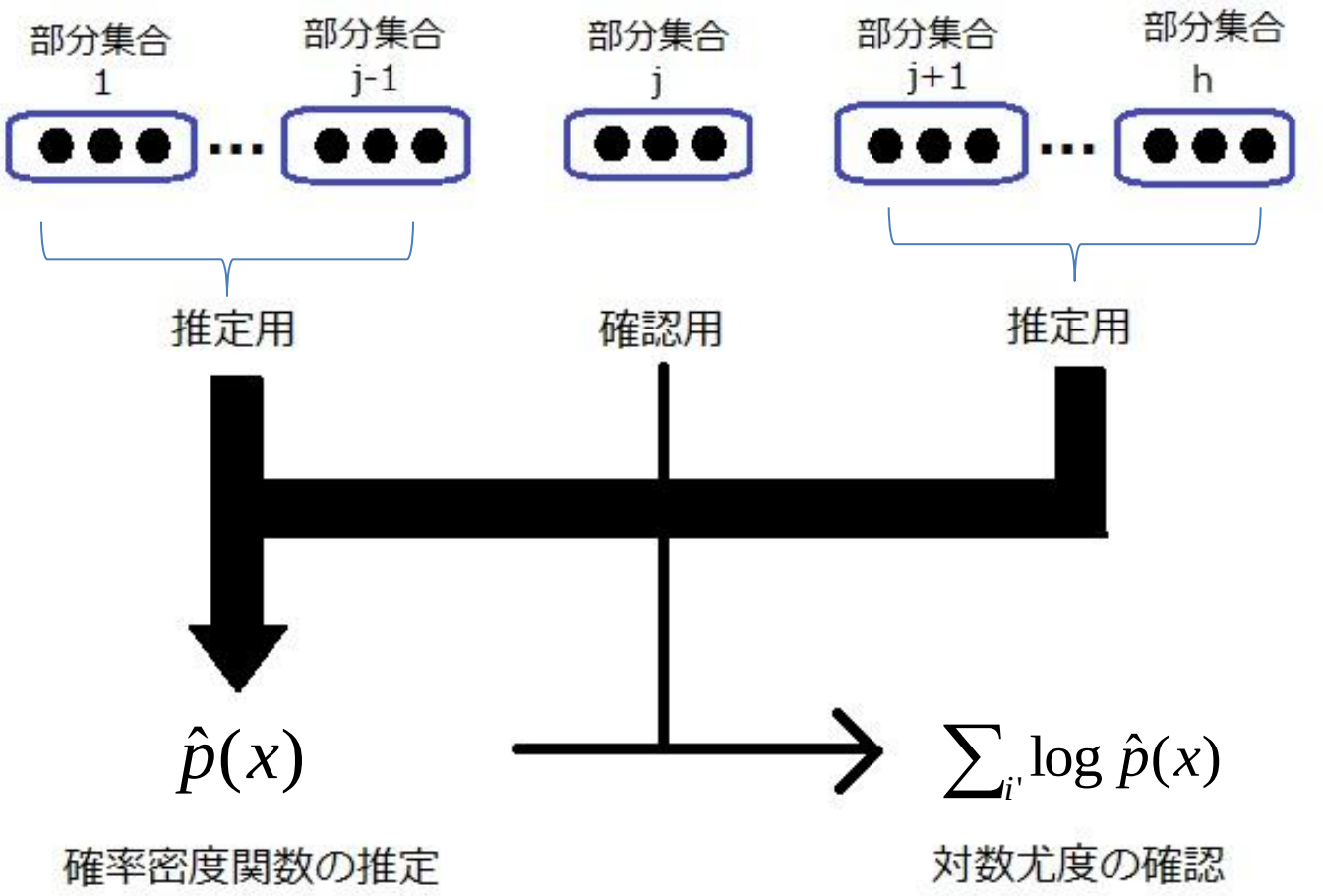
1.訓練標本を、h 個の重なりのない同じ大きさの部分集合に分割する

2. $(h-1)$個の標本集合を「推定用」とし、残りの1つを「確認用」とする。

3.全ての組み合わせに対して尤度の評価を行う

（ $(h-1)$個の標本集合の組み合わせは h 個）

尤度交差確認法(3)


部分集合
1
部分集合
j-1
部分集合
j
部分集合
j+1
部分集合
h
推定用
確認用
推定用
$\hat{p}(x)$
$\sum_{i'} \log \hat{p}(x)$
確率密度関数の推定
対数尤度の確認

尤度交差確認法（4）

- アルゴリズム

  1.バンド幅行列の候補をいくつか用意する: $\{\mathbf{B}_j\}_j$

  2.訓練標本$\{\mathbf{x}_i\}_{i=1}^n$ をh 個の重なりのない同じ大きさの部分集合に分割する: : $\{\mathbf{x}_i\}_{i=1}^n = \{\chi_j\}_{j=1}^h$

  3.全てのモデル $\mathbf{B}_j$ に対して以下を繰り返す

  (a)$\ell = 1,\ldots,h$ に対して以下を繰り返す

  i. ℓ 番目の部分集合を除いた訓練標本$\{\chi_i\}_{i\neq\ell}$ を用いて、モデル $\mathbf{B}_j$ に対する確率密度関数の推定 $\hat{p}_j^{(\ell)}(\mathbf{x})$ を求める。

  ii. ℓ 番目の部分集合の訓練標本 χ_ℓ を用いて、$\hat{p}_j^{(\ell)}(\mathbf{x})$ の対数尤度 $J_j^{(\ell)}$ を評価する

$$J_j^{(\ell)} := \sum_{\mathbf{x}' \in \chi_\ell} \log \hat{p}_j^{(\ell)}(\mathbf{x}')$$

尤度交差確認法(５)

(b)モデル $\mathbf{B}_j$ に対する対数尤度の評価値の平均を求める

$$J_j := \frac{1}{h}\sum_{\ell=1}^{h} J_j^{(\ell)}$$

4.対数尤度の平均評価値が最大のモデル $B_{\hat{j}}$ を選ぶ。

$$\hat{j} := \arg\max_j J_j$$

5.選ばれたモデル $B_{\hat{j}}$ に対してすべての訓練標本$\{\mathbf{x}_i\}_{i=1}^n$を用いて確率密度関数の推定 $\hat{p}(\mathbf{x})$ を求める

尤度交差確認法の評価（1）

- カルバック・ライブラー情報量を考える

$$KL(p\|\hat{p}) := E_x\left[\log\frac{p(\mathbf{x})}{\hat{p}(\mathbf{x})}\right]$$
$$= E_x[\log p(\mathbf{x})] - E_x[\log \hat{p}(\mathbf{x})]$$

E_xは$\mathbf{x}$に対する期待値を表す

$$E_x[f(\mathbf{x})] := \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

第1項目は定数なので無視し、第2行目をJで表す。

$$J := -E_x[\log \hat{p}(\mathbf{x})]$$

  - 結果、モデル選択はカルバック・ライブラー情報量を近似的に最小にするモデルを選ぶことに対応

尤度交差確認法の評価(2)

- 結論
  - 「交差確認法」と「赤池の情報量基準」は訓練標本数が十分に多いと等価である
  - 訓練標本数が少ない場合は「交差確認法」の方が良い結果が出る
  - 「交差確認法」は任意の距離尺度、任意の推定法にも適応可能

    ↳ 「汎用的なモデル選択手法」である！

  - 交差確認法は繰り返し推定を行うため、計算時間が多くかかる