

情報認識

「最近傍密度推定法(第13章)」

- 担当教員： 杉山 将（計算工学専攻）
- 居室： W8E-505
- 電子メール： sugi@cs.titech.ac.jp

「情報認識」の全体構成

168

- 識別関数のよさを測る規準
- 条件付き確率の推定
 - パラメトリック法
 - 最尤推定法, EMアルゴリズム
 - ベイズ推定法, 最大事後確率推定法
 - ノンパラメトリック法
 - カーネル密度推定法
 - 最近傍密度推定法
- 手書き文字認識の計算機実習

ノンパラメトリック法の表記 169

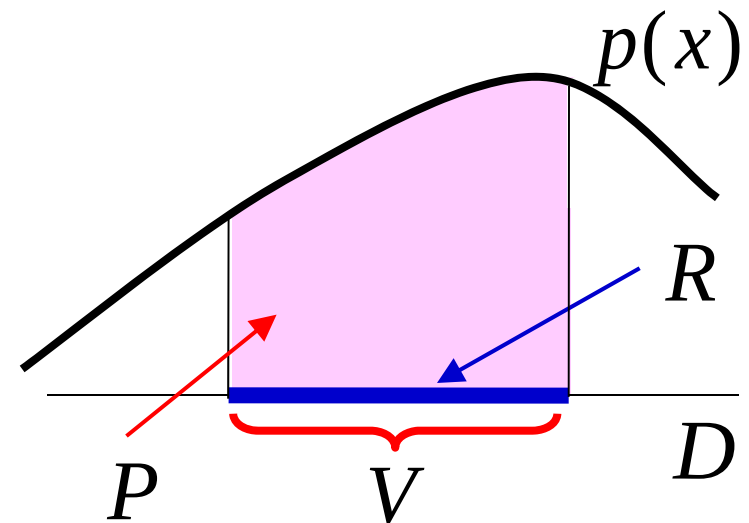
- ある注目点 x' での確率密度 $p(x')$ を推定する
- $R: x'$ を含むパターン空間 D 内のある領域(region)
- $V: R$ の体積(volume)

$$V = \int_R dx$$

- $P: \text{あるパターン } x \text{ が } R \text{ に入る確率}$

$$P = \int_R p(x) dx$$

- $k: n$ 個の訓練標本のうち
 R に入っている個数



ノンパラメトリック法の基礎

170

- 確率 P を二つの方法で近似する.

A) k, n を用いれば,

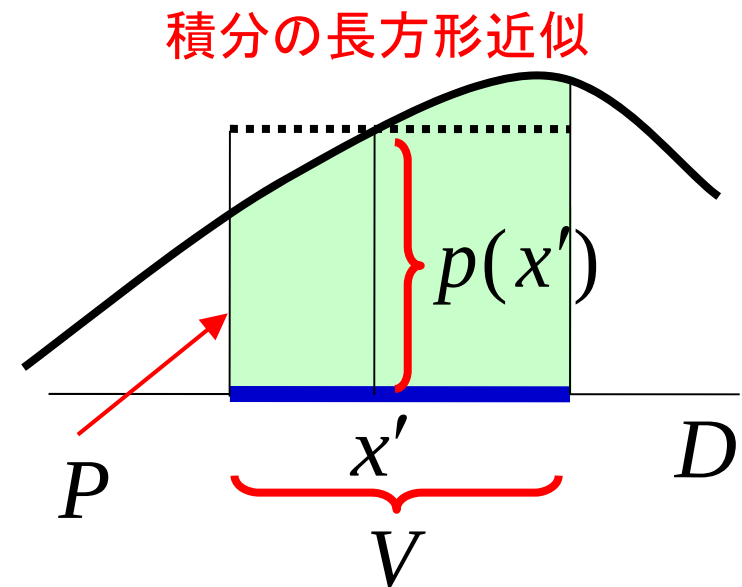
$$P \approx k/n$$

B) 領域 R 内のある点 x' を用いれば,

$$P \approx Vp(x')$$

- これらより

$$p(x) \approx \frac{k}{nV}$$



$$p(x) \approx \frac{k}{nV}$$

■ 訓練標本を用いて領域 R を決める.

- パーゼン窓法, カーネル密度推定法:

R の形を決め V を固定したもとで k を標本から決定

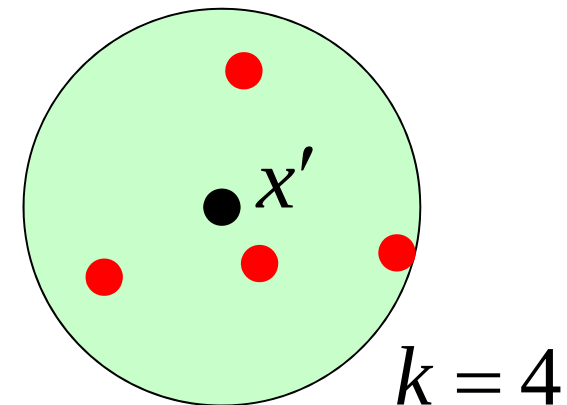
- 最近傍密度推定法:

R の形を決め k を固定したもとで V を標本から決定

■ k-最近傍密度推定法 (k-nearest neighbor density estimation):

- 領域 R として, ある点 x' を中心とする超球 (hypersphere) を用いる.
- 超球の半径 r : 訓練標本が k 個含まれる最小の大きさに設定
- 超球の体積:

$$V = \frac{\pi^{\frac{d}{2}} r^d}{\Gamma(\frac{d}{2} + 1)}$$



- $$\hat{p}(x) = \frac{k\Gamma(\frac{d}{2} + 1)}{n\pi^{\frac{d}{2}} r^d}$$

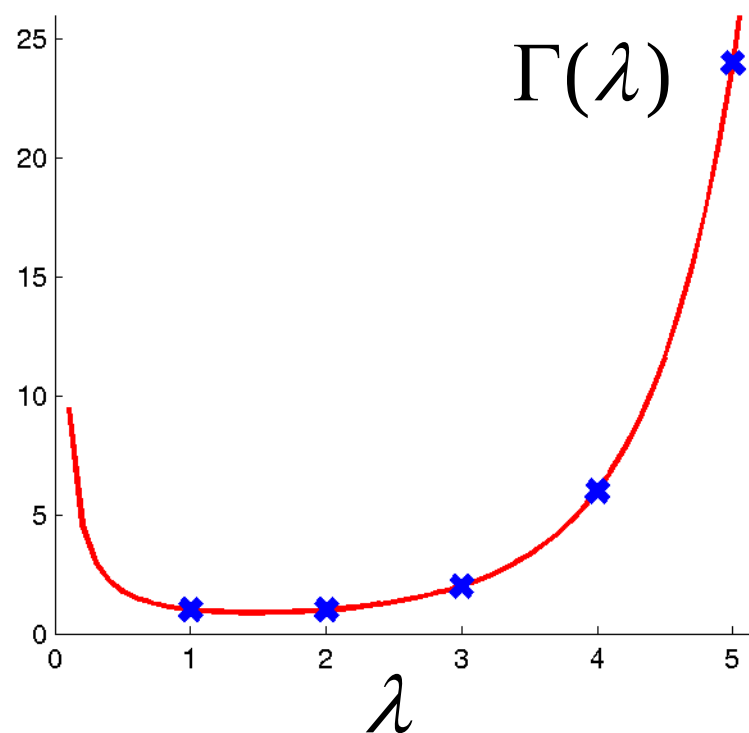
$$p(x) \approx \frac{k}{nV}$$

ガンマ関数

173

■ ガンマ関数(gamma function):

$$\Gamma(\lambda) = \int_0^{\infty} x^{\lambda-1} e^{-x} dx$$



ガンマ関数の性質

174

$$\Gamma(\lambda) = \int_0^{\infty} x^{\lambda-1} e^{-x} dx$$

■ 階乗の一般化: 正の整数 n に対して

$$\Gamma(n+1) = n!$$

■ その他の性質

- $\Gamma(t) = (t-1)\Gamma(t-1), \forall t \in \mathbf{R}$
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

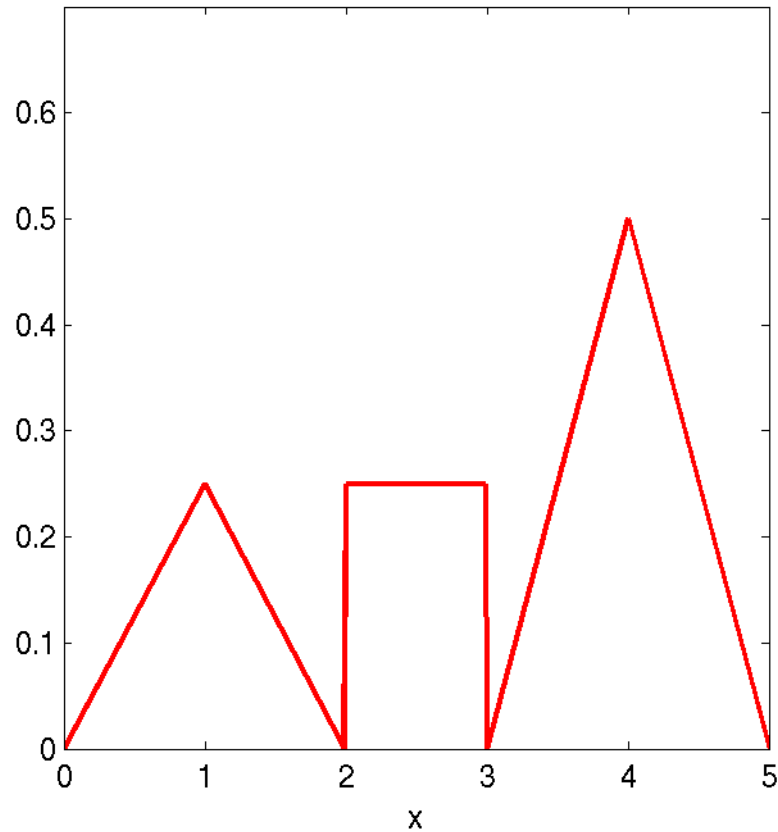
■ 演習:

- $d = 2$ のとき $V = \pi r^2$
- $d = 3$ のとき $V = \frac{4}{3} \pi r^3$

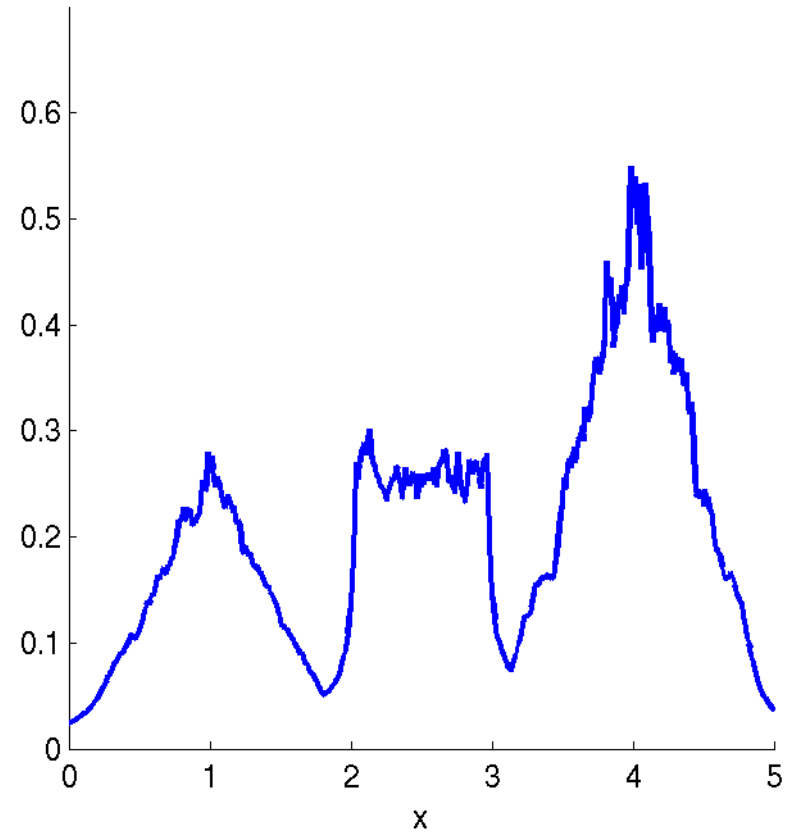
$$V = \frac{\pi^{\frac{d}{2}} r^d}{\Gamma(\frac{d}{2} + 1)}$$

最近傍密度推定法の例

175



真の確率密度関数



1-最近傍密度推定法で
推定した確率密度関数

ノンパラメトリック法のまとめ 176

■カーネル密度推定法

- 滑らかなカーネルを使えば, 滑らかな確率密度推定量が得られる
- 計算が簡単

■最近傍密度推定法

- 近傍の標本を見つけるためには距離をソートする必要があり, 大規模データに対しては計算時間がかかる
- 得られる確率密度推定量は比較的ギザギザしている?
- パターン認識との相性がよい(次ページ参照)

条件付き確率の推定

177

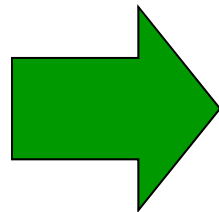
- 各カテゴリに対して、条件付き確率 $p(x|y)$ を 1-最近傍密度推定法により推定.

$$\hat{p}(x|y) = \frac{\Gamma(\frac{d}{2} + 1)}{n_y \pi^{\frac{d}{2}} r_y^d}$$

r_y : カテゴリ y に属する標本のうち x に最も近いものと, x との距離

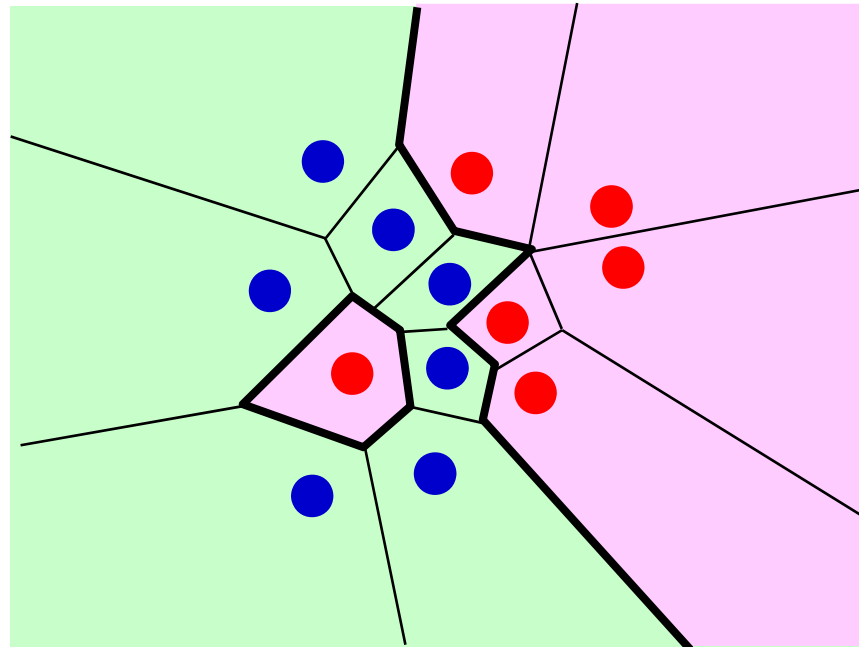
- $p(y) \approx n_y / n$ と事後確率 $p(y|x)$ は,

$$p(y|x) \propto p(x|y)p(y) \approx \frac{\Gamma(\frac{d}{2} + 1)}{n_y \pi^{\frac{d}{2}} r_y^d} \frac{n_y}{n} \propto \frac{1}{r_y^d}$$



r_y が小さいほうが
事後確率が大きい！

- 事後確率が最大のカテゴリ
= x に一番近い訓練標本が属するカテゴリ
- このような識別法を, 最近傍識別器(nearest neighbor classifier)とよぶ.



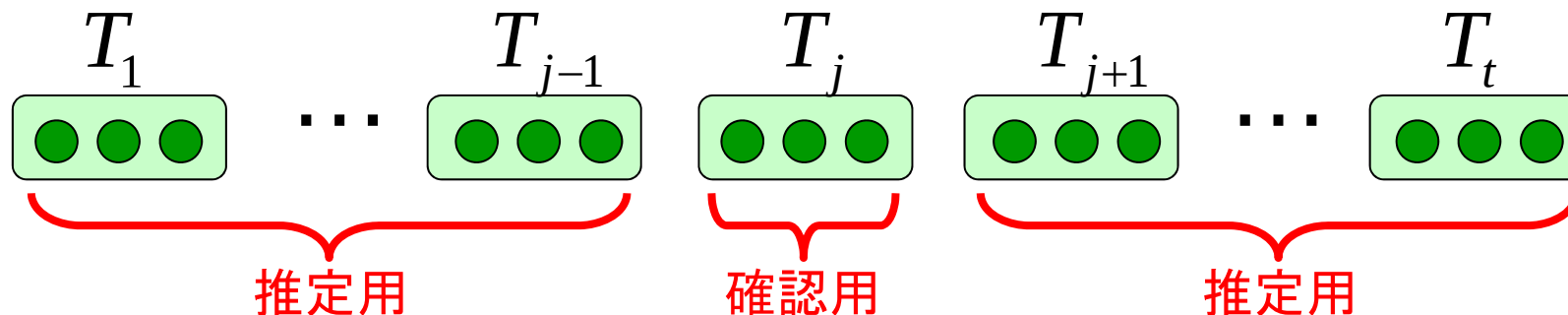
- 実用的には, x の近傍 k 個の訓練標本が属するカテゴリの多数決で, x の属するカテゴリを決める k -最近傍識別器(k-nearest neighbor classifier)がよく用いられる.

- k -最近傍識別器では近傍数 k を適切に決める必要がある.
 1. 値の候補を用意する. 例えば, $k = 1, 2, \dots, 10$
 2. それぞれのモデルに対して, **パターンの誤認識率**を推定する.
 3. 誤認識率の推定値を最小にするモデルを選ぶ.

- どうやってパターンの誤認識率を推定するか？

■ 交差確認法(cross validation)

- 訓練標本 $\{x_i\}_{i=1}^n$ を t 個の重なりの無い, (ほぼ) 同じ大きさの部分集合 $\{T_i\}_{i=1}^t$ に分ける.
- j 番目の部分集合 T_j に含まれる訓練標本を使わずに識別器を学習する.
- T_j に含まれる標本の誤認識率を計算する(訓練標本なので答えを知っている!).
- これを全ての j に対して繰り返し平均する.



今後の予定

182

- 今週は宿題なし
- 12月12日: 計算機演習
 - ガウス混合モデルに対するEMアルゴリズム
- 12月19日: 計算機演習
 - ノンパラメトリック法を用いた手書き文字認識
- 計算機演習は出欠を取らないので, 各自好きな時間に演習を行なっても良い. 課題はOCWにて事前に公開する
- 1月16日: 通常の講義
 - ベイズ推定