情報認識
「最近傍密度推定法」

- 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

ノンパラメトリック法の基礎

- 推定したい確率密度関数 $p(x)$ が R 内でほぼ一定値を取るとき，確率 P は領域 R 内のある点 x を用いて

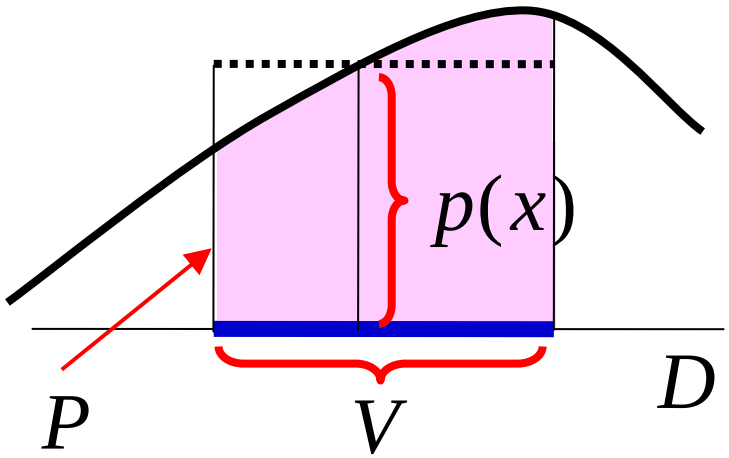
$$P \approx Vp(x)$$

- k ：n個の訓練標本のうち R に入っている個数

$$P \approx k/n$$

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

積分の長方形近似



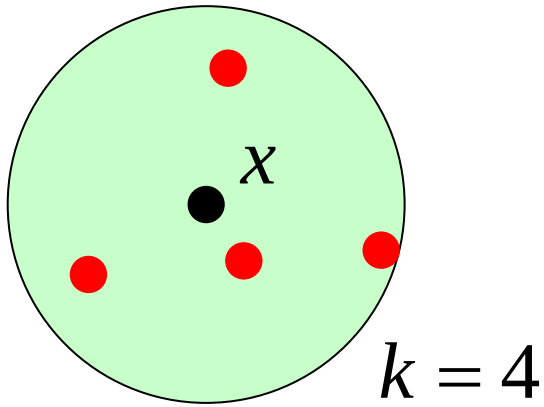
ノンパラメトリック法

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

- 訓練標本を用いて領域 R を決める.
  - パーゼン窓法, 核密度推定法:
    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:
    k を固定したもとで V を標本から決定

超球

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



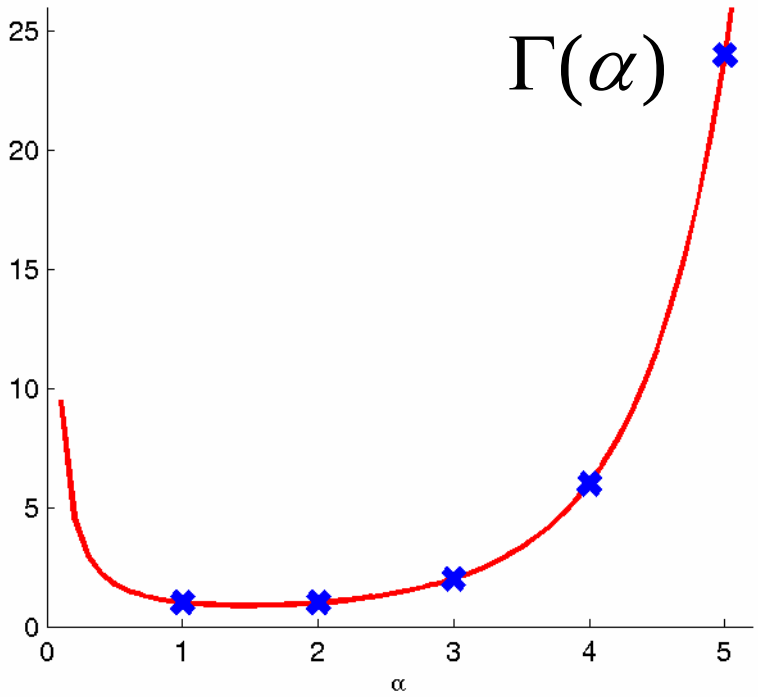
- 超球の体積：

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

ガンマ関数

■ **ガンマ関数**(gamma function)：

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



ガンマ関数の性質

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

- **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**
  - $\Gamma(z) = (z-1)\Gamma(z-1)$
  - $\Gamma(\frac{1}{2}) = \sqrt{\pi}$
- **これより，**

$d = 2$ **のとき** $V = \pi h^2$

$d = 3$ **のとき** $V = \frac{4}{3}\pi h^3$

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

最近傍密度推定法

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

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

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

kの決定

- k は領域 R に含まれる標本数なので，標本全体の数 n が変化したとき，k も相対的に変えるのが自然.
- 標本数 n が無限に増えていったとき，最近傍密度推定法によって推定した $\hat{p}(x)$ が本当の $p(x)$ に収束するための十分条件：

$$\lim_{n\to\infty} k = \infty \quad \text{かつ} \quad \lim_{n\to\infty} \frac{k}{n} = 0, \qquad \text{例えば} \quad k = \sqrt{n}$$

- しかし実際には，**最小二乗交差確認法**などを用いて，データから k を決めることが多い.

条件付き確率の推定

- 全てのカテゴリに対して，条件付き確率 $\hat{p}(x \mid \omega_i)$ を $k=1$ で推定．

$$\hat{p}(x \mid \omega_i) = \frac{\Gamma(\frac{d}{2}+1)}{n_i \pi^{\frac{d}{2}} h_i^d}$$

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

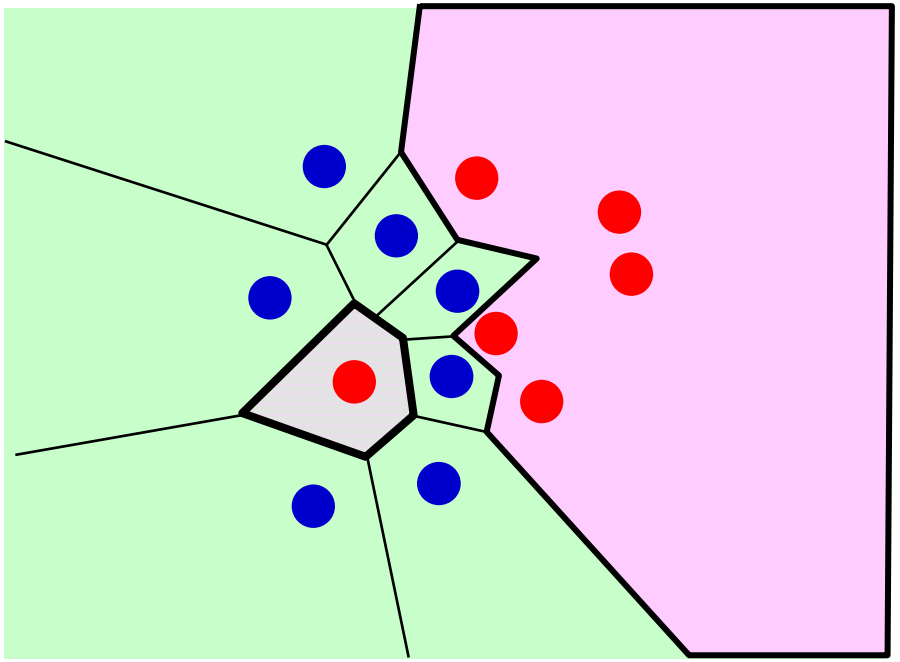
- $p(\omega_i \mid x) \propto p(x \mid \omega_i) p(\omega_i)$ と $p(\omega_i) \approx n_i / n$ より，事後確率は近似的に次式に比例

$$\frac{\Gamma(\frac{d}{2}+1)}{n_i \pi^{\frac{d}{2}} h_i^d} \frac{n_i}{n} \propto \frac{1}{h_i^d}$$

- h_i が小さいほうが事後確率が大きい！

最近傍識別機

- 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，**最近傍識別機**(nearest neighbor classifier)とよぶ．



k-最近傍識別機

- 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 -最近傍識別機(k-nearest neighbor classifier)がよく用いられる.

小レポート：計算機実験

- 最近傍密度推定法を実装し，適当なデータを用いて確率密度関数を推定せよ．
- 最小二乗交差確認法を用いて近傍数を決定せよ．
- データ標本数，真の確率分布，近傍数などの条件を変化させたとき，どのように推定結果が変わるかを考察せよ．
- 核密度推定法の推定結果と比較せよ．
- メ切：試験の日(2月6日) の試験開始前

生成モデル vs. 識別モデル

- 訓練標本は事前確率や条件付き確率に従って生成されているため，訓練標本からこれらの確率を推定する方法は**生成モデル**(generative model)**に基づく方法**とよばれている．
- 一方，生成モデルは直接推定せずに，決定境界を直接モデル化する方法もあり，これは**識別モデル**(discriminative model)に基づく方法とよばれている．

Vapnikの法則

- ある問題を解くとき，その問題よりも難しい問題を途中段階で解いてはならない
- 生成モデルの推定 (実世界を知る)
- 識別モデルの構成 (予測する)
- Transduction：与えられたテストパターンのカテゴリだけを予測する
- Selection：与えられたテストパターンの中から，あるクラスに属するパターンを選ぶ

試験(2月6日3，4限) の予告

- 専門用語の英語名
- 授業で学んだ内容に関する記述問題，例えば
  - 識別関数のよさを測る規準
  - 最尤推定法，モデル選択
  - ベイズ推定法，経験ベイズ法
  - ノンパラメトリック法
- 半年間パターン認識について学んだ経験に基づいた，自分独自のパターン認識の応用例(第1回目のレポート) についての再考