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

- 確率 P を二つの方法で近似する.

A) k, n を用いれば,

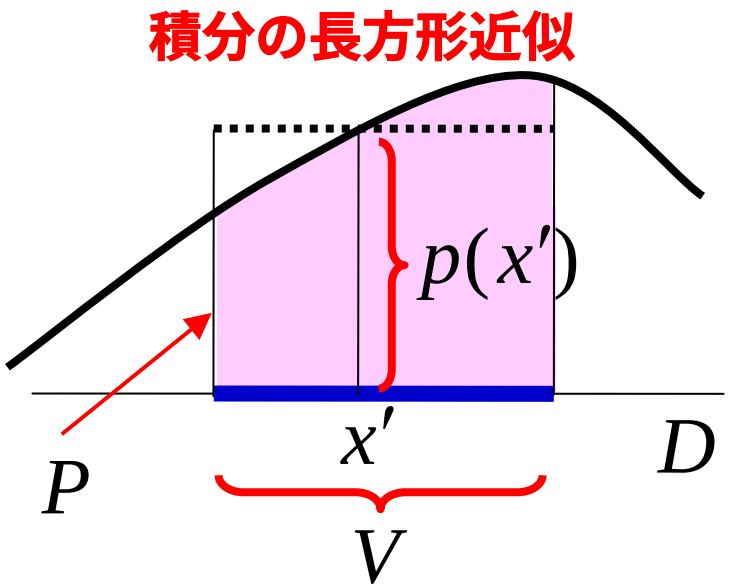
$$P \approx k/n$$

B) 領域 R 内のある点 x' を用いれば,

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

- これらより

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



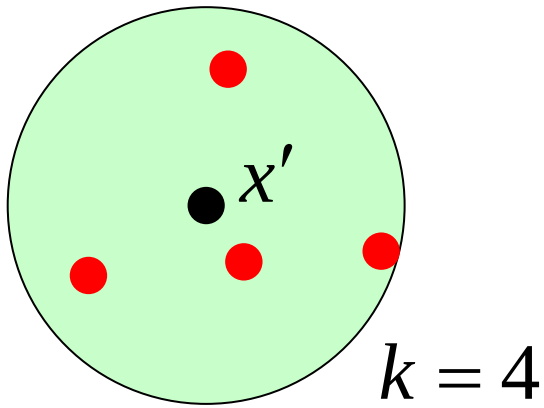
ノンパラメトリック法

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

- 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



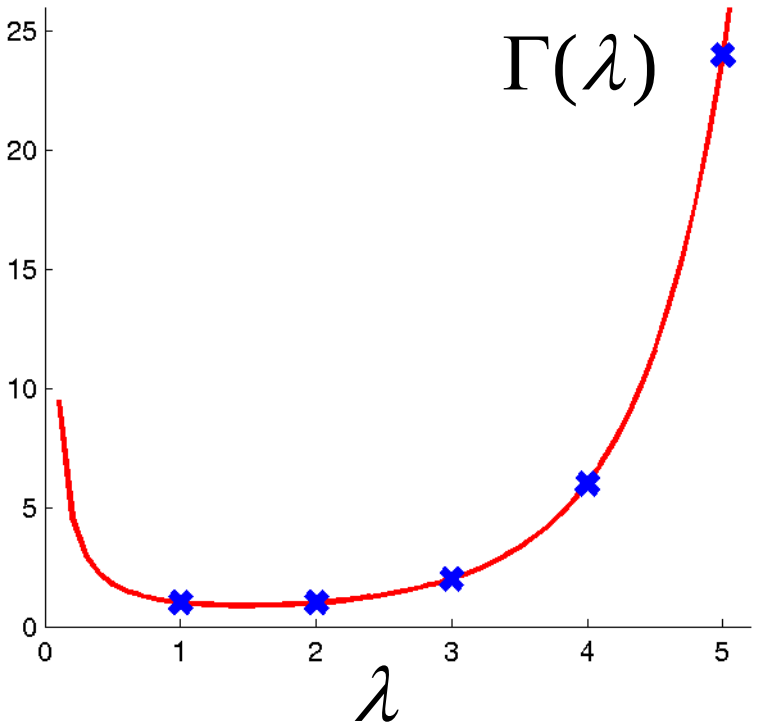
- 超球の体積：

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

ガンマ関数

■ **ガンマ関数**(gamma function)：

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



ガンマ関数の性質

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

- **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**
  - $\Gamma(z) = (z-1)\Gamma(z-1)$
  - $\Gamma(\frac{1}{2}) = \sqrt{\pi}$
- **演習：**
  - $d = 2$ **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)：

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

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

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 $\hat{p}(x \mid y)$ を $k=1$ で推定．

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

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

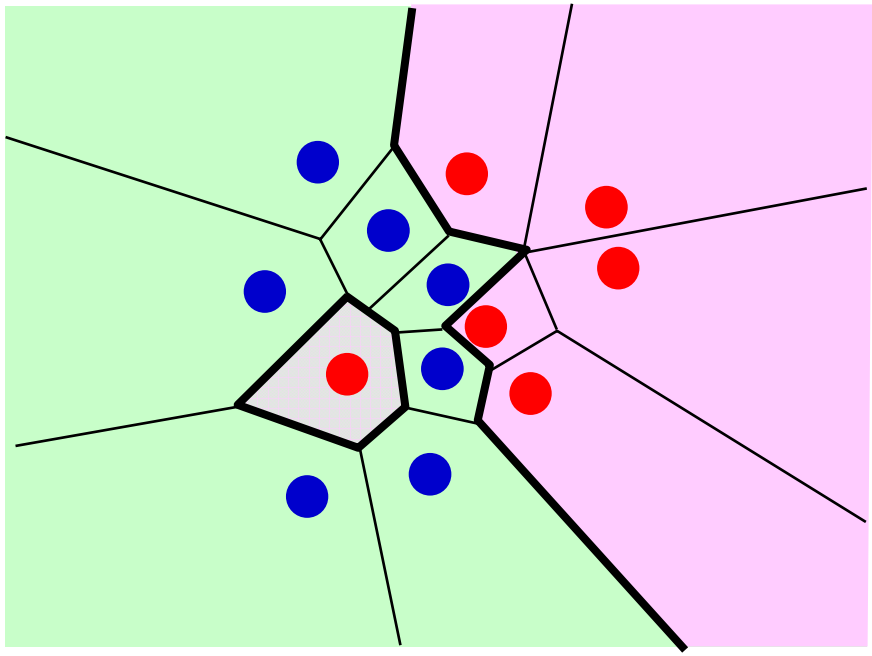
- $p(y \mid x) \propto p(x \mid y) p(y)$ と $p(y) \approx n_y / n$ より，

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

h_y が小さいほうが事後確率が大きい！

最近傍識別器

- 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．



k-最近傍識別器

- 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.

近傍数kの決め方

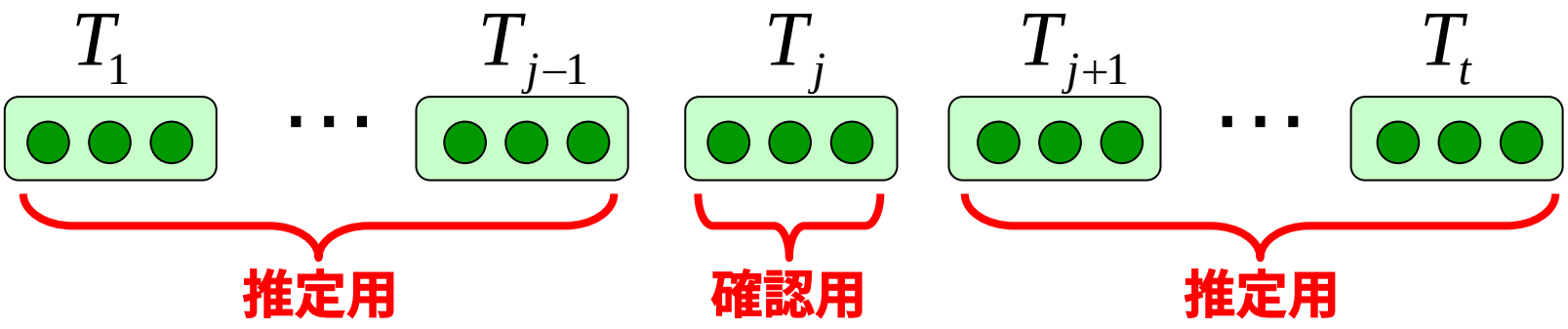
- k-最近傍識別器では近傍数kを適切に決める必要がある.

1. 値の候補を用意する. 例えば, $k = 1,2,\ldots,10$
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

- 訓練標本 $\{x_i\}_{i=1}^{n}$ を t 個の重なりの無い部分集合 $\{T_i\}_{i=1}^{t}$ 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 に対して行ない平均する.



連絡事項

- 来週（12月3日），再来週（12月10日）：出張のため講義はなし．代わりに第8回のレポート課題の計算機実習を各自で行なうこと．レポートは12月17日の講義の最初に提出すること．

小レポート (第8回)

- 1次元の入力に対する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 (第7回小レポート参照) .
- データ標本数，真の確率分布，核関数，幅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.

- 余力のある学生は，入力が2次元の場合に対しても同様の実験を行え．また，次元が増えたことによりどのような変化が生じたかを考察せよ.