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,

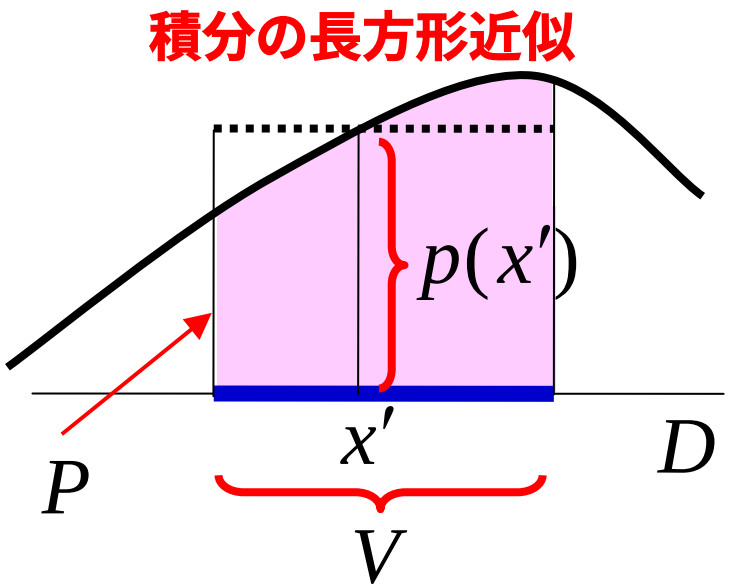
$$P \approx k/n$$

B) 領域 R 内のある点 x' を用いれば,

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

- これらより

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



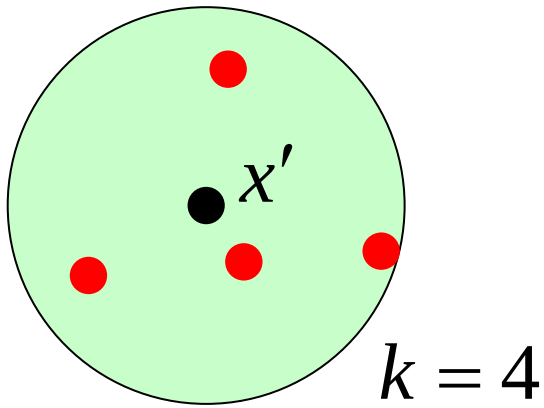
ノンパラメトリック法

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

- 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



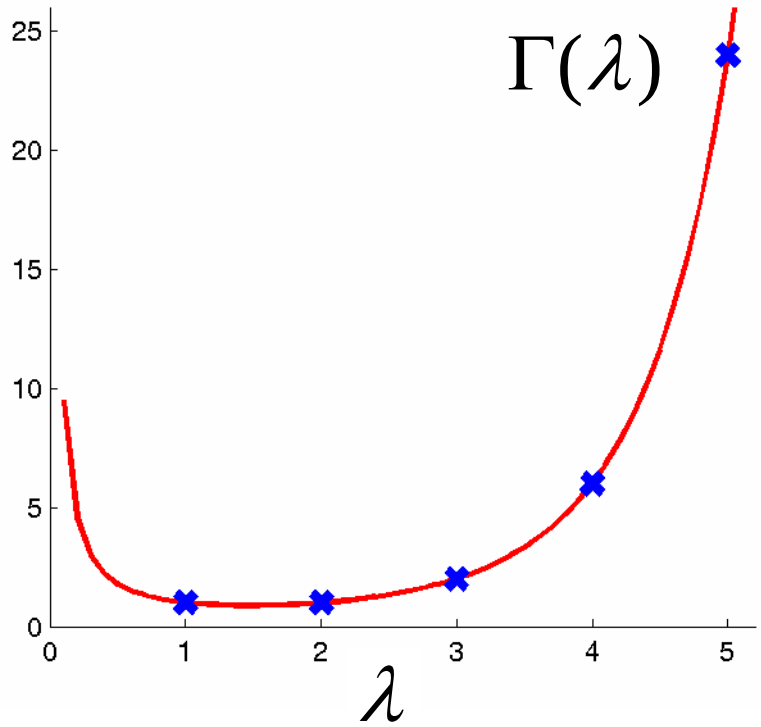
- 超球の体積：

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

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 $\hat{p}(x \mid \omega_i)$ 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

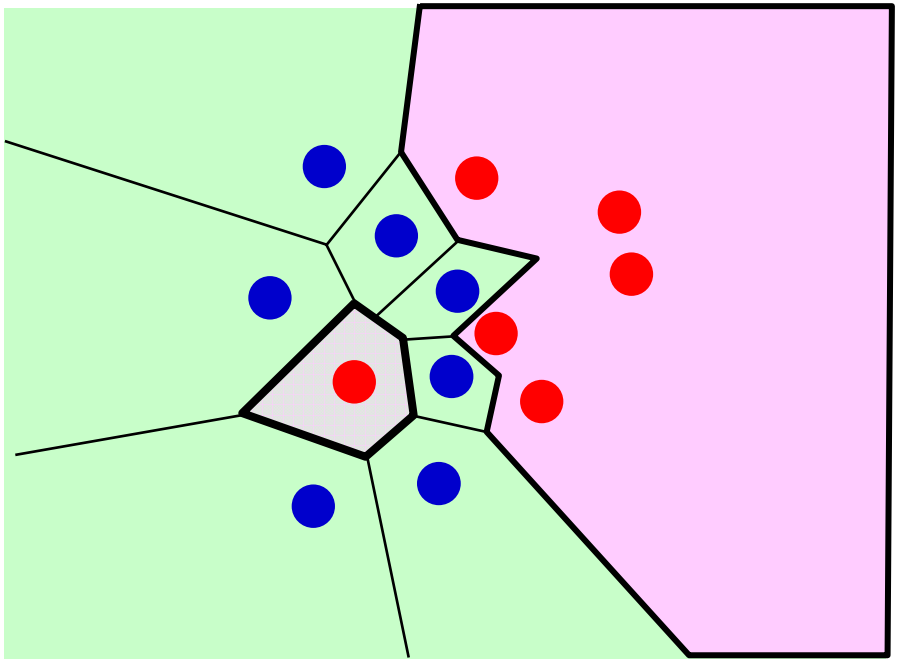
- $p(\omega_i \mid x) \propto p(x \mid \omega_i) p(\omega_i)$ と $p(\omega_i) \approx n_i / n$ より，

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

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.

hやkの決め方

- 核密度推定法では幅hを，最近傍密度推定法では近傍数kを適切に決める必要がある．
- $\hat{p}(x)$ のよさを測る規準：**積分二乗誤差** (integrated squared error)

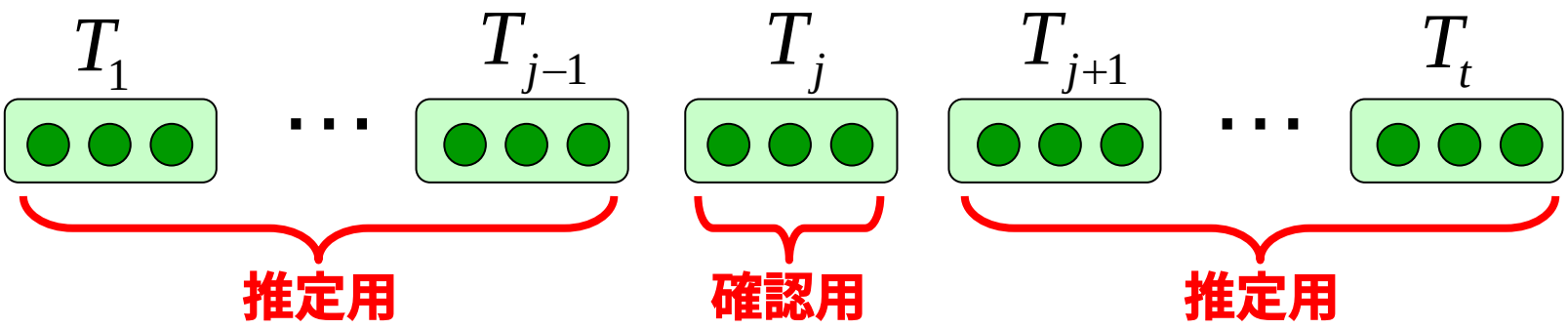
$$ISE(h) = \int_D \left(\hat{p}(x) - p(x)\right)^2 dx$$
$$= \underbrace{\int_D \left(\hat{p}(x)\right)^2 dx}_{\text{計算可}} - 2\int_D \hat{p}(x)p(x)dx + \underbrace{\int_D \left(p(x)\right)^2 dx}_{\text{定数なので無視}}$$

- 第二項目 $Z = \int_D \hat{p}(x)p(x)dx$ を推定する．

Zの推定

- 訓練標本を t 個の重なりの無い部分集合 $\{T_i\}_{i=1}^{t}$ に分ける（通常は $t=5$ か 10）.
- j 番目の部分集合 T_j に含まれる訓練標本を使わずに確率密度関数を推定： $\hat{p}_j(x)$
- 残った標本で推定のよさを確認：

$$\hat{Z}_j = \frac{1}{|T_j|}\sum_{i \in T_j} \hat{p}_j(x_i)$$



Zの推定 (続き)

- $n \to \infty$のとき，$\hat{Z}_j$は Z に収束する！
- しかし，推定結果は j に依存する．
- $\hat{Z}_j$ を全ての jについて平均する：

$$\hat{Z} = \frac{1}{t}\sum_{j=1}^{t}\hat{Z}_j$$

- この方法を**最小二乗交差確認法**(least-squares cross-validation method)という．

最小二乗交差確認法の使い方

- 幅や近傍数の**候補**をいくつか用意
- それぞれの候補に対して最小二乗交差確認法を用いて**積分二乗誤差**を推定し，最小のものを選ぶ

- 最小二乗交差確認法は，実はパラメトリックモデルのモデル選択にも使える！

まとめ

- **最近傍密度推定法**：標本数を固定し，体積をデータから決定する
- 核密度推定法から，**最近傍識別器**が得られる．
- **最小二乗交差確認法**により，
  - 核密度関数法における核関数の幅
  - 最近傍密度推定法における近傍数
  - パラメトリック法におけるパラメトリックモデル

  を決定できる

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

小レポート (第11回)

- k-最近傍識別器を用いて手書き文字の認識を行なえ (サンプルプログラム参照) .
- kの値を変えたときに, 認識率にどのような変化が生じるかを考察せよ.

- 余力のある学生は, k の値を交差確認法を用いて決定せよ.

交差確認法は, ISEだけでなく認識率に対しても適用できる. 即ち, T_j に含まれる訓練標本を用いずに認識器を学習させ, 確認用のデータで認識率を推定する.

Octaveのサンプルプログラム

ex11.m

```
clear all
load digit.mat X T

[d,n,nc]=size(X); nt=size(T,2);
k=100; XX=[]; Y=[];
for c=1:nc
  XX=[XX X(:,:,c)];
  Y=[Y c*ones(1,n)];
end

for ct=1:nc
  Uh=knn(XX,Y,T(:,:,ct),k);
  for c=1:nc
    C(ct,c)=sum(Uh==c);
  end
end
C
```

knn.m

```
function Uh=knn(X,Y,T,k)

n_train=size(X,2); n_test=size(T,2);
d=repmat(sum(X.^2,1)',[1 n_test]) ...
  +repmat(sum(T.^2,1),[n_train 1])-2*X'*T;
[dum,nn_index]=sort(d,1);

for i=1:n_test          %majority vote
  Ynn=Y(nn_index(1:k,i));
  Ys=sort(Ynn);
  ch_flag=(Ys(1:end-1)~=Ys(2:end));
  ch=find(horzcat(1,ch_flag));
  co=horzcat(ch(2:end),k+1)-ch;
  [maxco,index]= max(co);
  Uh(1,i)=Ys(ch(index));
end
```

第8回小レポートの添付資料も参照せよ.