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

- 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

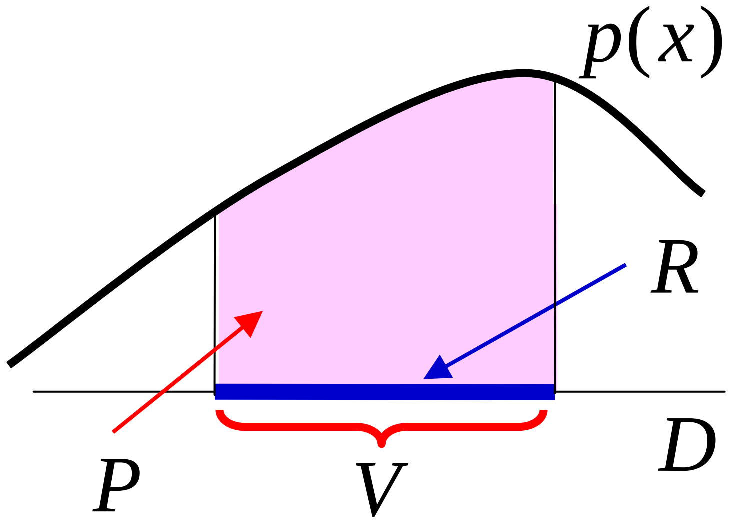
- ある注目点 x' での確率密度 $p(x')$ を推定する
- $R : x'$ を含むパターン空間 D 内のある領域(region)
- V: R の体積(volume)

$$V = \int_R dx$$

- P: あるパターン x が R 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

- 確率 P を二つの方法で近似する.

A) k, n を用いれば,

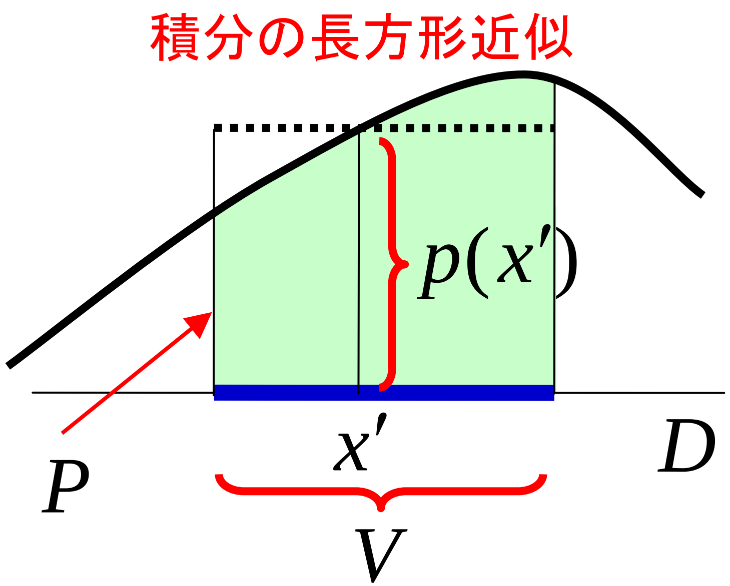
$$P \approx k/n$$

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

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

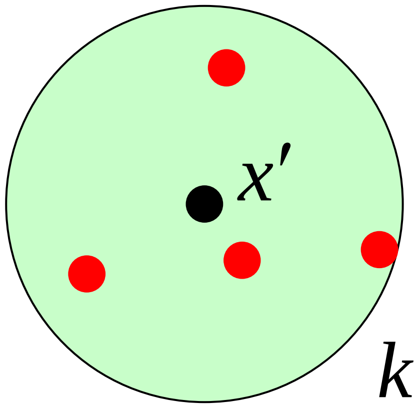
- 訓練標本を用いて領域 R を決める.
  - パーゼン窓法, カーネル密度推定法:
    R の形を決め V を固定したもとで k を標本から決定
  - 最近傍密度推定法:
    R の形を決め k を固定したもとで V を標本から決定

最近傍密度推定法

■ 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：

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



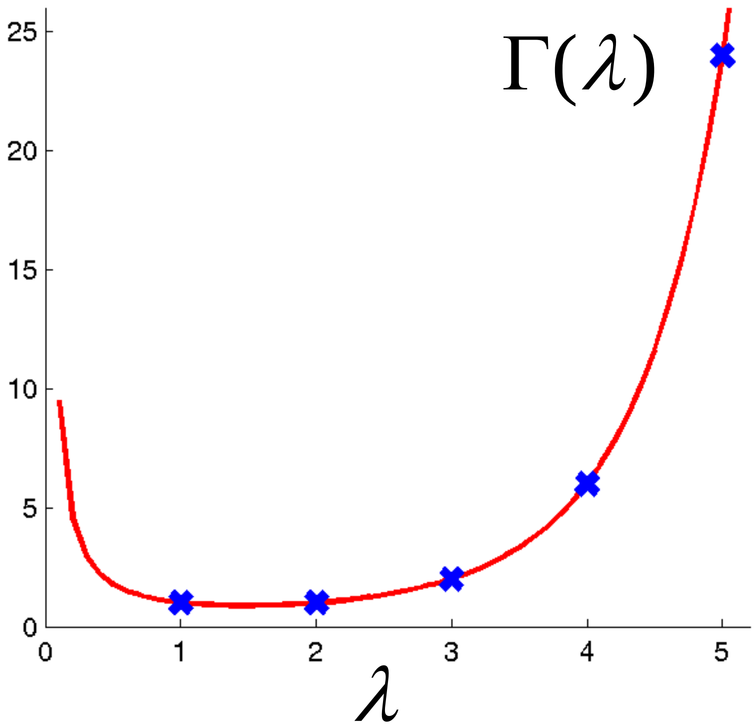
- $$\hat{p}(x) = \frac{k\Gamma(\frac{d}{2}+1)}{n\pi^{\frac{d}{2}} r^d}$$

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

ガンマ関数

- 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

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

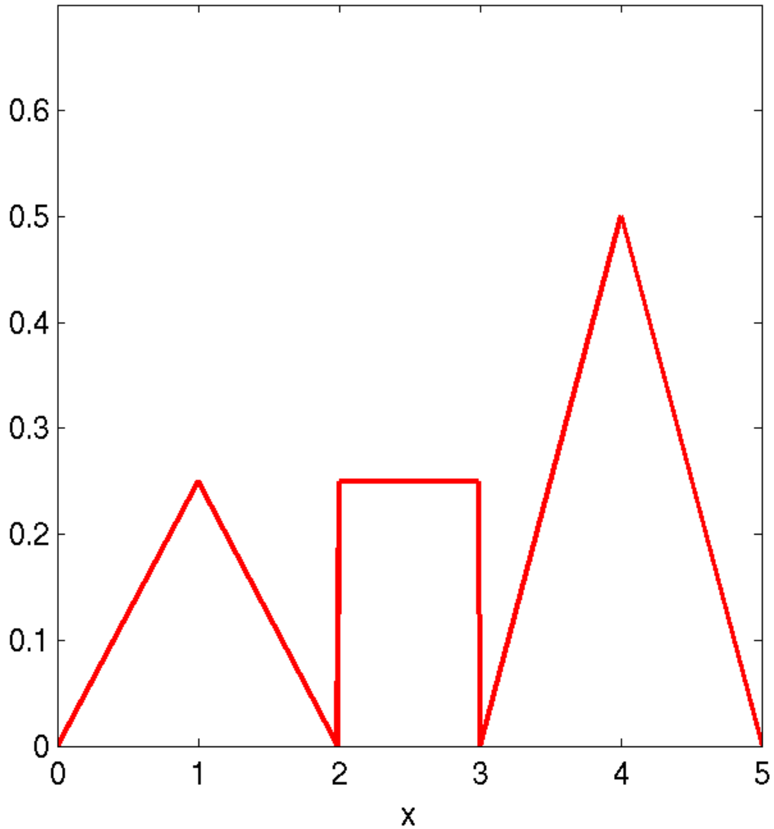
- 階乗の一般化：正の整数 n に対して

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

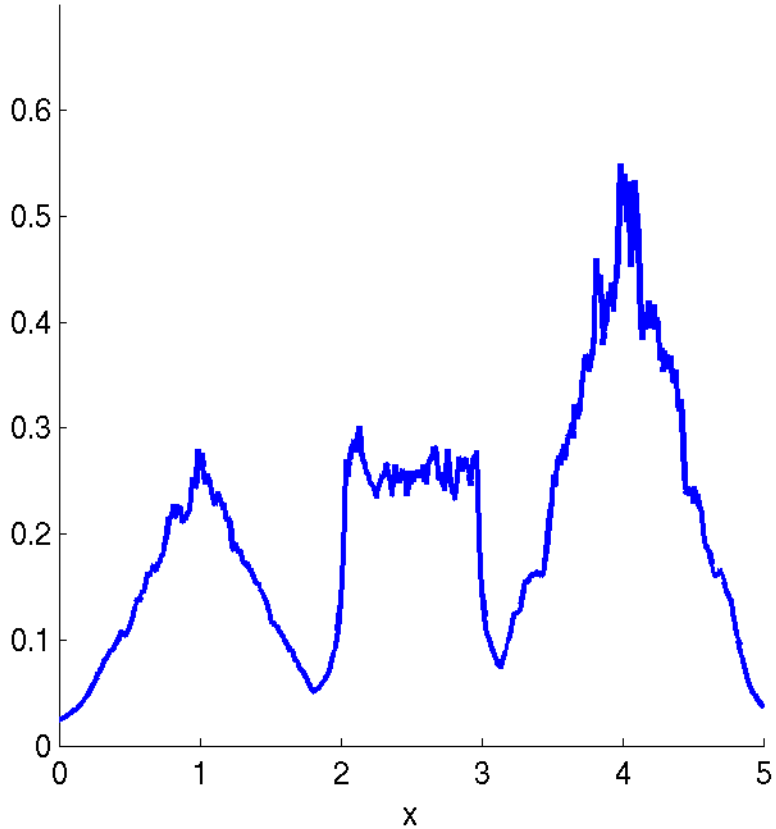
- その他の性質
  - $\Gamma(t) = (t-1)\Gamma(t-1), \ \forall t \in \mathbf{R}$
  - $\Gamma(\frac{1}{2}) = \sqrt{\pi}$
- 演習：
  - $d = 2$ のとき $V = \pi r^2$
  - $d = 3$ のとき $V = \frac{4}{3}\pi r^3$

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

最近傍密度推定法の例



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

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

- 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

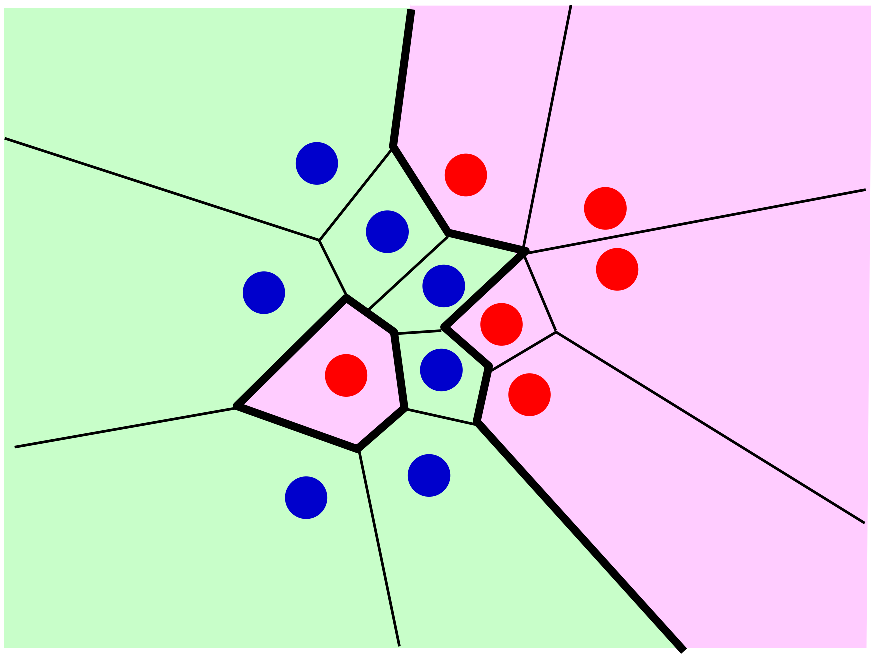
- $p(y) \approx n_y / n$と事後確率 $p(y\,|\,x)$ は,

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

⇒ r_y が小さいほうが事後確率が大きい！

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

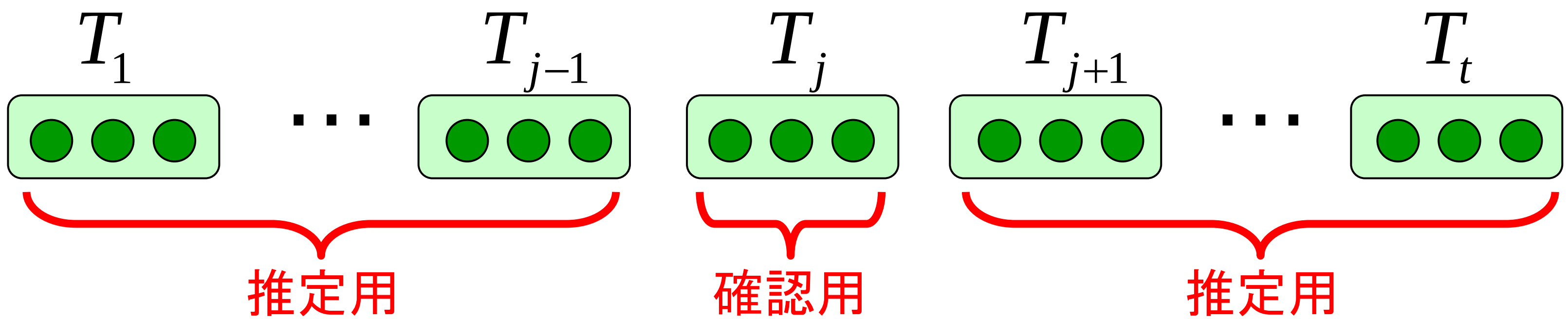
- k-最近傍識別器では近傍数kを適切に決める必要がある.
  1. 値の候補を用意する. 例えば, $k = 1,2,\ldots,10$
  2. それぞれのモデルに対して, パターンの誤認識率を推定する.
  3. 誤認識率の推定値を最小にするモデルを選ぶ.

- 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

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.



連絡事項

- 来週，再来週（12月7日，14日）：
  - 情報工学科計算機室にて計算機実習を行う.
  - 資料・実習課題は事前にウェブに公開する.
  - 出席は取らないので各自で好きな時間に　実習を行ってもよいが，当日はTAの学生が計算機室に常駐するため，質問がある学生は授業時間に演習を行うとよい.

小レポート(第7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.
- 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.

- 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.

Octaveのサンプルプログラム

nnde.m

```
clear all

n=10000; x=myrand(n); k=200;
xx=0:0.01:5; m=length(xx);
dist=repmat(xx',[1 n])-repmat(x,[m 1]);
sort_dist=sort(abs(dist),2);
r=sort_dist(:,k)';
pxh=k*gamma(3/2)./(n*pi^ (1/2)*r);

figure(1); clf;
hist(x,0:0.1:5,10);
figure(2); clf;
plot(xx,pxh,'r-');
legend('true','estimated')
print -depsc nnde.eps
```

myrand.m

```
function x=myrand(n)

x=zeros(1,n);
u=rand(1,n);

flag=(0<=u & u<1/8);
x(flag)=sqrt(8*u(flag));
flag=(1/8<=u & u<1/4);
x(flag)=2-sqrt(2-8*u(flag));
flag=(1/4<=u & u<1/2);
x(flag)=1+4*u(flag);
flag=(1/2<=u & u<3/4);
x(flag)=3+sqrt(4*u(flag)-2);
flag=(3/4<=u & u<=1);
x(flag)=5-sqrt(4-4*u(flag));
```