# 13章　最近傍密度推定法

13.1　最近傍密度推定法
13.2　最近傍識別器
13.3　k-最近傍識別器

# 最近傍密度推定法

領域 $R$ の体積を$V$とする

$$V := \int_R dx$$

$n$ 個の訓練標本 $\{x_i\}_{i=1}^n$ のうち、領域 $R$ 内に入っている個数を$k$とする

このときの注目点 $x'$ における確率密度 $p(x')$ は

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

# 最近傍密度推定法

ある点 $x$ を中心とする超球を領域 $R$ を考える

領域 $R$ の体積 $V$ は次式で与えられる

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

$\Gamma(\bullet)$ はガンマ関数である

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

# 最近傍密度推定法

超球の半径 $r$ を訓練標本が $k$ 個含まれる最小の大きさに設定すると、次式が得られる

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

この方法を $k$-最近傍推定法という

$k=1$ の場合をとくに最近傍推定法という

# 最近傍密度推定法

真の確率密度関数 $p(x)$ によく近似させるために、訓練標本 $k$ を適切に決定する必要がある

標本 $n$ が増えていくとき、$\hat{p}_{KNN}(x)$ が $p(x)$に収束するためには、次を満たす必要がある

$$\lim_{n\to\infty} k = \infty \quad , \quad \lim_{n\to\infty} \frac{k}{n} = 0$$

# 最近傍識別器

各カテゴリ $y$ に対する条件付き確率 $p(x \mid y)$ を最近傍密度推定法によって推定する

# 最近傍識別器

ベイズの定理から、事後確率 $p(y \mid x)$ は

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{\sum_{y'=1}^{c} p(x \mid y')p(y')} \propto p(x \mid y)p(y)$$

# 最近傍識別器

近傍数 $k=1$ である最近傍密度推定法により
カテゴリ $y$ の条件付き確率 $p(x \mid y)$ は

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

$r_y$ : カテゴリ $y$ に属する訓練標本のうち $x$ に
最も近いものと、パターン $x$ との距離
$n_y$ : カテゴリ $y$ に属する訓練標本の数

# 最近傍識別器

事前確率 $p(y)$ を次のように近似する

$$p(y) \approx \frac{n_y}{n}$$

これにより、次の式が得られる

$$p(y \mid x) \approx \frac{\Gamma\left(\frac{d}{2}+1\right)}{n_y \pi^{\frac{d}{2}} r_y^d} \cdot \frac{n_y}{n} \propto r_y^{-d}$$

# 最近傍識別器

先の式より、最近傍密度推定法によるパターン認識は、$r_y^{-d}$ だけ見ればよいことが分かる

$r_y^{-d}$ を最大にするカテゴリ $y$ は $r_y$ が一番小さいカテゴリであるので、パターン $x$ に最も近いカテゴリを選択する

このパターン識別法を最近傍識別器という

# k-最近傍識別器

最近傍識別器は外れ値に弱く、雑音等による影響を受けることがある

この問題を回避するため、近傍数を $k > 1$ に対するk-最近傍識別器を用いる

# k-最近傍識別器

カテゴリ $y$ に属する訓練標本の数が $n_y$ で
訓練標本を $k$ 個含む最少の超球の体積を $V_y$
とする

条件付き確率の推定 $\hat{p}(x \mid y)$ は次で与えられる

$$\hat{p}(x \mid y) = \frac{k}{n_y V_y}$$

# k-最近傍識別器

カテゴリ毎ではなく全てのカテゴリの訓練標本に対し、訓練標本を $k$ 個含む最小の超球を構成する

確率 $p(x)$を次のように推定する

$$\hat{p}(x) = \frac{k}{n\,V}$$

# k-最近傍識別器

確率 $p(x)$を次のように分解する

$$p(x) = \sum_{y=1}^{c} p(x \mid y) p(y)$$

さらに、事前確率 $p(y)$ を次のように推定する

$$\hat{p}(y) = \frac{n_y}{n}$$

# k-最近傍識別器

これにより、次式が得られる

$$\frac{k}{nV} \approx \sum_{y=1}^{c} p(x \mid y) \frac{n_y}{n}$$

ここから、事後確率 $\hat{p}(x \mid y)$ を次式のように設定すればよいことがわかる

($k_y : k$ 個の訓練標本のうち $y$ に属するものの数)

$$\hat{p}(x \mid y) = \frac{k_y}{n_y V}$$

# k-最近傍識別器

こうした方法により、簡単にパターン識別器を構成することができる

# k-最近傍識別器

k-最近傍識別器は次のアルゴリズムに基づく

1. 入力されたパターン $x$ の近傍にある $k$ 個の訓練標本を取り出す。すなわち、全ての訓練標本とパターン $x$ との距離を計算し、最も距離の近い $k$ 個の訓練標本を取り出す

2. その $k$ 個の訓練標本が属するカテゴリを数え上げ、一番数が多いカテゴリにパターン $x$ を分類する

# 近傍数kを決定するにあたって

最近傍密度推定法
→尤度交差確認法で決定できる

最近傍識別器
→尤度交差確認法で決定できる
→パターン認識の問題を解決する場合、
　直接パターンの誤認識率を交差確認法で推定する