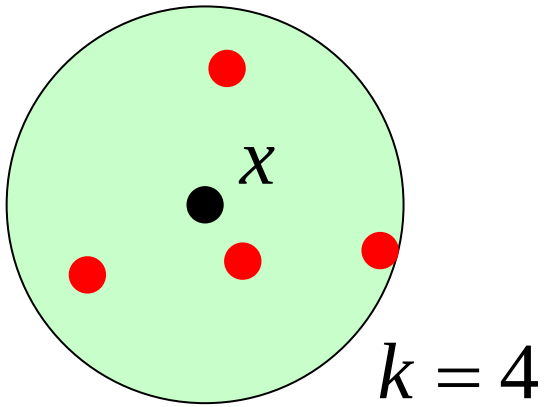
超級

- 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．
- 超級の半径 h：訓練標本が k 個含まれる最小の大きさに設定



超級の体積

- ガンマ関数(gamma function)

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

- 正の整数 n に対して

$$\Gamma(n+1) = n!$$ (階乗の一般化)

- その他の性質: $\Gamma(z) = (z-1)\Gamma(z-1)$

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$

- d 次元の球の体積:

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

$d = 2$ $\quad V = \pi h^2$

$d = 3$ $\quad V = \frac{4}{3}\pi h^3$

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：x とカテゴリ ω_i に属する標本のうち x に最も近いものとの距離

- $p(\omega_i \mid x) \propto p(x \mid \omega_i) p(\omega_i)$ 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．
- 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-最近傍識別機がよく用いられる．