知能機械と自然言語処理
知能機械部　第11回

ソフトウェア情報学部

Goutam Chakraborty

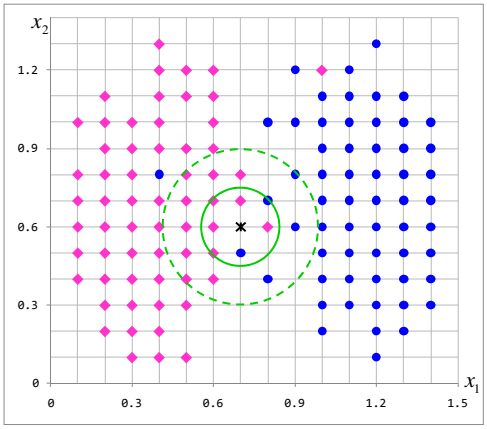
k近傍法（1）

- k-Nearest Neighbor
  – 教師ありデータを扱う

1. Classの数：m
   - $k = m \times n + 1$
2. k個のSample Dataが入るために必要な$Volume$
   - $C_1 \rightarrow V_1$
   - $C_2 \rightarrow V_2$

$V_1 < V_2$ のとき、Sample Dataは、 C_1 のものになる

k近傍法（2）



$V_2 > V_1$

$\therefore$ Class C_1

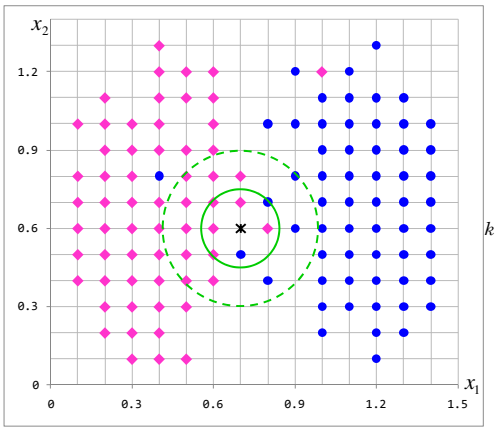
k近傍法(k-Nearest Neibor) *k*-NN

- k-Nearest Neighbor（k近傍法）
  – 教師ありデータを扱う

1. Classの数：m
   - $k = m \times n + 1$ （*nは任意の整数です*）
2. k個のサンプルデータを入るために必要な範囲(hypersphere – 多次元の範囲はhypersphereと言います）は下記の V_1と V_2です。
   - クラス C_1の時必要な範囲は V_1です
   - クラス C_2の時必要な範囲は V_2です

$V_1 < V_2$ のとき、Sample Dataは、 C_1 のクラスになります。

k近傍法の例



2クラスですから
$m = 2$ です

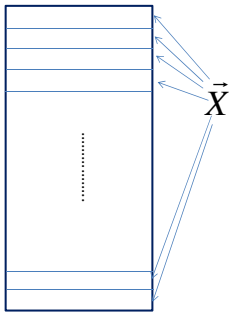
n は2の時、

$k = m \times n + 1$
= 2 × 2 + 1=5 です

$V_2 > V_1$

$\therefore$ Class C_1

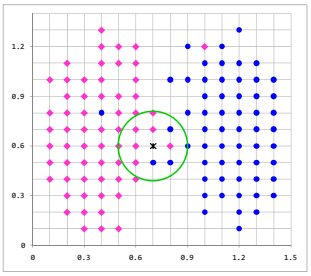
k近傍法の計算量



$\vec{X}$ から全てのデータの距離を求めて、近い順にソートする。クラス毎に上から *k 番目*のデータを取り出します。そうする事で $\vec{X}$ を一番近い取り出したデータのクラスに分類することが出来ます。

近傍法(Nearest Neighbor) NN

- Classの数：m の場合、
  - $k = m \times n + 1$ となる。
  - nは任意の整数、
    $m = 2, n = 5$ のとき、$k = 2 \times 5 + 1$
  - 11個のデータが入る範囲をつくる

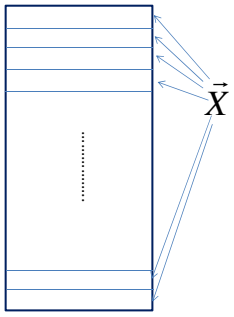


11個のデータが入る範囲を作ると、

- $C_1 \rightarrow 7$個
- $C_2 \rightarrow 4$個

よって、＊は C_1 の
クラスのものになる

近傍法(NN)の計算量

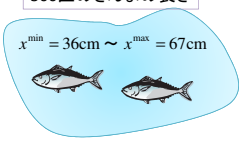


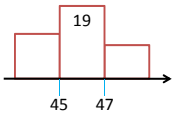
$\vec{X}$ から全てのデータの距離を求めて、近い順にソートする。上から k 個のデータを取り出します。そうする事で $\vec{X}$ を一番多い取り出したデータのクラスに分類することが出来ます。

パルザン窓(1)[一次元]

（Parzen Window）

500匹のさんまの長さ





関係がある！

長さ45～47cmの確率

$$P(45 \le length < 47 \mid C_{\vec{x}}) = \frac{19}{500}$$

長さ46cmの確率密度

$$p(45 \le length < 47 \mid C_{\vec{x}}) = \frac{1}{2} \cdot \frac{19}{500}$$

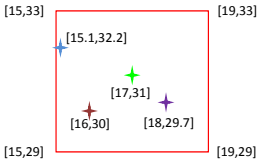
パルザン窓(2)[二次元]

（Parzen Window）

$\vec{x} = \begin{bmatrix} 17 \\ 31 \end{bmatrix}$, $b = 4$ のとき、この窓の中にあるデータはどれか？

たとえば、以下のようなデータがあれば・・・？

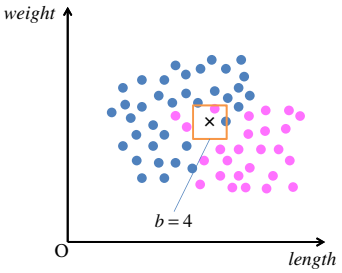
$$\vec{x^1} = \begin{bmatrix} 16.0 \\ 30.0 \end{bmatrix} \quad \vec{x^2} = \begin{bmatrix} 15.1 \\ 32.2 \end{bmatrix} \quad \vec{x^3} = \begin{bmatrix} 18.0 \\ 42.3 \end{bmatrix} \quad \vec{x^4} = \begin{bmatrix} 14.5 \\ 28.3 \end{bmatrix} \quad \cdots \quad \vec{x^N} = \begin{bmatrix} x_1^N \\ x_2^N \end{bmatrix}$$



確率密度関数(1)

$\vec{x}$ の付近の確率密度は、

$$p(\vec{x}) = \frac{1}{b^d}\left(\frac{1}{N}\sum_{i=1}^{N}\phi\left(\frac{|\vec{x^i} - \vec{x}|}{b}\right)\right) \qquad \cdots\cdots(1)$$



$$\phi(x) = \begin{cases} 1 & \left(x \le \frac{1}{2}\right) \\ 0 & \left(x > \frac{1}{2}\right) \end{cases}$$

確率密度関数(2)

$$\phi\left(\frac{|\vec{x^i} - \vec{x}|}{b}\right) = \phi\left(\frac{|x_1^i - x_1|}{b}\right) \cap \phi\left(\frac{|x_2^i - x_2|}{b}\right)$$

i=1 のとき、

$$\phi\left(\frac{|\vec{x^1} - \vec{x}|}{b}\right) = \phi\left(\frac{|16-17|}{4}\right) \cap \phi\left(\frac{|30-31|}{4}\right)$$

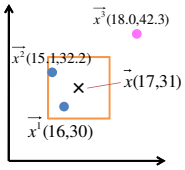
$$= \phi\left(\frac{1}{4}\right) \cap \phi\left(\frac{1}{4}\right) = 1 \cap 1 = 1$$

同様に$i = 2 \cdots N$ のとき、

$$\phi\left(\frac{|\vec{x^2} - \vec{x}|}{b}\right) = 1, \ \phi\left(\frac{|\vec{x^3} - \vec{x}|}{b}\right) = 0, \ \phi\left(\frac{|\vec{x^4} - \vec{x}|}{b}\right) = 0, \cdots, \ \phi\left(\frac{|\vec{x^N} - \vec{x}|}{b}\right) = 0$$

確率密度関数(3)

$\vec{x^i}: i = 1 \sim N$ の中で窓の中に入るものが、
$\vec{x^1}, \vec{x^2}$ だけであれば、確率密度の値は、
以下のようになる

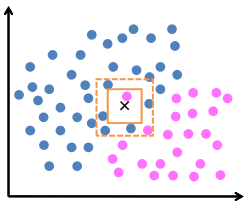


$$p(\vec{x}) = \frac{1}{4^2}\left(\frac{1}{N}(1+1+0+0+\cdots)\right) = \frac{1}{16}\cdot\frac{2}{N}$$

パルザン窓(3)

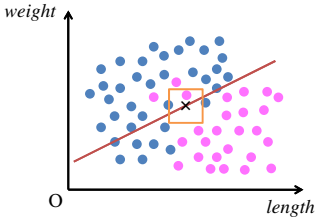
(Parzen Window)

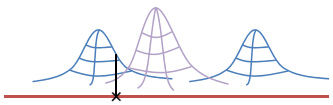
- Parzen Windowは、bの大きさで結果が変わってしまう。



- 窓が 小さい とき、●のクラスになってしまう。
- 窓を 大きく したとき、●のクラスになる。
- 最適な窓のサイズを決めるのは、難しい。

Kernel Function(1)





$N(0,1)$

$$p(\vec{x}) = \frac{1}{N b^d}\cdot\frac{1}{(2\pi)^{\frac{d}{2}}}\sum_{i=1}^{N}\exp\left(-\frac{(\vec{x^i}-\vec{x})^T(\vec{x^i}-\vec{x})}{2b^2}\right) \quad \cdots\cdots(2)$$

正規分布の広がりの問題が残る

Kernel Function(2)

$$p(\vec{x}) = \frac{1}{b^d}\left(\frac{1}{N}\sum_{i=1}^{N}\phi\left(\frac{|\vec{x^i}-\vec{x}|}{b}\right)\right) \quad \cdots\cdots(1)$$

$$p(\vec{x}) = \frac{1}{N b^d}\cdot\frac{1}{(2\pi)^{\frac{d}{2}}}\sum_{i=1}^{N}\exp\left(-\frac{(\vec{x^i}-\vec{x})^T(\vec{x^i}-\vec{x})}{2b^2}\right) \quad \cdots\cdots(2)$$

- 上の2つの式を見比べると、式(2)のとき、Sampleの位置は $\vec{x}$ より遠ければ遠いほど影響が少なくなる。