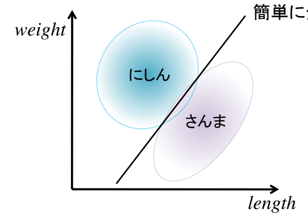
知能機械と自然言語処理

知能機械部　第8回

ソフトウェア情報学部

Goutam Chakraborty

簡単な分類の仕方



境界線が分かれば、簡単に分類できる！

$C_{さ} \rightarrow \overrightarrow{\mu_{さ}},\ \Sigma_{さ}$

$C_{に} \rightarrow \overrightarrow{\mu_{に}},\ \Sigma_{に}$

境界線を求める

（境界線の上では両方の確率が同じ）

多次元の確率密度関数

Class C_1のデータの平均$\overrightarrow{\mu_1}$, 相関行列Σ_1の場合、条件付き確率密度関数は

$$f_1(\vec{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma_1|}} \exp\left(-\frac{1}{2}\left(\vec{x}-\overrightarrow{\mu_1}\right)^T \Sigma_1^{-1}\left(\vec{x}-\overrightarrow{\mu_1}\right)\right) \quad \cdots\cdots ①$$

Class C_2のデータの平均$\overrightarrow{\mu_2}$, 相関行列Σ_2の場合、条件付き確率密度関数は

$$f_2(\vec{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma_2|}} \exp\left(-\frac{1}{2}\left(\vec{x}-\overrightarrow{\mu_2}\right)^T \Sigma_2^{-1}\left(\vec{x}-\overrightarrow{\mu_2}\right)\right) \quad \cdots\cdots ②$$

logarithm（対数）の復習

$$\log_{10} x = \log x$$
$$\log_e x = \ln x$$
$$\log_2 x = \lg x$$

$$e = 1 + \frac{1}{1} + \frac{1}{2!} + \cdots \infty$$
$$= 1 + \sum_{i=1}^{\infty} \frac{1}{i!} = 2.781828\cdots$$

$$\log_a b \times \log_b c \times \log_c a = 1$$
$$\log_a a = 1$$
$$\log_2 8a = \log_2 2^3 + \log_2 a = 3 + \log_2 a$$

$\ln 3.3 = ?$ $\qquad \log_e 3.3 \times \log_{10} e \times \frac{1}{\log_{10} 3.3} = 1$

対数表から求める

境界線の求め方（1）

①, ②より、$f_1(\vec{x}) = f_2(\vec{x}),\ P(C_1) = P(C_2)$ ならば、「$\vec{x}$の軌跡は、2つのクラス境界線」となる。これに $\log_e$ を両辺にとると

$$-\ln\sqrt{|\Sigma_1|} + \left[-\frac{1}{2}\left(\vec{x}-\overrightarrow{\mu_1}\right)^T \Sigma_1^{-1}\left(\vec{x}-\overrightarrow{\mu_1}\right)\right] = -\ln\sqrt{|\Sigma_2|} + \left[-\frac{1}{2}\left(\vec{x}-\overrightarrow{\mu_2}\right)^T \Sigma_2^{-1}\left(\vec{x}-\overrightarrow{\mu_2}\right)\right]$$

$$-\frac{1}{2}\ln|\Sigma_1| - \frac{1}{2}\left(\vec{x}^T\Sigma_1^{-1}\vec{x} - \vec{x}^T\Sigma_1^{-1}\overrightarrow{\mu_1} - \overrightarrow{\mu_1}^T\Sigma_1^{-1}\vec{x} + \overrightarrow{\mu_1}^T\Sigma_1^{-1}\overrightarrow{\mu_1}\right)$$
$$= -\frac{1}{2}\ln|\Sigma_2| - \frac{1}{2}\left(\vec{x}^T\Sigma_2^{-1}\vec{x} - \vec{x}^T\Sigma_2^{-1}\overrightarrow{\mu_2} - \overrightarrow{\mu_2}^T\Sigma_2^{-1}\vec{x} + \overrightarrow{\mu_2}^T\Sigma_2^{-1}\overrightarrow{\mu_2}\right)$$

$$\ln|\Sigma_1| + \left(\vec{x}^T\Sigma_1^{-1}\vec{x} - \vec{x}^T\Sigma_1^{-1}\overrightarrow{\mu_1} - \overrightarrow{\mu_1}^T\Sigma_1^{-1}\vec{x} + \overrightarrow{\mu_1}^T\Sigma_1^{-1}\overrightarrow{\mu_1}\right)$$
$$= \ln|\Sigma_2| + \left(\vec{x}^T\Sigma_2^{-1}\vec{x} - \vec{x}^T\Sigma_2^{-1}\overrightarrow{\mu_2} - \overrightarrow{\mu_2}^T\Sigma_2^{-1}\vec{x} + \overrightarrow{\mu_2}^T\Sigma_2^{-1}\overrightarrow{\mu_2}\right)$$

$\Sigma_1 = \Sigma_2$のとき、$\vec{x}^T \cdot \vec{x}$ の項がなくなり、境界線は直線になる。

境界線の求め方（2）

これから、2つのクラスの問題について考える。

C_1 の平均と相関行列は $\overrightarrow{\mu_1}, \Sigma_1$ 、C_2 は $\overrightarrow{\mu_2}, \Sigma_2$ 。

データは d 次元のデータで、特色の数は d 。

ベイズの定理から境界線の式は以下の式で求められる。

$$\underbrace{\frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma_1|}} e^{-\frac{1}{2}\left[\left(\vec{x}-\overrightarrow{\mu_1}\right)^T \Sigma_1^{-1}\left(\vec{x}-\overrightarrow{\mu_1}\right)\right]}}_{P(C_1 | \vec{x})} \times P(C_1)$$
$$= \underbrace{\frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma_2|}} e^{-\frac{1}{2}\left[\left(\vec{x}-\overrightarrow{\mu_2}\right)^T \Sigma_2^{-1}\left(\vec{x}-\overrightarrow{\mu_2}\right)\right]}}_{P(C_2 | \vec{x})} \times P(C_2)$$

境界線の求め方(3)

$\log_e$ を両辺にとると

$$-\frac{1}{2}\ln|\Sigma_1| - \frac{1}{2}\left[\left(\vec{x}-\vec{\mu_1}\right)^T\ \Sigma_1^{-1}\left(\vec{x}-\vec{\mu_1}\right)\right] + \ln P(C_1)$$

$$= -\frac{1}{2}\ln|\Sigma_2| - \frac{1}{2}\left[\left(\vec{x}-\vec{\mu_2}\right)^T\ \Sigma_2^{-1}\left(\vec{x}-\vec{\mu_2}\right)\right] + \ln P(C_2)$$

$$\left[\left(\vec{x}-\vec{\mu_2}\right)^T\ \Sigma_2^{-1}\left(\vec{x}-\vec{\mu_2}\right)\right] - \left[\left(\vec{x}-\vec{\mu_1}\right)^T\ \Sigma_1^{-1}\left(\vec{x}-\vec{\mu_1}\right)\right] = \ln|\Sigma_1| - \ln|\Sigma_2| + 2\ln P(C_2) - 2\ln P(C_1)$$

境界線の求め方(4)

ここで、実際の数値を入れて計算してみる。ただし、 $P(C_1)=P(C_2)$ であるとする。

$$\vec{\mu_1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 0.30 & 0 \\ 0 & 0.35 \end{bmatrix}$$

$$|\Sigma_1| = 0.105$$

$$\Sigma_1^{-1} = \begin{bmatrix} 3.3 & 0 \\ 0 & 2.8 \end{bmatrix}$$

$$\vec{\mu_2} = \begin{bmatrix} 4 \\ 0 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1.20 & 0 \\ 0 & 1.85 \end{bmatrix}$$

$$|\Sigma_2| = 2.22$$

$$\Sigma_2^{-1} = \begin{bmatrix} 0.8 & 0 \\ 0 & 0.5 \end{bmatrix}$$

境界線の求め方(5)

$$[x_1-4 \quad x_2]\begin{bmatrix} 0.8 & 0 \\ 0 & 0.5 \end{bmatrix}\begin{bmatrix} x_1-4 \\ x_2 \end{bmatrix} - [x_1 \quad x_2]\begin{bmatrix} 3.3 & 0 \\ 0 & 2.8 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \ln|\Sigma_1| - \ln|\Sigma_2|$$

$$[0.8(x_1-4) \quad 0.5x_2]\begin{bmatrix} x_1-4 \\ x_2 \end{bmatrix} - [3.3x_1 \quad 2.8x_2]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = -2.2-0.8$$

$$\left[0.8(x_1-4)^2 + 0.5x_2^{\,2}\right] - \left[3.3x_1^{\,2} + 2.8x_2^{\,2}\right] = -3.0$$

$$\left[0.8x_1^{\,2} - 6.4x_1 + 12.8 + 0.5x_2^{\,2}\right] - \left[3.3x_1^{\,2} + 2.8x_2^{\,2}\right] = -3.0$$

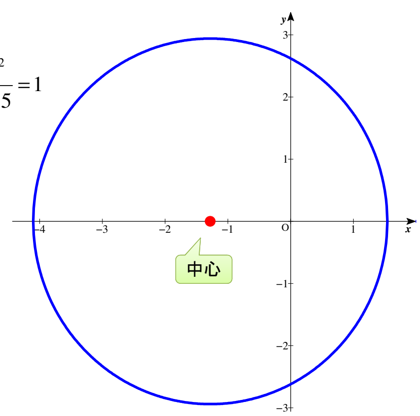
$$-2.5x_1^{\,2} - 6.4x_1 - 2.3x_2^{\,2} = -15.8$$

$$\frac{(x_1+1.3)^2}{2.3} + \frac{x_2^{\,2}}{2.5} = 3.5$$

この境界線は楕円になる

境界線の求め方(6)

$$\frac{(x_1+1.3)^2}{8.05} + \frac{x_2^{\,2}}{8.75} = 1$$



境界線の種類

- $\Sigma_1 = \Sigma_2$ のとき、境界線は直線になる
- $\Sigma_1 \neq \Sigma_2$ のとき、境界線は曲線(Conics)になる
  - 円 (Circle)
  - 楕円 (Ellipse)
  - 放物線(Parabola)
  - 双曲線(Hyperbola)

行列の積の復習

$\Sigma = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix}$ のとき、 $\vec{x}^T\Sigma^{-1}\vec{x}$ を求めてみよう！

$$\Sigma^{-1} = \frac{1}{0.3\times0.3}\begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix}$$

$$\vec{x}^T\Sigma^{-1}\vec{x} = [x_1 \quad x_2]\times\frac{1}{0.3\times0.3}\begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= \frac{100}{9}[x_1 \quad x_2]\begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= 11\times[0.3x_1 \quad 0.3x_2]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 3.3x_1^{\,2} + 3.3x_2^{\,2}$$

曲線の基礎

- $ax_1^2 + bx_2^2 + cx_1x_2 + dx_1 + ex_2 + f = 0$
  - $a \neq 0, b \neq 0$ ➔ 円（$a = b$），楕円
  - $a = 0, b \neq 0$ または $a \neq 0, b = 0$ ➔ 放物線
  - $a = b = c = 0$ ➔ 直線 $dx_1 + ex_2 + f = 0$
    - $\Sigma_1 = \Sigma_2$ のとき、成立する

課題

- Nishin2D-300.txt, Sanma2D-500.txtのデータの最初の200のデータを使って、境界線の式を求めよ。（データを正規化する必要あり）
  - つまり、 $P(C_1) = P(C_2)$ となる。
  - $\overrightarrow{\mu_1}, \overrightarrow{\mu_2}, \Sigma_1, \Sigma_2$ を求める必要がある

境界線の問題（例題）〔1〕

C_1	$\Sigma_1 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$	$\overrightarrow{\mu_1} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$
C_2	$\Sigma_2 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$	$\overrightarrow{\mu_2} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$

$P(C_1) = P(C_2)$

このときの境界線の
式を求めてみよう！

$$\left[\left(\vec{x} - \overrightarrow{\mu_2}\right)^T \Sigma_2^{-1} \left(\vec{x} - \overrightarrow{\mu_2}\right)\right] - \left[\left(\vec{x} - \overrightarrow{\mu_1}\right)^T \Sigma_1^{-1} \left(\vec{x} - \overrightarrow{\mu_1}\right)\right] = \cancel{\ln|\Sigma_1|} - \cancel{\ln|\Sigma_2|} + \cancel{2\ln P(C_2)} - \cancel{2\ln P(C_1)}$$

$$\therefore \left[\left(\vec{x} - \overrightarrow{\mu_1}\right)^T \Sigma^{-1} \left(\vec{x} - \overrightarrow{\mu_1}\right)\right] = \left[\left(\vec{x} - \overrightarrow{\mu_2}\right)^T \Sigma^{-1} \left(\vec{x} - \overrightarrow{\mu_2}\right)\right]$$

境界線の問題（例題）〔2〕

$$|\Sigma| = 1 \cdot 2 - 1 \cdot 1 = 1$$

$$\Sigma^{-1} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$$

$$\left[\left(\vec{x} - \overrightarrow{\mu_1}\right)^T \Sigma^{-1} \left(\vec{x} - \overrightarrow{\mu_1}\right)\right] = \begin{bmatrix} x_1 - 2 & x_2 - 4 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 - 2 \\ x_2 - 4 \end{bmatrix}$$

$$=$$