Bayes' theorem

Asst. Prof. Dr. Waranyu Wongseree

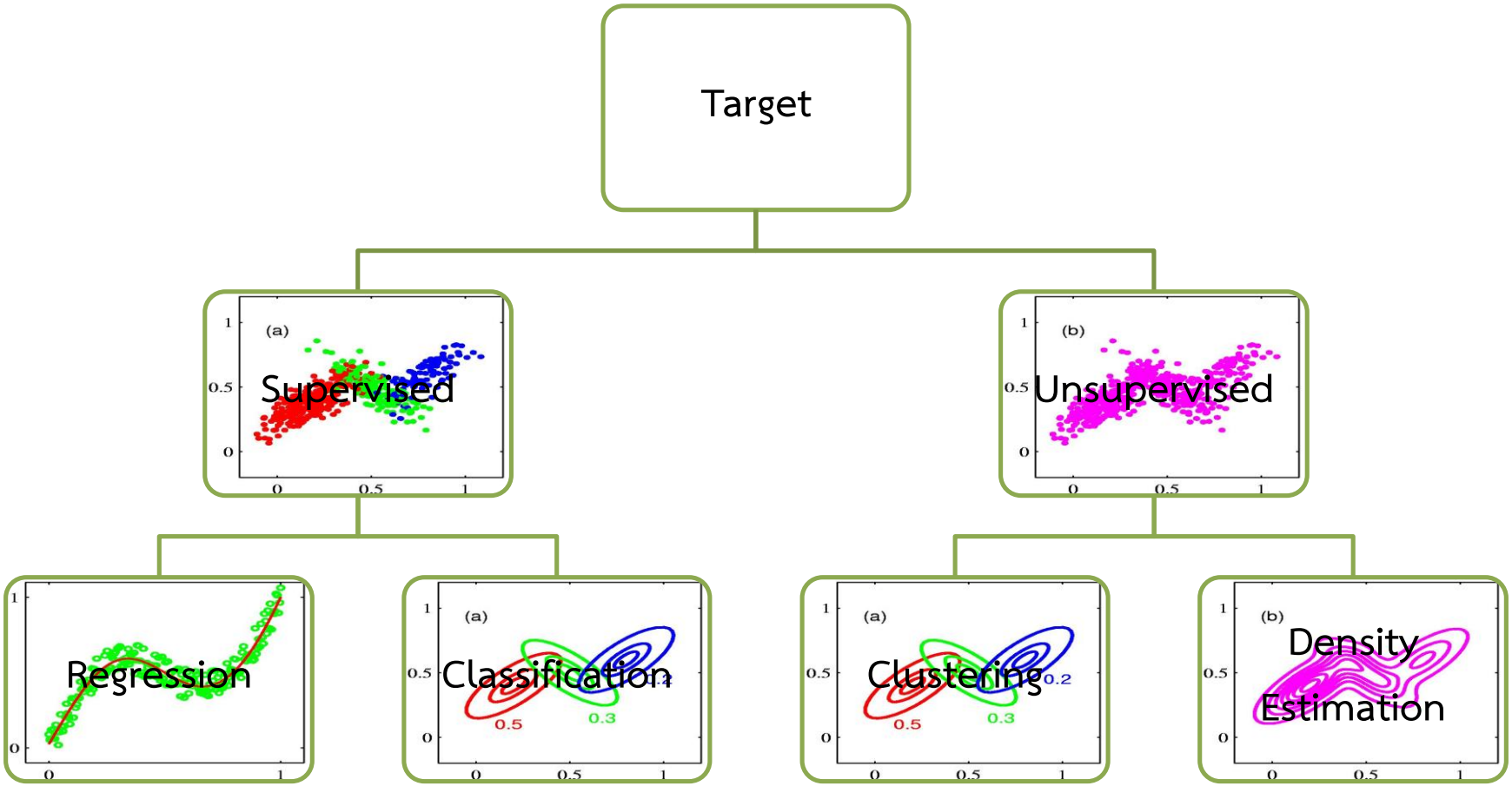
Faculty of Digital Technology

Chitralada Technology Institute

การจำแนก

คือการสร้างแบบจำลองทางคณิตศาสตร์สำหรับการรู้จำแบบรูป (Pattern) ตัวอย่าง (Sample) สัญญาณ (Signal) หรือรูปภาพ (Image) โดยอัตโนมัติจากลักษณะประจำ (Attribute หรือ Feature) จากการวัด หรือการสังเกต

การแบ่งประเภทรู้จำแบบ



ทฤษฎีการตัดสินใจแบบเบส์

PSA (x)	Class (y)
4.1	Cancer
3.4	Cancer
2.9	Cancer
2.8	Cancer
2.7	Cancer
2.1	Cancer
1.6	Cancer

PSA (x)	Class (y)
2.5	Healthy
2.0	Healthy
1.7	Healthy
1.4	Healthy
1.2	Healthy
0.9	Healthy
0.8	Healthy

ความน่าจะเป็นของเหตุการณ์ y เมื่อให้หลักฐาน x

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

ทฤษฎีการตัดสินใจแบบเบส์

PSA (x)	Class (y)
4.1	Cancer
3.4	Cancer
2.9	Cancer
2.8	Cancer
2.7	Cancer
2.1	Cancer
1.6	Cancer
2.5	Healthy
2.0	Healthy
1.7	Healthy
1.4	Healthy
1.2	Healthy
0.9	Healthy
0.8	Healthy

ความน่าจะเป็นของเหตุการณ์ y เมื่อให้หลักฐาน x

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

$P(y)$ คือ ความน่าจะเป็นก่อน (Prior probability)

$$P(\text{Class} = \text{Cancer}) = 0.5$$

$$P(\text{Class} = \text{Healthy}) = 0.5$$

ความน่าจะเป็นก่อนสามารถประมาณจากสัดส่วนของคลาสจากตัวอย่าง

ทฤษฎีการตัดสินใจแบบเบส์

PSA (x)	Class (y)
4.1	Cancer
3.4	Cancer
2.9	Cancer
2.8	Cancer
2.7	Cancer
2.1	Cancer
1.6	Cancer
2.5	Healthy
2.0	Healthy
1.7	Healthy
1.4	Healthy
1.2	Healthy
0.9	Healthy
0.8	Healthy

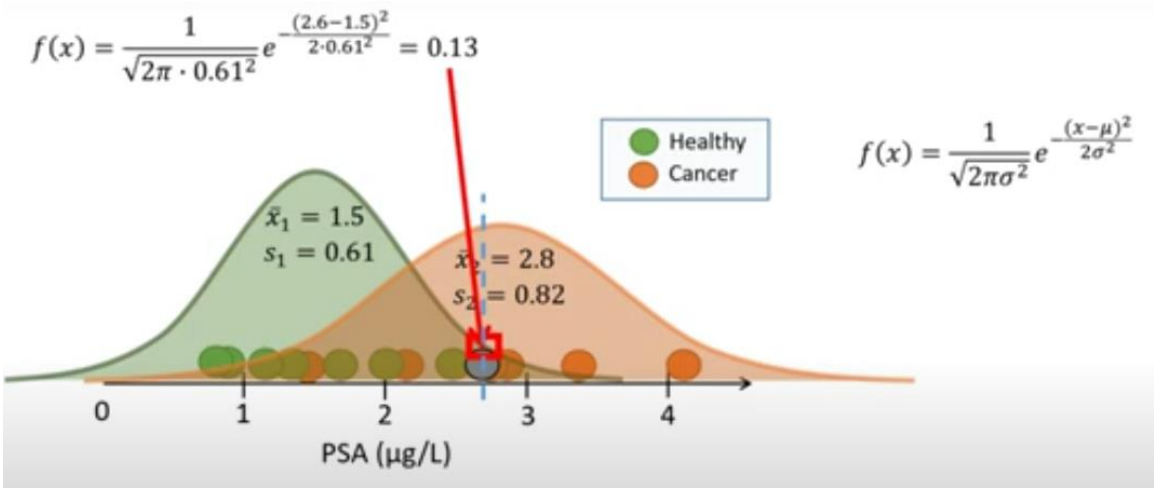
ความน่าจะเป็นของเหตุการณ์ y เมื่อให้หลักฐาน x

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

$P(x|y)$ คือ ความน่าจะเป็นมีเงื่อนไขคลาส (Class-Conditional Probability)

ความน่าจะเป็นมีเงื่อนไขคลาส (Class-Conditional Probability)

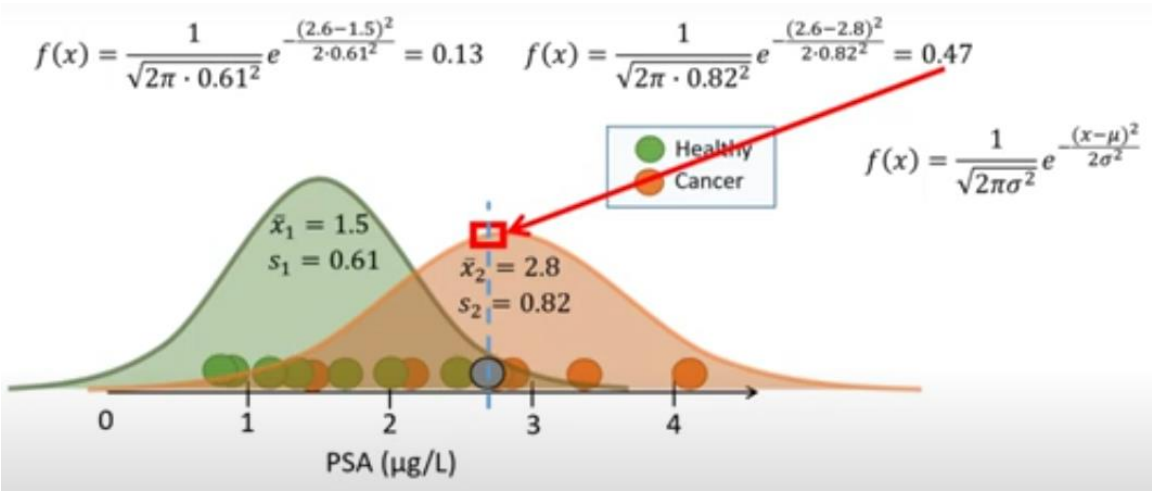
PSA (x)	Class (y)	Statistics
4.1	Cancer	$\bar{x}_2 = 2.8$ $s_2 = 0.82$
3.4	Cancer	
2.9	Cancer	
2.8	Cancer	
2.7	Cancer	
2.1	Cancer	
1.6	Cancer	
2.5	Healthy	$\bar{x}_1 = 1.5$ $s_1 = 0.61$
2.0	Healthy	
1.7	Healthy	
1.4	Healthy	
1.2	Healthy	
0.9	Healthy	
0.8	Healthy	



$$P(\text{PSA} = 2.6|\text{Healthy}) = 0.13$$

ความน่าจะเป็นมีเงื่อนไขคลาส (Class-Conditional Probability)

PSA (x)	Class (y)	Statistics
4.1	Cancer	$\bar{x}_2 = 2.8$ $s_2 = 0.82$
3.4	Cancer	
2.9	Cancer	
2.8	Cancer	
2.7	Cancer	
2.1	Cancer	
1.6	Cancer	
2.5	Healthy	$\bar{x}_1 = 1.5$ $s_1 = 0.61$
2.0	Healthy	
1.7	Healthy	
1.4	Healthy	
1.2	Healthy	
0.9	Healthy	
0.8	Healthy	



$$P(\text{PSA} = 2.6|\text{Cancer}) = 0.47$$

ความน่าจะเป็นมีเงื่อนไขคลาส (Class-Conditional Probability)

PSA (x)	Class (y)
4.1	Cancer
3.4	Cancer
2.9	Cancer
2.8	Cancer
2.7	Cancer
2.1	Cancer
1.6	Cancer
2.5	Healthy
2.0	Healthy
1.7	Healthy
1.4	Healthy
1.2	Healthy
0.9	Healthy
0.8	Healthy

ความน่าจะเป็น (Likelihood) ที่ PSA = 2.6 เมื่อพิจารณากลุ่ม Cancer

$$P(\text{PSA} = 2.6|\text{Cancer}) = 0.47$$

ความน่าจะเป็น (Likelihood) ที่ PSA = 2.6 เมื่อพิจารณากลุ่ม Healthy

$$P(\text{PSA} = 2.6|\text{Healthy}) = 0.13$$

ความน่าจะเป็นมีเงื่อนไขคลาส (Class-Conditional Probability)

PSA (x)	Class (y)
4.1	Cancer
3.4	Cancer
2.9	Cancer
2.8	Cancer
2.7	Cancer
2.1	Cancer
1.6	Cancer
2.5	Healthy
2.0	Healthy
1.7	Healthy
1.4	Healthy
1.2	Healthy
0.9	Healthy
0.8	Healthy

ความน่าจะเป็น (Likelihood) ที่ PSA = 2.6 เมื่อพิจารณากลุ่ม Cancer

$$P(\text{PSA} = 2.6|\text{Cancer}) = 0.47$$

ความน่าจะเป็น (Likelihood) ที่ PSA = 2.6 เมื่อพิจารณากลุ่ม Healthy

$$P(\text{PSA} = 2.6|\text{Healthy}) = 0.13$$

ความน่าจะเป็นภายหลัง (Posteriori probability)

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

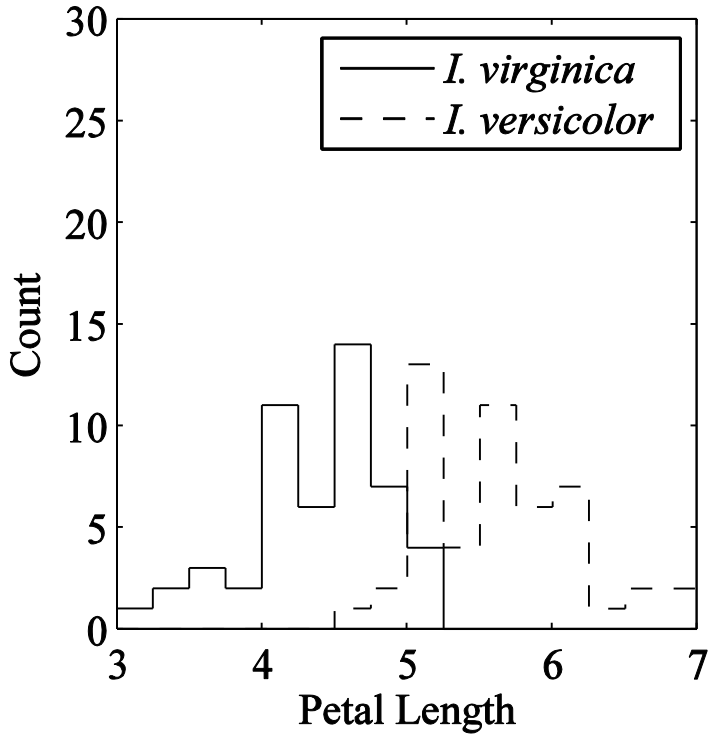
$$P(\text{Class} = \text{Cancer}|\text{PSA} = 2.6)$$
$$= \frac{P(\text{PSA} = 2.6|\text{Class} = \text{Cancer})\, P(\text{Class} = \text{Cancer})}{P(x)}$$
$$= \frac{0.5 \cdot 0.47}{0.5 \cdot 0.13 + 0.5 \cdot 0.47} = 0.79$$

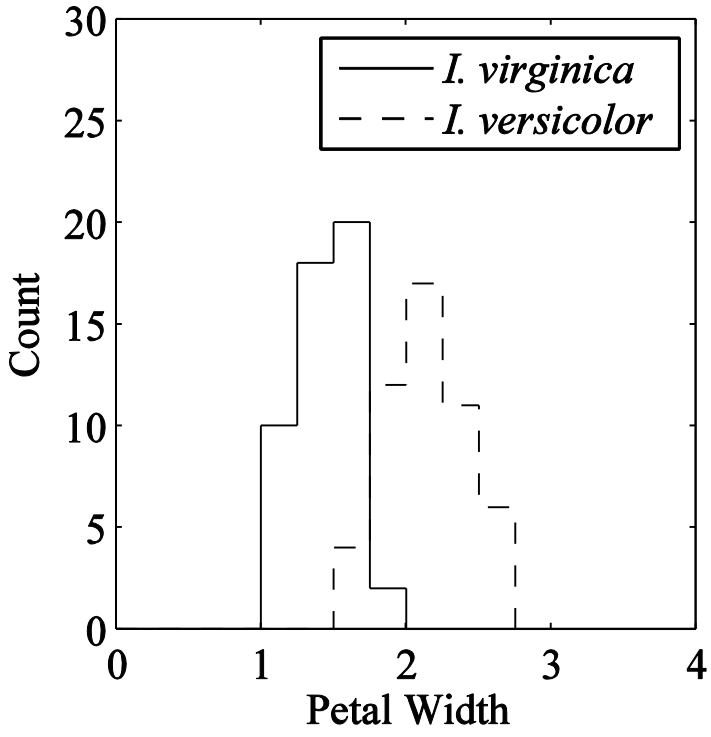
$$P(\text{Class} = \text{Healthy}|\text{PSA} = 2.6)$$
$$= \frac{P(\text{PSA} = 2.6|\text{Class} = \text{Healthy})\, P(\text{Class} = \text{Healthy})}{P(x)}$$
$$= \frac{0.5 \cdot 0.13}{0.5 \cdot 0.13 + 0.5 \cdot 0.47} = 0.21$$

ตัวอย่างการจำแนก

การจำแนกพันธุ์ไอริสโดยใช้ข้อมูลเข้าหรือลักษณะประจำเป็นความยาวกลีบเลี้ยง (Sepal Length) ความกว้างกลีบเลี้ยง (Sepal Width) ความยาวกลีบดอก (Petal Length) และความกว้างกลีบดอก (Petal Width)

ฮิสโทแกรมของความยาวและความกว้างกลีบดอกไอริส





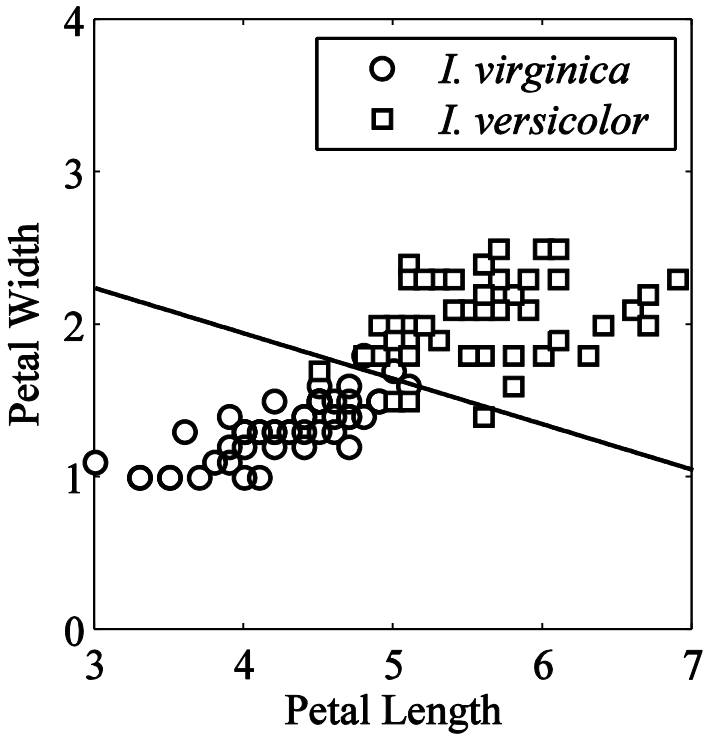
เลือกความยาวหรือความกว้างกลีบดอกค่าหนึ่งเพื่อใช้เป็นขีดเริ่มเปลี่ยน (Threshold) สำหรับการสร้างขอบตัดสินใจ (Decision Boundary)

ตัวจำแนกเชิงเส้น (Linear Discriminant)

การใช้ลักษณะประจำเดียวไม่เพียงพอสำหรับการเลือกขีดเริ่มเปลี่ยนเพื่อการตัดสินใจ

ปรากฏการณ์ดังกล่าวเรียกว่าการเข้ากับตัวอย่างน้อยกว่าพอเหมาะ (Underfitting)

ค่าคลาดเคลื่อนการจำแนกมีค่าต่ำกว่าตัวจำแนกเชิงเส้นตัวแปรเดียว

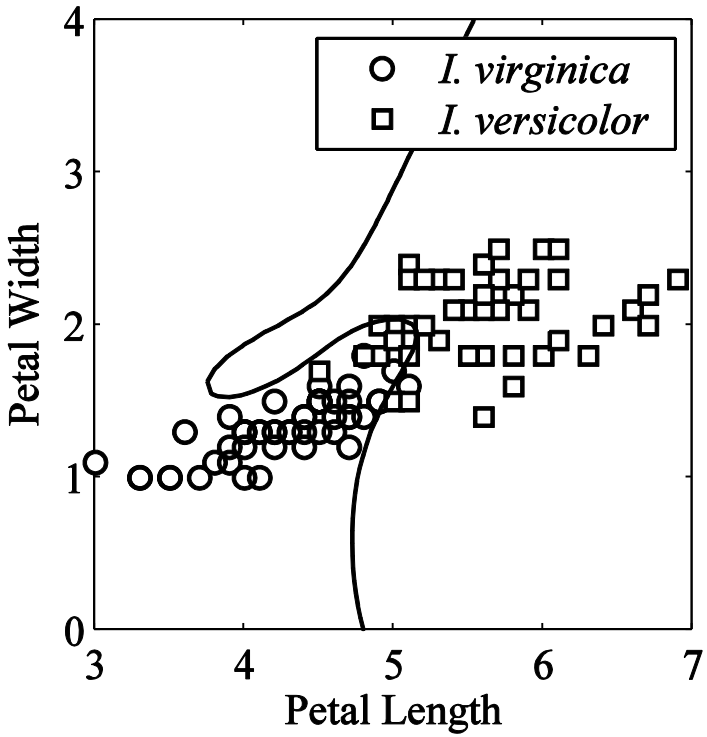


ตัวจำแนกไม่เชิงเส้น (Non-linear Discriminant)

ขอบตัดสินใจมีลักษณะซับซ้อนเกินไป ส่งผลให้ตัวจำแนกไม่สามารถจำแนกพันธุ์ไอริสที่ไม่เคยใช้สร้างตัวจำแนกได้

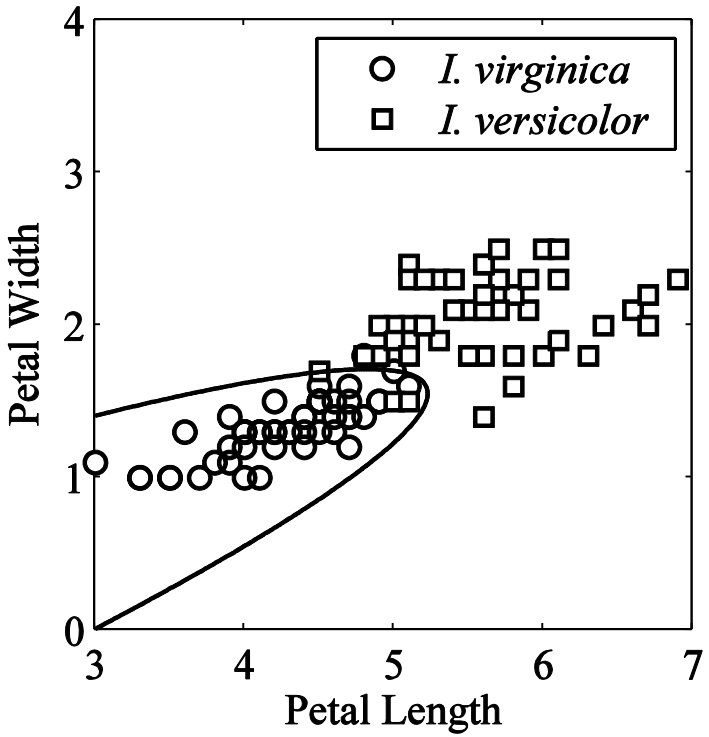
ปรากฏการณ์ดังกล่าวเรียกว่าการเข้ากับตัวอย่างเกินพอเหมาะ (Overfitting)

แก้ปัญหาโดยการเพิ่มจำนวนตัวอย่าง แต่ในทางปฏิบัติมักมีจำนวนตัวอย่างไม่เพียงพอ



ตัวจำแนกที่เหมาะสม

การสร้างตัวจำแนกจึงต้องมีความสมดุลระหว่างสมรรถนะการจำแนกและความซับซ้อนของตัวจำแนก



การจำแนกโดยใช้ทฤษฎีการตัดสินใจแบบเบส์

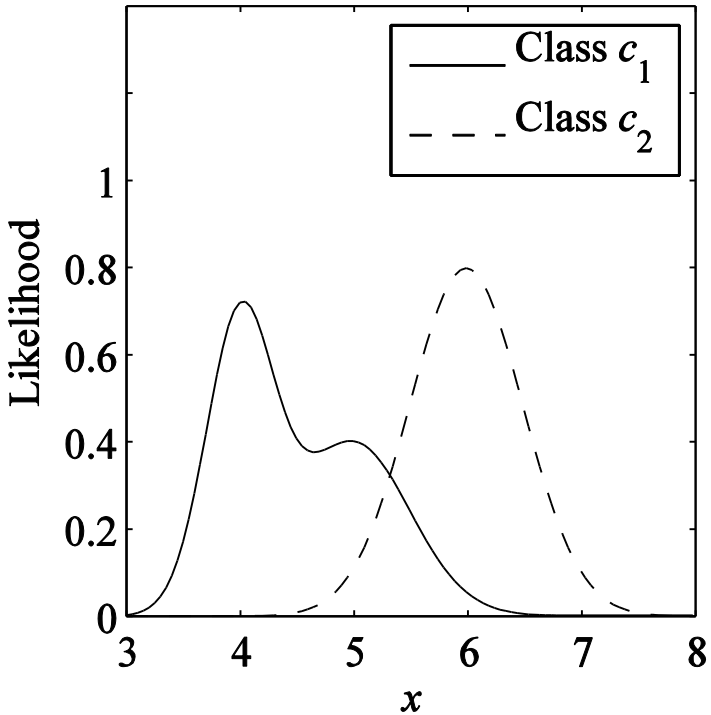
- $p(c_1)$ แทนความน่าจะเป็นที่จะพบ *Iris versicolor* และ
- $p(c_2)$ แทนความน่าจะเป็นที่จะพบ *Iris virginica*
- ถ้าไม่มีไอริสพันธุ์อื่น แล้วผลรวมของความน่าจะเป็นจะมีค่าเท่ากับหนึ่ง
- ความน่าจะเป็นนี้เรียกว่าความน่าจะเป็นก่อน (Prior Probability)
- ความน่าจะเป็นก่อนสามารถประมาณจากสัดส่วนของคลาสจากเซตตัวอย่าง (Sample Set)

การจำแนกโดยใช้ทฤษฎีการตัดสินใจแบบเบส์

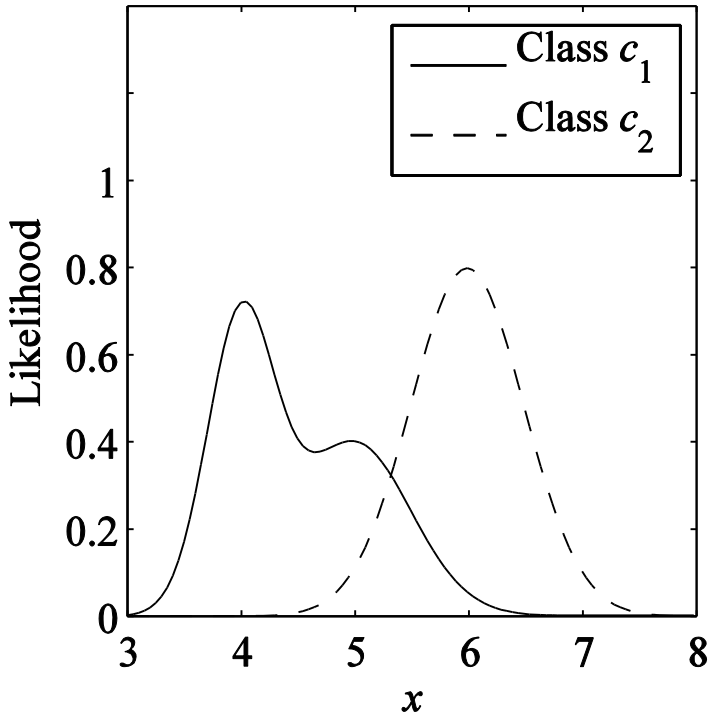
$p(x|c_i)$ ความน่าจะเป็นมีเงื่อนไขคลาส (Class-Conditional Probability) อธิบายการแจกแจงตัวอย่างเข้า

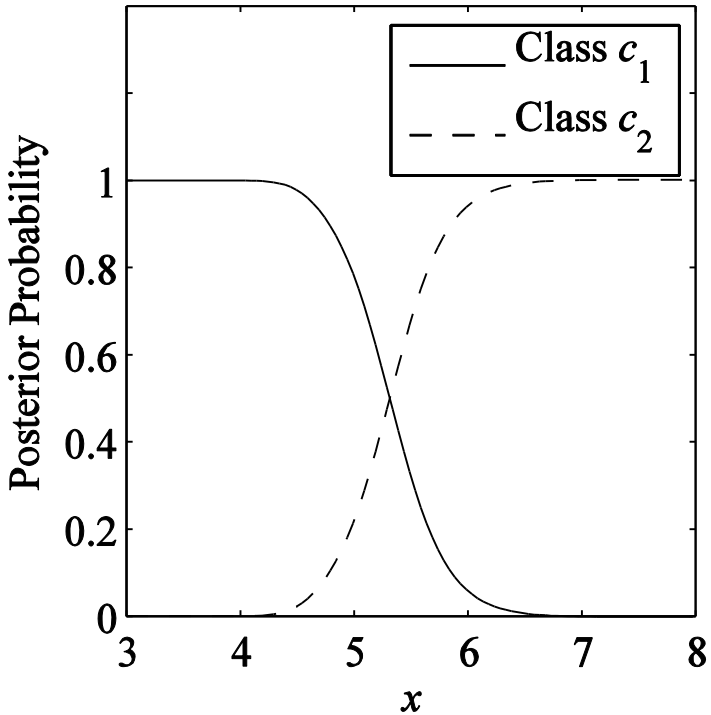
$$p(c_i \mid x) = \frac{p(x \mid c_i)\, p(c_i)}{p(x)}$$

$$p(x) = \sum_{i=1}^{2} p(x \mid c_i)\, p(c_i)$$

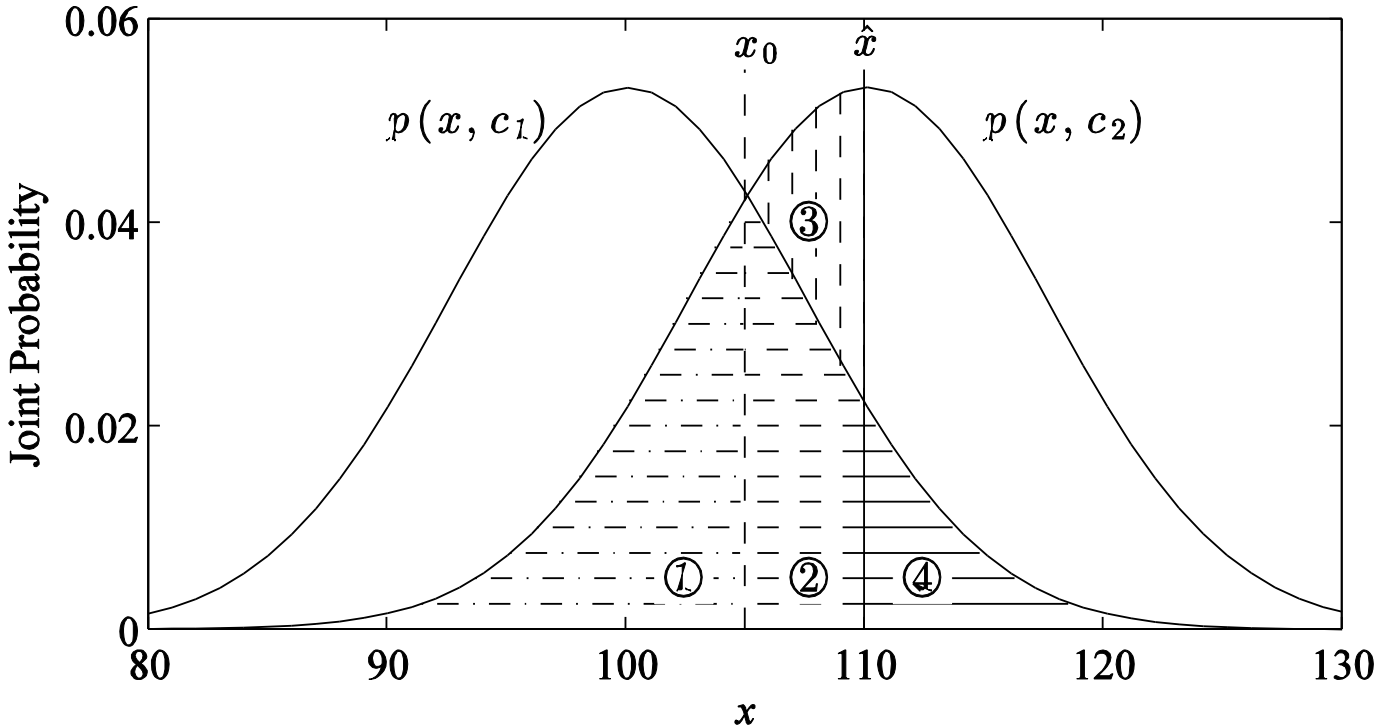


การจำแนกโดยใช้ทฤษฎีการตัดสินใจแบบเบส์





การจำแนกโดยใช้ทฤษฎีการตัดสินใจแบบเบส์



กฎตัดสินใจ (Decision Rule)

- ถ้าตัวจำแนกตัดสินใจให้ตัวอย่างเข้ามาจากคลาสที่ความน่าจะเป็นภายหลังสูงสุด
- แล้วจะทำให้ค่าคลาดเคลื่อนมีเงื่อนไขตัวอย่างเข้ามีค่าต่ำสุด
- ดังนั้นกฎตัดสินใจจึงสามารถอธิบายได้โดย

If $p(c_1 \mid x) > p(c_2 \mid x)$ then classify x as belonging to c_1, otherwise classify x as belonging to c_2.

ฟังก์ชันจำแนก (Discriminant Function)

- เนื่องจากความน่าจะเป็นภายหลังมักอธิบายโดยใช้ฟังก์ชันเลขชี้กำลัง
- การใช้ประยุกต์ใช้ฟังก์ชันเพิ่มทางเดียวกับความน่าจะเป็นภายหลังจะให้ผลการวิเคราะห์ตรงกัน และจะเรียกว่าฟังก์ชันจำแนก

$$
\begin{aligned}
g_i(x) &= \ln(p(c_i \mid x)) \\
&= \ln\left(\frac{p(x \mid c_i)\,p(c_i)}{p(x)}\right)
\end{aligned}
$$

- และกฎตัดสินใจสามารถอธิบายได้โดย

If $g_1(x) > g_2(x)$ then classify x as belonging to c_1, otherwise classify x as belonging to c_2.

ขอบตัดสินใจ (Decision Boundary)

$$g(x) = g_1(x) - g_2(x) = 0$$

หรือ

$$\begin{aligned}
\ln(p(c_1 \mid x)) - \ln(p(c_2 \mid x)) &= \ln\left(\frac{p(c_1 \mid x)}{p(c_2 \mid x)}\right) \\
&= \ln\left(\frac{p(x \mid c_1)p(c_1)}{p(x)} \times \frac{p(x)}{p(x \mid c_2)p(c_2)}\right) \\
&= \ln\left(\frac{p(x \mid c_1)}{p(x \mid c_2}\right) + \ln\left(\frac{p(c_1)}{p(c_2)}\right) \\
&= 0
\end{aligned}$$

กฎตัดสินใจและขอบตัดสินใจกรณีหลายคลาส

กรณีปัญหาการจำแนกประกอบด้วยหลายคลาส กฎตัดสินใจสามารถอธิบายได้โดย

$$\boxed{\text{If } g_i(x) > g_j(x), \forall j \neq i \text{ then classify } x \text{ as belonging to } c_i.}$$

และขอบตัดสินใจสามารถอธิบายได้โดย

$$g_{ij}(x) = g_i(x) - g_j(x) = 0, \quad \forall j \neq i \tag{1-10}$$

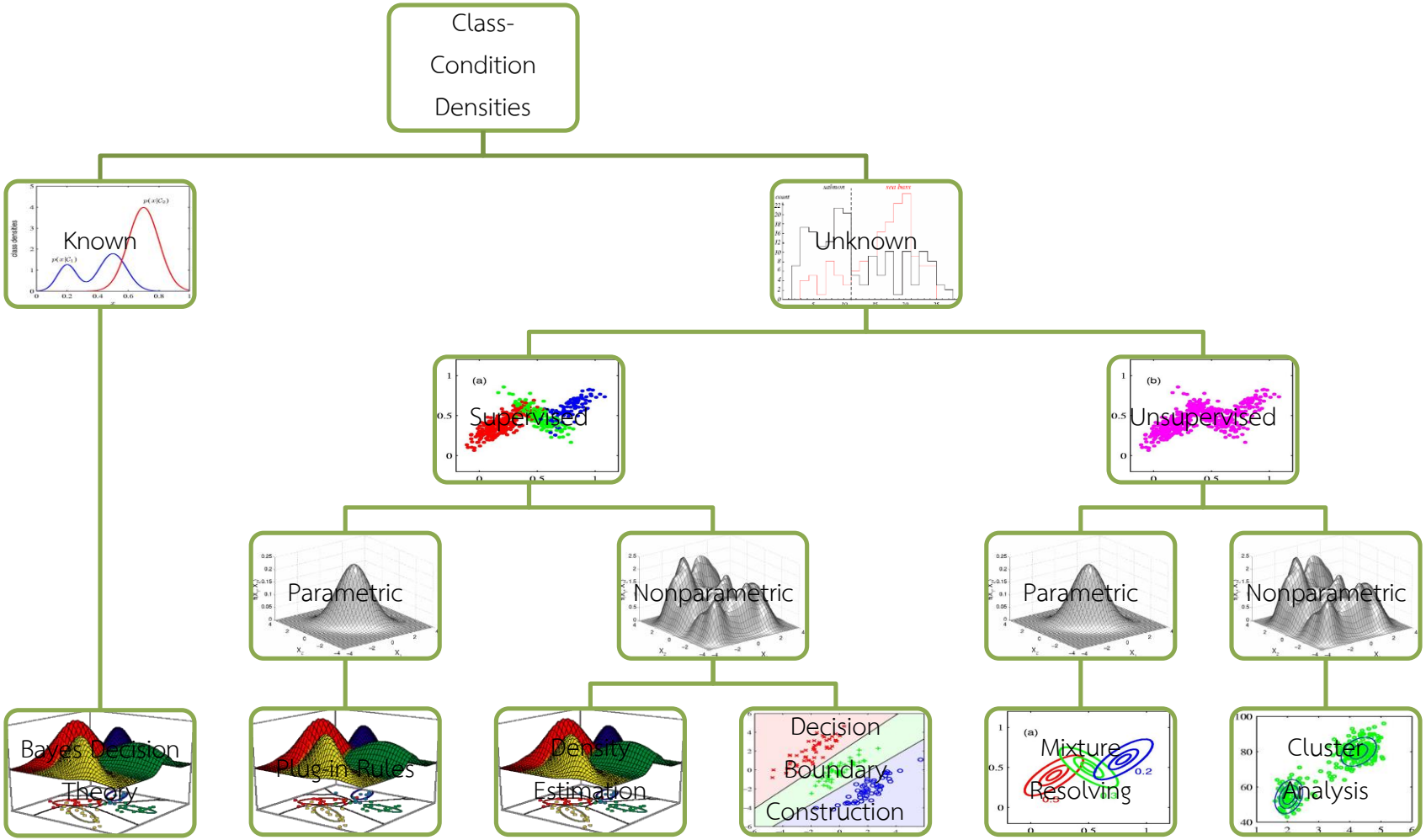
ทฤษฎีการตัดสินใจแบบเบส์

ตัวจำแนกมีสองส่วนประกอบสำคัญ คือ ความน่าจะเป็นมีเงื่อนไขคลาส และความน่าจะเป็นก่อน

กรณีที่ทราบฟังก์ชันความน่าจะเป็นมีเงื่อนไขคลาส การแก้ปัญหาการจำแนกจะเป็นเพียงการประมาณค่าพารามิเตอร์ของการแจกแจง และจะเรียกวิธีการแก้ปัญหานี้ว่าเทคนิคอิงพารามิเตอร์

ถ้าไม่ทราบฟังก์ชันความน่าจะเป็นมีเงื่อนไขคลาส จะต้องประมาณฟังก์ชันความน่าจะเป็นมีเงื่อนไขคลาสหรือสร้างขอบตัดสินใจโดยการประมาณความน่าจะเป็นภายหลัง ทั้งสองวิธีแก้ปัญหานี้เรียกว่าเทคนิคไม่อิงพารามิเตอร์

ทฤษฎีการตัดสินใจแบบเบส์



การแก้ปัญหาการจำแนกโดยใช้เทคนิคอิงพารามิเตอร์

การแจกแจงปรกติซึ่งเหมาะสมสำหรับกรณีตัวอย่างเข้าประกอบด้วยลักษณะประจำค่าต่อเนื่อง (Continuous-Valued Attribute)

และการแจกแจงแบร์นูลลี (Bernoulli Distribution) ซึ่งเหมาะสมสำหรับกรณีตัวอย่างเข้าประกอบด้วยลักษณะประจำค่าวิยุต (Discrete-Valued Attribute)

การแจกแจงปรกติตัวแปรเดียว

$$p(r) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2} \left(\frac{r-\mu}{\sigma} \right)^2 \right]$$

เมื่อ

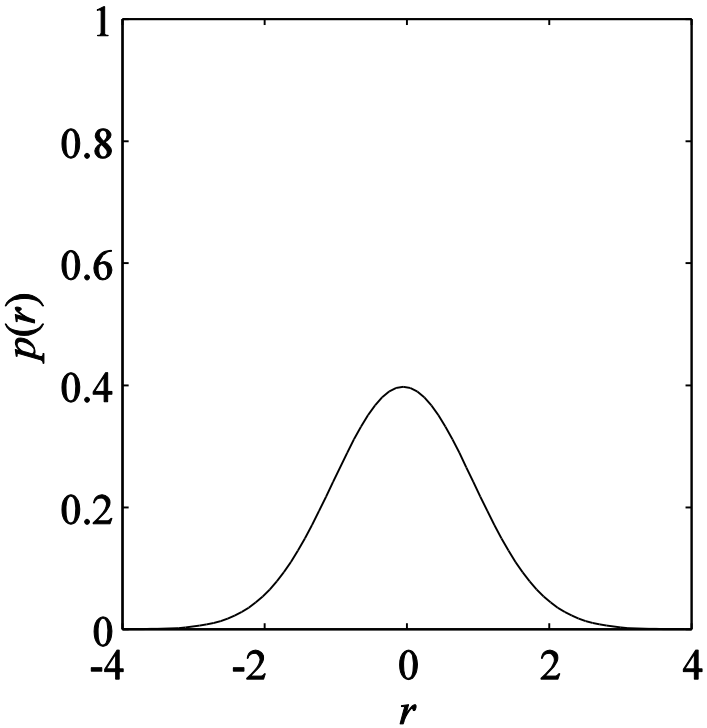
$$\mu = E[r]$$

$$= \int_{-\infty}^{\infty} r p(r) dr$$

และ

$$\sigma^2 = E[(r-\mu)^2]$$

$$= \int_{-\infty}^{\infty} (r-\mu)^2 p(r) dr$$



ตัวจำแนกแบบเบส์สำหรับการแจกแจงปรกติตัวแปรเดียว

เมื่อใช้การแจกแจงปรกติเป็นฟังก์ชันความหนาแน่นความน่าจะเป็นมีเงื่อนไขคลาส

$$p(r) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2} \left(\frac{r-\mu}{\sigma} \right)^2 \right]$$

ฟังก์ชันจำแนกจะเป็นฟังก์ชันกำลังสอง (Quadratic Function)

$$g_i(x) = -\frac{(x-\mu_{c_i})^2}{2\sigma_{c_i}^2} - \frac{1}{2}\ln 2\pi - \ln \sigma_{c_i} + \ln p(c_i) - \ln p(x)$$

Single Threshold

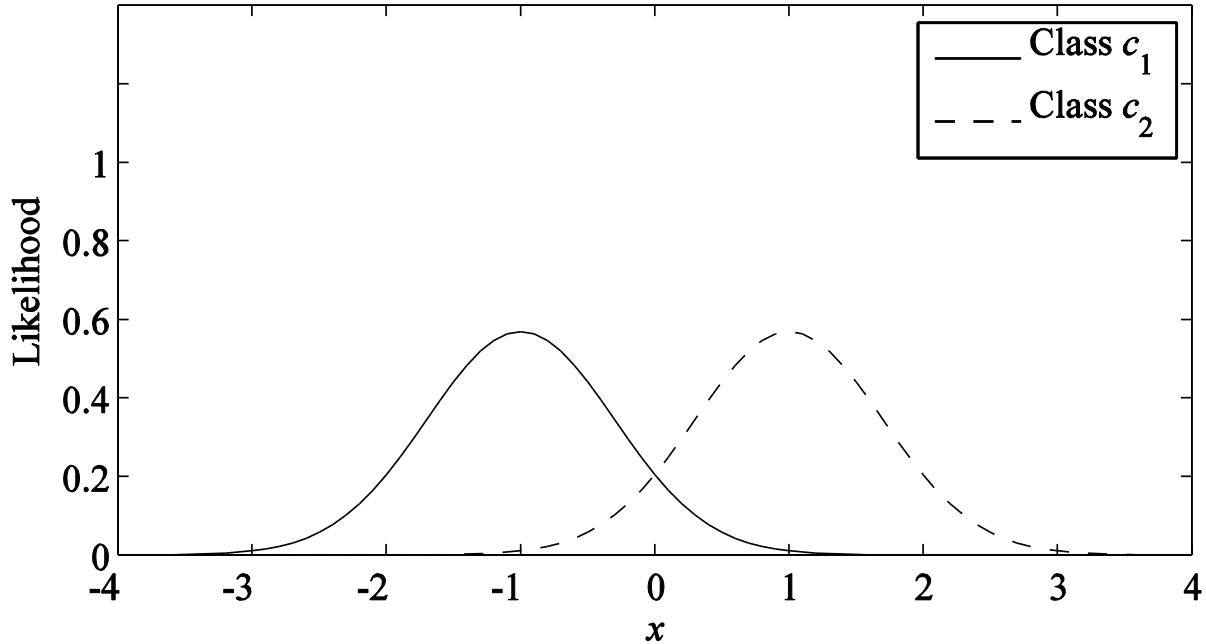
$$g_i(x) = -\frac{(x-\mu_{c_i})^2}{2\sigma_{c_i}^2} - \frac{1}{2}\ln 2\pi - \ln \sigma_{c_i} + \ln p(c_i) - \ln p(x)$$

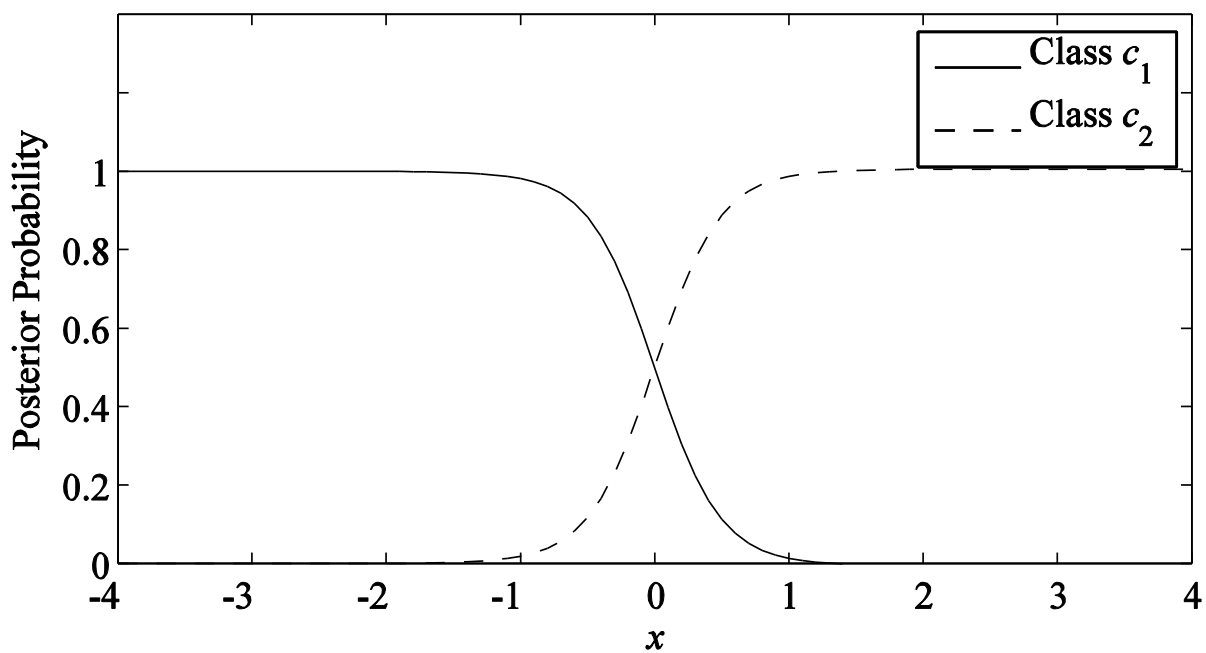
ถ้าสมมติให้ความน่าจะเป็นก่อนและความแปรปรวนของทั้งสองคลาสมีค่าเท่ากัน แล้วจะได้ขอบตัดสินใจเป็น

$$\begin{aligned} g(x) &= g_1(x) - g_2(x) \\ &= -\frac{(x-\mu_{c_1})^2}{2\sigma_c^2} + \frac{(x-\mu_{c_2})^2}{2\sigma_c^2} \\ &= 0 \end{aligned}$$

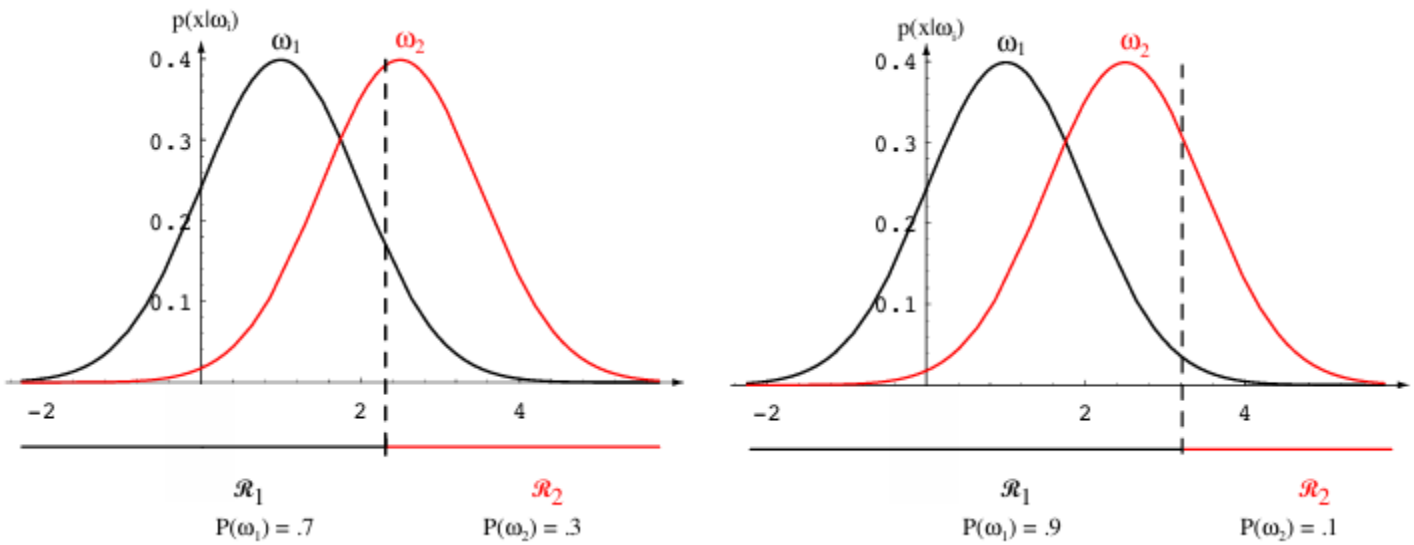
นั่นคือ

$$x = \frac{\mu_{c_1} + \mu_{c_2}}{2}$$


Class c_1
Class c_2
Likelihood
1
0.8
0.6
0.4
0.2
0
-4
-3
-2
-1
0
1
2
3
4
x


Class c_1
Class c_2
Posterior Probability
1
0.8
0.6
0.4
0.2
0
-4
-3
-2
-1
0
1
2
3
4
x

ผลจากค่า Prior Probability



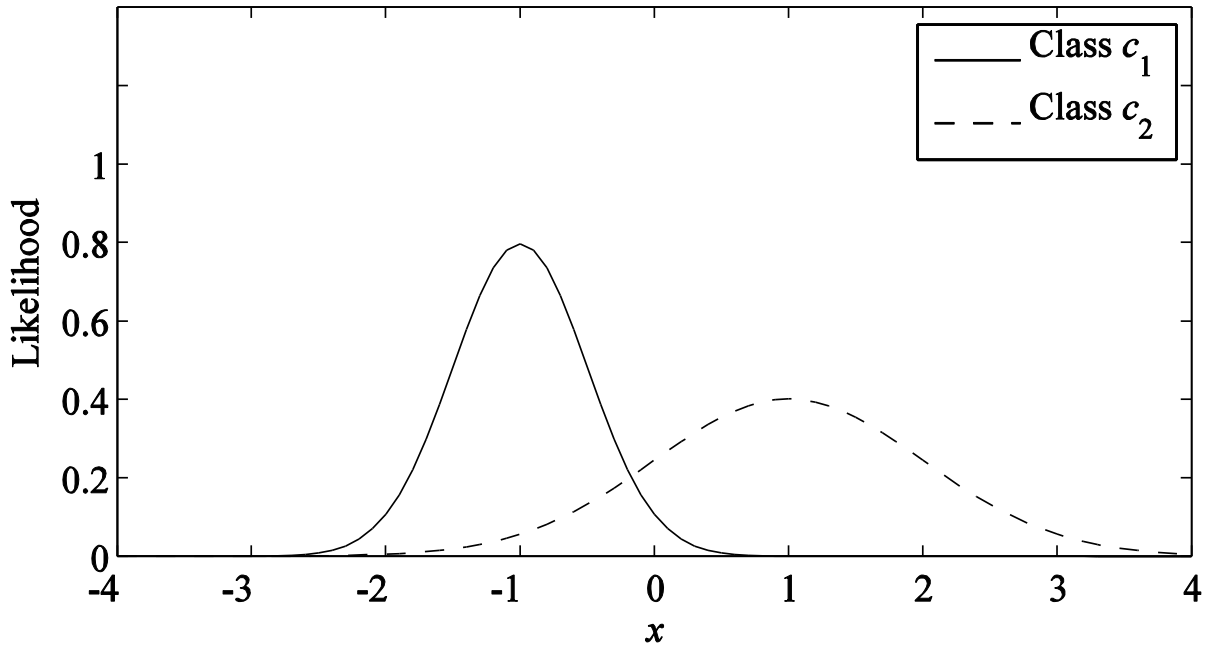
Double Threshold

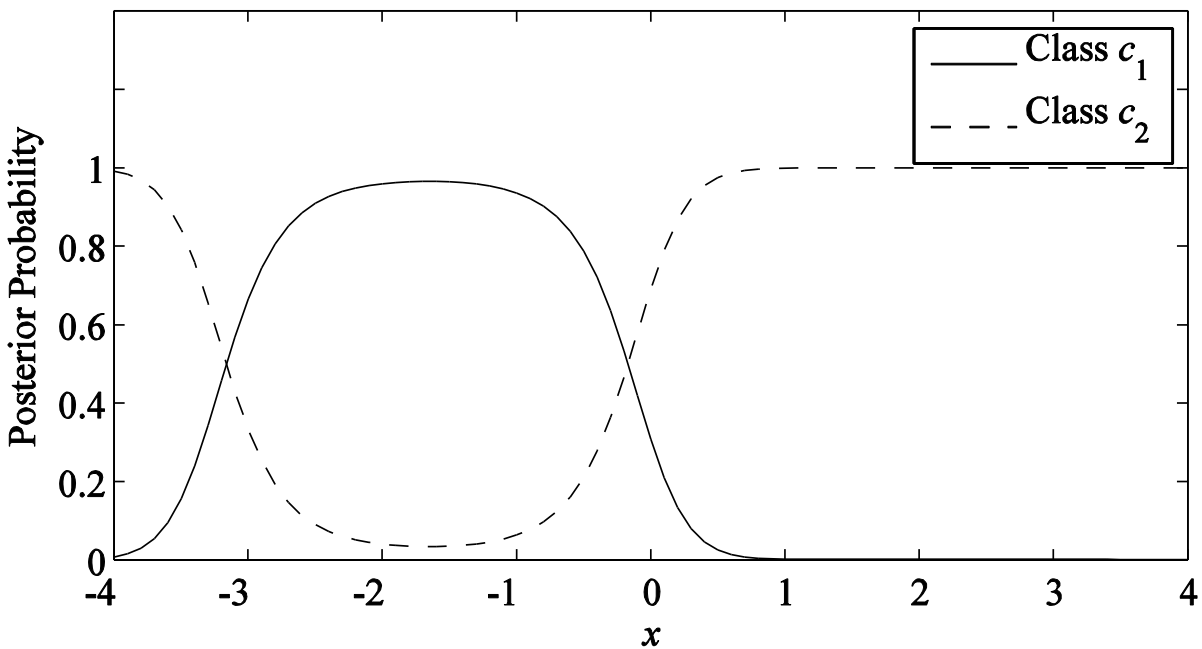
$$g_i(x) = -\frac{(x-\mu_{c_i})^2}{2\sigma_{c_i}^2} - \frac{1}{2}\ln 2\pi - \ln\sigma_{c_i} + \ln p(c_i) - \ln p(x)$$

ถ้าสมมติให้ความน่าจะเป็นก่อนของทั้งสองคลาสมีค่าเท่ากัน แต่ความแปรปรวนของทั้งคลาสมีค่าไม่เท่ากัน แล้วจะได้ขอบตัดสินใจเป็น

$$\begin{aligned} g(x) &= g_1(x) - g_2(x) \\ &= -\frac{(x-\mu_{c_1})^2}{2\sigma_{c_1}^2} - \ln\sigma_{c_1} + \frac{(x-\mu_{c_2})^2}{2\sigma_{c_2}^2} + \ln\sigma_{c_2} \\ &= 0 \end{aligned}$$

ขอบตัดสินใจจะเปลี่ยนเป็นสองขีดเริ่มเปลี่ยน


Class c_1
Class c_2
Likelihood
1
0.8
0.6
0.4
0.2
0
-4
-3
-2
-1
0
1
2
3
4
x


Class c_1
Class c_2
Posterior Probability
1
0.8
0.6
0.4
0.2
0
-4
-3
-2
-1
0
1
2
3
4
x

ตัวจำแนกแบบเบส์สำหรับการแจกแจงปรกติหลายตัวแปร

$$p(\mathbf{r}) = \frac{1}{(2\pi)^{D/2}\left|\mathbf{\Sigma}\right|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{r}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{r}-\boldsymbol{\mu})\right]$$

เมื่อ

$$\boldsymbol{\mu} = E[\mathbf{r}] = \int_{-\infty}^{\infty} \mathbf{r}p(\mathbf{r})d\mathbf{r}$$

และ

$$\mathbf{\Sigma} = E[(\mathbf{r}-\boldsymbol{\mu})(\mathbf{r}-\boldsymbol{\mu})^T] = \int^{\infty} (\mathbf{r}-\boldsymbol{\mu})(\mathbf{r}-\boldsymbol{\mu})^T p(\mathbf{r})d\mathbf{r}$$

ความแปรปรวนร่วมเกี่ยว (Covariance)

พิจารณากรณีสองตัวแปร

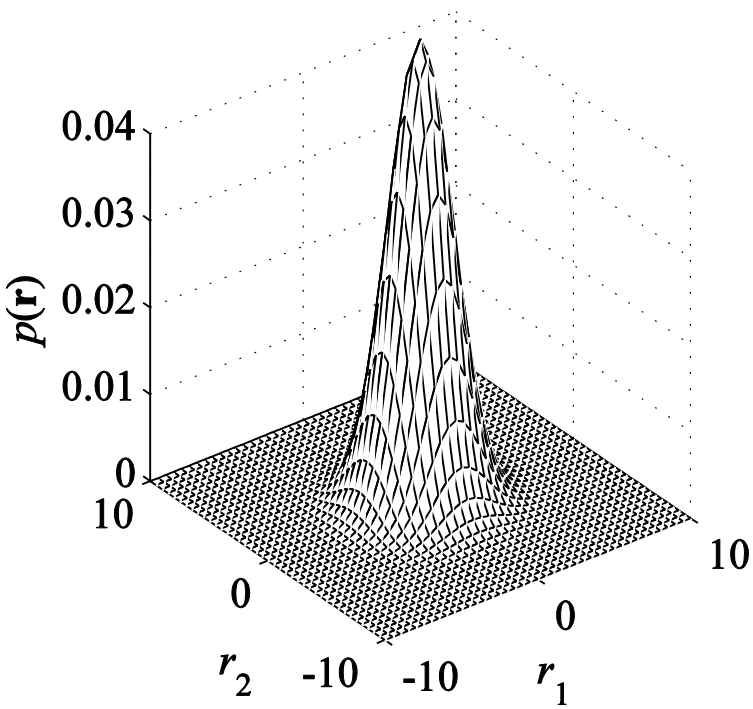
$$\mathbf{\Sigma} = E\left[\begin{bmatrix} r_1 - \mu_1 \\ r_2 - \mu_2 \end{bmatrix}\begin{bmatrix} r_1 - \mu_1 & r_2 - \mu_2 \end{bmatrix}\right]$$
$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

ความแปรปรวนร่วมเกี่ยวระหว่างตัวแปรสุ่มแสดงถึงสหสัมพันธ์ (Correlation) ระหว่างตัวแปรสุ่ม ถ้าตัวแปรสุ่มเป็นอิสระต่อกันทางสถิติ แล้วความแปรปรวนร่วมเกี่ยวจะมีค่าเท่ากับศูนย์

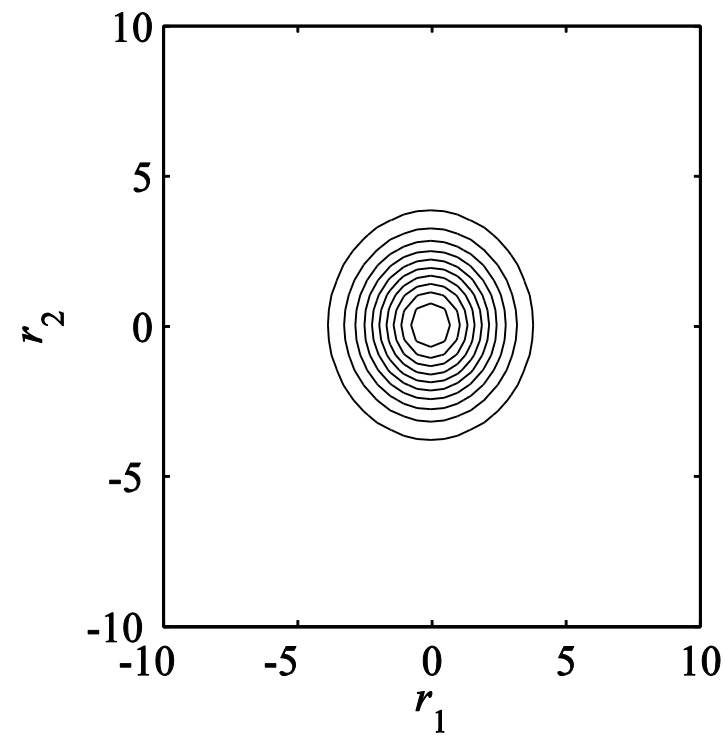
ความแปรปรวนร่วมเกี่ยวจะเป็นความแปรปรวนร่วมเกี่ยวทรงกลม (Spherical Covariance)

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$

ฟังก์ชันความหนาแน่นความน่าจะเป็น



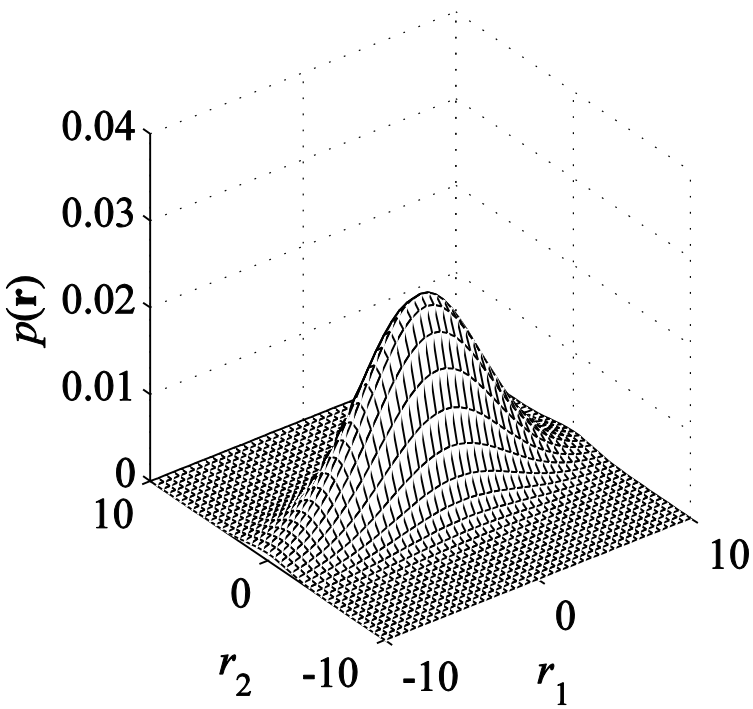
เส้นโค้งค่าเสมอ (Isovalue Curve)



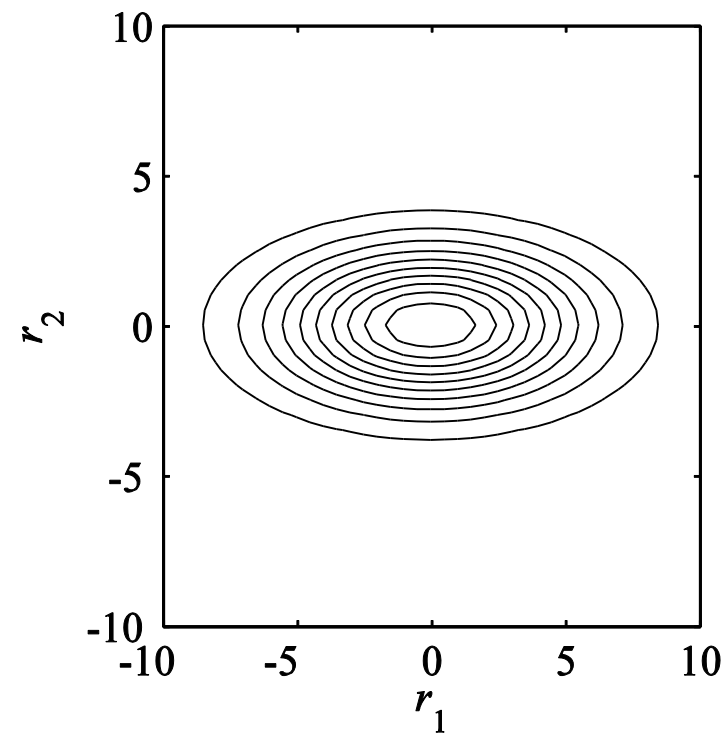
ความแปรปรวนร่วมเกี่ยวจะเป็นความแปรปรวนร่วมเกี่ยวทรงรี (Ellipsoid Covariance)

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

ฟังก์ชันความหนาแน่นความน่าจะเป็น

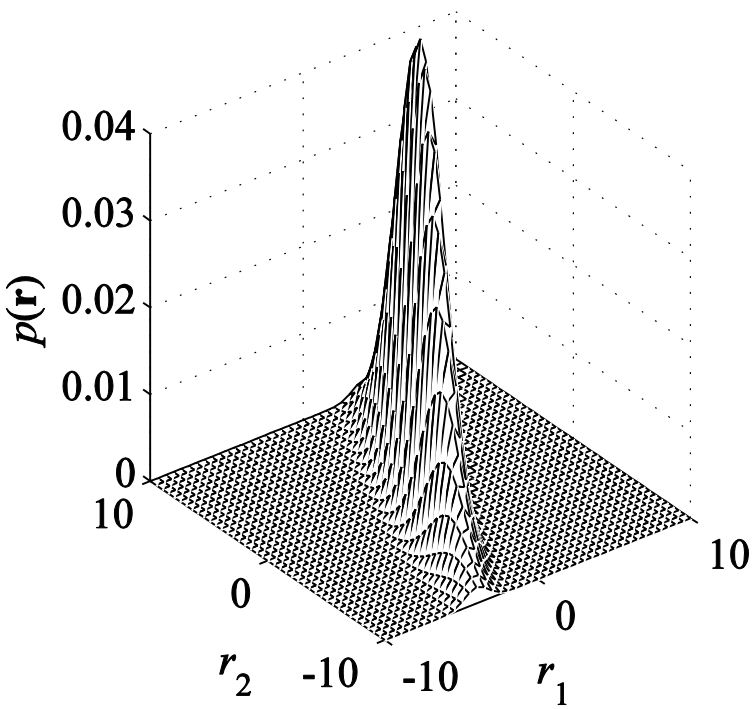


เส้นโค้งค่าเสมอ (Isovalue Curve)

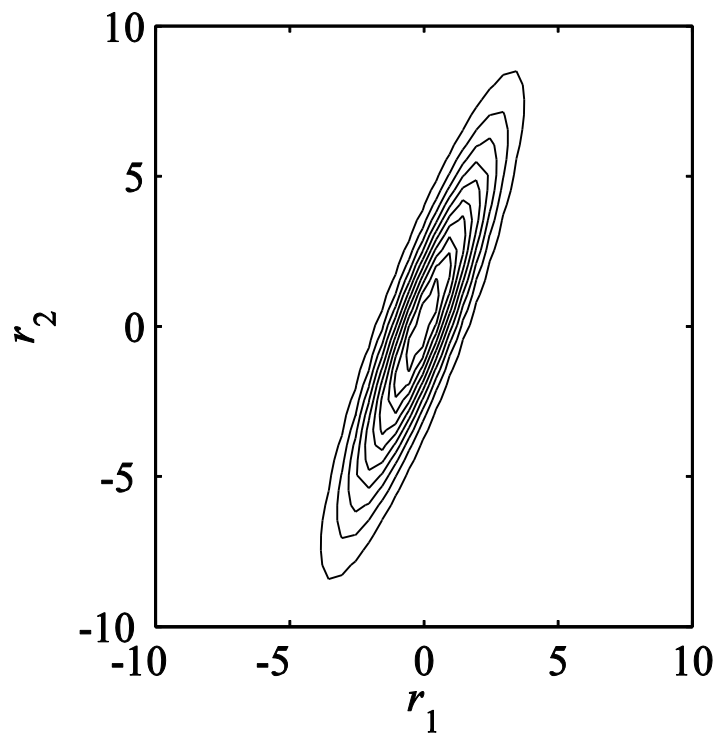


ความแปรปรวนร่วมเกี่ยวเป็นความแปรปรวนร่วมเกี่ยวเต็มอัตรา (Full Covariance)

ฟังก์ชันความหนาแน่นความน่าจะเป็น



เส้นโค้งค่าเสมอ (Isovalue Curve)



ตัวจำแนกกำลังสอง (Quadratic Discriminant)

$$p(\mathbf{r}) = \frac{1}{(2\pi)^{D/2}\left|\mathbf{\Sigma}\right|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{r}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{r}-\boldsymbol{\mu})\right]$$

เมื่อใช้การแจกแจงปรกติหลายตัวแปรเป็นฟังก์ชันความหนาแน่นความน่าจะเป็นมีเงื่อนไขคลาสจะได้ฟังก์ชันจำแนก

$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{c_i})^T \mathbf{\Sigma}_{c_i}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{c_i}) - \frac{|Attribute|}{2}\ln 2\pi - \frac{1}{2}\ln\left|\mathbf{\Sigma}_{c_i}\right| + \ln p(c_i) \\ &\quad -\ln p(\mathbf{x}) \\ &= -\frac{1}{2}\mathbf{x}^T \mathbf{\Sigma}_{c_i}^{-1}\mathbf{x} + \boldsymbol{\mu}_{c_i}^T \mathbf{\Sigma}_{c_i}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_{c_i}^T \mathbf{\Sigma}_{c_i}^{-1}\boldsymbol{\mu}_{c_i} - \frac{|Attribute|}{2}\ln 2\pi - \frac{1}{2}\ln\left|\mathbf{\Sigma}_{c_i}\right| + \ln p(c_i) \\ &\quad -\ln p(\mathbf{x}) \end{aligned}$$

เพื่อแสดงขั้นตอนการคำนวณหาขอบตัดสินใจของตัวจำแนกกำลังสอง พิจารณาปัญหาการจำแนกซึ่งประกอบด้วยสองคลาสและตัวอย่างเข้า $\mathbf{x}$ ประกอบด้วยสองลักษณะประจำ กำหนดให้ความน่าจะเป็นก่อน $p(c_1) = p(c_2)$, ค่าเฉลี่ยของตัวอย่างเข้าจากคลาส c_1 คือ $\mu_{c_1} = \begin{bmatrix} 2 & 0 \end{bmatrix}^T$, ค่าเฉลี่ยของตัวอย่างเข้าจากคลาส c_2 คือ $\mu_{c_2} = \begin{bmatrix} 1 & 0 \end{bmatrix}^T$, เมทริกซ์ความแปรปรวนร่วมเกี่ยวของตัวอย่างเข้าจากคลาส c_1 คือ $\mathbf{\Sigma}_{c_1} = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.25 \end{bmatrix}$ และเมทริกซ์ความแปรปรวนร่วมเกี่ยวของตัวอย่างเข้าจากคลาส c_2 คือ $\mathbf{\Sigma}_{c_2} = \begin{bmatrix} 0.25 & 0 \\ 0 & 0.5 \end{bmatrix}$ จะได้ฟังก์ชันจำแนกของคลาส c_1 หรือ $g_1(\mathbf{x})$ ดังนี้

$$g_1(\mathbf{x}) = -\frac{1}{2}\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 2 & 0 \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$-\frac{1}{2}\begin{bmatrix} 2 & 0 \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}\begin{bmatrix} 2 \\ 0 \end{bmatrix} - \ln 2\pi - \frac{1}{2}\ln\begin{vmatrix} 0.5 & 0 \\ 0 & 0.25 \end{vmatrix} + \ln 0.5 - \ln p(\mathbf{x}) \qquad (1\text{-}25)$$

$$= -x_1^2 - 2x_2^2 + 4x_1 - 4 - \ln 2\pi - \frac{1}{2}\ln 0.125 + \ln 0.5 - \ln p(\mathbf{x})$$

$$g_2(\mathbf{x}) = -\frac{1}{2}\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \end{bmatrix}\begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$-\frac{1}{2}\begin{bmatrix} 1 & 0 \end{bmatrix}\begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \ln 2\pi - \frac{1}{2}\ln\begin{vmatrix} 0.25 & 0 \\ 0 & 0.5 \end{vmatrix} + \ln 0.5 - \ln p(\mathbf{x}) \tag{1-26}$$

$$= -2x_1^2 - x_2^2 + 4x_1 - 2 - \ln 2\pi - \frac{1}{2}\ln 0.125 + \ln 0.5 - \ln p(\mathbf{x})$$

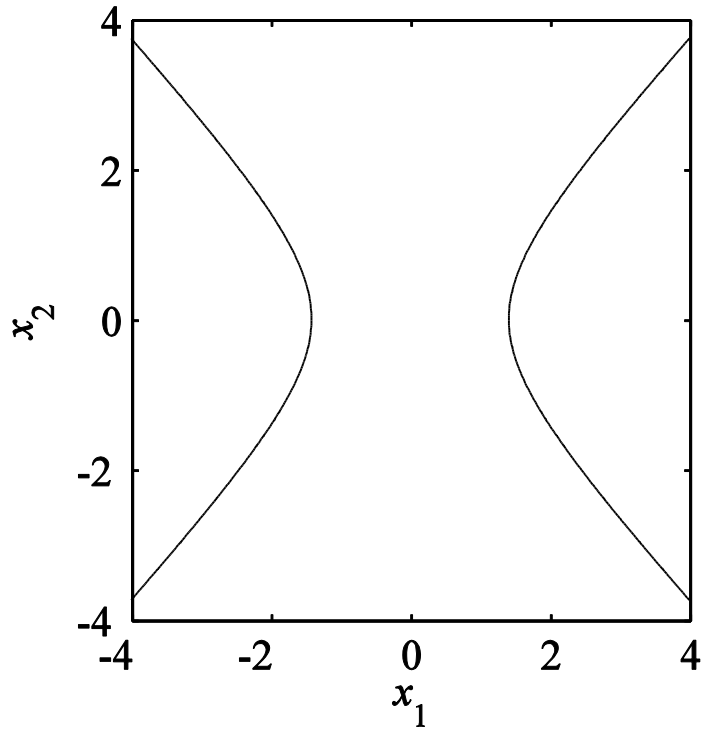
และจะได้ขอบตัดสินใจเป็น

$$\begin{aligned} g(\mathbf{x}) &= g_1(\mathbf{x}) - g_2(\mathbf{x}) \\ &= x_1^2 - x_2^2 - 2 \\ &= 0 \end{aligned} \tag{1-27}$$

หรือ

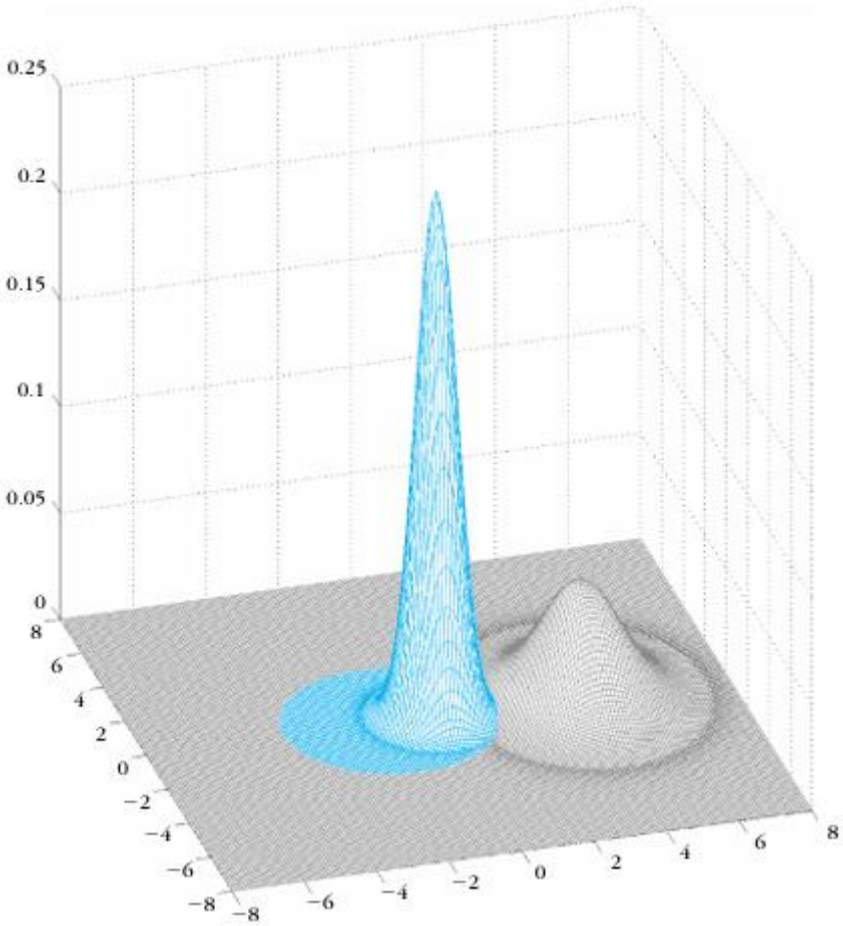
$$\frac{x_1^2}{2} - \frac{x_2^2}{2} = 1 \tag{1-28}$$

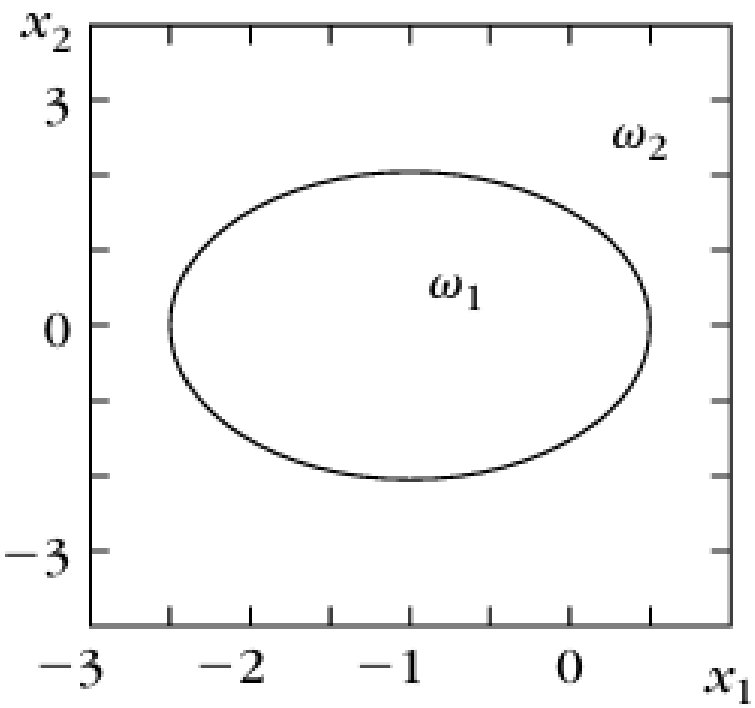
ขอบตัดสินใจของตัวจำแนกกำลังสอง



$P(\omega_1) = P(\omega_2)$, $\boldsymbol{\mu}_1 = [0, 0]^T$ and $\boldsymbol{\mu}_2 = [4, 0]^T$. The covariance matrices for the two classes are

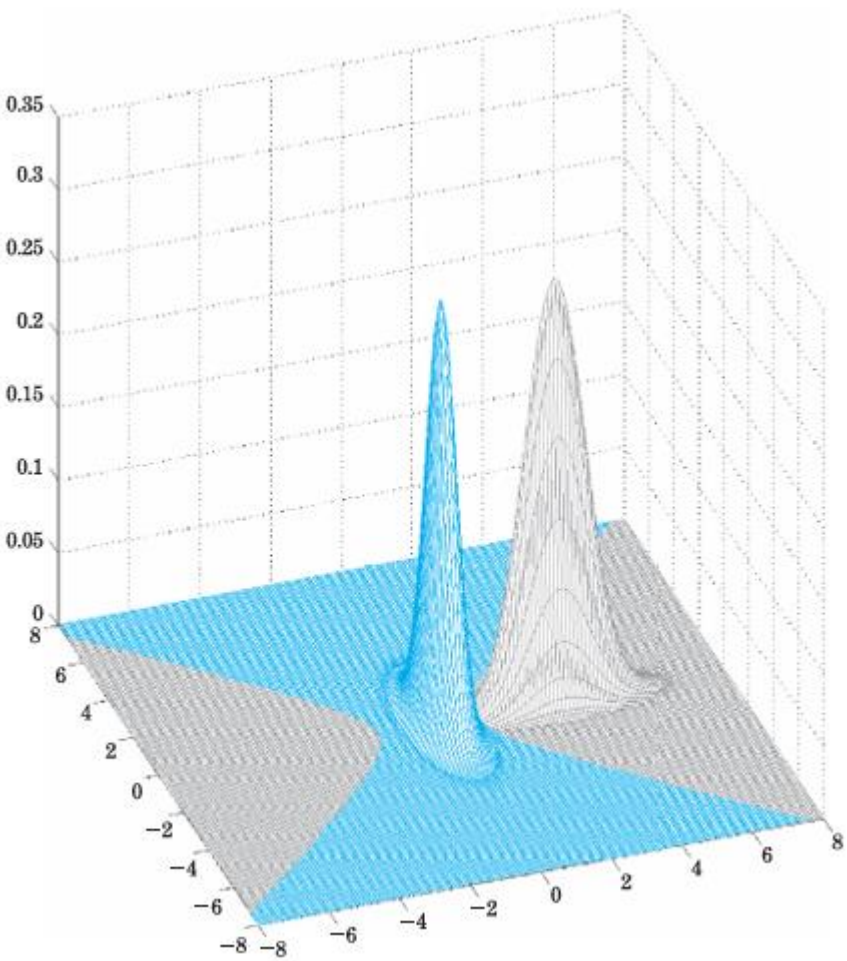
$$\Sigma_1 = \begin{bmatrix} 0.3 & 0.0 \\ 0.0 & 0.35 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1.2 & 0.0 \\ 0.0 & 1.85 \end{bmatrix}$$

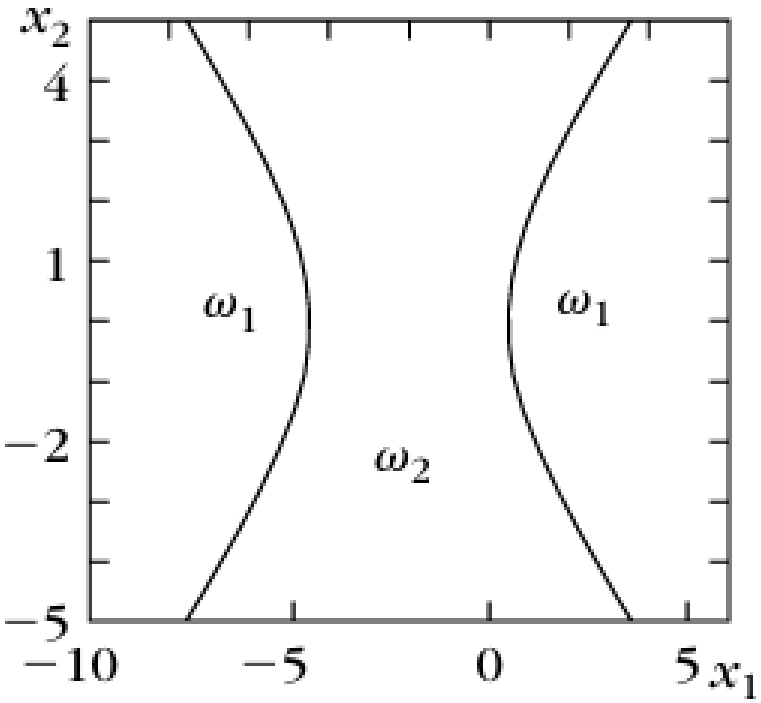




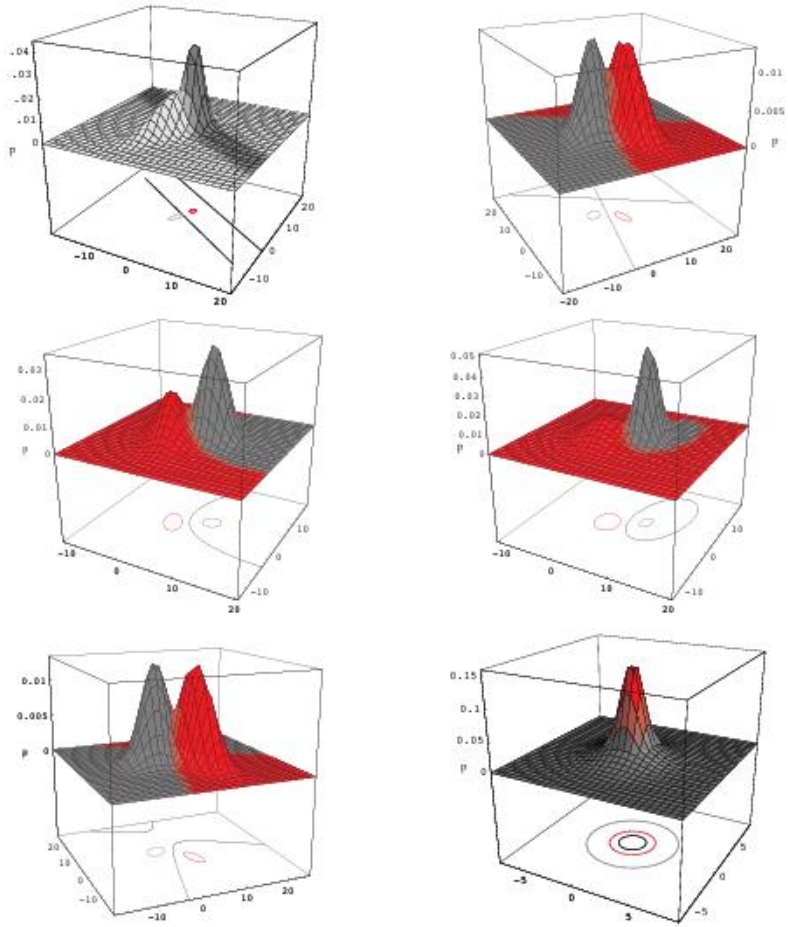
For the case of Figure 2.7b the classes are also equiprobable with $\boldsymbol{\mu}_1 = [0, 0]^T$, $\boldsymbol{\mu}_2 = [3.2, 0]^T$ and covariance matrices

$$\Sigma_1 = \begin{bmatrix} 0.1 & 0.0 \\ 0.0 & 0.75 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 0.75 & 0.0 \\ 0.0 & 0.1 \end{bmatrix}$$





Arbitrary Gaussian distributions



ปัญหา

เมื่อปัญหาการจำแนกประกอบด้วยหลายคลาส ถ้ามีจำนวนลักษณะประจำมากแต่มีจำนวนตัวอย่างน้อย แล้วเมทริกซ์ความแปรปรวนร่วมเกี่ยวของตัวอย่างเข้าจากแต่ละคลาสจะเป็นเมทริกซ์เอกฐาน (Singular Matrix) ส่งผลให้ไม่สามารถสร้างเมทริกซ์ผกผัน (Inverse Matrix)

ปัญหานี้สามารถแก้โดยการรวมตัวอย่างจากทุกคลาสและกำหนดให้เมทริกซ์ความแปรปรวนร่วมเกี่ยวของตัวอย่างเข้าจากแต่ละคลาสเป็นเมทริกซ์ความแปรปรวนร่วมเกี่ยวของตัวอย่างเข้าจากทุกคลาส ส่งผลให้จำนวนพารามิเตอร์ที่ต้องประมาณค่าลดลงและกำจัดปัญหาการสร้างเมทริกซ์ผกผันข้างต้น

ตัวจำแนกเชิงเส้น (Linear Discriminant)

$$p(\mathbf{r}) = \frac{1}{(2\pi)^{D/2}\left|\mathbf{\Sigma}\right|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{r}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{r}-\boldsymbol{\mu})\right]$$

เมื่อเมทริกซ์ความแปรปรวนร่วมเกี่ยวของตัวอย่างเข้าจากแต่ละคลาสได้รับการสมมติให้เป็นเมทริกซ์เดียวกัน ฟังก์ชันจำแนกจะสามารถเขียนได้ดังนี้

$$g_i(x) = -\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_c^{-1}\mathbf{x} + \boldsymbol{\mu}_{c_i}^T\mathbf{\Sigma}_c^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_{c_i}^T\mathbf{\Sigma}_c^{-1}\boldsymbol{\mu}_{c_i} - \frac{|Attribute|}{2}\ln 2\pi - \frac{1}{2}\ln\left|\mathbf{\Sigma}_c\right|$$
$$+\ln p(c_i) - \ln p(\mathbf{x})$$

ดังนั้นขอบตัดสินใจสามารถอธิบายได้โดย

$$g_{ij}(\mathbf{x}) = g_i(\mathbf{x}) - g_j(\mathbf{x})$$
$$= (\boldsymbol{\mu}_{c_i} - \boldsymbol{\mu}_{c_j})^T\mathbf{\Sigma}_c^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_{c_i} - \boldsymbol{\mu}_{c_j})^T\mathbf{\Sigma}_c^{-1}(\boldsymbol{\mu}_{c_i} + \boldsymbol{\mu}_{c_j}) + \ln\left(\frac{p(c_i)}{p(c_j)}\right)$$
$$= 0$$

เพื่อแสดงขั้นตอนการคำนวณหาขอบตัดสินใจของตัวจำแนกเชิงเส้น พิจารณาปัญหาการจำแนกซึ่งประกอบด้วยสองคลาสและตัวอย่างเข้า $\mathbf{x}$ ประกอบด้วยสองลักษณะประจำ กำหนดให้ความน่าจะเป็นก่อน $p(c_1) = p(c_2)$, ค่าเฉลี่ยของตัวอย่างเข้าจากคลาส c_1 คือ $\mu_{c_1} = \begin{bmatrix} -1 & -1 \end{bmatrix}^T$, ค่าเฉลี่ยของตัวอย่างเข้าจากคลาส c_2 คือ $\mu_{c_2} = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$ และเมทริกซ์ความแปรปรวนร่วมเกี่ยวของตัวอย่างเข้าจากแต่ละคลาสคือ $\mathbf{\Sigma}_c = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ จะได้ฟังก์ชันจำแนกของคลาส c_1 หรือ $g_1(\mathbf{x})$ ดังนี้

$$
\begin{aligned}
g_1(\mathbf{x}) &= -\frac{1}{2}\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} -1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\
&\quad -\frac{1}{2}\begin{bmatrix} -1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} -1 \\ -1 \end{bmatrix} - \ln 2\pi - \frac{1}{2}\ln\begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} + \ln 0.5 - \ln p(\mathbf{x}) \qquad (1\text{-}31) \\
&= -\frac{1}{2}x_1^2 - \frac{1}{2}x_2^2 - x_1 - x_2 - 1 - \ln 2\pi + \ln 0.5 - \ln p(\mathbf{x})
\end{aligned}
$$

จะได้ฟังก์ชันจำแนกของคลาส c_2 หรือ $g_2(\mathbf{x})$ ดังนี้

$$g_2(\mathbf{x}) = -\frac{1}{2}\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$-\frac{1}{2}\begin{bmatrix} 1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} - \ln 2\pi - \frac{1}{2}\ln\begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} + \ln 0.5 - \ln p(\mathbf{x}) \qquad (1\text{-}32)$$

$$= -\frac{1}{2}x_1^2 - \frac{1}{2}x_2^2 + x_1 + x_2 - 1 - \ln 2\pi + \ln 0.5 - \ln p(\mathbf{x})$$

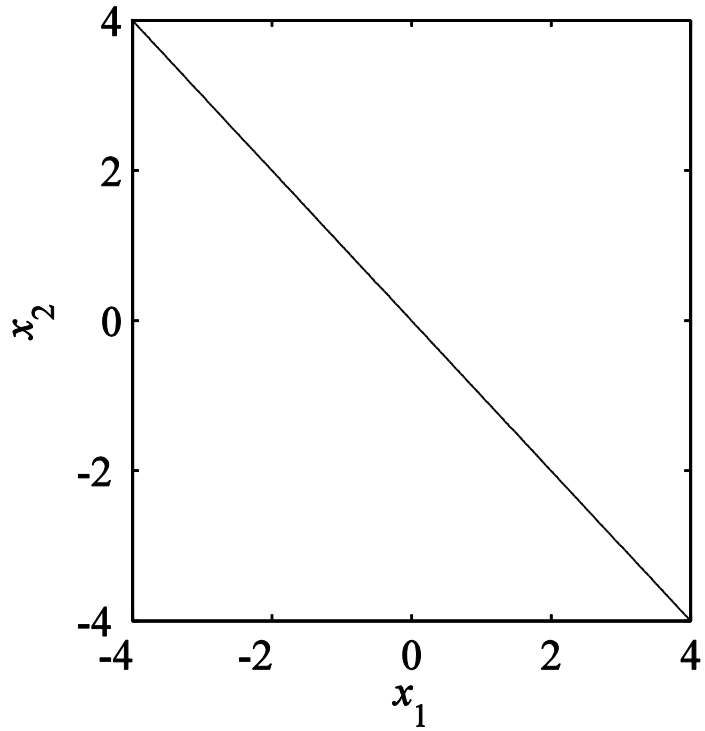
และจะได้ขอบตัดสินใจเป็น

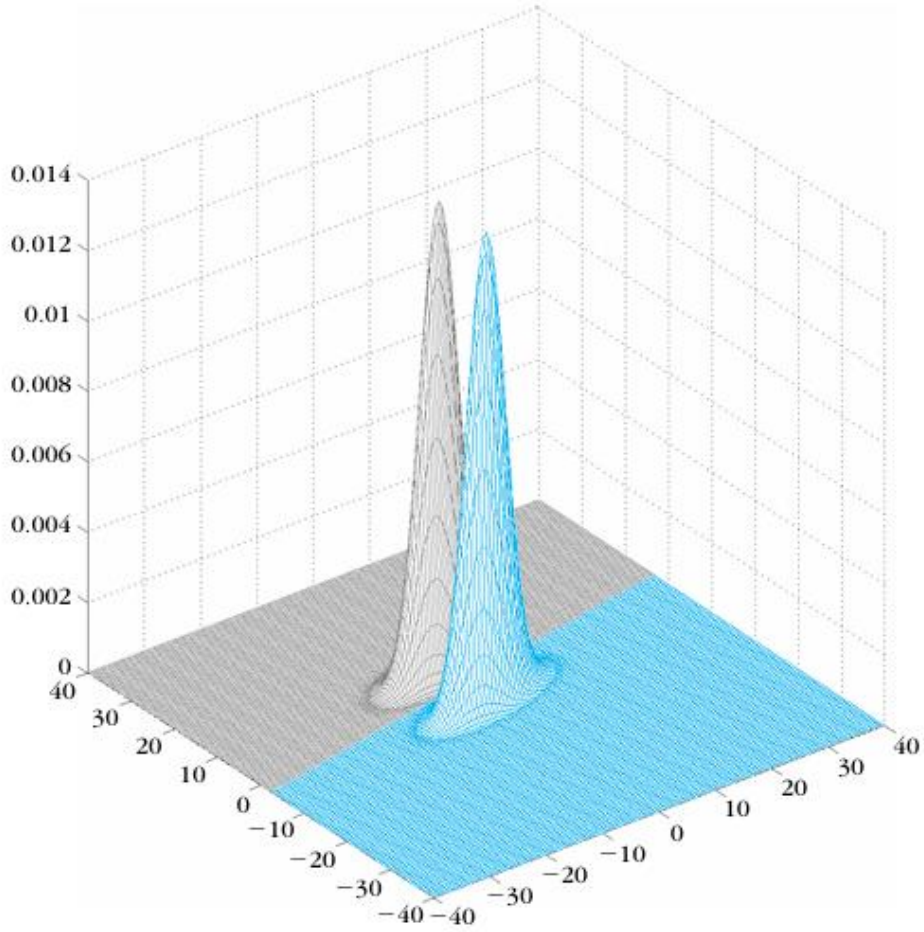
$$\begin{aligned} g(\mathbf{x}) &= g_1(\mathbf{x}) - g_2(\mathbf{x}) \\ &= -2x_1 - 2x_2 \\ &= 0 \end{aligned} \qquad (1\text{-}33)$$

หรือ

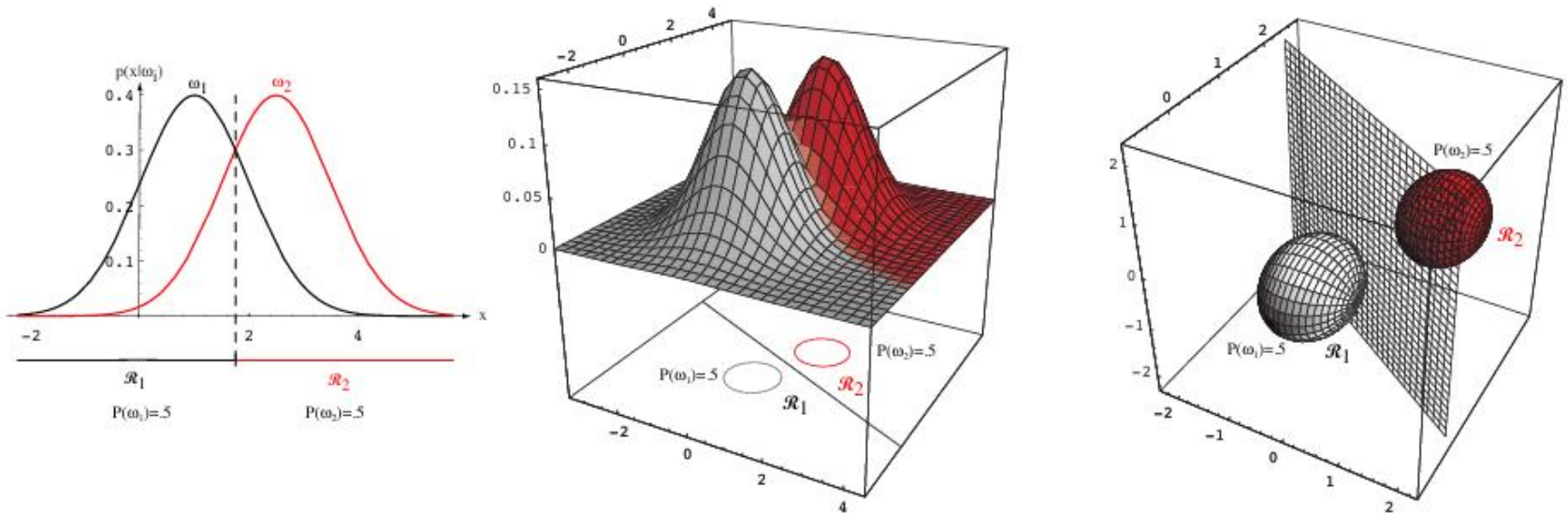
$$x_2 = -x_1 \qquad (1\text{-}34)$$

ขอบตัดสินใจของตัวจำแนกเชิงเส้น

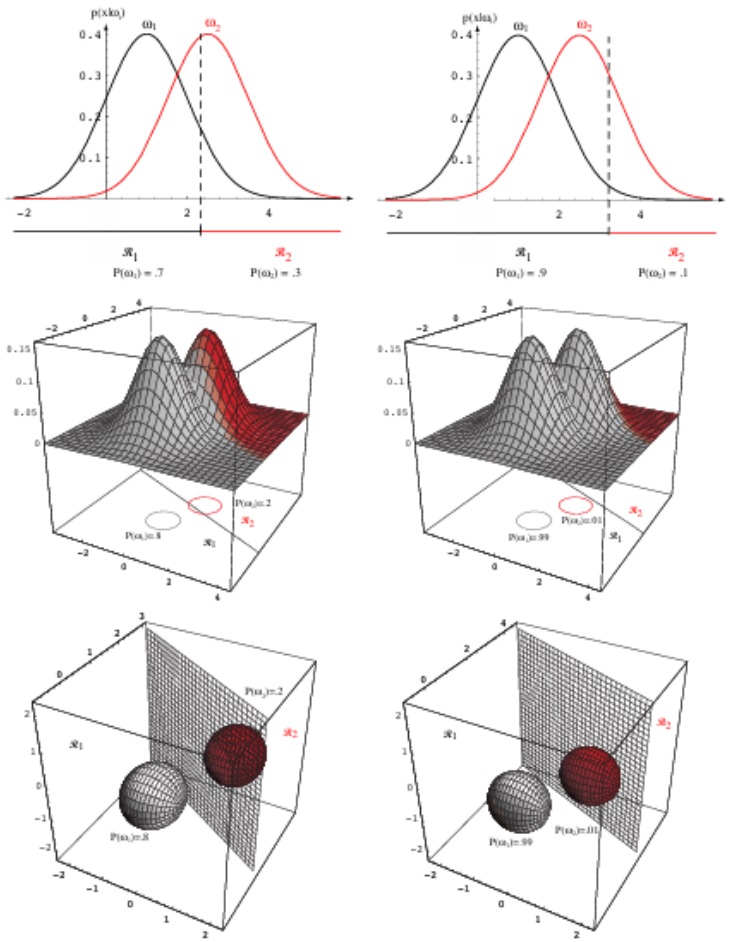



0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
40
30
20
10
0
−10
−20
−30
−40
−40
−30
−20
−10
0
10
20
30
40

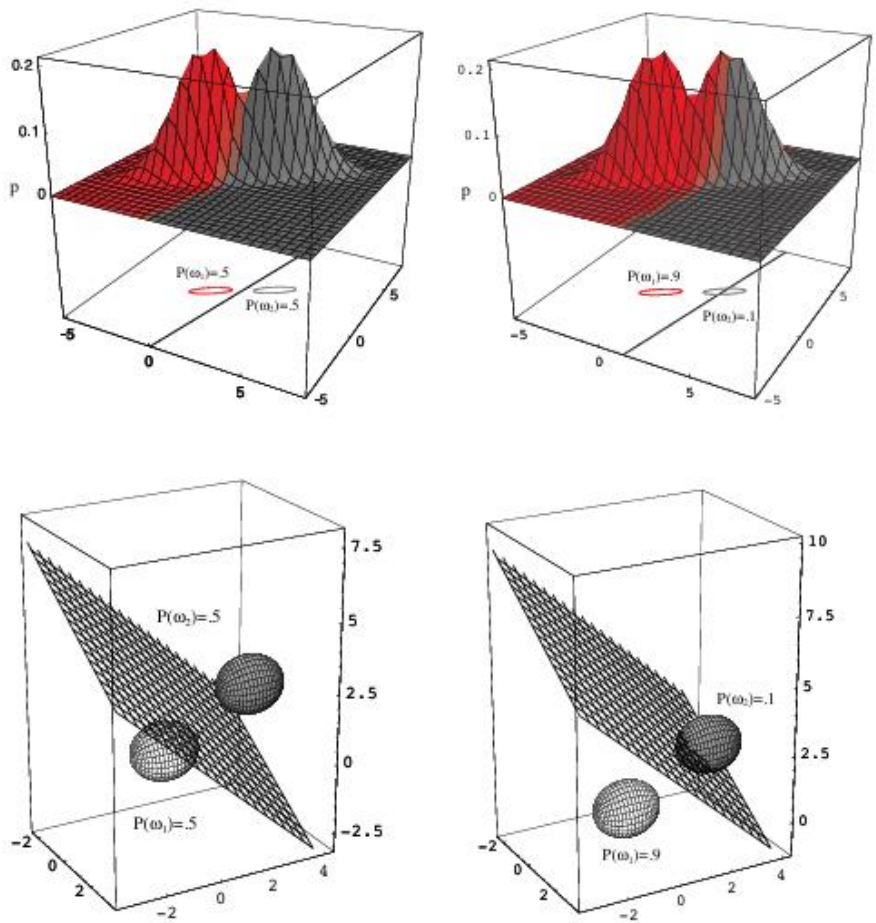
Covariances of two distributions are equal and proportional to the identity matrix



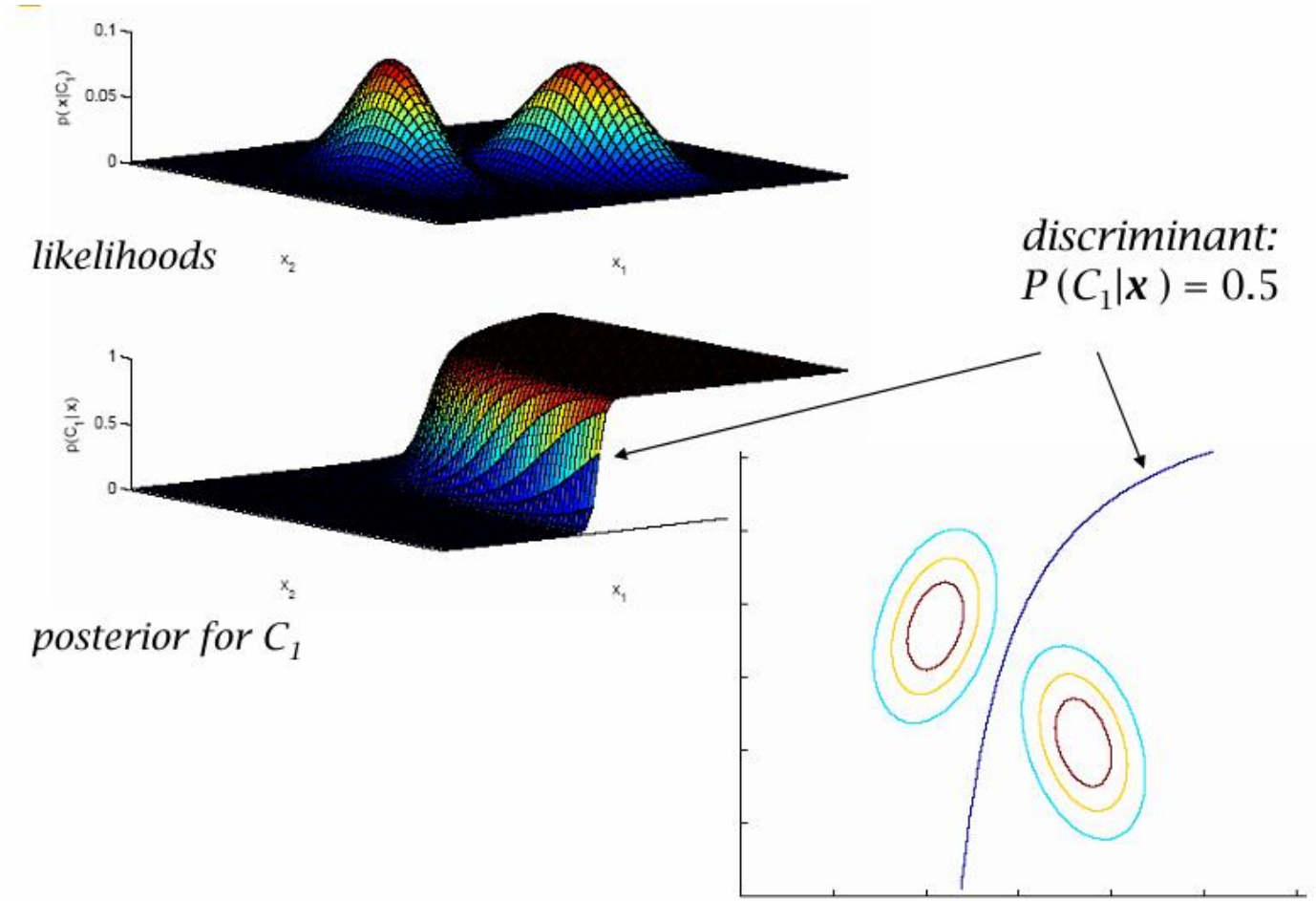
Priors are changed



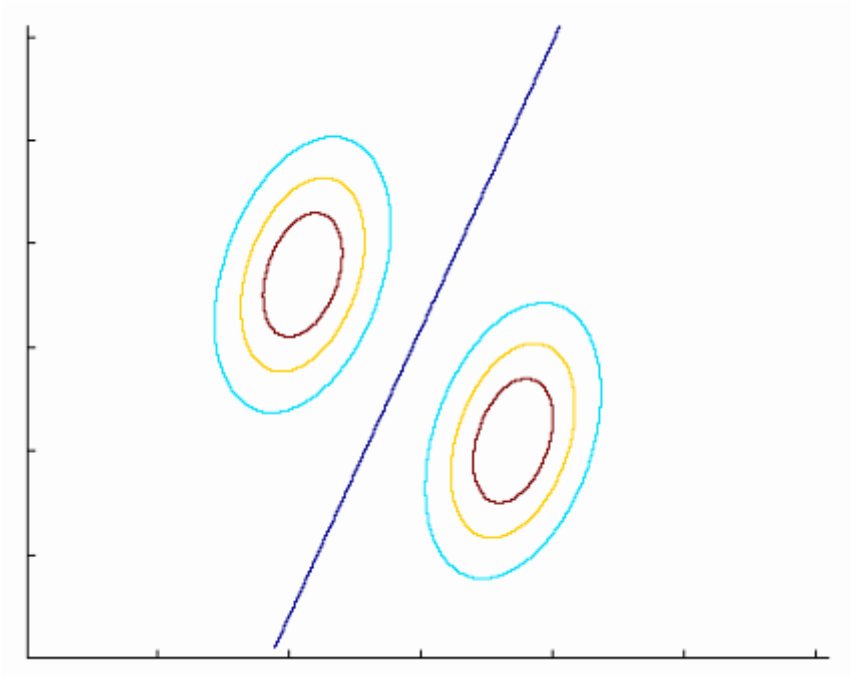
Asymmetric Gaussian distributions



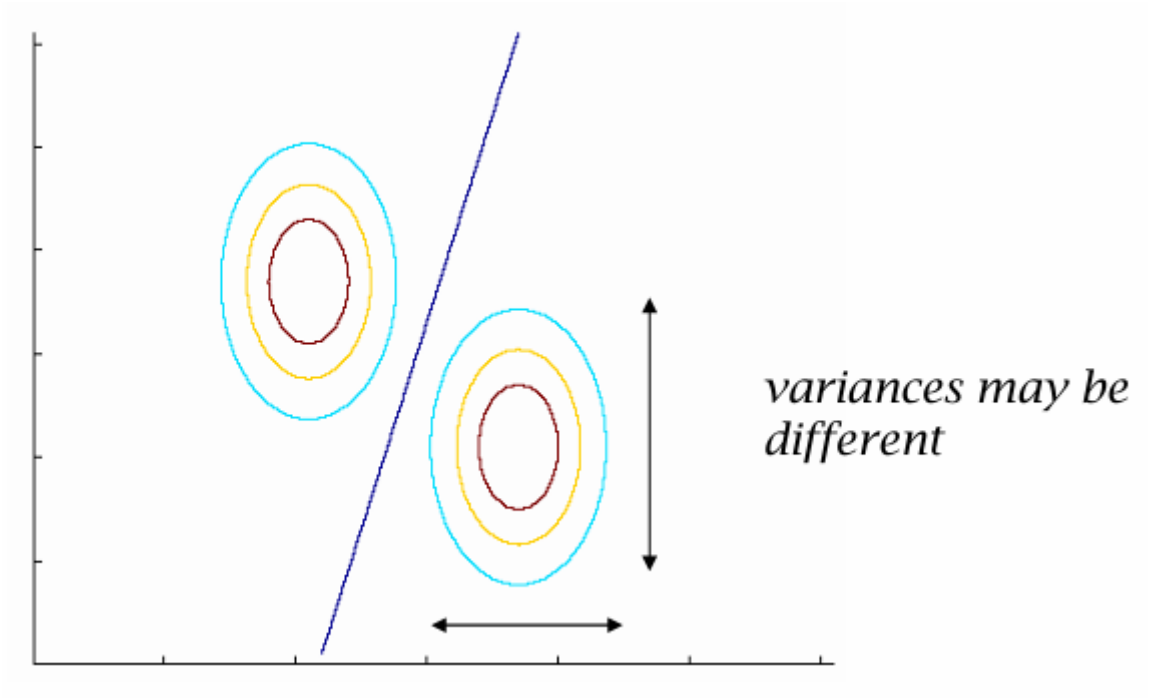
Different cases of the covariance matrices



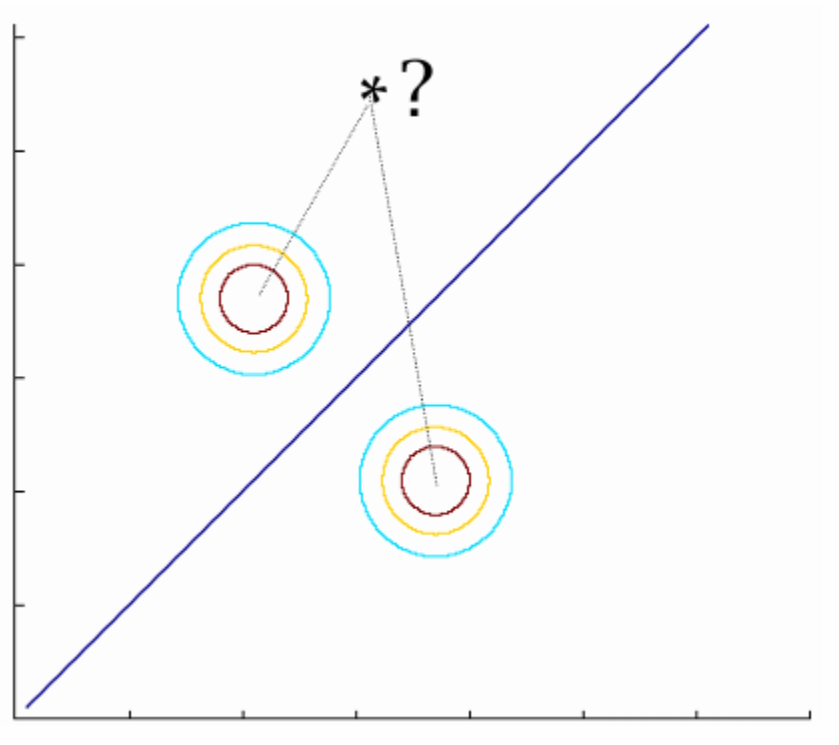
Common Covariance

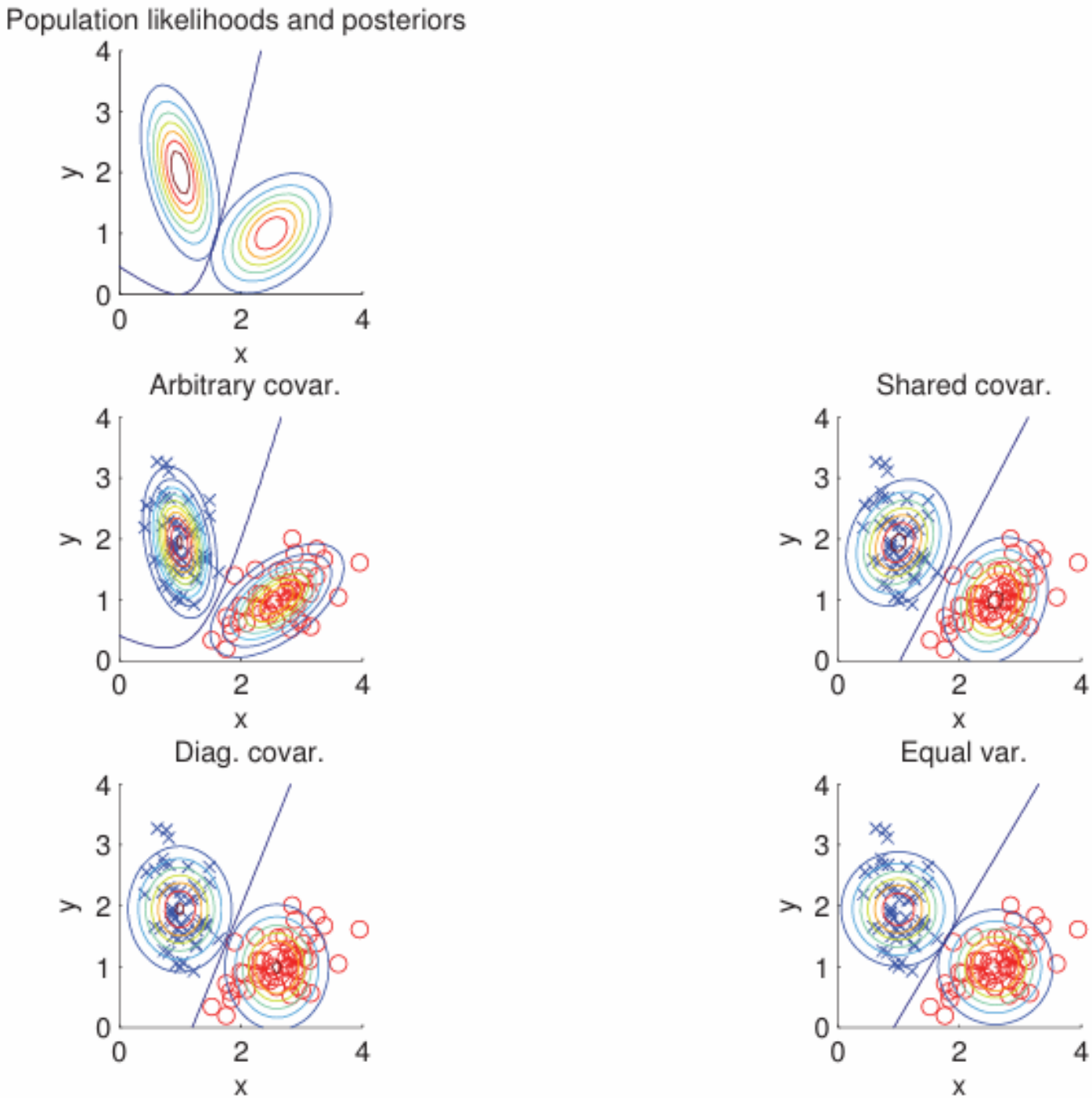


Diagonal Covariance



Diagonal and Equal Covariance




Population likelihoods and posteriors
Arbitrary covar.
Shared covar.
Diag. covar.
Equal var.
x
y
0
1
2
3
4

การประมาณค่าพารามิเตอร์ของการแจกแจง

ถ้าการสร้างตัวจำแนกใช้การสมมติรูปแบบการแจกแจงให้กับฟังก์ชันความหนาแน่นความน่าจะเป็นมีเงื่อนไขคลาสแล้วการประมาณฟังก์ชันความหนาแน่นความน่าจะเป็นมีเงื่อนไขคลาส จะมีความซับซ้อนลดลงและเหลือเพียงการประมาณค่าพารามิเตอร์ของการแจกแจงเท่านั้น

การประมาณความควรจะเป็นสูงสุด (Maximum Likelihood Estimation)

การประมาณความควรจะเป็นสูงสุด

กำหนดให้ $R = \{r_1, r_2, \ldots, r_{|R|}\}$ เป็นเซตตัวแปรสุ่มที่มีการแจกแจงเดียวกันและเป็นอิสระต่อกัน (Independent and Identical Distribution) และให้ตัวแปรเหล่านี้ได้รับการสุ่มโดยใช้ฟังก์ชันควรจะเป็นตัวแปรเดียว $l(r_i \mid \theta)$ ซึ่งสามารถนิยามโดย

$$l(r_i \mid \theta) = p(r_i \mid \theta) \tag{1-43}$$

โดยที่ $p(r_i \mid \theta)$ คือความน่าจะเป็นที่ตัวแปร r_i ได้รับการสุ่มโดยอิงจากการแจกแจงซึ่งมีพารามิเตอร์เป็น θ เนื่องจากแต่ละตัวแปรสุ่มเป็นอิสระต่อกัน ดังนั้นฟังก์ชันควรจะเป็นมีเงื่อนไขตัวแปรสุ่ม $l(\theta \mid R)$ คือผลคูณของความน่าจะเป็นของแต่ละตัวแปรสุ่ม หรือ

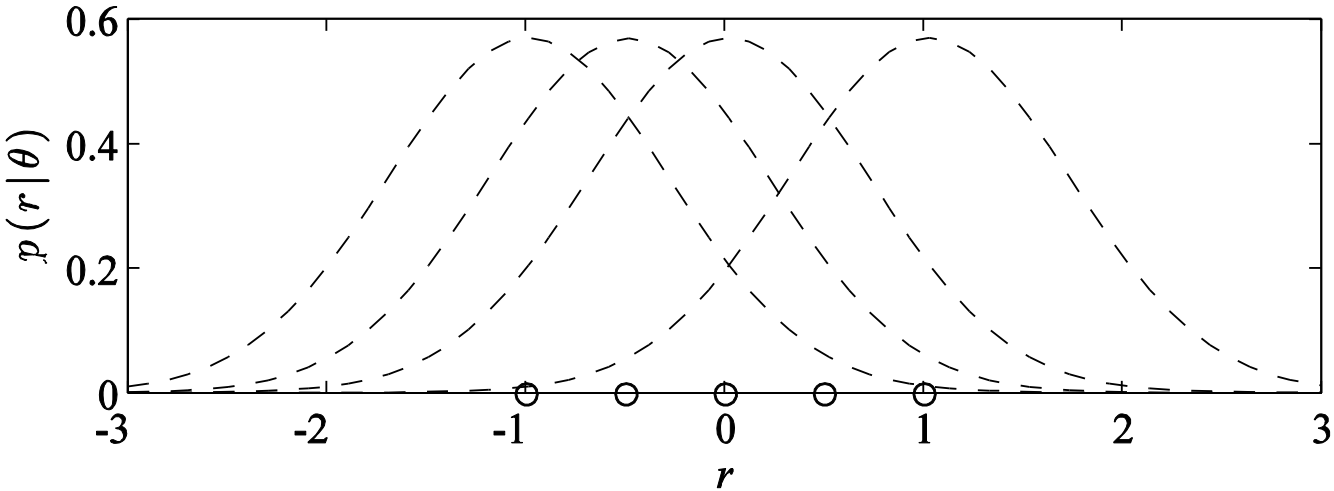
$$l(\theta \mid R) = \prod_{i=1}^{|R|} p(r_i \mid \theta) \tag{1-44}$$

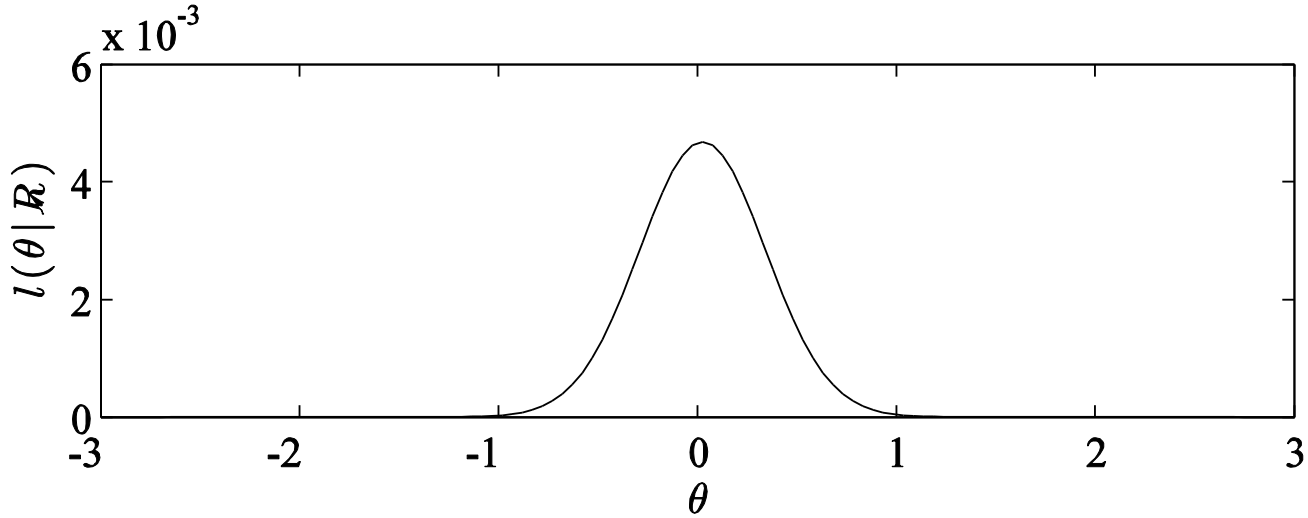
การประมาณความควรจะเป็นสูงสุด

เมื่อใช้ฟังก์ชันลอการิทึม (Logarithm Function) กับสมการที่ (1-44) ฟังก์ชันลอกควรจะเป็น (Log-Likelihood Function) มีเงื่อนไขตัวแปรสุ่ม $L(\theta \mid R)$ จะสามารถอธิบายได้โดย

$$L(\theta \mid R) = \log(l(\theta \mid R)) = \sum_{i=1}^{|R|} \log p(r_i \mid \theta) \qquad (1\text{-}45)$$

ฟังก์ชันควรจะเป็นมีเงื่อนไขตัวแปรสุ่ม





การประมาณค่าพารามิเตอร์ของการแจกแจงปรกติโดยการประมาณความควรจะเป็นสูงสุด

ฟังก์ชันลอกควรจะเป็นมีเงื่อนไขตัวอย่างเข้าจะสามารถอธิบายได้โดย

$$
\begin{aligned}
L(\mu,\sigma^2 \mid S) &= \ln \prod_{i=1}^{|S|} p(x_i \mid \mu,\sigma^2) \\
&= \ln \prod_{i=1}^{|S|} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] \\
&= -\frac{|S|}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{|S|}(x_i-\mu)^2
\end{aligned}
$$

ค่าเฉลี่ยของตัวอย่างเข้า

$$-\frac{|S|}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{|S|}(x_i-\mu)^2$$

ค่าเฉลี่ยที่ทำให้ฟังก์ชันลอกควรจะเป็นมีเงื่อนไขมีค่าสูงสุดคือค่าเฉลี่ยที่ทำให้อนุพันธ์ (Derivative) ของฟังก์ชันลอกควรจะเป็นมีเงื่อนไขมีค่าเป็นศูนย์ นั่นคือ

$$\frac{\partial L(\mu,\sigma^2\mid S)}{\partial\mu}=\frac{1}{\sigma^2}\sum_{i=1}^{|S|}(x_i-\mu)=0$$

ส่งผลให้

$$\mu=\frac{1}{|S|}\sum_{i=1}^{|S|}x_i$$

ความแปรปรวนของตัวอย่างเข้า

$$-\frac{|S|}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{|S|}(x_i-\mu)^2$$

ความแปรปรวนที่ทำให้ฟังก์ชันลอกควรจะเป็นมีเงื่อนไขมีค่าสูงสุดคือ ความแปรปรวนที่ทำให้อนุพันธ์ของฟังก์ชันลอกควรจะเป็นมีเงื่อนไขมีค่าเป็นศูนย์ นั่นคือ

$$\frac{\partial L(\mu,\sigma^2 \mid S)}{\partial \sigma^2}=-\frac{|S|}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{|S|}(x_i-\mu)^2=0$$

ส่งผลให้

$$\sigma^2=\frac{1}{|S|}\sum_{i=1}^{|S|}(x_i-\mu)^2$$

การประมาณค่าพารามิเตอร์ของการแจกแจงปรกติหลายตัวแปร

ฟังก์ชันลอกควรจะเป็นมีเงื่อนไขตัวอย่างเข้าจะสามารถอธิบายได้โดย

$$
\begin{aligned}
L(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid S) &= \ln \prod_{i=1}^{|S|} p(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \ln \prod_{i=1}^{|S|} \frac{1}{(2\pi)^{|Attribute|/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \\
&= -\frac{|S|}{2} \ln((2\pi)^{|Attribute|} |\boldsymbol{\Sigma}|) - \frac{1}{2} \sum_{i=1}^{|S|} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})
\end{aligned}
$$

เวกเตอร์ค่าเฉลี่ยและเมทริกซ์ความแปรปรวนร่วมเกี่ยว

เช่นเดียวกับกรณีการแจกแจงปรกติตัวแปรเดียว เวกเตอร์ค่าเฉลี่ยของตัวอย่างเข้า $\boldsymbol{\mu}$ ที่ทำให้ฟังก์ชันลอกควรจะเป็นมีเงื่อนไขตัวอย่างเข้า $L(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid S)$ มีค่าสูงสุดคือ

$$\boldsymbol{\mu} = \frac{1}{|S|} \sum_{i=1}^{|S|} \mathbf{x}_i \tag{1-52}$$

และเมทริกซ์ความแปรปรวนร่วมเกี่ยวของตัวอย่างเข้า $\boldsymbol{\Sigma}$ ที่ทำให้ฟังก์ชันลอกควรจะเป็นมีเงื่อนไขตัวอย่างเข้า $L(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid S)$ มีค่าสูงสุดคือ

$$\boldsymbol{\Sigma} = \frac{1}{|S|} \sum_{i=1}^{|S|} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \tag{1-53}$$

เบย์อย่างง่าย

- การแสดงแบบจำลอง
  - ใช้ลักษณะประจำทุกตัว
  - ลักษณะประจำแต่ละตัวอิสระต่อกัน
- เกณฑ์ประเมินแบบจำลอง
  - ความน่าจะเป็นภายหลัง (Posterior probability)
- วิธีการค้นหา
  - ไม่ต้องค้นหา

ข้อมูลสภาพภูมิอากาศ

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

ความน่าจะเป็นสำหรับข้อมูลสภาพภูมิอากาศ

Outlook			Temperature			Humidity			Windy			Play	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								

Outlook			Temperature			Humidity			Windy			Play	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

การจำแนกวันใหม่

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

ผลคูณของความน่าจะเป็นของทั้งสองคลาส

- “yes” = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053
- “no” = 3/5 × 1/5 × 4/5 × 3/5 × 5/14 = 0.0206

แปลงให้อยู่ในรูปของความน่าจะเป็น

- P(“yes”) = 0.0053 / (0.0053 + 0.0206) = 0.205
- P(“no”) = 0.0206 / (0.0053 + 0.0206) = 0.795

กฎของเบย์

- ความน่าจะเป็นของเหตุการณ์ (Event) H เมื่อให้หลักฐาน (Evidence) E

$$\Pr[H|E] = \frac{\Pr[E|H]\Pr[H]}{\Pr[E]}$$

- ความน่าจะเป็นก่อนหน้า (Priori probability) ของ H ($\Pr[H]$)
  คือ ความน่าจะเป็นของเหตุการณ์ก่อนที่พบหลักฐาน
- ความน่าจะเป็นภายหลัง (Posteriori probability) of H ($\Pr[H|E]$)
  คือ ความน่าจะเป็นของเหตุการณ์หลังจากพบหลักฐาน

การจำแนกประเภทด้วยเบย์อย่างง่าย

- ความน่าจะเป็นของคลาสเมื่อมีตัวอย่าง
  - หลักฐาน *E* คือ ตัวอย่าง
  - เหตุการณ์ *H* คือ คลาสของตัวอย่าง
- สมมติฐานของเบย์อย่างง่าย คือ หลักฐานแบ่งออกเป็นหลายส่วนที่อิสระต่อกัน (ลักษณะประจำ)

$$\Pr[H|E] = \frac{\Pr[E_1|H]\Pr[E_2|H]\ldots\Pr[E_n|H]\Pr[H]}{\Pr[E]}$$

ปัญหาความถี่ศูนย์

- จะเกิดอะไรขึ้นถ้าค่าของลักษณะประจำไม่ปรากฏในคลาส (เช่น “Humidity = high” สำหรับคลาส “yes”)
  - ความน่าจะเป็นจะมีค่าเป็นศูนย์ Pr[*Humidity* = *High*| *yes*] = 0
  - ความน่าจะเป็นภายหลังจะมีค่าเป็นศูนย์ Pr[*yes*| *E*] = 0
- วิธีแก้ไข
  - บวก 1 ในการนับความถี่ของค่าลักษณะประจำ (Laplace estimator)
- ผลลัพธ์
  - ความน่าจะเป็นจะไม่เป็นศูนย์

การปรับปรุงการประมาณค่าความน่าจะเป็น

- บางกรณีการบวกค่าคงที่ที่ไม่เท่ากับ 1 อาจจะเหมาะสมกว่า
- ตัวอย่าง ลักษณะประจำ outlook สำหรับคลาส yes

Sunny	Overcast	Rainy
$\frac{2+\mu/3}{9+\mu}$	$\frac{4+\mu/3}{9+\mu}$	$\frac{3+\mu/3}{9+\mu}$

- ค่าถ่วงน้ำหนักไม่จำเป็นต้องเท่ากันเสมอไปแต่ต้องรวมกันเท่ากับ 1

$\frac{2+\mu p_1}{9+\mu}$	$\frac{4+\mu p_2}{9+\mu}$	$\frac{3+\mu p_1}{9+\mu}$

ค่าสูญหาย (Missing values)

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

การเรียนรู้: ตัวอย่างนี้จะไม่ถูกรวมเข้าไปในการนับความถี่

การจำแนก: ลักษณะประจำนี้จะไม่นำไปคำนวณ

- Likelihood of “yes” = 3/9 × 3/9 × 3/9 × 9/14 = 0.0238
- Likelihood of “no” = 1/5 × 4/5 × 3/5 × 5/14 = 0.0343
- P(“yes”) = 0.0238 / (0.0238 + 0.0343) = 41%
- P(“no”) = 0.0343 / (0.0238 + 0.0343) = 59%

ลักษณะประจำเชิงตัวเลข

- สมมติฐาน คือ ลักษณะประจำมีการกระจายตัวแบบปกติ
- ฟังก์ชันความหนาแน่นความน่าจะเป็น (Probability density function) สำหรับการแจกแจงแบบปกตินิยามโดยพารามิเตอร์สองตัว:
  - ค่าเฉลี่ย

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - ส่วนเบี่ยงเบนมาตรฐาน

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2$$

  - ฟังก์ชันความหนาแน่น

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

ความน่าจะเป็นสำหรับข้อมูลสภาพภูมิอากาศ

Outlook			Temperature		Humidity		Windy			Play	
	Yes	No	Yes	No	Yes	No		Yes	No	Yes	No
Sunny	2	3	64, 68,	65, 71,	65, 70,	70, 85,	False	6	2	9	5
Overcast	4	0	69, 70,	72, 80,	70, 75,	90, 91,	True	3	3		
Rainy	3	2	72, ...	85, ...	80, ...	95, ...					

Outlook			Temperature		Humidity		Windy			Play	
	Yes	No	Yes	No	Yes	No		Yes	No	Yes	No
Sunny	2/9	3/5	μ =73	μ =75	μ =79	μ =86	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	σ =6.2	σ =7.9	σ =10.2	σ =9.7	True	3/9	3/5		
Rainy	3/9	2/5									

การจำแนกวันใหม่

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	True	?

Likelihood of “yes” = 2/9 × 0.0340 × 0.0221 × 3/9 × 9/14 = 0.000036
Likelihood of “no” = 3/5 × 0.0291 × 0.0380 × 3/5 × 5/14 = 0.000136

P(“yes”) = 0.000036 / (0.000036 + 0. 000136) = 20.9%
P(“no”) = 0.000136 / (0.000036 + 0. 000136) = 79.1%

การแจกแจงแบร์นูลลี

$$p(r) = p_{Success}^{r}(1 - p_{Success})^{1-r}$$

ค่าคาดหมายและความแปรปรวนของตัวแปรสุ่ม r สามารถอธิบายได้โดย

$$E[r] = \sum_{r} rp(r) = 1 \times p_{Success} + 0 \times (1 - p_{Success}) = p_{Success}$$

และ

$$\mathrm{Var}(r) = \sum_{r} (r - E[r])^2 p(r) = p_{Success}(1 - p_{Success})$$

ความหนาแน่นความน่าจะเป็นมีเงื่อนไขคลาส

ถ้าตัวอย่างเข้าประกอบด้วยหลายลักษณะประจำค่าวิยุต แล้วจำนวนค่าความน่าจะเป็นสำเร็จ จะมีการเพิ่มแบบเลขชี้กำลัง (Exponential Growth) ตามจำนวนลักษณะประจำ

ดังนั้นจึงมักสมมติให้แต่ละลักษณะประจำค่าวิยุตเป็นอิสระต่อกัน

เมื่อใช้การแจกแจงแบร์นูลลีหลายตัวแปรซึ่งเป็นอิสระต่อกันเป็นฟังก์ชันความหนาแน่นความน่าจะเป็นมีเงื่อนไขคลาส

$$p(\mathbf{x} \mid c_i) = \prod_{j=1}^{|Attribute|} p_{c_{ij}}^{x_j} (1 - p_{c_{ij}})^{(1-x_j)}$$

ตัวจำแนกเบส์สามัญ (Naïve Bayes Classifier)

ฟังก์ชันจำแนก

$$g_i(\mathbf{x}) = \sum_{j=1}^{|Attribute|} (x_j \ln p_{c_{ij}} + (1-x_j)\ln(1-p_{c_{ij}})) + \ln p(c_i) - \ln p(\mathbf{x})$$

ขอบตัดสินใจ

$$\begin{aligned} g(\mathbf{x}) &= g_1(\mathbf{x}) - g_2(\mathbf{x}) \\ &= \sum_{j=1}^{|Attribute|} \left(x_j \ln\left(\frac{p_{c_{1j}}}{p_{c_{2j}}}\right) + (1-x_j)\ln\left(\frac{1-p_{c_{1j}}}{1-p_{c_{2j}}}\right)\right) + \ln\left(\frac{p(c_1)}{p(c_2)}\right) \\ &= \sum_{j=1}^{|Attribute|} x_j \ln\left(\frac{p_{c_{1j}}(1-p_{c_{2j}})}{p_{c_{2j}}(1-p_{c_{1j}})}\right) + \sum_{j=1}^{|Attribute|} \ln\left(\frac{1-p_{c_{1j}}}{1-p_{c_{2j}}}\right) + \ln\left(\frac{p(c_1)}{p(c_2)}\right) \\ &= 0 \end{aligned}$$

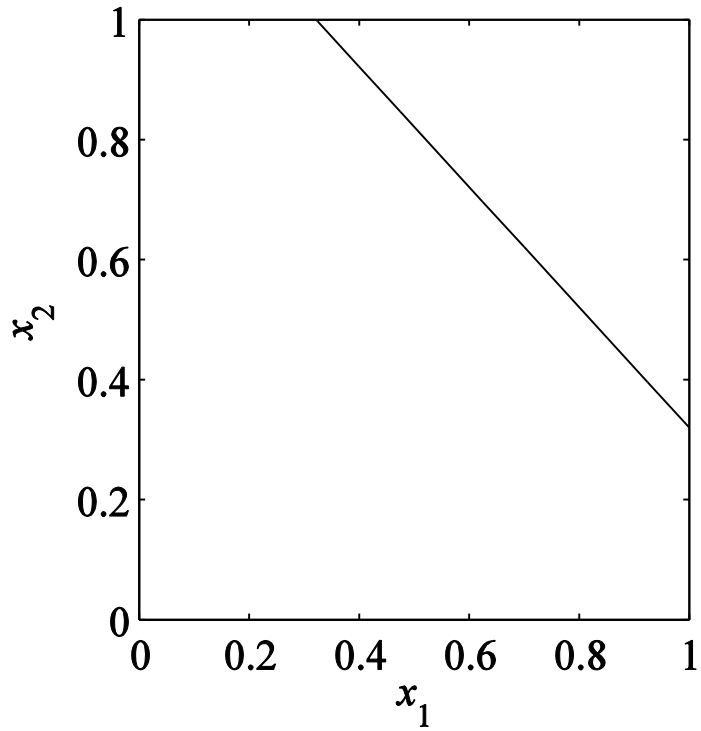
เพื่อแสดงขั้นตอนการคำนวณหาขอบตัดสินใจของตัวจำแนกเบส์สามัญ พิจารณาปัญหาการจำแนกซึ่งประกอบด้วยสองคลาสและตัวอย่างเข้า $\mathbf{x}$ ประกอบด้วยสองลักษณะประจำ กำหนดให้ความน่าจะเป็นก่อน $p(c_1) = p(c_2)$ และความน่าจะเป็นสำเร็จ $p_{c_{11}} = p_{c_{12}} = 0.8$ และ $p_{c_{21}} = p_{c_{22}} = 0.5$ จะได้ขอบตัดสินใจดังนี้

$$
\begin{aligned}
g(\mathbf{x}) &= x_1 \ln\left(\frac{0.8(1-0.5)}{0.5(1-0.8)}\right) + x_2 \ln\left(\frac{0.8(1-0.5)}{0.5(1-0.8)}\right) + 2\ln\left(\frac{1-0.8}{1-0.5}\right) + \ln\left(\frac{0.5}{0.5}\right) \\
&= 1.3863x_1 + 1.3863x_2 - 1.8326 \qquad (1\text{-}41) \\
&= 0
\end{aligned}
$$

หรือ

$$x_2 = -x_1 + 1.3219 \qquad (1\text{-}42)$$

ขอบตัดสินใจของตัวจำแนกเบส์สามัญ



การประมาณค่าพารามิเตอร์ของการแจกแจงแบร์นูลลีโดยการประมาณความควรจะเป็นสูงสุด

ฟังก์ชันลอกควรจะเป็นมีเงื่อนไขตัวอย่างเข้าจะสามารถอธิบายได้โดย

$$L(p_{Success} \mid S) = \ln \prod_{i=1}^{|S|} p_{Success}^{x_i} (1 - p_{Success})^{(1-x_i)}$$

$$= \sum_{i=1}^{|S|} (x_i \ln p_{Success} + (1 - x_i) \ln(1 - p_{Success}))$$

ความน่าจะเป็นสำเร็จที่ทำให้ฟังก์ชันลอกควรจะเป็นมีเงื่อนไข มีค่าสูงสุด คือความน่าจะเป็นสำเร็จที่ทำให้อนุพันธ์ของฟังก์ชันลอกควรจะเป็นมีเงื่อนไขมีค่าเป็นศูนย์ นั่นคือ

$$\frac{dL(p_{success} \mid S)}{dp_{Success}} = \frac{1}{p_{Success}} \sum_{i=1}^{|S|} x_i - \frac{1}{1 - p_{Success}} \left(|S| - \sum_{i=1}^{|S|} x_i \right) = 0$$

ส่งผลให้

$$p_{Successs} = \frac{1}{|S|} \sum_{i=1}^{|S|} x_i$$

การประมาณค่าพารามิเตอร์ของการแจกแจงแบร์นูลลีหลายตัวแปร

ฟังก์ชันลอกควรจะเป็นมีเงื่อนไขตัวอย่างเข้า จะสามารถอธิบายได้โดย

$$L(p_{Success_1},\ldots,p_{Success_{|Attribute|}} \mid S) = \ln \prod_{i=1}^{|S|} \prod_{i=1}^{|Attribute|} p_{Success_j}^{x_{ij}} (1-p_{Success_j})^{(1-x_{ij})}$$

เช่นเดียวกับกรณีการแจกแจงแบร์นูลลีตัวแปรเดียวข้างต้น ความน่าจะเป็นสำเร็จที่ทำให้ฟังก์ชันลอกควรจะเป็นมีเงื่อนไขตัวอย่างเข้า มีค่าสูงสุดคือ

$$p_{Success_j} = \frac{1}{|S|} \sum_{i=1}^{|S|} x_{ij}$$