บทที่ 3
ทฤษฎีที่เกี่ยวข้อง

3.1 บทนำ

การรู้จำรูปแบบ (Pattern Recognition) [17] เป็นศาสตร์ที่ว่าด้วยกระบวนการตัดสินใจที่เกี่ยวกับการจำแนกกลุ่ม (Classification) การจัดกลุ่ม (Clustering) การรู้จำ (Recognition) ศึกษาถึงความแนวคิดต่างๆให้คอมพิวเตอร์สามารถทำงานเหล่านี้ได้โดยใช้เหตุผลหรือคณิตศาสตร์เพื่อหารูปแบบ (Pattern) ซึ่งอาจได้แก่เซตของ การวัด, ข้อสังเกต, หรือคำอธิบายของวัตถุใดๆ โดยจะใช้ความรู้ด้านอื่นๆ เช่น โครงข่ายประสาทเทียม (Neural Network), ทฤษฎีฟัซซี่ (Fuzzy Theory) มาช่วยในการวิเคราะห์เป็นวิทยาการที่สามารถประยุกต์ใช้ได้กับงานทุกสาขาและเป็นพื้นฐานสำคัญสำหรับงานวิจัยในด้านปัญญาประดิษฐ์หรือการสร้างความฉลาดให้คอมพิวเตอร์ ตัวอย่างปัญหาในงานด้านนี้ได้แก่ การทำให้คอมพิวเตอร์รู้ว่าภาพที่เข้ามาเป็นอักษรอะไร เสียงที่เข้ามาเป็นเสียงตัวเลขอะไรหรือคำพูดอะไร ภาพใบหน้าคนเป็นภาพของใคร กระบวนการเหล่านี้เป็นพื้นฐานที่สำคัญของความฉลาดของมนุษย์ซึ่งติดตัวมาตั้งแต่แรกเกิดและยังคงเป็นปัญหาที่ยังท้าทายนักวิจัยอยู่ถึงปัจจุบันและสามารถประยุกต์ใช้ในสาขาอื่นได้อีกมาก

การรู้จำรูปแบบสามารถแบ่งได้เป็น

3.1.1 การรู้จำรูปแบบทางสถิติ (Statistic Pattern Recognition) หรือทฤษฎีการตัดสิน (Decision Theory) โดยจะใช้พื้นฐานของทฤษฎีความน่าจะเป็นในการวิเคราะห์

3.1.2 การรู้จำรูปแบบสังเคราะห์ (Syntactic Pattern Recognition) หรือ Structural Pattern Recognition (Linguistic Method) โดยจะใช้อัลกอริทึมอื่นๆมาวิเคราะห์

สำหรับขั้นตอนการทำงานของกระบวนการ สามารถแบ่งออกได้เป็นสามส่วนใหญ่ดังนี้

1) การเก็บข้อมูล (Data Collection)

2) การประมวลผลข้อมูลเบื้องต้น (Data Pre-Processing) ซึ่งแบ่งออกเป็นสองส่วนย่อย คือ การสร้างและค้นหาคุณลักษณะเด่น (Feature Extraction) และการคัดเลือกคุณลักษณะเด่น (Feature Selection)

3) การจำแนกประเภทข้อมูล (Classification)

ซึ่งแต่ละขั้นตอนจะมีวิธีการที่แตกต่างกันไปขึ้นอยู่กับงานที่นำไปประยุกต์ใช้ว่าวิธีการใดจะเหมาะสมและให้ผลลัพธ์ที่ดีที่สุด

3.2 การเก็บข้อมูล (Data Collection)

การเก็บข้อมูลเป็นแฟ้มข้อมูลเอ็ชทีเอ็มแอลเพื่อนำมาวิเคราะห์และเพื่อนำไปใช้สำหรับงานวิจัย โดยจะแยกแต่ละส่วนตามจุดประสงค์และขอบเขตของงานวิจัยที่ได้ทำการออกแบบโครงสร้าง

3.3 การจัดเตรียมข้อมูล (Data Pre-Processing) ประกอบด้วย 5 ส่วนหลักดังนี้

3.3.1 การสร้างและค้นหาลักษณะ (Feature Extraction)

เป็นการนำข้อมูลดิบที่ได้มาจัดรูปแบบให้อยู่ในค่าหรือลักษณะที่เหมาะสม โดยลักษณะหรือคุณลักษณะเด่นนั้นจะเป็นเวกเตอร์ของคุณลักษณะเด่นของวัตถุ เช่น คนหนึ่งคน อาจกำหนดคุณลักษณะเด่นที่ใช้เป็น น้ำหนัก, ส่วนสูง, อายุ หรือ ชนชาติ เก็บมาเป็นเวกเตอร์ ซึ่งอาจจะเป็นตัวเลข ตัวอักษร หรือ ถูก/ผิด ก็ได้ โดยหลักการในการเลือกคุณลักษณะเด่นจากข้อมูลดิบคือ

1) สามารถปรับเปลี่ยนหรือคำนวณได้
2) สามารถนำไปจำแนกประเภทได้ดี
3) ยังคงมีคุณค่าของข้อมูลเดิมอยู่

3.3.2 การตัดส่วนที่เป็นค่าผิดพลาด (Outlier Removal) สามารถทำได้โดย

1) สร้างระยะจุดเปลี่ยน (Threshold Distance) สำหรับข้อมูลที่เป็นค่าผิดพลาด
2) เลือกข้อมูลที่มีค่าไม่เกินสองหรือสามเท่าของส่วนเบี่ยงเบนมาตรฐาน

3.3.3 การทำข้อมูลให้เป็นบรรทัดฐาน (Data Normalization)

เป็นการจัดการเพื่อให้ค่าของข้อมูลหรือคุณลักษณะเด่นมาอยู่บนบรรทัดฐานเดียวกัน ซึ่งวิธีที่ใช้กันอย่างแพร่หลายคือการแปลงเป็นค่ามาตรฐาน โดยคิดจากค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของข้อมูล แต่สำหรับการประมวลผลเวลาจริง (Real Time) นั้นการหาค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของข้อมูลทำได้ยาก จึงต้องหาวิธีที่เหมาะสมต่อไป

3.3.4 การจัดการข้อมูลที่ขาดหาย (Missing Data)

สำหรับข้อมูลที่ขาดหายหรือไม่ครบ อันจะทำให้ไม่สามารถประมวลผลข้อมูลชุดนั้นได้ วิธีจัดการข้อมูลที่ขาดหายขึ้นอยู่กับความสำคัญของข้อมูลชุดนั้นต่อผลการจำแนก ตัวอย่างวิธีที่ใช้กันโดยทั่วไปเช่น

1) แทนข้อมูลนั้นด้วยค่าทางสถิติของข้อมูลที่เหลือ เช่น ค่าเฉลี่ยของข้อมูล, ค่าต่ำสุดหรือค่าสูงสุด
2) ตัดชุดข้อมูลนั้นออกไปจากระบบ

3.3.5 การคัดเลือกคุณลักษณะเด่น (Feature Selection)

เป็นส่วนการทำงานที่เลือกคุณลักษณะเด่นที่ได้จากการสร้างและค้นหาคุณลักษณะเด่น เพื่อหาลักษณะที่เหมาะสมที่สุดสำหรับแต่ละงาน ซึ่งส่วนใหญ่จะเป็นการหาจำนวนลักษณะที่น้อยที่สุด เพื่อให้ความซับซ้อนของการคำนวณน้อย แต่ให้ผลการจำแนกประเภทข้อมูลได้ผลดีที่สุดโดยวิธีที่เลือกใช้เพื่อเปรียบเทียบผลของค่า

1) ค่าระยะห่าง (Distance) ของข้อมูลระหว่างสองคลาส

การเลือกคุณสมบัติ โดยใช้ค่าระยะห่างระหว่างสองคลาสนั้น การหาค่าระยะห่างสามารถหาได้หลายวิธี และวิธีที่งานวิจัยนี้เลือกมาใช้คือการหาผลต่างระหว่างค่าเฉลี่ยของข้อมูลที่ทำเป็นค่ามาตรฐานแล้วจะได้ค่าคุณลักษณะเด่นของทั้งสองคลาสอยู่บนฐานของข้อมูลชุดเดียวกัน ซึ่งจะเรียกว่าค่าระยะห่างสัมพัทธ์ของค่าเฉลี่ย โดยมีวิธีการดังนี้

1.1 จากข้อมูลที่ผ่านการเปลี่ยนเป็นค่ามาตรฐาน

1.2 นำค่าของแต่ละคุณลักษณะเด่นมาหาค่าความแตกต่าง

1.3 นำค่าแตกต่างที่ได้นำมาเทียบ(หาร)กับค่าของคลาสที่เป็น “Non-Fault”

1.4 ทำการเลือกคุณลักษณะเด่นที่มีค่าเริ่มจากค่าน้อยที่สุดเพื่อหาจำนวนคุณลักษณะเด่นที่เหมาะสม

2) การวิเคราะห์แกนหลัก (Principal Component Analysis : PCA)

เป็นวิธีการลดมิติของชุดข้อมูลให้น้อยลงก่อนนำไปวิเคราะห์ ซึ่งอาจเรียกว่า Karhunen Loeve Transform โดยเป็นวิธีการแปลงของ Harold Hotelling และ PCA นี้เป็นที่นิยมใช้อย่างกว้างขวางในการวิเคราะห์ข้อมูลและสร้างแบบจำลองพยากรณ์ โดยจะใช้การคำนวณค่าลักษณะเฉพาะของเมทริกซ์ความแปรปรวนของข้อมูลและผลของ PCA มักจะแสดงถึงความสำคัญของแต่ละข้อมูล

ในมุมมองทางคณิตศาสตร์นั้น PCA คือการแปลงเชิงเส้นเชิงตั้งฉาก (Orthogonal Linear Transform) ซึ่งเป็นการแปลงชุดข้อมูลไปยังระบบพิกัดใหม่ โดยแปลงข้อมูลความแปรปรวนมากที่สุดไปยังพิกัดที่หนึ่ง (ซึ่งเรียกว่าแกนหลักที่ 1) และแปลงข้อมูลค่าความแปรปรวนลำดับสองไปเป็นแกนหลักที่สอง ตามลำดับ ดังนั้นข้อมูลที่มีความสำคัญมากที่สุดจะถูกแปลงให้ไปอยู่บนแกนหลักที่หนึ่ง และความสำคัญจะลดลงตามลำดับแกนหลักที่เพิ่มขึ้น ดังนั้น PCA จึงเป็นการแปลงเชิงเส้นที่ดีที่สุดแต่ใช้การคำนวณที่ซับซ้อน เนื่องจากเป็นการ แปลงที่ไม่มีเวกเตอร์ฐานที่กำหนดตายตัว ต้องขึ้นอยู่กับงานที่นำไปใช้

การคำนวณ PCA จากเมทริกซ์ความแปรปรวนมีขั้นตอนดังนี้

2.1 ทำการแปลงข้อมูลเป็นค่าปกติที่ต้องการ (Normalization)

2.2 หาค่าเฉลี่ยของแต่ละคุณลักษณะเด่น (Feature)

2.3 หาค่าเมทริกซ์ความแปรปรวนของและคลาส

2.4 หาค่าลักษณะเฉพาะและเรียงลำดับ โดยที่ $\lambda_1 > \lambda_2 > \ldots$

2.5 หาเมทริกซ์ลักษณะเฉพาะที่สอดคล้องกับค่าลักษณะเฉพาะนั้นๆ

2.6 หาค่า K ที่ทำให้

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{N} \lambda_i} > Threshold \tag{3.1}$$

โดยจุดเปลี่ยนเริ่มต้น (Threshold) เป็น 0.95

หาคุณลักษณะเด่นใหม่ โดย

$$\begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_K \end{bmatrix} = \begin{bmatrix} u_1^T \\ u_2^T \\ \ldots \\ u_K^T \end{bmatrix} (x - \bar{x}) = U^T (x - \bar{x}) \tag{3.2}$$

และ ui คือเวกเตอร์ลักษณะเฉพาะที่ i

3.4 การจำแนกประเภทข้อมูล (Data Classification)

ตัวแยกประเภทแบบเบส์ (Bayes Classification) เป็นการจำแนกโดยใช้หลักความน่าจะเป็นและสถิติแบบง่าย

3.4.1 ทฤษฎีความน่าจะเป็น

อัลกอริทึมนี้มีพื้นฐานมาจากทฤษฎีของเบย์ (Bayes Theorem) ซึ่งสามารถทำนายค่าคำตอบได้จากการเรียนรู้จากค่าของความน่าจะเป็นของเหตุการณ์ที่เกิดขึ้นที่ได้สอนให้กับมัน คำว่าการเรียนรู้จากค่าความน่าจะเป็นคือ การคำนวณค่าความน่าจะเป็น (Probabilities) สำหรับสมมติฐาน (Hypothesis) ที่เกิดขึ้น โดยที่มีการเรียนรู้แบบ Incremental (Online Learning) คือแต่ละตัวอย่างที่ทำการสอนจะส่งผลให้ค่าของความน่าจะเป็นของการเกิดคำตอบเพิ่มขึ้น หรือลดลงได้ตามค่าที่ได้สอนไป โดยงานวิจัยที่มีการใช้อัลกอริทึมนี้มีหลายอย่าง เช่น การทำ Data Mining , การจัดกลุ่มเอกสาร, การแก้ไขความผิดพลาด (Error-Correction) รวมถึงการสืบค้นข้อมูล

$$P(A) = \frac{N(A)}{N(S)} \tag{3.3}$$

จากการศึกษาในข้างต้นจะสามารถหาค่าความน่าจะเป็นของเหตุการณ์ได้โดยใช้หลักของ "ความถี่สัมพัทธ์" นั่นคือ ถ้าภายใน Sample Space มีสมาชิกเท่ากันและเหตุการณ์ A เกิดขึ้นเท่ากับจำนวนครั้ง จะได้ความน่าจะเป็นที่จะเกิดเหตุการณ์ A เท่ากับ P(A) โดยค่าของความน่าจะเป็นจะอยู่ระหว่าง 0 และ 1 ซึ่งจะแสดงถึงโอกาสของการเกิดเหตุการณ์นั้นๆ

Sample space (S) คือ เซตที่มีสมาชิกเป็นผลการทดลอง หรือประชากรทั้งหมด

Event คือ เหตุการณ์ที่เกิดหรือสับเซตของ Sample space

ความน่าจะเป็นแบบมีเงื่อนไขและเหตุการณ์อิสระ

ให้ E1 เป็นเหตุการณ์ที่เกิดขึ้นก่อน

E2 เป็นเหตุการณ์ที่เกิดขึ้นทีหลัง

ถ้าการเกิดของเหตุการณ์ E2 ขึ้นอยู่กับการเกิดของเหตุการณ์ E1 แสดงว่า เหตุการณ์ทั้งสองมีเงื่อนไขต่อกัน (Conditional Event) ความน่าจะเป็นในการเกิดเหตุการณ์ E2 เมื่อเกิดเหตุการณ์ E1 (E2 given E1) คือ $P(E2/E1)$ โดย

$$P(E2/E1) = \frac{P(E1 \cap E2)}{P(E1)} \tag{3.4}$$

ถ้า E1 และ E2 เป็นเหตุการณ์ที่เป็นอิสระต่อกัน (Independent Events) คือเหตุการณ์ซึ่งถ้าเหตุการณ์หนึ่งเกิดขึ้นแล้วจะไม่มีผลต่อความน่าจะเป็นของอีกเหตุการณ์หนึ่ง

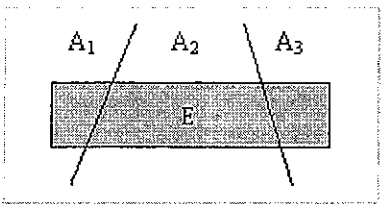
$$P(E1/E2) = P(E1) \tag{3.5}$$

$$P(E2/E1) = P(E2) \tag{3.6}$$

$$P(E1 \cap E2) = P(E1) \bullet P(E2) \tag{3.7}$$

3.4.2 ทฤษฎีพื้นฐานทั่วไปของเบย์ (Bayes theorem)

สมมุติให้ A1,A2,...,An แทนเหตุการณ์ที่ไม่เกิดร่วมกันซึ่งเป็นเหตุการณ์ที่เกิดขึ้นใน Sample space (S) ทั้งหมด E เป็นเหตุการณ์หนึ่งใน Sample space และเป็นส่วนหนึ่งของ $A_i (i = 1,2,3,...,k) by P(E) >= 0$



จะได้ว่า

$$P(E) = \sum_{i=1}^{k} P(A_i) \bullet P(E/Ai) \tag{3.8}$$

$$P(A_i/E) = \frac{P(E/A_i) \bullet P(A_i)}{\sum_{i=1}^{k} P(E/Ai) \bullet P(A_i)} \tag{3.9}$$

โดย

1.P(Ai) แทนความน่าจะเป็นของเหตุการณ์ที่เป็นไปได้ก่อนทราบข้อมูล = ความน่าจะเป็นโดยหลักเกณฑ์ (prior probability)

2. P(Ai/E) แทนความน่าจะเป็นของเหตุการณ์ที่เป็นไปได้หลังทราบข้อมูล = ความน่าจะเป็นโดยประสบการณ์ (posterior probability)

3.P(E/Ai) แทนความน่าจะเป็นของเหตุการณ์ E ภายใต้ข้อสมมติว่าส่วนย่อย A1 เกิดขึ้น

Bayes Probability theorem classifier

กรณี 2 Class โดยที่ w1 เป็น Class ที่ 1 และ w2 เป็น Class ที่ 2 ของ pattern ทั้งหมด โดยที่ต้องทราบค่า P(w1) ซึ่งเป็น A priori probabilities และทราบ conditional probability density function ที่แน่นอนดังนั้น

$$P(w_i/\vec{X}) = \frac{f(\vec{x}/w_i)P(w_i)}{f(\vec{x})} \tag{3.10}$$

$$f(\vec{x}) = \sum_{k=1}^{n} f(\vec{x}/w_k)P(w_k) \tag{3.11}$$

Bayes Classification

$$\begin{aligned} &if \ \frac{f(\vec{x}/w_1)}{f(\vec{x}/w_2)} > \frac{P(w_1)}{P(w_2)} \quad then \quad \vec{x} \in w_1 \\ &if \ \frac{f(\vec{x}/w_1)}{f(\vec{x}/w_2)} < \frac{P(w_1)}{P(w_2)} \quad then \quad \vec{x} \in w_2 \\ &if \ \frac{f(\vec{x}/w_1)}{f(\vec{x}/w_2)} = \frac{P(w_1)}{P(w_2)} \quad then \quad random class \ for \ \vec{x} \end{aligned} \tag{3.12}$$

ถ้ากำหนดให้

$$\frac{f(\vec{x}/w_1)}{f(\vec{x}/w_2)} \quad is \quad Likelihood \ function \Rightarrow \quad L(\vec{x})$$
$$\frac{P(w_2)}{P(w_1)} \qquad\qquad \Rightarrow \quad \eta_{MAP} \tag{3.13}$$

ดังนั้น

$$\begin{aligned} &if \quad L(\vec{x}) \quad > \quad \eta_{MAP} \quad then \quad \vec{x} \in w_1 \\ &if \quad L(\vec{x}) \quad < \quad \eta_{MAP} \quad then \quad \vec{x} \in w_2 \\ &if \quad L(\vec{x}) \quad = \quad \eta_{MAP} \quad then \quad randomclass \ for \ \vec{x} \end{aligned} \tag{3.14}$$

กรณีที่มีทั้งหมด M-Class

$$\begin{aligned} &if \quad P(w_i/\vec{x}) \ > \ P(w_j/\vec{x}) \qquad ;\forall i \neq j \quad then \quad \vec{x} \in w_i \\ &or \\ &if \quad f(\vec{x}/w_i)P(w_i) > f(\vec{x}/w_j)P(w_j) \quad ;\forall i \neq j \quad then \quad \vec{x} \in w_i \\ &or \\ &if \quad L_i(\vec{x})P(w_i) > \ L_j(\vec{x})P(w_j) \qquad ;\forall i \neq j \quad then \quad \vec{x} \in w_i \\ &L_i(\vec{x}) = \frac{f(\vec{x}/w_i)}{f(\vec{x}/w_m)} \quad and \quad L_m(\vec{x}) = 1 \end{aligned} \tag{3.15}$$

Normal Density Function

$$X \sim N(\mu, \delta^2) \tag{3.16}$$

โดย

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \tag{3.17}$$

ที่ซึ่ง

μ คือค่า Mean ซึ่งก็จะคือค่า Expect Value

δ^2 คือค่า Covariant

$$E(x) = \mu = \int_{-\infty}^{\infty} x f(x) dx \tag{3.18}$$

$$\delta^2 = E\left[(x-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx \tag{3.19}$$

1-Dimension feature vector

$$f(x/w_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-(x_i-\mu_i)^2}{2\sigma_i^2}\right) \tag{3.20}$$

Marginal density of i-th component of x

$$f(x_i) = \int\int\cdots f(\vec{x})dx_1dx_2\cdots dx_{i-1}dx_{i+1}\cdots dx_n \tag{3.21}$$

Covariant matrix

$$\begin{aligned}\overline{\overline{\Sigma}} &= E\left\{(\vec{x}-\vec{m})(\vec{x}-\vec{m})^T\right\} \\ &= e\left\{\begin{bmatrix} x_1-m_1 \\ x_2-m_2 \\ \vdots \\ x_n-m_n \end{bmatrix}\begin{bmatrix} x_1-m_1 & x_2-m_2 & \cdots & x_n-m_n \end{bmatrix}\right\} \\ &= E[\vec{x}\vec{x}^T - \vec{x}\vec{m}^T - \vec{m}\vec{x}^T + \vec{m}\vec{m}^T] \\ &= E\left[\vec{x}\vec{x}^T\right] - \vec{m}\vec{m}^T\end{aligned} \tag{3.22}$$

สมการหา mean และ covariance matrix

$$\begin{aligned}\vec{m} &= \frac{1}{N}\sum_{k=\vec{a}}^{N}\vec{x}_k \\ \overline{\overline{\Sigma}} &= \frac{1}{N-1}\sum_{k=1}^{N}(\vec{x}_k-\vec{m})(\vec{x}_x-\vec{m})^T\end{aligned} \tag{3.23}$$

Normal pdf สำหรับ feature vector ที่ มากกว่า 1-dimension

จะใช้สมการดังต่อไปนี้ซึ่งในการทดลองก็จะใช้สมการนี้ในการหา

$$N(\vec{x},\vec{m},\overline{\overline{\Sigma}}) = (2\pi)^{\frac{-n}{2}}\left|\overline{\overline{\Sigma}}\right|^{\frac{-1}{2}}\exp\left\{-\frac{1}{2}d_m^2(\vec{x},\vec{m},\overline{\overline{\Sigma}})\right\} \tag{3.24}$$

โดยที่

$$d_m^2(\vec{x},\vec{m},\overline{\overline{\Sigma}}) = \left((\vec{x}-\vec{m})^T\Sigma^{-1}(\vec{x}-\vec{m})\right) \tag{3.25}$$

3.4.3 การประมาณค่าพารามิเตอร์โดยใช้วิธีการของ Maximum Likelihood

คือวิธีการหาค่าตัวแปรที่ไม่ทราบค่าของ Probability density function โดยจะต้องมี training set เพื่อหา parameter ที่ทำให้ได้การจัดกลุ่มที่ดีที่สุด ซึ่งคุณสมบัติของ conditional probability density function (f(x/wi)) สำหรับแต่ละ Class i ที่เหมาะสมสำหรับวิธีการนี้คือ

f(x/wi) ต้องมีรูปแบบพารามิเตอร์ที่ 'nice' คือจะเป็น Form Parameter ที่ดี เช่น Gaussian
f(x/wi) ไม่มีผลกับ f(x/wj) เมื่อ j ≠ i และ $\vec{\theta}_i$ เป็นอิสระเชิงเส้นกับ $\vec{\theta}_j$
โดยที่ $\vec{\theta}_i$ เป็น parameter vector ของ class i

สำหรับ Supervised training
โดย Xi เป็น sample vector ของ class i
$X_i = \{\vec{x}_1, \vec{x}_2, \vec{x}_3, ..., \vec{x}_n\}$ ซึ่งมีความเป็น Randomly และ เป็นอิสระเชิงเส้นสำหรับ f(x/wi)
สำหรับ Class i ใดๆ
Output : $\vec{\theta}$ ของแต่ละ Class ที่ให้ค่า conditional probability สูงที่สุด
Input : $X_i = \{\vec{x}_1, \vec{x}_2, \vec{x}_3, ..., \vec{x}_n\}$ของแต่ละ Class i
ถ้า f(x/wi) เป็น Gaussian distribution function

$$f(x/w_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-(x_i - \mu_i)^2}{2\sigma_i^2}\right) \tag{3.26}$$

$$\vec{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \mu_i \\ \sigma_i^2 \end{bmatrix} \tag{3.27}$$

ดังนั้นแต่ละ Class

นั่นคือถ้า $\vec{X} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ แล้ว จะมีค่า Parameter ที่เกี่ยวข้องคือ $\vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$

และ

$$\vec{\vec{\varepsilon}} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix} \tag{3.28}$$

ทำให้มี Unknown parameter ทั้งหมดคือ

$$\vec{\theta} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \sigma_{11}^2 \\ \sigma_{12}^2 \\ \sigma_{22}^2 \end{bmatrix} \tag{3.29}$$

เนื่องจาก $\sigma_{12}^2 = \sigma_{21}^2$

$$\begin{aligned} X_i &= \{\vec{x}_1, \vec{x}_2, \vec{x}_3, ..., \vec{x}_n\} \quad ; \vec{x}_R \in \vec{X} \\ \therefore f(\vec{X}/\vec{\theta}) &= f(\vec{x}_1, \vec{x}_2, \vec{x}_3, ..., \vec{x}_n/\vec{\theta}) \\ &= f(\vec{x}_1/\vec{\theta}) f(\vec{x}_2/\vec{\theta}) ..... f(\vec{x}_n/\vec{\theta}) \\ &= \prod_{k=1}^{n} f(\vec{x}_k/\vec{\theta}) \end{aligned} \tag{3.30}$$

หาค่า $\vec{\theta}$ ที่ทำให้ Conditional probability สูงสุด นั่นคือ

$$\hat{\vec{\theta}} = \arg\max_{\theta} \prod_{k=1}^{n} f(\vec{x}_k/\vec{\theta}) \tag{3.31}$$

ดังนั้น

$$\frac{\partial(\prod_{k=1}^{n} f(\vec{x}_k/\vec{\theta}))}{\partial\theta} = \vec{\theta} \tag{3.32}$$

จากคุณสมบัติของ Logarithmic function ที่เป็น monotonic increasing ดังนั้นเมื่อนิยาม log likelihood function เป็น

$$\begin{aligned} L(\vec{\theta}) &= L(\vec{x}/\vec{\theta}) = \ln(f(\vec{x}/\vec{\theta})) \\ \therefore \ln(f(X/\vec{\theta})) &= L(X/\vec{\theta}) = \sum_{i=1}^{n} \ln(f(\vec{x}_i/\vec{\theta})) \end{aligned} \tag{3.33}$$

และ เมื่อ $\nabla_{\vec{\theta}} = \begin{bmatrix} \frac{\partial}{\partial\theta_1} \\ \frac{\partial}{\partial\theta_2} \\ \vdots \\ \frac{\partial}{\partial\theta_n} \end{bmatrix}$ ทำให้สมการ เป็นดังนี้

$$\frac{\partial f(X/\vec{\theta})}{\partial\vec{\theta}} = \nabla_{\vec{\theta}} L(X/\vec{\theta})$$

$$0 = \sum_{i=1}^{n} \nabla_{\vec{\theta}} L(\vec{x}_i/\vec{\theta})$$

จากนั้นจึงแก้สมการเพื่อหาคำตอบต่อไป

ตัวแยกประเภทแบบเบส์ (Bayes Classification) เป็นการจำแนกโดยใช้หลักความน่าจะเป็นและสถิติแบบง่าย โดยใช้พื้นฐานของทฤษฎีของเบส์ดังนี้

ถ้ามีประเภทข้อมูล (Class) $\omega 1, \omega 2, ..., \omega 3$ ทั้งหมด c กลุ่ม ที่เป็นเหตุการณ์หรือกลุ่มข้อมูลที่เป็นอิสระต่อกันอย่างสิ้นเชิง นั่นคือ $\omega_i \cap \omega_j = \phi; i \neq j$ และมีเวกเตอร์คุณคุณลักษณะเด่น $\vec{x}$ ที่มีความน่าจะเป็นที่จะเกิดมากกว่า 0 จะหา A Posteriori Probability, $P(\omega_j / \overrightarrow{X})$ ได้ดังนี้

$$P(w_i / \vec{X}) = \frac{f(\vec{x} / w_i)P(w_i)}{f(\vec{x})} \qquad ; f(\vec{x}) = \sum_{k=1}^{n} f(\vec{x} / w_k)P(w_k) \tag{3.34}$$

3.4.4 ฟังก์ชันความหนาแน่นของความน่าจะเป็นแบบปกติ (Normal Density Function)

การแจกแจงของข้อมูลในธรรมชาติโดยทั่วไป จะเป็นการแจกแจงแบบปกติ ซึ่งจะขึ้นอยู่กับค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของข้อมูล สามารถเขียนรูปแบบดังนี้

$$X \sim N(\mu, \delta^2) \tag{3.35}$$

โดยที่ μ คือค่าเฉลี่ยหาได้จาก

$$E(x) = \mu = \int_{-\infty}^{\infty} x f(x) dx \tag{3.36}$$

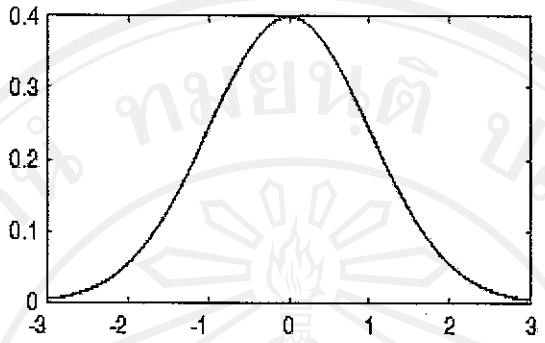
δ^2 คือค่าความแปรปรวน ซึ่งหาได้จาก

$$\delta^2 = E\left[(x-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx \tag{3.37}$$

กรณีที่เวกเตอร์คุณลักษณะเด่นแบบหนึ่งมิติ ฟังก์ชันความหนาแน่นแบบปกติสามารถเขียนในรูปสมการดังนี้

$$f(x / w_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-(x_i - \mu_i)^2}{2\sigma_i^2}\right) \tag{3.38}$$

ตัวอย่างกราฟดังรูปที่ 3.1 มีค่าพารามิเตอร์เป็น $\mu = 0$ and $\delta = 1$



รูปที่ 3.1 กราฟการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็นศูนย์และส่วนเบี่ยงเบนมาตรฐานเป็นหนึ่ง

กรณีที่เวกเตอร์คุณลักษณะเด่นมีหลายมิติ เมทริกซ์ค่าความแปรปรวนหาได้จาก

$$\begin{aligned}\overline{\overline{\Sigma}} &= E\left\{(\vec{x}-\vec{m})(\vec{x}-\vec{m})^T\right\} \\ &= e\left\{\begin{bmatrix} x_1-m_1 \\ x_2-m_2 \\ \vdots \\ x_n-m_n \end{bmatrix}\begin{bmatrix} x_1-m_1 & x_2-m_2 & \cdots & x_n-m_n \end{bmatrix}\right\} \\ &= E[\vec{x}\vec{x}^T - \vec{x}\vec{m}^T - \vec{m}\vec{x}^T + \vec{m}\vec{m}^T] \\ &= E\left[\vec{x}\vec{x}^T\right] - \vec{m}\vec{m}^T\end{aligned} \tag{3.39}$$

โดยที่ $\vec{m}$ คือเวกเตอร์ค่าเฉลี่ย สามารถคำนวณได้จาก

$$\vec{m} = \frac{1}{N}\sum_{k=1}^{N}\vec{x}_k \tag{3.40}$$

ดังนั้นค่าความแปรปรวนหาได้จาก

$$\overline{\overline{\Sigma}} = \frac{1}{N-1}\sum_{k=1}^{N}(\vec{x}_k - \vec{m})(\vec{x}_x - \vec{m})^T \tag{3.41}$$

โดยฟังก์ชันความหนาแน่นแบบปกติหาได้จาก

$$N(\vec{x}, \vec{m}, \overline{\overline{\Sigma}}) = (2\pi)^{\frac{-n}{2}}\left|\overline{\overline{\Sigma}}\right|^{\frac{-1}{2}}\exp\left\{-\frac{1}{2}d_m^2(\vec{x}, \vec{m}, \overline{\overline{\Sigma}})\right\} \tag{3.42}$$

โดยที่ $d_m^2(\vec{x},\vec{m},\overline{\overline{\Sigma}})$ จะคำนวณตามรูปแบบของเมทริกซ์ความแปรปรวนดังนี้ ถ้าเมทริกซ์ความแปรปรวนที่เท่ากันสำหรับทั้งสองคลาสที่มีค่าเฉพาะแนวทแยง นั่นคือ $\overline{\overline{\Sigma}} = \delta^2 \overline{I}$ แล้วทำให้ $d_m^2(\vec{x},\vec{m},\overline{\overline{\Sigma}})$ เป็นระยะแบบยูคลิด ดังนี้

$$d_m^2(\vec{x},\vec{m},\overline{\overline{\Sigma}}) = d_d^2(x,m,\varepsilon) = \left\|\vec{x}-\vec{m}\right\|^2 \tag{3.43}$$

ถ้าเมทริกซ์ความแปรปรวนมีค่าเท่ากันทั้งสองคลาสและไม่ใช่เมทริกซ์แนวทแยงจะได้ค่า $d_m^2(\vec{x},\vec{m},\overline{\overline{\Sigma}})$ เป็น Mahalanobis Distance

$$d_m^2(\vec{x},\vec{m},\overline{\overline{\Sigma}}) = \left((\vec{x}-\vec{m})^T \Sigma^{-1}(\vec{x}-\vec{m})\right) \tag{3.44}$$

3.4.5 การประมาณค่าพารามิเตอร์ โดยใช้ Maximum Likelihood

ในกรณีที่รู้ว่าฟังก์ชันความหนาแน่นของความน่าจะเป็นแบบมีเงื่อนไข f(x/wi) เป็นฟังก์ชันใดที่แน่นอนแล้ว การประมาณค่าพารามิเตอร์โดยใช้ Maximum Likelihood เป็นวิธีการหนึ่งในการหาค่าตัวแปรที่ไม่ทราบค่าของฟังก์ชันความหนาแน่นของความน่าจะเป็น โดยจะต้องมีชุดข้อมูลทดสอบเพื่อหาพารามิเตอร์ที่ทำให้ได้การจัดกลุ่มที่ดีที่สุด ซึ่งคุณสมบัติของฟังก์ชันความหนาแน่นของความน่าจะเป็นแบบมีเงื่อนไข f(x/wi) สำหรับแต่ละคลาส i ที่เหมาะสมสำหรับวิธีการนี้คือ

f(x/wi) ต้องมีรูปแบบพารามิเตอร์ที่เหมาะสม

f(x/wi) ไม่มีผลกับ f(x/wj) เมื่อ j ≠ i และ $\vec{\theta}_i$ เป็นอิสระเชิงเส้นกับ $\vec{\theta}_j$ โดยที่ $\vec{\theta}_i$ เป็นเวกเตอร์พารามิเตอร์ของคลาส i

สำหรับ Supervised training โดย Xi เป็นเวกเตอร์กลุ่มตัวอย่างของคลาส i $X_i = \{\vec{x}_1, \vec{x}_2, \vec{x}_3, ..., \vec{x}_n\}$ $X_i = \{\vec{x}_1,\}$ ซึ่งมีคุณสมบัติการสุ่มและเป็นอิสระเชิงเส้นสำหรับ f(x/wi)

สำหรับคลาส i ใดๆ $X_i = \{\vec{x}_1, \vec{x}_2, \vec{x}_3, ..., \vec{x}_n\}$

เอาท์พุต $\vec{\theta}$ ของแต่ละคลาสที่ให้ค่าความน่าจะเป็นแบบมีเงื่อนไขสูงที่สุด

อินพุต $X_i = \{\vec{x}_1, \vec{x}_2, \vec{x}_3, ..., \vec{x}_n\}$ ของแต่ละ Class i

ถ้า f(X/wi) เป็นฟังก์ชันการแจกแจงปกติ โดย

$$f(x/w_i)=\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp(\frac{-(x_i-\mu_i)^2}{2\sigma_i^2}) \tag{3.45}$$

ดังนั้น แต่ละคลาสจะมีเวกเตอร์พารามิเตอร์คือ $\vec{\theta}=\begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix}=\begin{bmatrix}\mu_i\\ \sigma_i^2\end{bmatrix}$

นั่นคือ $\vec{X}=\begin{bmatrix}x_1\\ x_2\end{bmatrix}$ ถ้า $\vec{\mu}=\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}$ แล้วจะมีพารามิเตอร์ที่เกี่ยวข้องคือ

$\vec{\vec{\varepsilon}}=\begin{bmatrix}\sigma_{11}^2 & \sigma_{12}^2\\ \sigma_{21}^2 & \sigma_{22}^2\end{bmatrix}$

ทำให้มีพารามิเตอร์ที่ไม่ทราบค่าทั้งหมดคือ

$\vec{\theta}=\begin{bmatrix}\mu_1\\ \mu_2\\ \sigma_{11}^2\\ \sigma_{12}^2\\ \sigma_{22}^2\end{bmatrix}$ เนื่องจาก $\sigma_{12}^2=\sigma_{21}^2$

วิธีการ

กำหนดให้แต่ละสมาชิกของเวกเตอร์กลุ่มตัวอย่างเป็นอิสระเชิงเส้นต่อกัน

$$\begin{aligned}X_i&=\{\vec{x}_1,\vec{x}_2,\vec{x}_3,...,\vec{x}_n\}\quad ;\vec{x}_R\in\vec{X}\\ \therefore f(\vec{X}/\vec{\theta})&=f(\vec{x}_1,\vec{x}_2,\vec{x}_3,...,\vec{x}_n/\vec{\theta})\\ &=f(\vec{x}_1/\vec{\theta})f(\vec{x}_2/\vec{\theta})....f(\vec{x}_n/\vec{\theta})\\ &=\prod_{k=1}^{n}f(\vec{x}_k/\vec{\theta})\end{aligned} \tag{3.46}$$

หาค่า $\vec{\theta}$ ที่ทำให้ค่าความน่าจะเป็นแบบมีเงื่อนไขสูงสุด นั่นคือ

$$\hat{\vec{\theta}}=\arg\max_{\theta}\prod_{k=1}^{n}f(\vec{x}_k/\vec{\theta}) \tag{3.47}$$

ดังนั้น

$$\frac{\partial(\prod_{k=1}^{n} f(\vec{x}_k / \vec{\theta}))}{\partial\theta} = \vec{\theta} \tag{3.48}$$

จากคุณสมบัติของฟังก์ชันลอการิทึมที่เป็นฟังก์ชันเพิ่ม ดังนั้นเมื่อนิยาม log-likelihood function เป็น

$$\begin{aligned} L(\vec{\theta}) &= L(\vec{x}/\vec{\theta}) = \ln(f(\vec{x}/\vec{\theta})) \\ \therefore\ \ln(f(X/\vec{\theta})) &= L(X/\vec{\theta}) = \sum_{i=1}^{n} \ln(f(\vec{x}_i/\vec{\theta})) \end{aligned} \tag{3.49}$$

และเมื่อ $\nabla_{\vec{\theta}} = \begin{bmatrix} \frac{\partial}{\partial\theta_1} \\ \frac{\partial}{\partial\theta_2} \\ \vdots \\ \frac{\partial}{\partial\theta_n} \end{bmatrix}$ ทำให้สมการที่สามเป็นดังนี้

$$\begin{aligned} \frac{\partial f(X/\vec{\theta})}{\partial\vec{\theta}} &= \nabla_{\vec{\theta}} L(X/\vec{\theta}) \\ 0 &= \sum_{i=1}^{n} \nabla_{\vec{\theta}} L(\vec{x}_i/\vec{\theta}) \end{aligned} \tag{3.50}$$

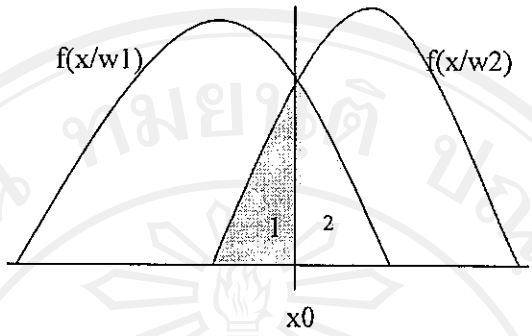
จากนั้นจึงแก้สมการหาคำตอบของเวกเตอร์พารามิเตอร์

3.4.6 การประเมินค่าความผิดพลาด (Error assessment)

Bayes Classifier จะเป็นวิธีที่หาจุด Boundary ที่ทำให้เกิด Error ในการจัดกลุ่มน้อยที่สุด โดยสามารถหาความน่าจะเป็นที่จะเกิด Error ในการจัดกลุ่มได้ ดังนี้

P(error) = P(x is assigned to the wrong class)

โดยพื้นที่แสดงการจำแนกคลาสผิดนั้นแสดงดังรูปที่ 3.2



รูปที่ 3.2 พื้นที่การเกิดความผิดพลาดในการจำแนกคลาส

โดยที่ x0 คือ Decision boundary

f(x/w1) เป็น pdf ของ Class 1, f(x/w2) เป็น pdf ของ Class 2

พื้นที่ในส่วนที่ 1 เป็น error ที่เกิดจากการจำแนกให้อยู่ Class 1แต่เป็นข้อมูลที่อยู่ Class 2

พื้นที่ในส่วนที่ 2 เป็น error ที่เกิดจากการจำแนกให้อยู่ Class 2แต่เป็นข้อมูลที่อยู่ Class 1

P(error) = Pe

= P(assigned to w1/x=class w2)+P(assigned to class w2 /x=class w1)

= P (x is in region 2, x is w1) + P(x is in region 1, x is w2)

= พื้นที่ในส่วนที่ 1+พื้นที่ในส่วนที่ 2

$$P(\vec{x} \in R2, w1)P(w1) + P(\vec{x} \in R1, w2)P(w2) \tag{3.51}$$

$$\therefore P_e = P(w1)\int_{R2} f(x/w1)dx + P(w2)\int_{R1} f(x/w2)dx \tag{3.52}$$

จากรูปที่ 3.2 จะได้ว่า

$$\therefore P_e = P(w1)\int_{x0}^{\infty} f(x/w1)dx + P(w2)\int_{0}^{x0} f(x/w2)dx \tag{3.53}$$

และจาก Bayes Theorem

$$P(w_i/\vec{X}) = \frac{f(\vec{x}/w_i)P(w_i)}{f(\vec{x})} \quad ; f(\vec{x}) = \sum_{k=1}^{n} f(\vec{x}/w_k)P(w_k) \tag{3.54}$$

$$P_e = \int_{R2} P(w1/x)f(x)dx + \int_{R1} P(w2/x)f(x)dx \tag{3.55}$$

ซึ่งจะ Optimized เมื่อ Pe มีค่าน้อยที่สุดตามหลักการของการจำแนกประเภทแบบเบส์ นั่นคือ

เลือก R1 เมื่อ $P(w1/x) > P(w2/x)$

เลือก R2 เมื่อ $P(w1/x) < P(w2/x)$

ดังนั้นการจำแนกประเภทแบบเบส์เป็นการจัดกลุ่มที่ทำให้ได้ค่าความผิดพลาดต่ำสุด