บทที่ 6

ซัพพอร์ตเวกเตอร์แมชชีนและนาอีฟเบย์

(Support Vector Machine and Naïve Bayes)

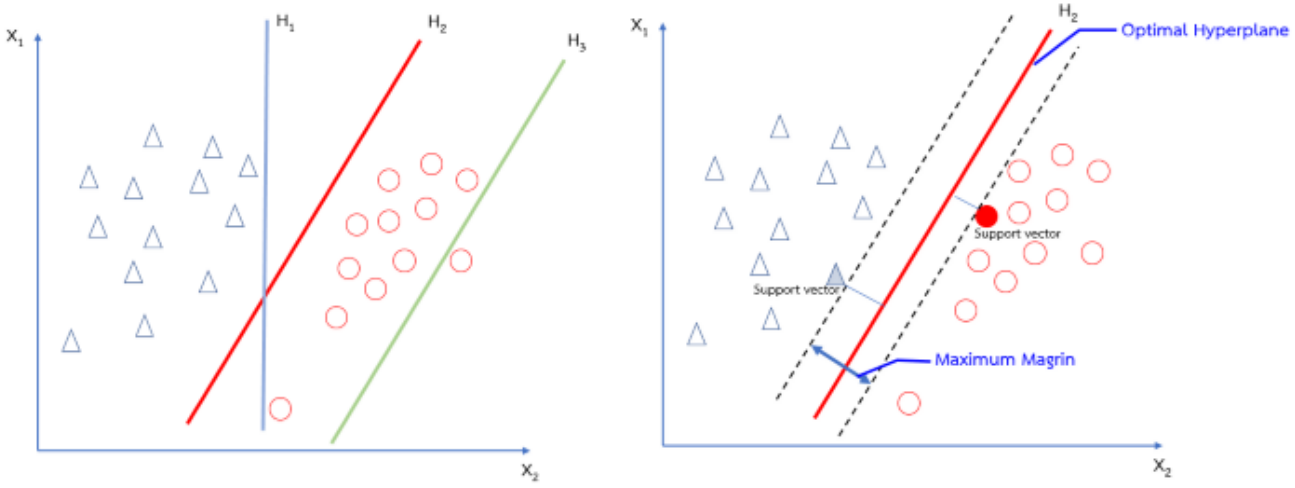
วัตถุประสงค์การเรียนรู้

1) ทราบถึงหลักการทำงานของวิธี Support Vector Machine (SVM)
2) เข้าใจถึงการปรับแต่งพารามิเตอร์และเปรียบเทียบผลลัพธ์ที่ได้จากการปรับแต่ง
3) เข้าใจหลักการทำงานของ Naïve Bayes

บทที่ 6
ซัพพอร์ตเวกเตอร์แมชชีนและนาอีฟเบย์
(Support Vector Machine and Naïve Bayes)

6.1 ซัพพอร์ตเวกเตอร์แมชชีน

ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) (Cortes et al., 1995) หรือ SVM เป็นอีกหนึ่งอัลกอริธึมในกลุ่มวิธีการเรียนรู้ของเครื่องแบบมีผู้สอน ที่สามารถนำมาช่วยแก้ปัญหาการจำแนกข้อมูลได้ โดยเฉพาะกับปัญหาที่มีขนาดของข้อมูลไม่ใหญ่มาก แต่คุณลักษณะ (features) ของข้อมูลมีเป็นจำนวนมาก SVM จะถือได้ว่าเป็นอัลกอริธึมที่ทำงานได้ค่อนข้างจะมีประสิทธิภาพมากๆ อัลกอริธึมหนึ่ง หลักการทำงานของ SVM จะอาศัยใช้การสร้างเส้นแบ่ง หรือไฮเปอร์เพลน (Hyperplane) ในการแบ่งแยกคลาสของข้อมูลออกจากกัน จากนั้นจะทำการหาว่าไฮเปอร์เพลนใดเป็นเส้นที่ใช้แยกคลาสของข้อมูลได้ดีที่สุด (Optimal hyperplane) ลักษณะดังรูปที่ 6.1



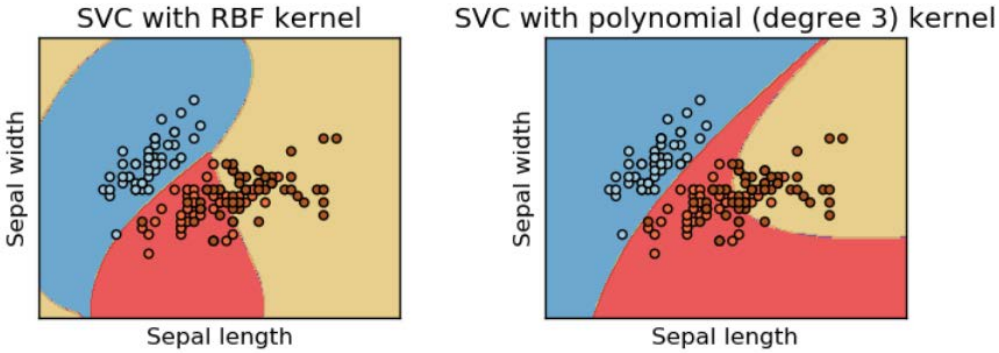
(ก) (ข)

รูปที่ 6.1 ตัวอย่าง SVM ใน 2 มิติ

ในรูปที่ 6.1 สมมติเราต้องการจำแนกข้อมูลออกเป็น 2 คลาส โดยใช้ไฮเปอร์เพลนที่เป็นเส้นตรง จะเห็นว่ามีเส้นตรงจำนวนมากที่สามารถแบ่งแยกข้อมูลออกจากกันได้ แต่เส้นตรงใดจะถูกพิจารณาให้เป็นเส้นที่ดี

ที่สุด (Optimal Line) นั้น จะพิจารณาจากไฮเปอร์เพลนที่มีผลรวมของระยะห่างระหว่างเส้นไฮเปอร์เพลนกับเส้นตรงที่ลากผ่านข้อมูลที่ใกล้ที่สุดและขนานกับเส้นไฮเปอร์เพลนของข้อมูลแต่ละกลุ่มที่มากสุด (Maximum Margin) ในรูปที่ 6.1 จะเห็นว่าไฮเปอร์เพลน H_1 และ H_2 สามารถใช้ในการแบ่งแยกคลาสของข้อมูลได้เหมือนกัน แต่ไฮเปอร์เพลน H_2 จะถูกให้เลือกให้เป็นไฮเปอร์เพลนที่ดีที่สุดของปัญหานี้ เนื่องจากมีระยะในการแบ่งจากไฮเปอร์เพลนไปถึงเส้นที่ลากผ่านข้อมูลที่ใกล้ที่สุดนั้นกว้างกว่า H_1 และจะเรียกข้อมูลที่อยู่บน margin นั้นว่า Support Vector

ตัวอย่างข้างต้นจะเป็นการสร้างไฮเปอร์เพลนที่เป็นเส้นตรง ซึ่งใช้ได้ดีกับปัญหาที่ลักษณะข้อมูลเป็นเชิงเส้น อย่างไรก็ตามเราสามารถใช้ SVM ในปัญหาที่ลักษณะข้อมูลไม่เป็นเชิงเส้นได้ ดังรูปที่ 6.2



รูปที่ 6.2 ตัวอย่างไฮเปอร์เพลนที่ใช้ในปัญหาที่ข้อมูลไม่เป็นเชิงเส้น

เราสามารถสร้างไฮเปอร์เพลนในลักษณะที่ไม่เป็นเส้นตรงได้ โดยอาศัยการปรับค่าสมการด้วยวิธีการที่เรียกว่า วิธีการเคอร์เนล (Kernel method) ในการหา Pattern and Relation เพื่อให้ได้ผลลัพธ์ที่มีความถูกต้องแม่นยำสูง Kernel Function มีอยู่เป็นจำนวนมากที่รู้จักกันดี เช่น Polynomial, RBF หรือ Sigmoid เป็นต้น

ดังนั้น SVM จะมีประสิทธิภาพสูงหรือต่ำนั้นขึ้นอยู่กับ Kernel ที่เหมาะสมกับลักษณะของชุดข้อมูลและพารามเตอร์ต่าง ๆ ดังนั้น นิสิตจำเป็นต้องทราบถึงพารามิเตอร์ที่สำคัญของ SVM ว่ามีอะไรบ้าง เช่น Penalty Parameter (C), Gamma (γ) หรอ Refining Gamma เป็นต้น

6.2 นาอีฟเบย์ (Naïve Bayes)

นาอีฟเบย์ เป็นวิธีที่ให้ผลการจำแนกได้ดีไม่แตกต่างวิธีการอื่นโดยมีอัลกอริธีมการทำงานที่ไม่ซับซ้อน การเรียนรู้ของนาอีฟเบย์จะเป็นการเรียนรู้โดยใช้หลักการของความน่าจะเป็น (Probability) ซึ่งมีพื้นฐานมาจากทฤษฎีของเบย์ (Bayes theorem) หรือทฤษฎีว่าด้วยโอกาสที่จะเกิดของเหตุการณ์ต่างๆ ซึ่งจะใช้การคำนวณความน่าจะเป็นแบบมีเงื่อนไขที่เรียกว่า Conditional Probability (Dietrich et al., 2015) แสดงได้ดังสมการ (6.1)

$$P(h \mid D) = \frac{P(D \mid h) \times P(h)}{P(D)} \tag{6.1}$$

โดยที่ $P(h)$ คือ ค่าความน่าจะเป็นของสมมติฐานที่คลาสเป็น h

$P(D)$ คือ ค่าความน่าจะเป็นของสมมติฐานที่ข้อมูลเป็น D

$P(h \mid D)$ คือ ค่าความน่าจะเป็นของสมมติฐานที่ข้อมูลเป็น D จะมีคลาสเป็น h

$P(D \mid h)$ คือ ค่าความน่าจะเป็นของสมมติฐานที่คลาสเป็น h จะมีข้อมูลเป็น D

ในปัญหาที่มีตัวแปรต้นหรือข้อมูลที่ต้องพิจารณามากกว่า 1 ค่าความน่าจะเป็นของเหตุการณ์ สามารถคำนวณได้จากผลคูณของความน่าจะเป็นของแต่ละข้อมูลที่มีคลาส h ดังสมการ (6.2)

$$P(h \mid D) = P(D_1 \mid h) \times P(D_2 \mid h) \times \ldots \times P(D_n \mid h) \times P(h) \tag{6.2}$$

ตัวอย่างเช่น สมมติเราต้องการสร้างแบบจำลองเพื่อจำแนกการเล่นกีฬา คือ เล่น (No) กับ ไม่เล่น (Yes โดยเรียนรู้จากชุดข้อมูลแสดงตารางที่ 1

ตารางที่ 1 ชุดข้อมูลตัวอย่างใช้เพื่อการเรียนรู้จำแนกจำแนกการเล่นกีฬา

No.	Outlook	Temperature	Humidity	Windy	Play
1	sunny	hot	high	false	No
2	sunny	hot	high	true	No
3	overcast	hot	high	false	Yes
4	rainy	mild	high	false	Yes
5	rainy	cool	normal	false	Yes
6	rainy	cool	normal	true	No

7	overcast	cool	normal	true	Yes
8	sunny	mild	high	false	No
9	sunny	cool	normal	false	Yes
10	rainy	mild	normal	false	Yes
11	sunny	mild	normal	true	Yes
12	overcast	mild	high	true	Yes
13	overcast	hot	normal	false	Yes
14	rainy	mild	high	true	No

ชุดข้อมูลในตารางที่ 1 กำหนดค่าตัวแปรในสมการได้ ดังนี้

h คือ คลาส ในที่นี้คือคอลัมน์ Play มี 2 ค่า ได้แก่ เล่น (Yes) และ ไม่เล่น (No)

D คือ ข้อมูลที่ใช้ในการพิจารณา ในที่นี้มีทั้งหมด 4 ข้อมูล ได้แก่ Outlook, Temperature, Humidity และ Windy

สมมติในกรณีที่ Outlook, Temperature, Humidity และ Windy มีค่าเป็น sunny, mild, normal และ false ตามลำดับ เราต้องการทราบว่าคำตอบที่ได้จะเป็น Yes หรือ No ด้วยวิธีการของนาอีฟเบย์ สามารถเริ่มต้นคำนวณ ได้ดังนี้

1. คำนวณค่า P(h) หรือค่าความน่าจะเป็นของแต่ละคลาส

$$P(play = Yes) = 9/14 = 0.64$$
$$P(play = No) = 5/14 = 0.36$$

2. คำนวณค่า P(D|h) หรือความน่าจะเป็นของแต่ละข้อมูล (D) ที่มีคลาสเป็น (h)

$\boldsymbol{D = Outlook}$

$P(sunny\|Yes)$	$= 2/9$	$P(sunny\|No)$	$= 3/5$
$P(overcast\|Yes)$	$= 4/9$	$P(overcast\|No)$	$= 0/5$
$P(rain\|Yes)$	$= 3/9$	$P(rain\|No)$	$= 2/5$

$\boldsymbol{D = Temperature}$

$P(hot\|Yes)$	$= 2/9$	$P(hot\|No)$	$= 2/5$
$P(mild\|Yes)$	$= 4/9$	$P(mild\|No)$	$= 2/5$
$P(cool\|Yes)$	$= 3/9$	$P(cool\|No)$	$= 1/5$

$\boldsymbol{D = Humidity}$

$P(high\|Yes)$	$= 3/9$	$P(high\|No)$	$= 4/5$
$P(normal\|Yes)$	$= 6/9$	$P(normal\|No)$	$= 1/5$

$\boldsymbol{D = Windy}$

$P(true\|Yes)$	$= 3/9$	$P(true\|No)$	$= 3/5$
$P(false\|Yes)$	$= 6/9$	$P(false\|No)$	$= 2/5$

3. นำค่าความน่าจะเป็นในขั้นตอนที่ (1) และ (2) แทนในสมการ เมื่อ Outlook, Temperature, Humidity และ Windy มีค่าเป็น sunny, mild, normal และ false ตามลำดับ

ค่าความน่าจะเป็นที่ play= Yes

$$\boldsymbol{P(\text{play} = \text{Yes}|D) = P(\text{sunny}|\text{Yes}) \times P(\text{mild}|\text{Yes}) \times P(\text{normal}|\text{Yes}) \times P(\text{false}|\text{Yes}) \times P(\text{Yes})}$$

$$\boldsymbol{= 2/9 \times 4/9 \times 6/9 \times 6/9 \times 9/14 = 0.028}$$

ค่าความน่าจะเป็นที่ play= No

$$\boldsymbol{P(\text{play} = \text{No}|D) = P(\text{sunny}|No) \times P(\text{mild}|No) \times P(\text{normal}|No) \times P(\text{false}|\text{No}) \times P(NO)}$$

$$\boldsymbol{= 3/5 \times 2/5 \times 1/5 \times 2/5 \times 5/14 = 0.0069}$$

ผลลัพธ์จากการคำนวณค่าความน่าว่าจะเล่นหรือไม่เล่นกีฬา จะเลือกจาก P(h|D) ที่สูงที่สุด ในที่นี้ผลของการจำแนกพบว่า ถ้าข้อมูล Outlook, Temperature, Humidity และ Windy มีค่าเป็น sunny, mild, normal และ false นั้นค่าคำตอบคือ เล่น (Yes)

6.3 อ้างอิง

Cortes, Corinna; Vapnik, Vladimir N. (1995). *Support-vector networks*. Machine Learning. 20 (3): 273–297. doi:10.1007/BF00994018. S2CID 206787478.

Dietrich, D., Heller, B., & Yang, B. (2015). *Data Science & Big Data Analytics Discovering, Analyzing, Visualizing and Presenting Data (pp. 420)*. John Wiley & Sons,Inc.