บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

แบบจำลองจักรกลเรียนรู้ (Machine Learning models) นั้นถูกใช้อย่างกว้างขวางในปัจจุบัน อย่างไรก็ตาม แบบจำลองใดๆ นั้นอาจมีความผิดพลาดต่อการทำการโจมตีประสงค์ร้าย (Adversarial attacks) เพื่อจงใจให้ผลลัพธ์ที่แบบจำลองนั้นคาดเดามีความผิดพลาดจากผลลัพธ์ที่ควรจะเป็น

ในการเรียนรู้เชิงตัวแปรเสริม (parameter-based learning) นั้น ตัวแปรเสริม (parameters) ค่าน้ำหนัก (weights) บนแบบจำลองการเรียนรู้เชิงลึก (deep Learning models) เป็นตัวกำหนดความฉลาดของแบบจำลอง อาจมีตัวแปรเสริมบางชุดที่ทำให้แบบจำลองมีช่องโหว่ต่อการโจมตีประสงค์ร้าย การโจมตีนั้นอาจเกิดจากการเพิ่มสัญญาณรบกวนซึ่งผ่านการคำนวน (calculated artefacts) เข้าสู่ข้อมูลรับเข้า (inputs) ซึ่งทำให้ความผิดพลาดของแบบจำลองในการพยากรณ์คำตอบนั้นเปลี่ยนไปอย่างชัดเจน

โครงงานวิศวกรรมคอมพิวเตอร์นี้มุ่งหวังจะนำตัวแปรเสริมบนแบบจำลองมาสร้างภาพแสดง (visualise) ถึงจุดโหว่ในการพยากรณ์ใดๆ ของแบบจำลอง เพื่อลดความเสียหายอันอาจเกิดขึ้นได้จากการโจมตีแบบจำลองขณะถูกใช้งานจริง

1.2 วัตถุประสงค์ของการศึกษา

โครงงานนี้มีวัตถุประสงค์และเป้าหมายดังนี้

1. สร้างแบบจำลองเชิงลึก (Deep Learning models) ซึ่งสามารถถูกโจมตีประสงค์ร้าย (Adversarial attacks) ได้
2. นำแบบจำลองในข้อ (1) มาสร้างเป็นรูปภาพแสดง (visualisation) เพื่อหาจุดโหว่ต่อการโจมตี รวมถึงคาดเดาแนวโน้มการโจมตีที่เป็นไปได้
3. ใช้ความรู้ในข้อ (2) สร้างแบบจำลองที่ทนทาน (prone) ต่อการโจมตีมากขึ้น

1.3 ขอบเขตของการทำโครงงาน

โครงงานนี้มีขอบเขตการดำเนินงานดังนี้

1. สร้างแบบจำลองเชิงลึก (Deep Learning models) ซึ่งสามารถถูกโจมตีประสงค์ร้าย (Adversarial attacks) ได้
2. นำแบบจำลองในข้อ (1) มาสร้างเป็นรูปภาพแสดง (visualisation) เพื่อหาจุดโหว่ต่อการโจมตี รวมถึงคาดเดาแนวโน้มการโจมตีที่เป็นไปได้
3. ใช้ความรู้ในข้อ (2) สร้างแบบจำลองที่ทนทาน (prone) ต่อการโจมตีมากขึ้น

1.4 ระยะเวลาและแผนดำเนินงาน

ในช่วงแรกของการทำโครงงาน แผนการดำเนินงานนั้นจะใช้ในรูปแบบของรบวนทวนซ้ำ (iteration) ตามกรรมวิธีการดำเนินงานแบบเอไจล์ (agile) ซึ่งประกอบไปด้วยขั้นตอนการวนทวนดังนี้...

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. เข้าใจถึงพื้นฐาน หลักการทำงาน และระบบจักรกลเรียนรู้แบบต่างๆ
2. เข้าใจถึงจุดอ่อนของระบบจักรกลเรียนรู้ในแต่ละกรณี
3. สามารถโจมตีระบบจักรกลเรียนรู้ เพื่อสร้างระบบจักรกลเรียนรู้ที่ทนทานต่อการโจมตีได้

1.6 คำนิยามศัพท์เฉพาะ

- **จักรกลเรียนรู้** (machine learning) คือระบบ หรือโค้ด หรือโปรแกรมคอมพิวเตอร์ที่เรียนรู้โครงสร้างของชุดคำถามและคำตอบโดยมิจำเป็นต้องทำการโปรแกรมลำดับการทำงานอย่างชัดแจ้ง (explicitly)
- **การเรียนรู้เชิงโจมตี** (adversarial learning) หมายถึงการศึกษาถึงการโจมตีแบบจำลอง (model) ของจักรกลเรียนรู้ (machine learning)

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 จักรกลเรียนรู้

ระบบจักรกลเรียนรู้ (machine learning) อาจนิยามได้ว่าเป็นระบบที่ไม่ต้องมีการป้อนข้อมูล หรือวิธีทำงาน เข้าไปยังโค้ดโปรแกรมอย่างชัดแจ้ง (explicitly) โดยระบบดังกล่าวจะถูกฝึกสอนด้วยชุดของข้อมูลหรือประสบการณ์ (experience) และปรับตัวเองให้ส่งออกคำตอบซึ่งอิงจากประสบการณ์ที่ตนเองเคยได้เรียนรู้มา

หากจะกล่าวให้ละเอียด เราสามารถนิยามโปรแกรมซึ่งสามารถทำการ*เรียน*ได้ดังนี้ [?]

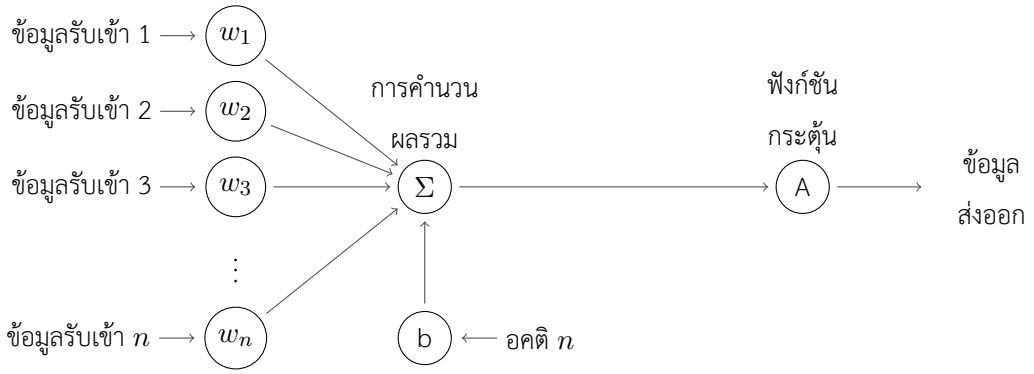
บทนิยาม 2.1.1. โปรแกรมใดๆ เรียน (learn) จากประสบการณ์ (experience) E บนงาน (task) T และการวัดประสิทธิผล (performance measurement) P หากประสิทธิผลบน T ซึ่งถูกวัดโดย P เพิ่มขึ้นตามประสบการณ์ E

2.2 การเรียนรู้เชิงลึก

การเรียนรู้เชิงลึก (Deep Learning) คือความพยายามในการจำลองเซลล์ประสามของมนุษย์ให้อยู่ในรูปโมเดลคณิตศาสตร์ด้วยความเชื่อทางหลักประสาทวิทยา (neurosciences) ว่าความฉลาดของสมองมนุษย์เกิดขึ้นได้จากโครงข่ายประสาทจำนวนมาก ที่เชื่อมเข้าถึงกัน [?]

2.2.1 เปอร์เซปตรอน (Perceptron)

เปอร์เซปตรอน (Perceptron) เป็นแบบจำลองทางคณิตศาสตร์ของเซลล์สมองหนึ่งเซลล์ โดยมีคุณสมบัติดังนี้



รูปที่ 2.1: เปอร์เซปตรอน

- รับเข้าข้อมูลมาในเซลล์จากหลายแหล่ง และให้น้ำหนักกับข้อมูลนั้นต่างกันไป
- ส่งออกข้อมูลเพียงค่าเดียว

ดังนั้น แบบจำลองทางคณิตศาสตร์สามารถเขียนออกมาจากหลักการสองข้อดังกล่าวได้ด้วยสมการ

$$y = f\left(W^T X + b\right)$$

เมื่อ W และ X เป็นเมทริกซ์ขนาด $1 \times n$ (โดย n เป็นจำนวนข้อมูลรับเข้า), b เป็นค่าสัมประสิทธิ์คงที่ (ไบแอส: bias) และ f เป็นฟังก์ชั่นกระตุ้น (activation function) ซึ่งอาจเขียนรูปร่างของเปอร์เซปตรอนให้มีลักษณะรูปคล้ายเซลล์สมองได้ในลักษณะรูปที่ 2.1

ยกตัวอย่างการใช้เปอร์เซปตรอนในการแก้ปัญหาอย่างง่ายได้ในที่นี้

การคาดเดาราคาอสังหาริมทรัพย์

หากสำรวจราคาอสังหาริมทรัพย์แล้วพบว่า

- ราคาอสังหาริมทรัพย์จะเพิ่มขึ้นตามที่ดิน โดยเพิ่มขึ้นทุก 10,000 บาทต่อตารางวา
- ราคาอสังหาริมทรัพย์จะเพิ่มขึ้นตามจำนวนห้องนอน โดยเพิ่มขึ้นทุก 200,000 บาทต่อห้องนอน
- ราคาอสังหาริมทรัพย์จะลดลงตามจำนวนอายุปี โดยลดลงทุก 7,000 บาทต่ออายุของอสังหาริมทัพย์

จะสามารถเขียนเปอร์เซปตรอนเพื่อคาดเดาราคาอสังหาริมทรัพย์ได้โดย

$$y = f\left(W^T X + b\right)$$

เมื่อ W ซึ่งเป็นค่าสัมประสิทธิ์แสดงถึงความสัมพันธ์ข้อมูลรับเข้า ซึ่งเขียนได้จากความสัมพันธ์ดังแสดงด้านล่าง

$$W^T = \begin{bmatrix} 10000 & 200000 & -7000 \end{bmatrix}$$

b เป็นไบแอส, จะสมมติให้ $b = 0$ (กล่าวโดยละเอียด ในที่นี้ b คือราคาตั้งต้นของบ้าน 0 ห้องนอน พื้นที่ 0 ตารางวา อายุ 0 ปี) และ $f(x) = x$ กล่าวคือเป็นฟังก์ชั่นเชิงเส้น

หากต้องการคาดเดาราคาบ้านที่มี 3 ห้องนอน เนื้อที่ 100 ตารางวา และมีอายุ 7 ปี จะสามารถเขียนเมทริกซ์ X ได้เป็น

$$X = \begin{bmatrix} 3 \\ 100 \\ 7 \end{bmatrix}$$

และผลการทำนายราคาบ้านคำนวนได้จาก

$$\begin{aligned} y &= f\left(W^T X + b\right) \\ &= f\left(\begin{bmatrix} 10000 & 200000 & -7000 \end{bmatrix} \times \begin{bmatrix} 3 \\ 100 \\ 7 \end{bmatrix}\right) \\ &= f(30000 + 20000000 + (-49000)) = f(19981000) \\ &= 19981000 \end{aligned}$$

การสร้างประตูสัญญาณตรรกะด้วยเปอร์เซปตรอน

เราสามารถสร้างประตูสัญญาณตรรกะ (logic gates) บางชนิดได้ด้วยเปอร์เซปตรอน เช่นการสร้าง AND และ OR gate

ยกตัวอย่างโครงสร้างของ AND gate ซึ่งสามารถสร้างได้ด้วยการกำหนดให้

- X เป็นเมทริกซ์ขนาด 1×2 กล่าวคือเมื่อรับค่า x_1, x_2 เป็นค่า 0 หรือ 1 แทนสัญญาณจริงหรือเท็จแล้ว

$$X = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

- กำหนดค่าของเมทริกซ์ W เป็น

$$W^T = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

- กำหนดค่าของไบแอส $b = -2$
- กำหนดฟังก์ชั่น $f(x)$ เป็น step function กล่าวคือ

$$f(x) = \begin{cases} 1; & x \geq 0 \\ 0; & \text{ในกรณีอื่น} \end{cases}$$

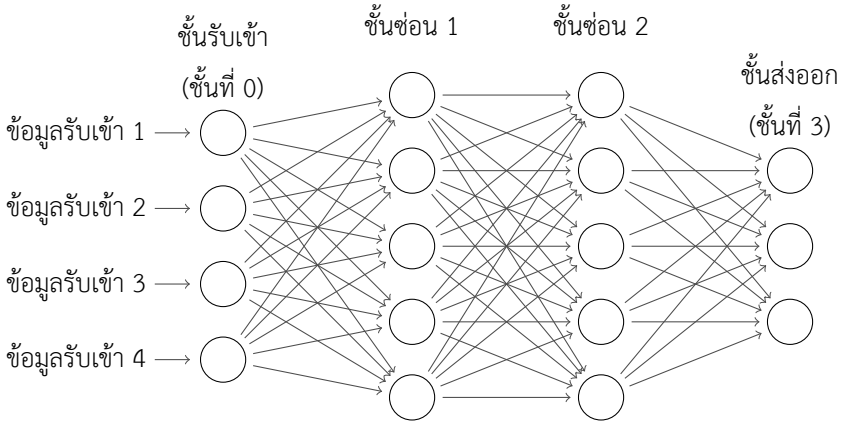
และการสร้าง OR gate สามารถทำได้ในลักษณะเดียวกันโดยเปลี่ยน b เป็น $b = -1$

2.2.2 เปอร์เซปตรอนแบบหลายชั้น (Multi Layer Perceptron)

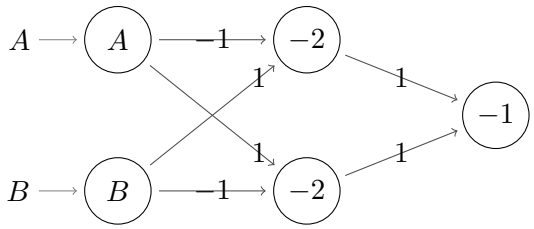
เราอาจสังเกตว่าเปอร์เซพตรอนหนึ่งตัวนั้นทำหน้าที่ได้เพียนแยก (classify) หรือถดถอย (regress) ปัญหาที่เป็นปัญหาเชิงเส้น (linear problems) ได้เท่านั้น อย่างไนก็ตามหากเรากำหนดให้ฟังก์ชั่น f เป็นฟังก์ชั่นที่ไม่ใช่ฟังก์ชั่นเส้นตรงแล้ว เราอาจสร้าง**เปอร์เซปตรอนแบบหลายชั้น** (Multi Layer Perceptron) ขึ้นมาได้โดยมีลักษณะดังรูปที่ 2.2

เราอาจเขียนแทนน้ำหนักของโครงข่ายจากเปอร์เซปตรอนชั้นที่ i ไปยังชั้นที่ j ($j = i + 1$) ได้เป็น

$$\boldsymbol{W}_{ij} = \begin{bmatrix} w_{11} & w_{21} & \dots & w_{n_i 1} \\ w_{12} & w_{22} & \dots & w_{n_i 2} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1 n_j} & w_{2 n_j} & \dots & w_{n_j n_i} \end{bmatrix}$$



รูปที่ 2.2: เปอร์เซปตรอนแบบหลายชั้น



รูปที่ 2.3: เปอร์เซปตรอนแบบหลายชั้นซึ่งทำหน้าที่เป็นประตูสัญญาณ XOR

เมื่อจำนวนเปอร์เซปตรอนในชั้นที่ k เขียนแทนด้วย n_k

ยกตัวอย่างเช่น เราจะสามารถสร้างประตูสัญญาณ XOR (XOR gate) ได้จากเปอร์เซปตรอนแบบหลายชั้นดังแสดงในรูปที่ 2.3 โดยเลขในแต่ละเปอร์เซปตรอนแทนค่าไบแอส (b) และเลขบนเส้นเชื่อมแทนค่าน้ำหนัก (w) และกำหนดให้ฟังก์ชั่นกระตุ้น f เป็นฟังก์ชั่นขั้นบันได (step function) กล่าวคือ

$$f(x) = \begin{cases} 1; & x \geq 0 \\ 0; & \text{ในกรณีอื่น} \end{cases}$$

เปอร์เซปตรอนดังกล่าว เมื่อรับค่า A และ B เป็น 0 หรือ 1 จะส่งออกค่า $A \oplus B$