

CHAPTER 1

ภาพรวมของวิศวกรรมคุณลักษณะ

ในปัจจุบันความรู้ทางด้านการเรียนรู้ของเครื่อง (Machine Learning) ถูกนำมาใช้ในการพัฒนาเทคโนโลยีทางด้านปัญญาประดิษฐ์ (Artificial Intelligent: AI) และการวิเคราะห์ข้อมูลในแนวทางของวิทยาการข้อมูล (Data Science) อย่างกว้างขวาง ในการกระบวนการเรียนรู้ของเครื่อง ข้อมูลดิบ (Raw Data) ซึ่งถือเป็นวัตถุดิบตั้งต้นหรือข้อมูลเข้า (Input) จะผ่านกระบวนการแปลงและวิเคราะห์ที่ซับซ้อนหลายขั้นตอนจนได้ผลลัพธ์ที่ต้องการ ซึ่งอาจอยู่ในรูปแบบแบบจำลอง (Model) หรือ ความเข้าใจอย่างลึกซึ้งในตัวข้อมูล วิศวกรรมคุณลักษณะ (Feature Engineering) เป็นเครื่องมือสำคัญในกระบวนการเรียนรู้ของเครื่องที่ช่วยแปลงข้อมูลเข้าที่มีความหลากหลายให้อยู่ในรูปแบบที่เหมาะสมต่อการนำไปวิเคราะห์ เพื่อให้ผลลัพธ์ที่มีประสิทธิภาพสูง ในบทนี้จะกล่าวถึงความรู้เบื้องต้นในกระบวนการเรียนรู้ของเครื่อง ความรู้พื้นฐานและนิยามศัพท์ที่เกี่ยวข้อง บทบาทของวิศวกรรมคุณลักษณะในกระบวนการเรียนรู้ของเครื่อง การแบ่งประเภทและการประเมินประสิทธิภาพของเทคนิคทางด้านวิศวกรรมคุณลักษณะ

1.1 กระบวนการเรียนรู้ของเครื่อง

การสร้างแบบจำลองด้วยวิธีการเรียนรู้ของเครื่อง เป็นการสอนแบบจำลองโดยนำเสนอข้อมูลจำนวนมากให้แบบจำลองทำการเรียนรู้ด้วยขั้นตอนวิธีการเรียนรู้ (Learning Algorithm) คล้ายกับการเรียนรู้ของมนุษย์ ข้อมูลจะถูกนำไปใช้ในกระบวนการเรียนรู้ตั้งแต่ขั้นตอนแรกจนกระทั่งได้มาซึ่งแบบจำลองที่มีประสิทธิภาพ กระบวนการสร้างแบบจำลองประกอบด้วยขั้นตอนหลัก 5 ขั้นตอนหลัก (แสดงดังภาพ 1.1) ดังนี้

1) การรวบรวมข้อมูล (Data Collection) การรวบรวมข้อมูลเป็นขั้นตอนแรกในการเตรียมข้อมูลสำหรับการสร้างแบบจำลอง ในขั้นตอนนี้ข้อมูลจะถูกรวบรวมจากแหล่งข้อมูลต่างๆ ซึ่งอาจอยู่ในรูปแบบที่หลากหลาย ทั้งเป็นข้อมูลโครงสร้าง (Structured Data) และข้อมูลไม่มีโครงสร้าง (Unstructured Data) การรวบรวมข้อมูลได้จำนวนมากและมีความหลากหลายนั้นสามารถนำไปสู่การสอนแบบจำลองที่มีประสิทธิภาพได้ เหมือนการเรียนรู้ของมนุษย์ เมื่อได้เห็นข้อมูลจำนวนมากและมีความหลากหลายครอบคลุมความเป็นไปได้แบบต่างๆ แล้ว ย่อมมีประสบการณ์มาก และนำไปสู่การตัดสินใจที่เหมาะสมได้

2) การเตรียมข้อมูล (Data Preparation) ข้อมูลดิบที่ได้มาจากระบบการรวบรวมข้อมูลนั้น บางครั้งอาจมีความไม่สะอาดและมีข้อมูลสูญหาย (Missing Data) โดยเฉพาะอย่างยิ่งในกรณีรวบรวมข้อมูลมาจากหลากหลายแหล่งข้อมูล ข้อมูลที่รวบรวมได้อาจมีรูปแบบในการจัดเก็บที่แตกต่างกัน ดังนั้นจึงต้องมีการเตรียมข้อมูลเพื่อให้ได้ข้อมูลที่มีคุณภาพ และจัดเก็บในรูปแบบที่เป็นมาตรฐานเดียวกัน ในขั้นตอน

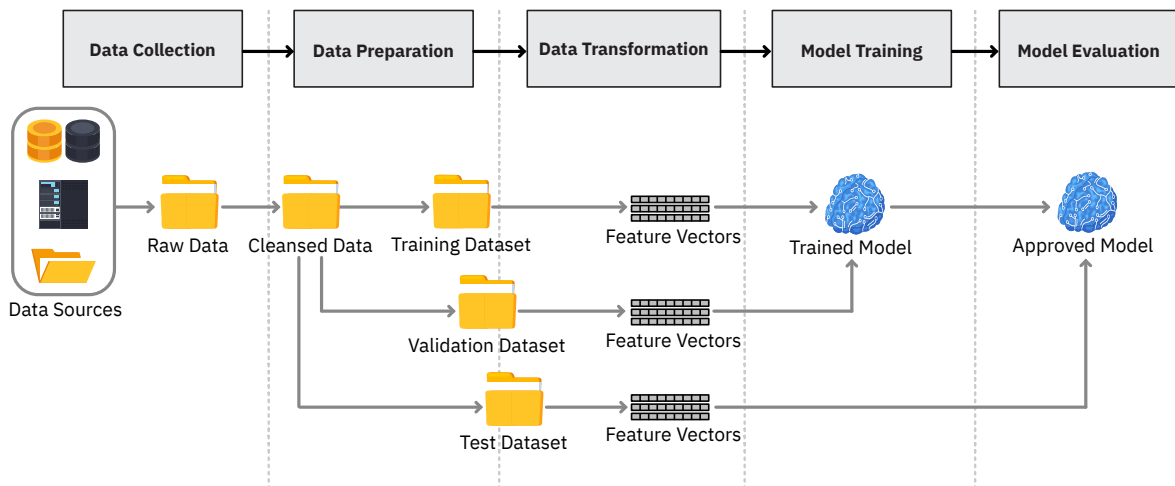


Figure 1.1. กระบวนการเรียนรู้ของเครื่อง

การเตรียมข้อมูลนี้ ข้อมูลจะผ่านการทำความสะอาดข้อมูล (Data Cleansing) การตรวจหาและเติมค่าข้อมูลสูญหาย การจัดโครงสร้างการจัดเก็บข้อมูล และการรวมข้อมูล (Data Integration) ให้เป็นชุดข้อมูล ข้อมูลจะถูกแบ่งเป็น 2 ชุดข้อมูล คือ ชุดข้อมูลฝึกสอน (Training Dataset) เป็นชุดข้อมูลที่จะถูกนำไปใช้ในการเรียนรู้ของแบบจำลอง และชุดข้อมูลทดสอบ (Test Dataset) เป็นชุดข้อมูลที่มีลักษณะเดียวกันกับข้อมูลในชุดข้อมูลฝึกสอน สำหรับนำมาใช้ในการประเมินประสิทธิภาพของแบบจำลองที่ผ่านการฝึกสอนแล้ว หากมีข้อมูลปริมาณมากเพียงพอชุดข้อมูลตรวจสอบ (Validation Dataset) เป็นอีกชุดข้อมูลหนึ่งที่มีถูกจัดทำขึ้นสำหรับการหาค่าไฮเปอร์พารามิเตอร์ (Hyper-parameter) ที่เหมาะสมของแบบจำลองในขั้นตอนการสอนแบบจำลอง โดยทั่วไปแล้วเราตั้งสมมติฐานว่าข้อมูลตัวอย่างในชุดข้อมูลมีความเป็นอิสระต่อกันและได้มาจากการสุ่มภายใต้การแจกแจงแบบเดียวกัน นั่นหมายถึงข้อมูลตัวอย่างในชุดข้อมูลสามารถเป็นตัวแทนของประชากรทั้งหมดภายใต้ขอบเขตที่สนใจได้

3) การแปลงข้อมูล (Data Transformation) เป็นขั้นตอนของการเปลี่ยนรูปแบบการแทนข้อมูลให้อยู่ในรูปแบบข้อมูลเข้า (Input) ที่เหมาะสมต่อแบบจำลอง มักมีความจำเพาะกับชนิดของแบบจำลองที่นำมาใช้ ซึ่งมีความแตกต่างกันไปในแต่ละชนิดของแบบจำลอง ตัวอย่างเช่น ต้นไม้ตัดสินใจ (Decision Tree) สามารถรับทั้งข้อมูลเชิงตัวเลข (Numerical Data) และข้อมูลจำแนกประเภท (Categorical Data) เป็นข้อมูลเข้าได้ ในขณะที่โครงข่ายประสาทเทียม (Artificial Neural Network: ANN) ต้องการข้อมูลเข้าในรูปแบบตัวเลขทั้งหมด จึงต้องแปลงข้อมูลที่ไม่ใช่ตัวเลขให้เป็นข้อมูลเชิงตัวเลขก่อนส่งเป็นข้อมูลเข้าของโครงข่ายประสาทเทียมด้วยวิธีการเข้ารหัสข้อมูล (Encoding) แบบจำลองการเรียนรู้เชิงลึก (Deep Learning) สามารถจัดการกับข้อมูลไม่มีโครงสร้าง (Unstructured Data) ได้โดยตรง แต่แบบจำลองแบบดั้งเดิมต้องการข้อมูลเข้ารูปแบบเวกเตอร์คุณลักษณะ (Feature Vector) ดังนั้น หากข้อมูลที่รวบรวมได้เป็นข้อมูลไม่มีโครงสร้างจะต้องสกัดคุณลักษณะ (Feature Extraction) แล้วแทนในรูปแบบเวกเตอร์คุณลักษณะก่อนนำไปใช้สร้างแบบจำลอง ในบางครั้งคุณลักษณะที่สกัดได้จากข้อมูลดิบมีจำนวนมาก นำไปสู่ปัญหาข้อมูลมิติสูงซึ่งเกิดขึ้นกับแบบจำลองบางประเภทโดยเฉพาะแบบจำลองที่ใช้การวัดระยะทางของข้อมูลในการตัดสินใจ นอกจากนี้ยังนำมาซึ่งการใช้เวลาและทรัพยากรในการประมวลผลมากด้วย การลดมิติข้อมูล (Dimensionality Reduction) จึงเป็นอีกการดำเนินงานหนึ่งที่มีักนำมาใช้ เพื่อการลดผลกระทบของปัญหาข้อมูลมิติสูงและลดระยะเวลาและการใช้ทรัพยากรในการประมวลผล อีกทั้งยังสามารถระบุคุณลักษณะที่มี

ศักยภาพในการนำมาให้แบบจำลองเรียนรู้เพื่อให้ผลลัพธ์ที่ถูกต้องด้วย

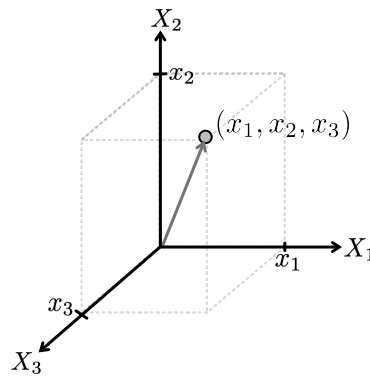
4) การสอนแบบจำลอง (Model Training) เมื่อข้อมูลถูกแทนในรูปแบบเวกเตอร์คุณลักษณะที่เหมาะสมแล้ว ในขั้นตอนนี้ข้อมูลในชุดข้อมูลฝึกสอนจะถูกนำมาฝึกสอนแบบจำลองให้มีความฉลาดด้วยขั้นตอนวิธีการเรียนรู้แบบมีผู้สอน (Supervised Learning) หรือการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) หรือการเรียนรู้แบบเสริมกำลัง (Reinforcement Learning) ก็ได้ ทั้งนี้ขึ้นอยู่กับทางเลือกแบบจำลองที่เหมาะสมกับปัญหาที่สนใจ ในที่นี้จะมุ่งเน้นรูปแบบปัญหาเชิงพยากรณ์ (Predictive Problem) อาจเป็นปัญหาการจำแนกประเภท (Classification Problem) หรือปัญหาการถดถอย (Regression Problem) ซึ่งจะใช้วิธีการเรียนรู้แบบมีผู้สอนในการหาค่าพารามิเตอร์ที่เหมาะสมของแบบจำลองโดยใช้ข้อมูลฝึกสอนที่มีผลเฉลย (Ground Truth)

5) การประเมินประสิทธิภาพแบบจำลอง (Model Evaluation) แบบจำลองที่ผ่านการสอนแล้วจะต้องถูกประเมินประสิทธิภาพโดยการนำข้อมูลในชุดข้อมูลทดสอบให้แบบจำลองทำนาย จากนั้นเปรียบเทียบการทำนายและผลเฉลย แล้วทำการวัดค่าประสิทธิภาพด้วยมาตรวัดแบบต่างๆ หากประสิทธิภาพของแบบจำลองยังไม่เพียงพอต่อการนำไปใช้งานจริง กระบวนการเรียนรู้ของเครื่องอาจต้องย้อนกลับไปดำเนินการรวบรวมข้อมูลฝึกสอนให้มีปริมาณมากขึ้น เพื่อให้ได้ข้อมูลที่มีความหลากหลายเพียงพอ หรืออาจต้องค้นหาหรือออกแบบคุณลักษณะเพิ่มเติมที่สามารถใช้ทำนายค่าตัวแปรที่สนใจได้อย่างมีประสิทธิภาพ เมื่อมีการเพิ่มข้อมูลฝึกสอนหรือเพิ่มคุณลักษณะใหม่จะต้องสอนแบบจำลองอีกครั้ง และทดสอบประสิทธิภาพเพื่อแสดงให้เห็นว่าแบบจำลองใหม่ที่ได้มีประสิทธิภาพเพิ่มสูงขึ้น

1.2 ความเกี่ยวข้องระหว่างข้อมูลและคุณลักษณะ

หากจะกล่าวถึงความเกี่ยวข้องระหว่าง “ข้อมูล” และ “คุณลักษณะ” อาจกล่าวได้ว่า สรรพสิ่ง หรือเหตุการณ์ คือ ข้อมูล และคุณลักษณะ คือ คำอธิบายข้อมูลนั้นในมุมมองต่าง ๆ ตัวอย่างเช่น ข้อมูลของคนหนึ่งคนสามารถอธิบายได้ด้วยคุณลักษณะตั้งแต่ระดับดีเอ็นเอ (Deoxyribonucleic acid: DNA) ยีน (Gene) จีโนไทป์ (Genotype) ฟีนไทป์ (Phenotype) ซึ่งแสดงออกมาผ่านทางลักษณะทางการภาพ เช่น สีตา สีผิว และสีผม ลักษณะใบหน้า รูปร่าง ชื่อ-สกุล อายุ อุปนิสัย ที่อยู่ ลำดับเครือญาติ การศึกษา ประวัติ สุขภาพ สถานะทางสังคม และความสัมพันธ์กับบุคคลอื่นๆ จะเห็นได้ว่าเราสามารถอธิบายคน ๆ หนึ่ง ด้วยคุณลักษณะที่หลากหลาย ทั้งนี้ในทางปฏิบัติเรามักสนใจข้อมูลในบางแง่มุมเท่านั้น และใช้คุณลักษณะบางประการในแง่มุมที่สนใจเพื่อแทนข้อมูลนั้น คุณลักษณะสามารถเรียกอีกอย่างว่า ตัวแปร (Variable) เนื่องจากค่าแปรเปลี่ยนไปตามข้อมูลที่แตกต่างกันไป

ในกระบวนการเรียนรู้ของเครื่อง ข้อมูลหนึ่งข้อมูลจะถูกแทนในรูปแบบเวกเตอร์ $\mathbf{x} = (x_1, x_2, x_3, \dots, x_d)^T$ ในปริภูมิคุณลักษณะ (Feature Space) d มิติ ซึ่งประกอบด้วยคุณลักษณะ d คุณลักษณะ เรียกเวกเตอร์นี้ว่า เวกเตอร์คุณลักษณะ (Feature Vector) ตำแหน่งของค่าคุณลักษณะแต่ละคุณลักษณะในเวกเตอร์มีความเฉพาะไม่สามารถสลับตำแหน่งได้ ภาพ 1.2 แสดงเวกเตอร์คุณลักษณะเวกเตอร์หนึ่งในปริภูมิคุณลักษณะ 3 มิติ เวกเตอร์นี้มีทิศออกจากจุดกำเนิด (Origin) ไปยังจุด (x_1, x_2, x_3) เรียกว่า จุดข้อมูล (Data Point)

Figure 1.2. เวกเตอร์คุณลักษณะ (x_1, x_2, x_3) ในปริภูมิคุณลักษณะ 3 มิติ

1.3 วิศวกรรมคุณลักษณะในการเรียนรู้ของเครื่อง

วิศวกรรมคุณลักษณะ เป็นกระบวนการเปลี่ยนแปลงข้อมูลเพื่อให้ได้มาซึ่งเวกเตอร์คุณลักษณะที่สามารถแทนข้อมูลได้ดีที่สุดภายใต้ขอบเขตของปัญหาที่สนใจ อันจะส่งผลให้แบบจำลองการเรียนรู้ของเครื่องมีประสิทธิภาพสูงขึ้น ซึ่งกระบวนการนี้อยู่ในขั้นตอนการเตรียมข้อมูลและการแปลงข้อมูลในกระบวนการเรียนรู้ของเครื่อง งานในวิศวกรรมคุณลักษณะแบ่งออกได้เป็น 5 ประเภท ได้แก่

1) **การปรับปรุงคุณลักษณะ (Feature Improvement)** การปรับปรุงคุณลักษณะเป็นการจัดการคุณลักษณะที่มีอยู่แล้วให้มีคุณภาพดีขึ้นหรือมีคุณสมบัติตามที่ต้องการโดยอาศัยการแปลงทางคณิตศาสตร์เพื่อเพิ่มศักยภาพการใช้ประโยชน์จากคุณลักษณะในกระบวนการเรียนรู้ของเครื่อง ประกอบด้วย การเติมข้อมูลสูญหาย (Imputation) การลดข้อมูลเสียหาย การทำข้อมูลให้เป็นบรรทัดฐาน (Data Normalization) การเข้ารหัสข้อมูล (Data Encoding) และการแปลงข้อมูล (Data Transformation)

2) **การสร้างคุณลักษณะ (Feature Construction)** การสร้างคุณลักษณะเพิ่มเติมจากคุณลักษณะที่มีอยู่เดิมเป็นอีกแนวทางหนึ่งที่ทำให้ได้คุณลักษณะใหม่ที่มีศักยภาพในการนำไปใช้ในการเรียนรู้ของเครื่องสูงขึ้น และยังคงคุณสมบัติในการตีความหมายได้ การสร้างคุณลักษณะใหม่สามารถทำได้โดยอาศัยการดำเนินการทางคณิตศาสตร์แปลงทางคณิตศาสตร์สถิติ ได้แก่ การรวมคุณลักษณะเดิม การขยายคุณลักษณะ และการผันคุณลักษณะ การดำเนินการเหล่านี้ต้องอาศัยความรู้ความเข้าใจในศาสตร์ที่เกี่ยวข้องกับข้อมูลเพื่อให้ได้คุณลักษณะใหม่ที่มีความหมายและตีความได้

3) **การสกัดคุณลักษณะ (Feature Extraction)** การสกัดคุณลักษณะเป็นอีกแนวทางหนึ่งในการสร้างคุณลักษณะใหม่จากข้อมูลหรือคุณลักษณะเดิม โดยการใช้วิธีการ (Method) หรือ ขั้นตอนวิธี (Algorithm) อัตโนมัติในการสกัดคุณลักษณะ คุณลักษณะใหม่ที่ได้อาจบางครั้งอาจไม่สามารถตีความได้โดยง่าย ทั้งนี้ขึ้นอยู่กับวิธีการที่เลือกใช้ในการสกัดคุณลักษณะและสมมติฐานที่เกี่ยวข้องกับข้อมูล การสกัดคุณลักษณะสามารถใช้ได้ทั้งกับข้อมูลโครงสร้างและข้อมูลไม่มีโครงสร้าง โดยส่วนใหม่มักดำเนินการกับข้อมูลไม่มีโครงสร้างเพื่อทำให้เกิดข้อมูลโครงสร้างซึ่งเป็นคุณลักษณะแทนข้อมูลเดิม ภาพ 1.3 แสดงตัวอย่างการสกัดฮิสโทแกรมของทิศทางของความชันของระดับสีในภาพเพื่อใช้เป็นเวกเตอร์คุณลักษณะแทนรูปภาพ

4) **การเรียนรู้คุณลักษณะ (Feature Learning)** การเรียนรู้คุณลักษณะเป็นการอาศัยความสามารถของการเรียนรู้เชิงลึก (Deep Learning) ในการสกัดคุณลักษณะอย่างอัตโนมัติจากข้อมูลโดยตรงโดยไม่จำเป็นต้องกำหนดสมมติฐานเกี่ยวกับข้อมูลและไม่ต้องใช้ผู้เชี่ยวชาญในการเลือกวิธีการสกัดคุณลักษณะ เพียงแต่อาศัยความสามารถของแบบจำลองการเรียนรู้เชิงลึกในการค้นหาคุณลักษณะที่มีศักยภาพจาก

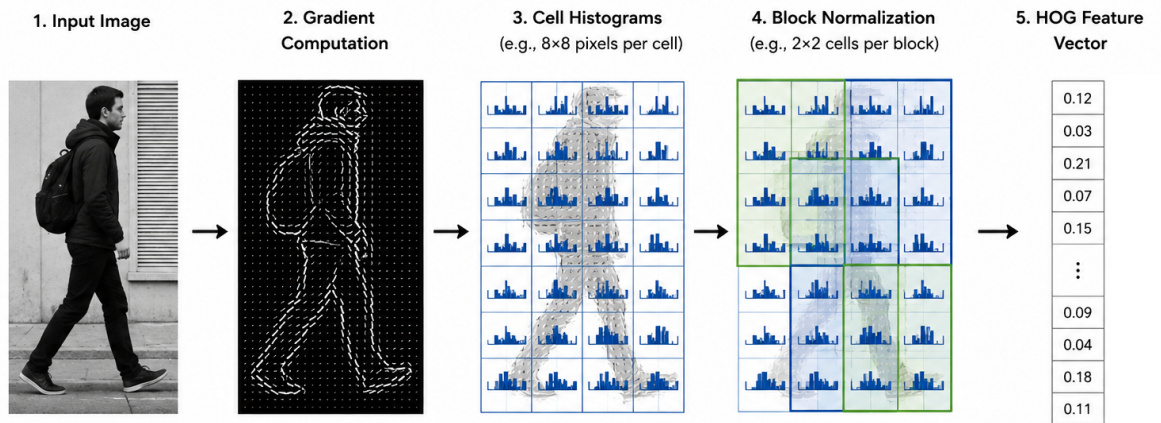


Figure 1.3. การสกัดคุณลักษณะจากภาพโดยใช้ฮิสโตแกรมของทิศทางของความชันของระดับสีในภาพแทนเวกเตอร์คุณลักษณะ (ภาพประกอบสร้างด้วย ChatGPT (GPT-5.5, OpenAI) และปรับแก้โดยผู้เขียน)

ข้อมูลจำนวนมากมหาศาลเท่านั้น คุณลักษณะที่ได้มักไม่สามารถตีความหมายได้ เนื่องจากแบบจำลองเชิงลึกที่นำมาใช้มีลักษณะเป็นกล่องดำ (Black Box) ภาพ 1.4 แสดงตำแหน่งของกลไกการสกัดคุณลักษณะในแบบจำลองโครงข่ายประสาทเทียมแบบสังวัตนาการ (Convolutional Neural Network) สถาปัตยกรรม LeNet-5 โดยในส่วนต้นของโครงสร้างโครงข่ายประสาทเทียม เป็นกระบวนการสร้างแผนที่คุณลักษณะโดยตรวจสอบรูปแบบที่ปรากฏบนภาพที่สอดคล้องกับตัวกรอง (Filter) ผ่านการดำเนินการสังวัตนาการ (Convolution Operation) โดยตัวกรองเหล่านี้สร้างผ่านกระบวนการเรียนรู้ของแบบจำลองอย่างอัตโนมัติและมีรูปแบบที่ทำให้ผลการทำนายของแบบจำลองมีความถูกต้องมากที่สุด

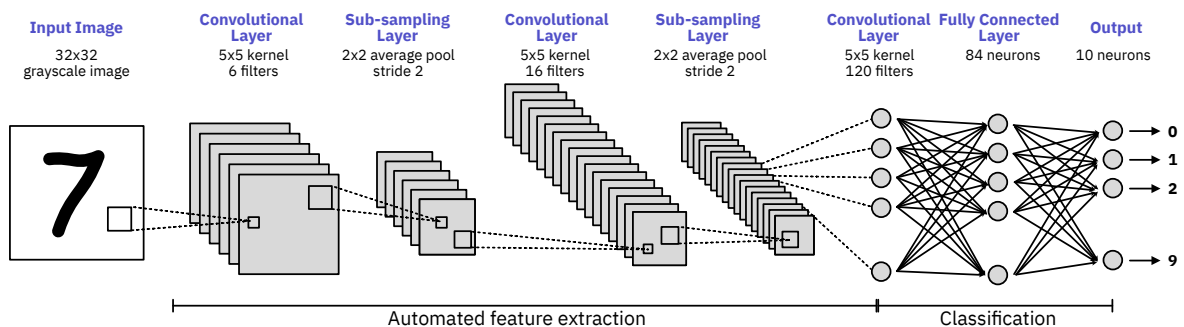


Figure 1.4. โครงข่ายประสาทเทียมแบบสังวัตนาการ (Convolutional Neural Network) สถาปัตยกรรม LeNet-5 ที่สามารถสกัดคุณลักษณะจากภาพโดยอัตโนมัติผ่านกระบวนการเรียนรู้ของแบบจำลอง

1.4 การประเมินประสิทธิภาพของวิธีการทางวิศวกรรมคุณลักษณะ

จากที่กล่าวมาแล้วในหัวข้อก่อนหน้านี้ถึงจุดประสงค์ของวิศวกรรมคุณลักษณะ จะเห็นได้ว่าไม่ว่าจะเลือกใช้วิธีการทางวิศวกรรมคุณลักษณะวิธีการใด วิธีการที่เลือกใช้นั้นจะต้องทำให้แบบจำลองการเรียนรู้ของเครื่องรู้มี

ประสิทธิภาพสูงขึ้น ทั้งนี้อาจเป็นประสิทธิภาพตามเป้าหมายของปัญหาหรือการใช้ทรัพยากรในการประมวลผลก็ได้ นั่นหมายความว่าเราจะประเมินประสิทธิภาพของวิธีการทางวิศวกรรมคุณลักษณะผ่านประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่อง ในมุมมองทางคณิตศาสตร์แบบจำลองการเรียนรู้ของเครื่อง คือ ฟังก์ชัน $f: \mathcal{X} \rightarrow \mathcal{Y}$ ฟังก์ชันหนึ่งที่ส่งเวกเตอร์ $\mathbf{x}$ ในปริภูมิ d มิติ ไปเป็นเวกเตอร์ $\mathbf{y}$ ในปริภูมิ q มิติ $\mathcal{Y}$ เวกเตอร์ $\mathbf{x}$ นี้ คือ เวกเตอร์คุณลักษณะในปริภูมิคุณลักษณะ $\mathcal{X}$ นั่นเอง ในขณะที่เวกเตอร์ $\mathbf{y}$ คือ ผลลัพธ์จากแบบจำลอง จะสังเกตได้ว่า ฟังก์ชัน f มีความจำเพาะต่อปริภูมิ $\mathcal{X}$ และ $\mathcal{Y}$ ในที่นี้เมื่อกล่าวถึงแบบจำลองจะหมายรวมทั้งตัวแบบจำลอง ปริภูมิของเวกเตอร์คุณลักษณะ และปริภูมิของเวกเตอร์ผลลัพธ์ด้วย เนื่องจากวิธีการทางวิศวกรรมคุณลักษณะแต่ละวิธีทำให้เกิดปริภูมิคุณลักษณะที่แตกต่างกันไป เมื่อนำข้อมูลในปริภูมิคุณลักษณะที่แตกต่างไปสอนแบบจำลองการเรียนรู้ของเครื่อง จะทำให้ได้แบบจำลองที่มีค่าพารามิเตอร์ที่แตกต่างกันด้วย ดังนั้น ในการประเมินประสิทธิภาพของวิธีการทางวิศวกรรมคุณลักษณะ m วิธีการ สำหรับปัญหาหนึ่ง เมื่อมีแบบจำลองการเรียนรู้ของเครื่องจำนวน n ชนิด จะได้แบบจำลองทั้งหมด $m \times n$ แบบจำลองที่ต้องทำการประเมินประสิทธิภาพ ตัวอย่างเช่น หากมีวิธีการทางวิศวกรรมคุณลักษณะ 3 วิธีการและแบบจำลองการเรียนรู้ของเครื่อง 4 ชนิด ดังนั้น จะมีแบบจำลองทั้งหมด 12 แบบจำลองที่ต้องทำการประเมินประสิทธิภาพ

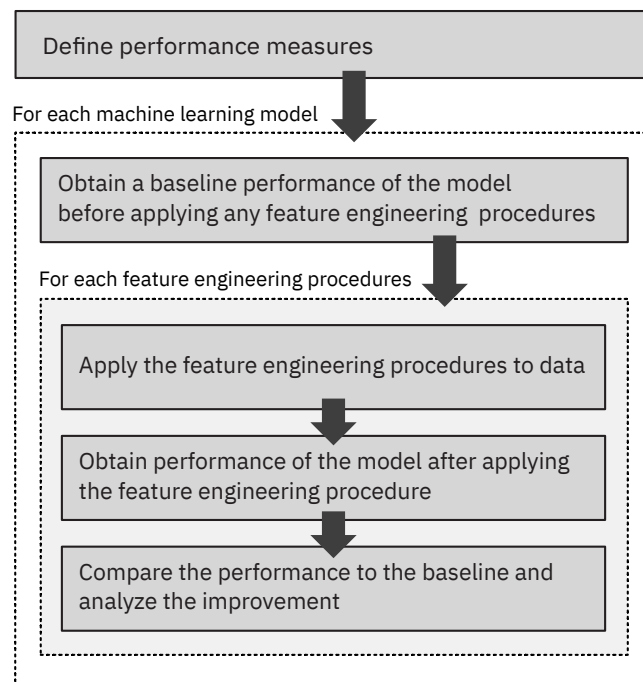


Figure 1.5. กระบวนการประเมินประสิทธิภาพของวิธีการทางวิศวกรรมคุณลักษณะ

ในการประเมินประสิทธิภาพแบบจำลองการเรียนรู้ของเครื่อง ชุดข้อมูลทดสอบซึ่งเป็นชุดข้อมูลที่มีข้อมูลไม่ซ้ำซ้อนกับชุดข้อมูลที่ใช้ในการฝึกสอนแบบจำลองจะถูกนำมาใช้สำหรับประเมิน เพื่อให้ได้ผลประเมินประสิทธิภาพสามารถบ่งบอกขึ้นความสามารถที่แท้จริงของแบบจำลองในการนำไปใช้งานในสถานการณ์จริงที่ข้อมูลใหม่ ๆ อาจเป็นข้อมูลที่แบบจำลองไม่เคยเห็นมาก่อน การทดลองเพื่อประเมินประสิทธิภาพของวิธีการทางวิศวกรรมคุณลักษณะ สามารถทำได้โดยขั้นตอน ดังแสดงในภาพ 1.5 เริ่มต้นด้วยการกำหนดตัวชี้วัดประสิทธิภาพ (Performance Metrics) แบบจำลองการเรียนรู้ของเครื่องที่เหมาะสม การเลือกตัวชี้วัดที่เหมาะสมสำหรับการประเมินประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องเป็นสิ่งสำคัญ ทั้งนี้

ขึ้นอยู่กับลักษณะของปัญหาที่สนใจและเป้าหมายของการแก้ปัญหานั้น ต่อมาทำการประเมินประสิทธิภาพของแบบจำลองอ้างอิง (Baseline) โดยใช้คุณลักษณะที่ไม่ผ่านวิธีการทางวิศวกรรมคุณลักษณะใด ๆ เป็นข้อมูลเข้าของแบบจำลองทุก ๆ ชนิดที่สนใจ ทำการฝึกสอนแบบจำลองด้วยชุดข้อมูลเรียนรู้และประเมินประสิทธิภาพด้วยชุดข้อมูลทดสอบ กำหนดค่าประสิทธิภาพของแบบจำลองที่ได้เป็นค่าประสิทธิภาพอ้างอิงของแบบจำลองแต่ละชนิด ต่อมาสำหรับแต่ละชุดการดำเนินการทางวิศวกรรมคุณลักษณะที่ศึกษา ใช้ชุดการดำเนินการนั้นกับข้อมูลดั้งเดิมและนำคุณลักษณะที่ผ่านการแปลงค่าแล้วเป็นข้อมูลเข้าของแบบจำลองทุก ๆ ชนิด ฝึกสอนและประเมินค่าประสิทธิภาพแบบจำลอง ดำเนินการเช่นนี้จนครบทุก ๆ ชุดการดำเนินการที่ศึกษา สุดท้ายเลือกใช้ชุดการดำเนินการทางวิศวกรรมคุณลักษณะที่ให้ค่าประสิทธิภาพสูงที่สุดสำหรับแบบจำลองการเรียนรู้ของเครื่องที่จะเลือกใช้ ในบาง สถานการณ์ขั้นตอนการประเมินประสิทธิภาพอ้างอิงได้สามารถละทิ้งได้ โดยเฉพาะในกรณีที่ข้อมูลดั้งเดิมไม่ได้อยู่ในรูปเวกเตอร์คุณลักษณะที่พร้อมสำหรับนำเข้าสู่แบบจำลอง จึงจำเป็นต้องสกัดคุณลักษณะจากข้อมูลดั้งเดิมก่อน ทำให้ไม่สามารถหาแบบจำลองอ้างอิงได้ ทั้งนี้การเลือกแบบจำลองยังคงสามารถเลือกแบบจำลองที่ให้ผลการประเมินประสิทธิภาพสูงที่สุดมาใช้งานข้อพึงระวังในการดำเนินการทางวิศวกรรมคุณลักษณะ คือ การใช้วิธีการทางวิศวกรรมคุณลักษณะที่ซับซ้อนมากเกินไป ความพยายามของวิศวกรที่ต้องการให้ประสิทธิภาพของแบบจำลองสูงขึ้น โดยเพิ่มการดำเนินการทางวิศวกรรมคุณลักษณะมากเกินไป อาจนำไปสู่ภาวะการเรียนรู้เกิน (Over Fitting) ซึ่งเป็นสถานการณ์ที่แบบจำลองจดจำรายละเอียดของข้อมูลฝึกสอนมากเกินไป ทำให้แบบจำลองทำนายข้อมูลชุดนั้นได้แม่นยำมาก แต่เมื่อนำไปเจอกับข้อมูลใหม่ ที่ไม่เคยเห็นมาก่อน กลับทำนายผิดพลาดและใช้งานจริงไม่ได้ การดำเนินการทางวิศวกรรมคุณลักษณะที่ดี จะต้องทำให้แบบจำลอง มีความสามารถในการนำสิ่งที่เรียนรู้จากข้อมูลชุดฝึกสอน ไปประยุกต์ใช้กับข้อมูลชุดใหม่ที่ไม่เคยเห็นมาก่อนได้อย่างแม่นยำ อันจะทำให้สามารถนำแบบจำลองนั้นไปใช้งานให้เกิดประโยชน์ในสถานการณ์จริงได้