บทที่ 1 บทนำ

หัวข้อหลัก

- วิทยาศาสตร์ข้อมูล คืออะไร
- ความสัมพันธ์ระหว่าง วิทยาศาสตร์ข้อมูล ปัญญาประดิษฐ์ และการเรียนรู้ของเครื่องจักร
- ทำไมจึงต้องมีวิทยาศาสตร์ข้อมูล
- งานหลักของวิทยาศาสตร์ข้อมูล
- อัลกอริทึมที่ใช้ในวิทยาศาสตร์ข้อมูล
- อธิบายแผนการเรียน

วิทยาศาสตร์ข้อมูล (data science)[1] คือการประยุกต์ใช้วิธีการทางวิทยาศาสตร์ อัลกอริทึม ระบบและกระบวนการต่าง ๆ ในการค้นหารูปแบบ ความเชื่อมโยง ความสัมพันธ์ หรือโครงสร้างต่าง ๆ ที่ซ่อนอยู่ในชุดข้อมูล แล้วนำมาใช้ให้เกิดประโยชน์ เช่น ช่วยในการตัดสินใจเลือกซื้อสินค้า, คัดกรองอีเมล์ขยะ, การกำหนดกลุ่มลูกค้าเป้าหมาย เป็นต้น

สาเหตุที่ทำให้ วิทยาศาสตร์ข้อมูล เป็นที่แพร่หลายในปัจจุบัน ก็คือความก้าวหน้าอย่างก้าวกระโดดของเทคโนโลยีสำหรับสร้าง จัดเก็บ และประมวลผลข้อมูล และการก้าวเข้าสู่ยุคของบิ๊กดาต้า (Big Data era) ซึ่งองค์กรและปัจเจกชนสามารถจัดเก็บข้อมูลปริมาณมากได้โดยมีค่าใช้จ่ายที่ไม่สูงมาก ทำให้ในปัจจุบัน องค์กรต่าง ๆ จำเป็นต้อง หาทางนำข้อมูลที่เก็บไว้มาใช้ในการสร้างมูลค่าเพิ่มให้กับธุรกิจของตน

1.1 ปัญญาประดิษฐ์ การเรียนรู้ของเครื่องจักร และ วิทยาศาสตร์ข้อมูล

ปัญญาประดิษฐ์ (artificial intelligence) การเรียนรู้ของเครื่องจักร (machine learning) และวิทยาศาสตร์ข้อมูล (data science) มีความสัมพันธ์ซึ่งกันและกัน และมักจะถูกใช้แทนที่กันในสื่อต่าง ๆ อย่างไรก็ตามทั้งสามสาขาวิชามีความแตกต่างกันขึ้นกับบริบท รูปที่ 1.1 แสดงความสัมพันธ์ระหว่างทั้งสามสาขา

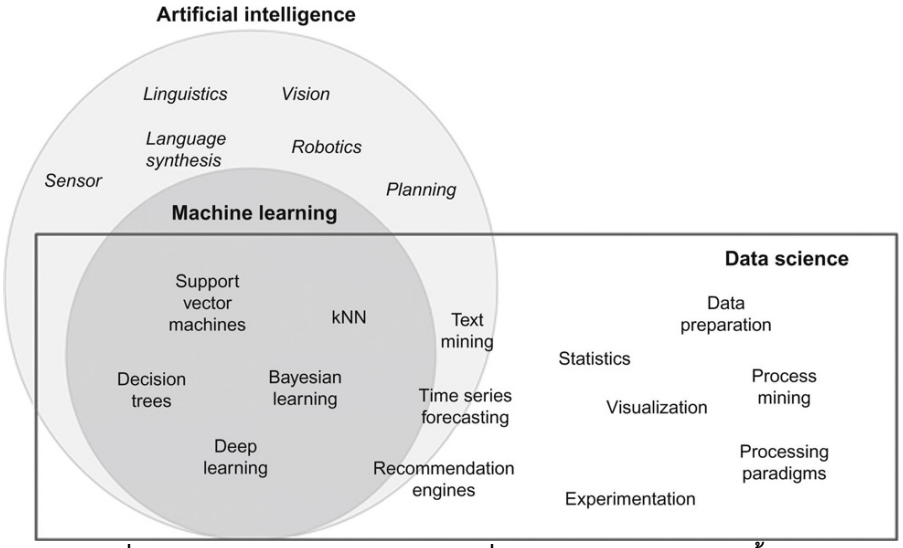
ปัญญาประดิษฐ์ (artificial intelligence) มีเป้าหมายคือการทำให้คอมพิวเตอร์สามารถเลียนแบบพฤติกรรมของมนุษย์ โดยเฉพาะอย่างยิ่งความสามารถในการคิดและทำความเข้าใจ (cognitive functions) เช่น การจดจำใบหน้า การขับขี่อัตโนมัติ การจำแนกจดหมายตามรหัสไปรษณีย์ เป็นต้น ตัวอย่างเทคนิคทางด้านปัญญาประดิษฐ์ ได้แก่ การประมวลผลภาษาธรรมชาติ การตัดสินใจ โรโบติกส์ การวางแผน การประมวลผลภาพ

การเรียนรู้ของเครื่องจักร (machine learning) เป็นสาขาหนึ่งของปัญญาประดิษฐ์ ที่ทำให้คอมพิวเตอร์(เครื่องจักร)สามารถเรียนรู้การแก้ปัญหาต่าง ๆ จากประสบการณ์ได้เอง[2] หากเปรียบเทียบกับการเขียนโปรแกรมเพื่อการแก้ไขปัญหา การเรียนรู้ของเครื่องจักรมีรูปแบบในการแก้ไขปัญหาที่แตกต่างกับการเขียนโปรแกรมอย่างสิ้นเชิง กล่าวคือ ในการเขียนโปรแกรม โปรแกรมเมอร์จะต้องออกแบบลำดับขั้นตอน (algorithms) ในการแก้ปัญหาแล้วแปลงให้อยู่ในรูปของโปรแกรมภาษาคอมพิวเตอร์ วิธีการแก้ปัญหาโดยการเขียนโปรแกรมนี้จะใช้ได้ก็ต่อเมื่อเราสามารถอธิบายวิธีการแก้ปัญหาออกมาเป็นขั้นตอนที่ชัดเจนได้ อย่างไรก็ตาม มีปัญหาบางประเภทที่เราไม่สามารถอธิบายวิธีการแก้ปัญหาออกมาเป็นขั้นตอนที่ชัดเจนได้ (หรืออาจจะได้แต่ยากมาก) เช่น การตรวจจับใบหน้าบนรูปถ่าย การรู้จำเสียงพูด ในกรณีนี้ วิธีการแก้ปัญหาที่เหมาะสมคือการเรียนรู้ด้วยเครื่องจักร โดยการใช้อัลกอริทึมการเรียนรู้ค้นหารูปแบบที่แฝงอยู่ในข้อมูล (เช่น รูปถ่าย และเสียงพูดภาษาไทย) เพื่อ

[1] https://en.wikipedia.org/wiki/Data_science
[2] ประสบการณ์ก็คือ ข้อมูล นั่นเอง

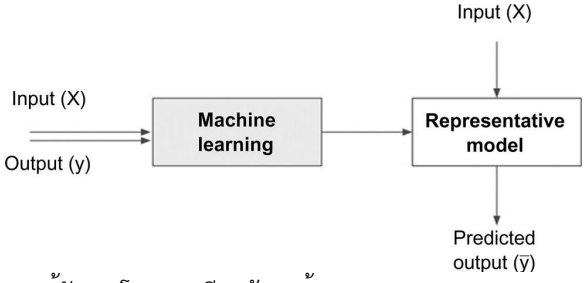
สร้างเป็นโปรแกรมแบบจำลองของความสัมพันธ์หรือรูปแบบแฝงในข้อมูล แล้วจึงนำแบบจำลองที่ได้ไปใช้ในการแก้ปัญหา (ตรวจจับใบหน้าบนรูปถ่าย และรู้จำเสียงพูด) ต่อไป รูปที่ 1.2 แสดงภาพเปรียบเทียบการแก้ปัญหาโดยการเขียนโปรแกรม กับการเรียนรู้ของเครื่องจักร

วิทยาศาสตร์ข้อมูล (data science) คือการประยุกต์ทฤษฎีและเทคนิคการเรียนรู้ของเครื่องจักร ปัญญาประดิษฐ์ สถิติ คณิตศาสตร์ ระบบฐานข้อมูล วิศวกรรมข้อมูล การวิเคราะห์ข้อมูล การทดลอง ปัญญาทางธุรกิจ (business intelligence) และการแสดงผลข้อมูล (visualization) เพื่อสร้างคุณค่าทางธุรกิจให้กับข้อมูล กล่าวโดยย่อ วิทยาศาสตร์ข้อมูล ก็คือสหวิทยาการ (interdisciplinary field) ที่มีเป้าหมายหลักอยู่ที่การสร้างคุณค่าให้กับข้อมูลที่ได้จัดเก็บไว้ โดยมีปัญญาประดิษฐ์และการเรียนรู้ของเครื่องจักร เป็นเครื่องมือหลักที่สำคัญ



รูปที่ 1.1 ปัญญาประดิษฐ์, การเรียนรู้เครื่องจักร, และวิทยาศาสตร์ข้อมูล

(ก) การแก้ปัญหาโดยการเขียนโปรแกรม (traditional programming approach)



(ข) การแก้ปัญหาโดยการเรียนรู้จากข้อมูล (machine learning approach)

รูปที่ 1.2 การเขียนโปรแกรม และ การเรียนรู้ของเครื่องจักร

1.2 วิทยาศาสตร์ข้อมูล คืออะไร

ในหัวข้อนี้ จะอธิบายความหมายของ วิทยาศาสตร์ข้อมูล โดยทำความเข้าใจวิทยาศาสตร์ข้อมูลจากมุมมองต่าง ๆ กัน ดังนี้คือ

- **การค้นหารูปแบบใหม่ที่มีนัยสำคัญ**

  เป้าหมายหลักของ วิทยาศาสตร์ข้อมูล ก็คือการค้นหารูปแบบหรือความสัมพันธ์ แบบใหม่ที่มีนัยสำคัญ และสามารถนำไปสู่การกระทำหรือการตัดสินใจอย่างหนึ่งอย่างใดได้ การค้นหารูปแบบดังกล่าว ใช้กระบวนการทางวิทยาศาสตร์ (นี่เป็นสาเหตุว่าทำไมจึงมีคำว่าวิทยาศาสตร์อยู่ในชื่อวิชานี้) ที่เริ่มจากการตั้งสมมติฐาน ทำการทดลอง สรุปผลและอนุมาน (inference) และกระทำซ้ำ (iteration) จนกว่าจะค้นพบรูปแบบใหม่ที่มีนัยสำคัญและนำไปสู่การกระทำอย่างหนึ่งอย่างใดที่มีประโยชน์ในเชิงธุรกิจได้

  สิ่งสำคัญที่ผู้ศึกษาต้องสังเกต ก็คือ กระบวนการของวิทยาศาสตร์ข้อมูลจะทำการค้นหารูปแบบจากข้อมูลเก่าที่จัดเก็บไว้ (historical data) แต่แบบจำลองหรือโปรแกรมที่ได้จะต้องสามารถนำไปใช้กับข้อมูลใหม่ที่ไม่เคยพบมาก่อนได้ (new unseen data) ดังนั้นอัลกอริทึมการเรียนรู้ของเครื่องจักรจะต้องสามารถ*วางนัยทั่วไปของรูปแบบจากชุดข้อมูลตัวอย่าง*ได้ (*generalization of patterns from a training dataset*) หมายความว่ารูปแบบที่ค้นพบจะต้องถูกต้องเป็นจริงทั้งในชุดข้อมูลตัวอย่าง และในข้อมูลใหม่ที่ไม่เคยพบมาก่อน

- **การสร้างแบบจำลองตัวแทน (representative models)**

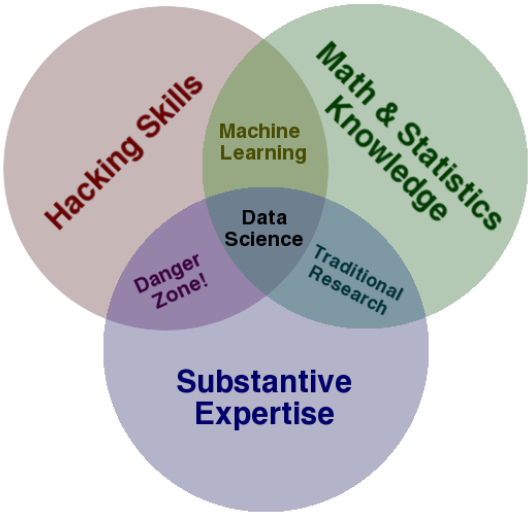
  ในวิชาสถิติ แบบจำลองหรือโมเดล คือ ตัวแทนของความสัมพันธ์ระหว่างตัวแปรต่าง ๆ ในชุดข้อมูล วิทยาศาสตร์ข้อมูล คือกระบวนการสร้างแบบจำลองตัวแทนที่เข้ากันได้กับชุดข้อมูลตัวอย่าง (a representative model that fits the training data) แบบจำลองที่ได้จะถูกนำไปใช้ในการทำนายค่าเอาท์พุทจากข้อมูลอินพุทที่ไม่เคยพบมาก่อน

  ตัวอย่างเช่น การสร้างแบบจำลองสำหรับทำนายราคาบ้าน จากชุดข้อมูลที่ประกอบด้วยอินพุทคือ (ขนาดของบ้านเป็นตารางเมตร, จำนวนห้องนอน, ตำแหน่งที่ตั้ง, ระยะทางจากแนวรถไฟฟ้า) และเอาท์พุทคือ ราคาบ้าน แบบจำลองที่ได้จากกระบวนการสร้างแบบจำลองจะสามารถนำมาใช้ในการทำนายราคาบ้านจากข้อมูลใหม่ที่ไม่เคยพบมาก่อนซึ่งประกอบด้วย (ขนาดของบ้านเป็นตารางเมตร, จำนวนห้องนอน, ตำแหน่งที่ตั้ง, ระยะทางจากแนวรถไฟฟ้า) ได้

- **การบรรจบกันของการเรียนรู้ของเครื่องจักร, คณิตศาสตร์ และความรู้ทางธุรกิจ**

  Drew Conway [3] ได้เสนอแผนภาพเวนน์ของวิทยาศาสตร์ข้อมูล เพื่ออธิบายทักษะแกนหลักของนักวิทยาศาสตร์ข้อมูล ดังรูปที่ 1.3 Conway เห็นว่า ผู้ที่จะเป็นนักวิทยาศาสตร์ข้อมูลที่เชี่ยวชาญได้นั้น จะต้องมีทักษะที่สำคัญ 3 อย่าง คือ (1) ความรู้ทางด้านคณิตศาสตร์และสถิติ (2) ทักษะในการเขียนโปรแกรมและการเรียนรู้ของเครื่องจักร และ (3) ความเข้าใจเกี่ยวกับโดเมนที่เกี่ยวข้อง เช่น หากเรานำวิทยาศาสตร์ข้อมูลไปใช้ในการทำนายราคาบ้าน นักวิทยาศาสตร์ข้อมูลก็จำเป็นต้องมีความเข้าใจเกี่ยวกับตลาดอสังหาริมทรัพย์ หรือ หากเราต้องการใช้วิทยาศาสตร์ข้อมูลสำหรับตรวจจับยีนที่ก่อให้เกิดโรคในสัตว์น้ำ นักวิทยาศาสตร์ข้อมูลก็จำเป็นต้องมีความเข้าใจเกี่ยวกับข้อมูลทางพันธุกรรม เป็นต้น

[3] http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram



รูปที่ 1.3 แผนภาพเวนน์ของวิทยาศาสตร์ข้อมูลโดย Drew Conway.

- **อัลกอริทึมการเรียนรู้ (learning algorithms)**

  การใช้อัลกอริทึมการเรียนรู้ เช่น การเรียนรู้เชิงลึก (deep learning), ตัวจำแนกแบบป่าสุ่ม (random forest classifier) สำหรับค้นหารูปแบบและโครงสร้างที่มีนัยสำคัญและน่าสนใจจากชุดข้อมูล ทำให้วิทยาศาสตร์ข้อมูลแตกต่างจากการวิเคราะห์ข้อมูลแบบดั้งเดิม (traditional data analysis)

  อัลกอริทึมการเรียนรู้ที่นำมาใช้ในวิทยาศาสตร์ข้อมูลส่วนใหญ่ ได้มาจากงานวิจัยในสาขาปัญญาประดิษฐ์และการเรียนรู้ของเครื่องจักร วิธีการทำงานของอัลกอริทึมเหล่านี้ จะใช้หลักการทำซ้ำโดยอัตโนมัติ (**automatic iterative methods**) ในการค้นหาค่าพารามิเตอร์หรือกฎสำหรับอธิบายรูปแบบที่แฝงอยู่ในชุดข้อมูล

  วิทยาศาสตร์ข้อมูล สามารถแบ่งได้ตามชนิดของงานการเรียนรู้หรือ learning tasks ได้ดังนี้คือ (1) การจำแนกประเภท และการวิเคราะห์การถดถอย (classification and regression analysis) (2) การวิเคราะห์ความสัมพันธ์ (association analysis) (3) การจัดกลุ่ม (clustering) และ (4) การตรวจจับความผิดปกติ (anomaly detection) งานแต่ละประเภทจะมีอัลกอริทึมเฉพาะ เช่น อัลกอริทึมสำหรับการจำแนกประเภท ได้แก่ ต้นไม้ของการตัดสินใจ (decision tree) เครือข่ายประสาทเทียม (neural network) และ k-nearest neighbors เป็นต้น

- **สาขาวิชาที่เกี่ยวข้อง**
  - *สถิติเชิงพรรณนา (Descriptive statistics)* การคำนวณค่ากลางของข้อมูล เช่น ค่าเฉลี่ย ค่าเบี่ยงเบนมาตรฐาน สหสัมพันธ์ ถูกนำไปใช้ในขั้นตอนการทำความเข้าใจข้อมูล
  - *การทำความเข้าใจข้อมูลโดยการทำให้เห็นภาพ (Exploratory visualization)* เช่น แผนภูมิแท่ง แผนภูมิวงกลม แผนภาพการกระจายข้อมูล ถูกนำไปใช้ในขั้นตอนการทำความเข้าใจข้อมูล และการนำเสนอผลการทดลอง
  - *การคิวรี่ข้อมูลเชิงมิติแบบต่าง ๆ* เช่น การ slice ข้อมูลตามช่วงเวลา หรือตามเขตพื้นที่ ถูกนำมาใช้ในการค้นหารูปแบบที่ซ่อนอยู่ในข้อมูล
  - *การทดสอบสมมติฐาน (Hypothesis testing)* วิทยาศาสตร์ข้อมูล คือ กระบวนการทำซ้ำของการตั้งสมมติฐานและการทดสอบสมมติฐาน
  - *วิศวกรรมข้อมูล (Data Engineering)* คือกระบวนการในการเก็บ บริหารจัดการ และเตรียมการเข้าถึงข้อมูล เพื่อให้การใช้และวิเคราะห์ข้อมูลเป็นไปอย่างมีประสิทธิภาพ

- *ปัญญาทางธุรกิจ (Business Intelligence: BI)* คือกระบวนการวิเคราะห์และนำเสนอข้อมูลที่ขับเคลื่อนด้วยเทคโนโลยี และส่งผ่านผลลัพธ์ไปยังผู้บริหาร เจ้าหน้าที่และผู้ใช้งานในองค์กร เพื่อช่วยให้ผู้ใช้งานสามารถตัดสินใจทางธุรกิจได้อย่างมีประสิทธิภาพดียิ่งขึ้น

1.3 ทำไมจึงต้องมีวิทยาศาสตร์ข้อมูล

การก้าวหน้าอย่างรวดเร็วของเทคโนโลยีการจัดเก็บข้อมูล การประมวลผลข้อมูล เครือข่ายคอมพิวเตอร์ ทำให้มนุษย์ สามารถสร้างและเก็บข้อมูลในรูปแบบดิจิตอลที่มีรูปแบบหลากหลายและมีความซับซ้อนในปริมาณมากได้ในราคาต่ำ และนำไปสู่ยุคของบิ๊กดาต้า ซึ่งมีคุณลักษณะที่สำคัญ 4 ประการ[4] ดังนี้คือ

- **Volume (ปริมาณมหาศาล)**
  มีการประมาณการว่า ในปี ค.ศ. 2020 ปริมาณข้อมูลที่จะถูกสร้างขึ้นคือ 40 ZETTABYTES (~ 10^{21} Byte)
- **Velocity (ถูกสร้างขึ้นด้วยความเร็วสูง)**
  1TB คือปริมาณข้อมูลที่เกิดขึ้นในแต่ละ trading session ของตลาดหุ้นนิวยอร์ค
- **Variety (มีรูปแบบที่หลากหลาย)**
  ข้อมูลในระบบคอมพิวเตอร์ อยู่ในรูปของข้อความ (text), รูปภาพ (image), วีดีโอ (video), เซ็นเซอร์ (sensor)
- **Veracity (มีระดับความถูกต้องแม่นยำหลายระดับ)**
  ข้อมูลส่วนใหญ่ไม่มีคุณภาพ เนื่องจากความไม่แม่นยำในการวัดค่า หรือความผิดพลาดอื่น ๆ ที่เกิดขึ้นในกระบวนการสร้างข้อมูล

แม้ว่าการวิเคราะห์ข้อมูลแบบดั้งเดิม เช่น สถิติพรรณนา การคิวรี่ข้อมูลเชิงมิติ การทดสอบสมมติฐาน จะสามารถใช้ในการทำความเข้าใจและค้นหาข้อมูลจากบิ๊กดาต้าได้ แต่ก็ไม่สามารถนำไปใช้ในการค้นหาความรู้เชิงลึกและรูปแบบอันซับซ้อนที่แฝงอยู่ภายในข้อมูลปริมาณมากได้ จึงเป็นแรงผลักดันให้เกิดวิทยาศาสตร์ข้อมูลขึ้นมา เพื่อใช้เป็นกรอบความคิดและเทคนิควิธีการ ในการจัดการและใช้ประโยชน์จากบิ๊กดาต้า

1.4 งานหลักของวิทยาศาสตร์ข้อมูล (Data Science Tasks)

ตารางที่ 1.1 แสดงคำอธิบายงานหลักของวิทยาศาสตร์ข้อมูล

ตารางที่ 1.1 Data Science Tasks

ประเภทงาน	คำอธิบาย	อัลกอริทึม	ตัวอย่าง
การจำแนกประเภท (Classification)	กำหนดคุณสมบัติต่างๆ ของอินพุท (input features) ทำนายว่าอินพุทดังกล่าวเป็นสมาชิกของหมวดหมู่ (class) ใด	ต้นไม้การตัดสินใจ (decision tree), เครือข่ายประสาทเทียม (neural network), เพื่อนบ้านที่ใกล้ที่สุด k แห่ง k-nearest neighbors	จำแนกอีเมล์ออกเป็น อีเมล์ขยะ และอีเมล์ปกติ จำแนกประเภทของผู้ลงคะแนนเสียงตามพรรคการเมืองที่เลือก

[4] ที่มา: https://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg

การวิเคราะห์การถดถอย (Regression)	กำหนดคุณสมบัติต่างๆ ของอินพุท (input features) ทำนายค่าของตัวแปรเป้าหมาย (target variable)	การวิเคราะห์การถดถอยเชิงเส้น (linear regression) การถดถอยโลจิสติกส์ (logistic regression)	การทำนายราคาบ้าน การทำนายอัตราการว่างงาน การทำนายค่าประกันภัย
การจัดกลุ่ม (Clustering)	ระบุกลุ่มของข้อมูลที่มีความคล้ายคลึงกัน โดยใช้คุณสมบัติที่แฝงอยู่ในชุดข้อมูล	k-Means, DBSCAN	การจัดกลุ่มลูกค้า (customer segments) ของบริษัท โดยใช้ข้อมูลประวัติการจับจ่าย
การวิเคราะห์ความสัมพันธ์ (association analysis)	การค้นหากลุ่มของข้อมูลที่มีความสัมพันธ์กัน โดยใช้ข้อมูลทรานแซกชัน (transaction data)	A Priori, FP-growth	การสร้างโอกาสทำ cross-selling โดยใช้อัลกอริทึมการวิเคราะห์ความสัมพันธ์ค้นหาสินค้าที่มักจะถูกซื้อพร้อมกัน เช่น ขนมปังกับนม เบียร์กับผ้าอ้อม เป็นต้น
การตรวจจับความผิดปกติ (Anomaly Detection)	กำหนดคุณสมบัติต่างๆ ของอินพุท (input features) ทำนายว่า ข้อมูลนั้น คือ outlier (ค่าผิดปกติ) หรือไม่	Distance-based Local Outlier Factor (LOF), Density-based Local Outlier Factor (LOF)	การตรวจจับทรานแซกชันบัตรเครดิตที่อาจมีปัญหา การตรวจจับการเจาะระบบเครือข่าย
การให้คำแนะนำ (Recommendation)	การแนะนำ item (สินค้า, หนังสือ, เพลง, ข่าว) ให้กับผู้ใช้	Collaborative Filtering, Content-based Filtering	รายการภาพยนตร์แนะนำใน Netflix รายการวีดีโอแนะนำใน YouTube

1.5 อัลกอริทึมที่ใช้ในวิทยาศาสตร์ข้อมูล (Data Science Algorithms)

อัลกอริทึม (Algorithm) คือ ลำดับขั้นตอนในการแก้ปัญหา ที่สามารถดำเนินการได้โดยคอมพิวเตอร์ อัลกอริทึมที่ใช้ในวิทยาศาสตร์ข้อมูล เป็นอัลกอริทึมที่คิดค้นขึ้นโดยนักวิจัยในสาขาปัญญาประดิษฐ์ การเรียนรู้ของเครื่องจักร การทำเหมืองข้อมูล และสถิติ อัลกอริทึมเหล่านี้สามารถจำแนกได้เป็น 4 ประเภทหลัก คือ

- การเรียนรู้แบบมีผู้สอน (Supervised Learning Algorithms)
- การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning Algorithms)
- การเรียนรู้แบบกึ่งมีผู้สอน (Semi-supervised Learning Algorithms)
- การเรียนรู้แบบเสริมกำลัง (Reinforcement Learning Algorithms)

1.5.1 อัลกอริทึมการเรียนรู้แบบมีผู้สอน (Supervised Learning Algorithms)

- อัลกอริทึมการเรียนรู้แบบมีผู้สอน จะหาความสัมพันธ์ระหว่างอินพุท (input features) กับ ค่าเอาท์พุทเป้าหมาย (target output) โดยการเรียนรู้จากชุดข้อมูลที่ประกอบด้วยคู่ของอินพุทและเอาท์พุท
- เนื่องจากอัลกอริทึมประเภทนี้ จำเป็นต้องใช้ตัวอย่างข้อมูลที่ประกอบด้วยทั้งค่าอินพุทและค่าเอาท์พุทที่ถูกต้องในการเรียนรู้ ซึ่งเปรียบได้กับนักศึกษาที่พยายามเรียนรู้วิชาโดยการศึกษาโจทย์ปัญหาที่อาจารย์ได้ให้คำตอบไว้ด้วย ดังนั้นจึงเรียกอัลกอริทึมในกลุ่มนี้ว่าเป็นการเรียนรู้แบบมีผู้สอน หรือ supervised learning
- จากมุมมองทางคณิตศาสตร์ อัลกอริทึมการเรียนรู้แบบมีผู้สอนก็คือการค้นหาฟังก์ชันที่อธิบายความสัมพันธ์ระหว่างค่าอินพุทฟีเจอร์ X กับค่าเอาท์พุทเป้าหมาย y (F: X -> y)
- ตัวอย่างอัลกอริทึมแบบมีผู้สอน
    - Linear regression
    - Logistic regression
    - Collaborative filtering
    - Decision tree
    - k-Nearest Neighbors
    - Neural Network

1.5.2 อัลกอริทึมการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning Algorithms)

- อัลกอริทึมแบบไม่มีผู้สอน จะใช้ชุดข้อมูลตัวอย่างที่ไม่มีการระบุค่าเอาท์พุทเป้าหมาย หรือ label ในการเรียนรู้เพื่อค้นหาโครงสร้างหรือรูปแบบที่แฝงอยู่ในชุดข้อมูล
- เนื่องจากอัลกอริทึมประเภทนี้ ไม่จำเป็นต้องใช้ค่าเอาท์พุทเป้าหมาย ซึ่งเปรียบได้กับการเรียนรู้โดยไม่มีอาจารย์มาคอยเฉลยคำตอบที่ถูกต้องให้ ดังนั้นจึงเรียกอัลกอริทึมในกลุ่มนี้ว่าเป็นการเรียนรู้แบบไม่มีผู้สอน หรือ unsupervised learning
- การเรียนรู้แบบไม่มีผู้สอน ถูกนำไปใช้ในการค้นหาและจัดกลุ่มที่แฝงอยู่ในชุดข้อมูล หรือเพื่อการสร้างแบบจำลองข้อมูลเชิงพรรณนา (descriptive modeling)
- ตัวอย่างอัลกอริทึมแบบไม่มีผู้สอน
    - k-Means
    - DBSCAN
    - A-Priori
    - FP-Growth

1.5.3 อัลกอริทึมการเรียนรู้แบบกึ่งมีผู้สอน (Semi-supervised Learning Algorithms)

- ปัญหาที่พบบ่อยในทางปฏิบัติ คือ การสร้างชุดข้อมูลสำหรับการเรียนรู้แบบมีผู้สอนที่มีทั้งอินพุท และค่าเอาท์พุทเป้าหมายหรือป้ายกำกับคลาส (class label) เป็นงานที่ใช้เวลาและค่าใช้จ่ายสูง เนื่องจากจำเป็นต้องให้ผู้เชี่ยวชาญเป็นผู้กำหนดป้ายกำกับคลาสให้ อัลกอริทึมการเรียนรู้แบบกึ่งมีผู้สอนถูกคิดค้นขึ้นเพื่อแก้ปัญหานี้ โดยการเรียนรู้เพื่อค้นหาความสัมพันธ์ระหว่างอินพุทและเอาท์พุทเป้าหมาย (F: X -> y) ที่ใช้ทั้งข้อมูลที่มีป้ายกำกับคลาสซึ่งมีปริมาณน้อย และข้อมูลที่ไม่มีป้ายกำกับที่มีปริมาณมาก ร่วมกัน
- ตัวอย่างอัลกอริทึมแบบกึ่งมีผู้สอน
    - Self-training
    - Co-training

1.5.4 อัลกอริทึมการเรียนรู้แบบเสริมกำลัง (Reinforcement Learning Algorithms)

- อัลกอริทึมเรียนรู้แบบเสริมกำลัง มีรูปแบบของการเรียนรู้ที่แตกต่างจากการเรียนรู้ทั้งสามแบบก่อนหน้า กล่าวคือ อัลกอริทึมในกลุ่มนี้จะเรียนรู้โดยการสร้างปฏิสัมพันธ์กับสิ่งแวดล้อมและนำผลลัพธ์ในรูปของรางวัล (rewards) หรือการลงโทษ (penalties) มาใช้ในการปรับค่าพารามิเตอร์ของแบบจำลองเพื่อเพิ่มโอกาสในการบรรลุเป้าหมายของการเรียนรู้ที่กำหนด
- ตัวอย่างการประยุกต์ใช้อัลกอริทึมแบบเสริมกำลัง ได้แก่ การควบคุมหุ่นยนต์ เกมคอมพิวเตอร์ต่าง ๆ เช่น หมากล้อม หมากรุก เป็นต้น
- ตัวอย่างอัลกอริทึมแบบเสริมกำลัง
    - Q-learning
    - Deep Q Network

1.6 แผนการเรียนสำหรับวิชาวิทยาศาสตร์ข้อมูลเบื้องต้น

จุดประสงค์หลักของรายวิชานี้ คือการศึกษากระบวนการของวิทยาศาสตร์ข้อมูล แนวคิดหลักของอัลกอริทึมที่ใช้ในงานหลักประเภทต่าง ๆ และการทดลองทางด้านวิทยาศาสตร์ข้อมูลโดยใช้ภาษา Python

เนื่องจากระยะเวลาที่จำกัด ประกอบกับรายวิชานี้เป็นวิชาแรกที่นักศึกษาได้เรียนเกี่ยวกับวิทยาศาสตร์ข้อมูล ดังนั้นรายวิชานี้ จึงเน้นที่ความเข้าใจแนวคิดพื้นฐานของอัลกอริทึม และศึกษาเฉพาะอัลกอริทึมที่สำคัญและถูกใช้งานอย่างแพร่หลายในทางปฏิบัติ ซึ่งส่วนใหญ่จะเป็นอัลกอริทึมการเรียนรู้แบบมีผู้สอน

สำหรับการเขียนโปรแกรมภาษา Python ในรายวิชานี้ มีวัตถุประสงค์เพื่อให้นักศึกษาได้สัมผัสกับกระบวนการของวิทยาศาสตร์ข้อมูล และเพื่อเสริมสร้างความเข้าใจอัลกอริทึมแต่ละตัวให้ลึกซึ้งยิ่งขึ้น แบบฝึกหัดและตัวอย่างโปรแกรมในรายวิชานี้ เหมาะสำหรับนักศึกษาที่มีความรู้พื้นฐานการเขียนโปรแกรมภาษา Python มาก่อน (สอบผ่านรายวิชา 03603111 หลักการโปรแกรมเบื้องต้น I แล้ว)

หัวข้อหลักของรายวิชา มีดังนี้คือ

- กระบวนการของวิทยาศาสตร์ข้อมูล
- วิทยาศาสตร์ข้อมูลด้วยภาษาไพธอน
- การทำความเข้าใจข้อมูลโดยใช้สถิติพรรณนาและ visualization แบบต่าง ๆ
- การจำแนกประเภทและการวิเคราะห์การถดถอย
    - Linear regression, Logistic regression, Neural network, Naïve Bayes, Decision trees
- การประเมินประสิทธิภาพของอัลกอริทึมการเรียนรู้
- การจัดกลุ่ม
    - k-Means, DBSCAN
- การตรวจจับความผิดปกติ
    - Local Outlier Factor Algorithm
- ระบบผู้แนะนำ (recommender systems)
    - Collaborative Filtering

แบบฝึกหัด

1. วิทยาศาสตร์ข้อมูล คืออะไร
2. จงอธิบายความแตกต่างของ การแก้ปัญหาโดยการเขียนโปรแกรม กับ การแก้ปัญหาโดยใช้การเรียนรู้ของเครื่องจักร
3. จงยกตัวอย่าง ปัญหาที่เหมาะกับการเรียนรู้ของเครื่องจักร (machine learning)
4. จงอธิบายความแตกต่างของ การเรียนรู้แบบมีผู้สอน กับ การเรียนรู้แบบไม่มีผู้สอน
5. สมมติว่า เราต้องการสร้างโปรแกรมสำหรับจำแนกประเภทข้อความทวีต (Tweets) ใน Twitter ออกเป็นหมวดหมู่ต่าง ๆ ได้แก่ การเมือง กีฬา ธุรกิจ บันเทิง การศึกษา และเทคโนโลยี
   (ก) เราควรใช้อัลกอริทึมการเรียนรู้ใด
   (ข) จงอธิบายเหตุผลที่ท่านเลือกใช้อัลกอริทึมในข้อ ก.

เอกสารอ้างอิง

[1] Bala Deshpande, Vijay Kotu. *Data Science*. 2nd Edition, Morgan Kaufmann, 2018.

[2] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly, 2017.