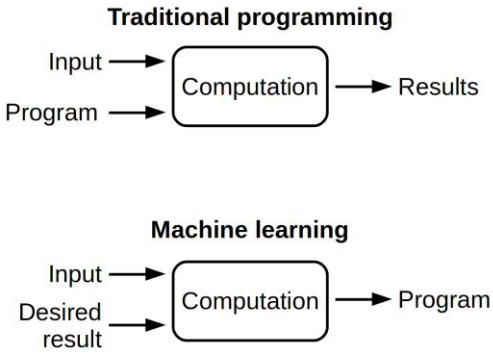
สถิติ กับ Machine Learning ความเหมือนที่แตกต่าง

โดย ณัฐรัฐ อังศุธรารักษ์

ในปี 2011 คอมพิวเตอร์ Watson ของ IBM ชนะการแข่งตอบปัญหากับแชมป์ของรายการ Jeopardy ที่เป็นมนุษย์ ต่อมาในปี 2016 AlphaGo ของ DeepMind Technologies สามารถเอาชนะแชมป์เกม Go หลายคน แสดงให้เห็นถึงศักยภาพของ AI และ Machine Learning ในการเรียนรู้และตัดสินใจที่เทียบเท่า หรืออาจดีกว่ามนุษย์ ซึ่งแน่นอนว่า การเรียนรู้และตัดสินใจดังกล่าว จะไม่ได้ถูกจำกัดไว้เพียงโลกของเกม แต่จะสามารถนำมาใช้ช่วยในการตัดสินใจแทนคน ทั้งการทำงาน การวางแผน การแก้ปัญหาในชีวิตจริงในที่สุด

จริงๆ แล้วคำว่า AI และ Machine Learning มักจะถูกใช้แทนกันเสมอๆ แต่จริงๆแล้ว AI มีความหมายกว้างกว่า Machine Learning ซึ่งเป็นเพียงหัวข้อหนึ่งภายใต้ AI เนื่องจากการเรียนรู้ของเครื่องจักรเป็นองค์ประกอบหนึ่งของการพัฒนาไปสู่ปัญญาประดิษฐ์ (AI) ที่แท้จริง และถึงแม้ว่า Machine Learning จะฟังดูเป็นนวัตกรรมใหม่ที่เพิ่งจะมีการพัฒนาในช่วงไม่กี่ปีที่ผ่านมา แต่ที่จริงแล้ว Arthur Samuel ผู้เชียวชาญด้านคอมพิวเตอร์และปัญญาประดิษฐ์ มีการพูดถึง Machine Learning มาตั้งแต่ปี 1959

โดยปกติแล้ว โปรแกรมเมอร์จะเป็นผู้กำหนดให้คอมพิวเตอร์ทำงานอัตโนมัติตามโปรแกรมที่เขียน แต่สำหรับ Machine Learning คอมพิวเตอร์จะสร้างโปรแกรมที่ตรงกับข้อมูลที่ป้อนให้นั้นขึ้นมาเอง แล้วคอมพิวเตอร์จึงนำโปรแกรมที่สร้างขึ้นมานั้นไปใช้ในการหาคำตอบต่อไปดังที่แสดงในแผนภาพที่ 1



แผนภาพที่ 1: การเปรียบเทียบระหว่าง Traditional Programming กับ Machine Learning

ทั้ง Machine Learning และ วิชาสถิติ ต่างมีเป้าหมายเดียวกัน คือ “การเรียนรู้จากข้อมูล” (Learning from data) โดยที่ Machine Learning จะใช้เทคนิคทางสถิติ (Statistics) ในการสร้างความสามารถให้คอมพิวเตอร์เกิดการ “เรียนรู้” จากข้อมูลโดยปราศจากการสร้างโปรแกรมให้แต่ต้น (without being explicitly programmed)

ประเภทของ Machine Learning สามารถแบ่งออกได้เป็น 3 กลุ่มใหญ่ๆคือ 1. Supervised learning ซึ่งเป็นการสอนคอมพิวเตอร์จากข้อมูลตัวอย่าง และ ผลลัพธ์ที่เรากำหนด (Label) เพื่อให้คอมพิวเตอร์สามารถตอบผลลัพธ์ของข้อมูลชุดใหม่จากตัวอย่างที่ให้ไป 2. Unsupervised learning ซึ่งเป็นการให้ข้อมูลไปจัดการโดยเราไม่ได้สอนผลลัพธ์แก่คอมพิวเตอร์ (ไม่มี Label หรือกฎที่ตายตัว) และ 3. Reinforcement learning ซึ่งเป็นชุดคำสั่ง (Algorithm) ที่ช่วยให้ AI ตัดสินใจตรงกับเป้าหมาย ที่เราตั้งเอาไว้ โดยการใช้รางวัลหรือแรงจูงใจมาช่วย ซึ่งแนวคิดจำนวนมากที่ Machine Learning ใช้จะมีพื้นฐานจากเทคนิคทางสถิติ เช่น Curve fitting โดยศาสตราจารย์ Robert Tibshirani จากมหาวิทยาลัยสแตนฟอร์ด พูดถึง Machine Learning ว่าเป็น “glorified statistics”

ดังนั้น คำถามสำคัญคือ Machine Learning แท้จริงแล้วมีความแตกต่างกันจาก สถิติ (Statistics) หรือไม่อย่างไร จริงๆ แล้วทั้งสองสิ่งนี้มีความเชื่อมโยงกันค่อนข้างมาก และมีกลไกภายใต้การทำงานที่คล้ายคลึงกัน แต่สิ่งที่ต่างกันคือ วัตถุประสงค์ การนำไปใช้งาน การประยุกต์ใช้และข้อควรระวังต่างๆ

ประเด็นที่สำคัญที่สุดคือ ความแตกต่างกันในเรื่องวัตถุประสงค์ของ Machine Learning กับ สถิติ ซึ่ง Machine Learning นั้น ถูกสร้างขึ้นมาเพื่อให้การพยากรณ์หรือคำทำนายจากข้อมูลมีความแม่นยำสูงสุด (Best possible accuracy) โดยไม่มีการตั้งสมมติฐานเกี่ยวกับความสัมพันธ์ระหว่างตัวแปรต่างๆ ซึ่งเราเพียงแค่ป้อนข้อมูลจำนวนมากเข้าไป ชุดคำสั่งคอมพิวเตอร์ก็จะประมวลผล และหารูปแบบของข้อมูลซึ่งเราสามารถนำไปใช้พยากรณ์ตัวแปรต่อไป โดยไม่สนใจการอธิบายความสัมพันธ์ต่างๆระหว่างตัวแปร ยกตัวอย่างเช่น ในกรณีการคิดเบี้ยประกันภัยรถยนต์ อาจจะมีการติดตั้งแอปพลิเคชั่นบนมือถือของผู้ขับขี่ ซึ่งจะเก็บข้อมูลตัวแปรต่างๆที่ได้จากเซ็นเซอร์ของมือถือโดยอัตโนมัติ ก่อนนำไปสร้างและปรับแบบจำลองที่จะคิดคำนวณราคาเบี้ยประกัน โดยไม่ได้พิจารณาถึงเหตุผลหรือความสัมพันธ์ของอัตราการเคลมกับค่าตัวแปรที่เก็บข้อมูลโดยโทรศัพท์

ในส่วนของแบบจำลองทางสถิติ ที่ถึงแม้จะสามารถทำการพยากรณ์ หรือให้คำทำนายโดยอาศัยข้อมูลเช่นเดียวกับ Machine Learning แต่โดยทั่วไปแล้ว แบบจำลองทางสถิติถูกออกแบบมาเพื่ออนุมาน หรือ

อธิบายความสัมพันธ์ระหว่างตัวแปลเป็นหลัก โดยนักสถิติจำเป็นต้องเข้าใจมิติต่างๆของข้อมูล ทั้งวิธีการเก็บข้อมูล คุณสมบัติทางสถิติของข้อมูล การกระจายตัวของกลุ่มตัวอย่าง ฯลฯ เพื่อให้มีความเข้าใจความสัมพันธ์ของตัวแปรแบบจำลอง ในขณะที่ความแม่นยำไม่ถือเป็นประเด็นสำคัญที่สุดของแบบจำลองทางสถิติ โดยจากตัวอย่างเรื่องการคิดเบี้ยประกันรถยนต์ที่ให้ไปข้างต้น แบบจำลองทางสถิติ จะต้องมีการกำหนดสมมติฐานของแต่ละตัวแปร ที่จะนำมาใช้ในการอนุมานความสัมพันธ์กับแนวโน้มในการเคลมประกันไว้อย่างชัดเจนตั้งแต่ต้นก่อนที่จะทำการเก็บข้อมูลและทดสอบแบบจำลองดังกล่าว

นอกจากนี้ อีกประเด็นหนึ่งคือ ความแตกต่างกันระหว่าง การพยากรณ์ ซึ่งเป็นการมองไปข้างหน้า (Forward looking) และการอธิบาย (Rearward looking) ความสัมพันธ์ระหว่างตัวแปร ซึ่งแน่นอนว่า Machine Learning ซึ่งมีการปรับแบบจำลองตามข้อมูลที่เข้ามาใหม่ตลอดเวลา เป็นการเน้นการมองไปข้างหน้า ในขณะที่แบบจำลองทางสถิติ เป็นการศึกษาข้อมูลและความสัมพันธ์ระหว่างตัวแปรในอดีต เพื่ออธิบายรูปแบบความสัมพันธ์ของข้อมูลในช่วงที่มีการเก็บรวบรวมซึ่งเป็นการเน้นการอธิบายสิ่งที่ได้เกิดขึ้น เช่น จากตัวอย่างเดิมเรื่องการคิดเบี้ยประกันภัยรถยนต์ ที่แบบจำลองจาก Machine Learning จะมีการปรับปรุงและเปลี่ยนแปลงอยู่ตลอดเวลาจากการเก็บข้อมูลเพิ่มเติมจากผู้ขับขี่รายใหม่ๆที่เข้าใช้งานระบบ ซึ่งต่างจากการศึกษาทางสถิติที่จะสร้างแบบจำลองความสัมพันธ์ระหว่างพฤติกรรมการขับขี่กับการเคลมประกันจากกลุ่มตัวอย่างที่เก็บข้อมูลมาแล้วเป็นหลัก

ความแตกต่างอีกด้านคือความต้องการด้านข้อมูล ซึ่ง Machine Learning มักจะต้องใช้ข้อมูลจำนวนมากในการศึกษา เพื่อให้ได้แบบจำลองที่สามารถทำงานและทำการพยากรณ์ได้ดี (ถึงแม้จะมีข้อยกเว้นในบางกรณี) เพราะการใช้ Machine Learning มีแนวโน้มที่จะเกิดการ Overfitting แบบจำลอง และข้อมูลจำนวนน้อยอาจจะมี Outliers คิดเป็นสัดส่วนที่มาก ส่งผลให้ความแม่นยำของแบบจำลองลดลง จึงมักจะต้องใช้กลุ่มตัวอย่างจำนวนหลักหมื่น จนถึงหลักล้าน การประยุกต์ใช้จึงเหมาะสำหรับอุตสาหกรรมที่มีการจัดเก็บข้อมูลจำนวนมากอยู่แล้ว เช่น ธนาคาร หรือข้อมูลที่เก็บได้จากอุปกรณ์ที่เก็บข้อมูลโดยอัตโนมัติ ซึ่งต่างจาก การวิเคราะห์ทางสถิติโดยทั่วไปที่มักจะสามารถทำการอนุมาน และพยากรณ์ ได้ค่อนข้างดีโดยใช้ข้อมูลที่มีกลุ่มตัวอย่างหลักสิบ หรือหลักร้อย ดังเช่น การออกแบบสำรวจข้อมูลเพื่อศึกษาเรื่องใดเรื่องหนึ่งเป็นการเฉพาะ โดยที่การเพิ่มจำนวนกลุ่มตัวอย่างไม่ได้ส่งผลมากนักต่อคุณภาพของแบบจำลอง เนื่องจาก สมมติฐานต่างๆที่ถูกกำหนดมาในแบบจำลองทางสถิติจะถูกใช้เป็นข้อมูลที่สามารถนำมาเติมเต็มช่องว่างในการวิเคราะห์ดังกล่าว

ดังนั้น อาจกล่าวได้ว่าแบบจำลองทางสถิติทั่วไปจะได้เปรียบกว่า Machine Learning ในกรณีที่ไม่มีข้อมูลจำนวนมาก

ในทำนองเดียวกันกับจำนวนตัวอย่าง หรือข้อมูล ในกรณีที่จำนวนตัวแปรต้น (Predictor variables) มีจำนวนมาก ตัวแปรมีความสัมพันธ์กันเอง (Multicollinearity) และมีลักษณะความสัมพันธ์ที่ไม่เป็นสมการเชิงเส้น (Nonlinear) การใช้ Machine Learning อาจจะมีประสิทธิภาพมากกว่าแบบจำลองสถิติแบบดั้งเดิม ดังนั้น หากเราต้องการความแม่นยำในการพยากรณ์ข้อมูลต่างๆ และ มีกลุ่มตัวอย่าง หรือตัวแปรจำนวนมากในแบบจำลอง การใช้ Machine Learning จะมีประสิทธิภาพที่สุด อย่างในการสร้างระบบ Image recognition และ Computer vision ซึ่งต้องใช้ข้อมูลรูปภาพในรูปแบบดิจิตัลจำนวนมากเพื่อให้คอมพิวเตอร์สามารถเรียนรู้ได้ ในทางกลับกัน หากสิ่งที่เราต้องการคือการอธิบายความสัมพันธ์ระหว่างตัวแปร และจำนวนกลุ่มตัวอย่างค่อนข้างจำกัด การใช้แบบจำลองทางสถิติทั่วไปอาจมีความเหมาะสมมากกว่า

สุดท้ายนี้ ต้นทุนที่ลดลงอย่างต่อเนื่องของพลังประมวลผลของคอมพิวเตอร์ และความสามารถในการเก็บข้อมูลปริมาณมหาศาล (Big data) จากทั้งราคาที่ลดลงของเทคโนโลยี การเชื่อมโยง IoT และ Platform คอมพิวเตอร์แบบ On-demand ต่างๆ อย่าง AWS Cloud หรือ Google Cloud เปิดโอกาสให้การเข้าถึง Machine Learning เป็นไปได้ง่ายยิ่งขึ้น จากการประหยัดการลงทุนทั้งด้าน Hardware และ Software อย่างไรก็ตาม การประยุกต์ใช้เทคโนโลยีนี้ ยังต้องอาศัยการวางแผนระยะยาว โดยเฉพาะในด้านแนวทางการเก็บข้อมูล ความเข้าใจในการทำงาน และข้อจำกัดของ Machine Learning รวมถึงความสัมพันธ์ระหว่างตัวแปร ซึ่งมีพื้นฐานมาจากวิชาสถิติ เพื่อให้เกิดประโยชน์สูงสุดในการนำไปใช้งาน