

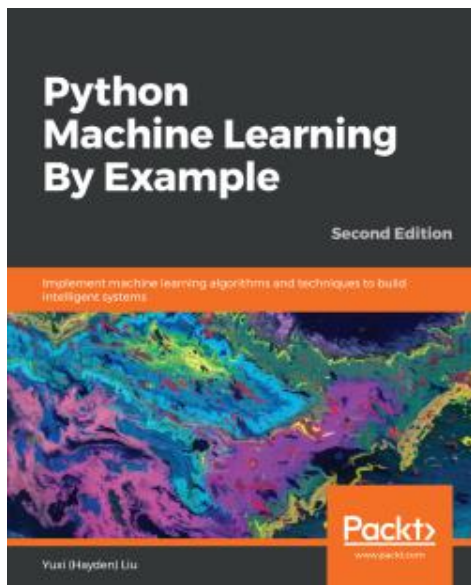
Book Review: Python Machine Learning By Example Second Edition

Author: Yuxi (Hayden) Liu

อนุสรณ์ เกิดศรี

Anusorn Koedsri

สำนักทะเบียนและวัดผล มหาวิทยาลัยสุโขทัยธรรมาธิราช



Python Machine Learning By Example ดี พิมพ์ ครั้งที่ สอง ในปี ค.ศ. 2019 เขียนโดย Yuxi (Hayden) Liu ซึ่งเป็นผู้เขียนหนังสือในชุดการเรียนรู้ของเครื่องหรือแมชชีนเลิร์นนิง (machine learning) หนังสือเล่มนี้ได้รับการตีพิมพ์ครั้งแรกในปี ค.ศ. 2017 และเป็นหนังสือขายดีอันดับ 1 ใน Amazon อินเดีย สองปีติดต่อกันคือในปี ค.ศ. 2017 และ ค.ศ. 2018 หนังสือเล่มนี้เหมาะกับผู้ที่มีความสนใจในเรื่องแมชชีนเลิร์นนิง การวิเคราะห์ข้อมูล วิศวกรรมข้อมูล และผู้ที่ต้องการเริ่มต้นเรียนรู้งานด้านแมชชีนเลิร์นนิง ผู้เขียนให้คำแนะนำว่าควรจะมีพื้นฐานความรู้เบื้องต้นเกี่ยวกับภาษาไพธอน (Python) และแนวคิดพื้นฐานทางสถิติ จะทำให้สามารถใช้ประโยชน์จากหนังสือเล่มนี้ได้สูงสุด

เนื้อหาในหนังสือจะเริ่มต้นด้วยการให้แนวคิดสำคัญเกี่ยวกับแมชชีนเลิร์นนิง วิทยาศาสตร์ข้อมูล และการฝึกปฏิบัติผ่านการใช้โปรแกรมภาษา

Python ในแต่ละบทของหนังสือ ผู้อ่านจะได้เรียนรู้แมชชีนเลิร์นนิง จากตัวอย่างการประยุกต์ใช้ในแอปพลิเคชัน โดยผู้เขียนจะอธิบายขั้นตอนการดำเนินงานอย่างละเอียดและชัดเจน สิ่งที่คุณจะได้เรียนรู้จากหนังสือเล่มนี้ คือ การจัดการปัญหาที่ขับเคลื่อนด้วยข้อมูลที่น่าไปสู่การแก้ปัญหาที่มีประสิทธิภาพผ่านการดำเนินการที่ง่ายด้วยภาษา Python โดยการใช้ไลบรารีเช่น TensorFlow¹, scikit-learn², gensim³ และ Keras⁴ ที่จะช่วยให้คุณเข้าใจขั้นตอนการทำงานของแมชชีนเลิร์นนิง การใช้ Python เพื่อสำรวจโลกของการทำเหมืองข้อมูล (data mining) และการฝึกแบบจำลอง (model training) โดยใช้ข้อมูลที่มีความซับซ้อนและมีประเภทที่หลากหลายด้วยเครื่องมือวิเคราะห์แบบรวมโอเพนซอร์สสำหรับการประมวลผลข้อมูลขนาดใหญ่ (Apache Spark) เพื่อเพิ่มขนาดการวิเคราะห์ รวมถึงการวิเคราะห์ข้อความเชิงลึกด้วยการประมวลผลภาษาธรรมชาติ (Natural Language Processing -- NLP) ซึ่งเป็นการทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์ผ่านการใช้ไลบรารี⁵ ของ Python เช่น NLTK และ gensim

¹ TensorFlow เป็นไลบรารีที่เปิดเผยโค้ดและให้ใช้ฟรีสำหรับการพัฒนาโปรแกรมด้วยการเรียนรู้ของเครื่อง โดยเฉพาะการสร้างโครงข่ายประสาทเทียมซึ่งเป็นผลงานการวิจัยและผลิตภัณฑ์ของกูเกิล

² scikit-learn เป็นไลบรารีที่เปิดเผยโค้ดของภาษา Python สำหรับการพัฒนาโปรแกรมโดยใช้การเรียนรู้ของเครื่อง จุดเด่น คือ มีฟังก์ชันในการจัดประเภทข้อมูล การแบ่งกลุ่มข้อมูล การวิเคราะห์การถดถอยหลายรูปแบบ

³ gensim เป็นไลบรารีที่เปิดเผยโค้ดและให้ใช้ฟรีสำหรับการสร้างแบบจำลองแบบ unsupervised learning และการประมวลผลภาษาธรรมชาติในแมชชีนเลิร์นนิง

⁴ Keras เป็นไลบรารีที่เปิดเผยโค้ดของภาษา Python สำหรับงานด้านการพัฒนาโครงข่ายประสาทเทียม และสามารถทำงานบน TensorFlow, Microsoft Cognitive Toolkit, R, Theano, หรือ PlaidML ได้

⁵ ไลบรารี (library) เป็นแหล่งที่เก็บชุดคำสั่งที่เป็นฟังก์ชันการทำงานที่เฉพาะทางเปรียบเสมือนโปรแกรมสำเร็จรูป ไลบรารีสำเร็จรูปที่มากับโปรแกรม Python ผู้ใช้งานไม่จำเป็นต้องสร้างชุดคำสั่งขึ้นมาใหม่เองทั้งหมด แต่สามารถนำไลบรารีที่มีอยู่แล้วมาใช้งานได้เลย

เป็นต้น การเลือกใช้และสร้างแบบจำลองในแมชชีนเลิร์นนิง การประเมินและการเพิ่มประสิทธิภาพของแบบจำลองในแมชชีนเลิร์นนิง ตั้งแต่เริ่มต้นผ่านตัวอย่างที่มีการนำไปใช้ในทางปฏิบัติจริงด้วย TensorFlow และ scikit-learn ซึ่งเป็นไลบรารีที่สามารถเรียกใช้งานได้ฟรีใน Python

เนื้อหาของหนังสือเล่มนี้มี 10 บท โดยแบ่งเป็น 3 ส่วน ได้แก่ ส่วนที่ 1 พื้นฐานของแมชชีนเลิร์นนิง เป็นเนื้อหาในบทที่ 1 ส่วนที่ 2 เรียนรู้แมชชีนเลิร์นนิงผ่านตัวอย่างที่ใช้งานจริง ประกอบด้วยเนื้อหาในบทที่ 2 ถึง บทที่ 9 และส่วนที่ 3 ตัวอย่างแนวปฏิบัติที่ดีของแมชชีนเลิร์นนิง เป็นเนื้อหาในบทที่ 10 ซึ่งเนื้อหาในแต่ละบทมีรายละเอียดโดยสังเขปดังนี้

บทที่ 1 : เริ่มต้นใช้งานแมชชีนเลิร์นนิง และ Python ภาพรวมของเทคโนโลยีแมชชีนเลิร์นนิง ประเภทของงานแมชชีนเลิร์นนิง ประวัติโดยย่อของการพัฒนาอัลกอริทึมการเรียนรู้ของแมชชีนเลิร์นนิงการติดตั้งซอฟต์แวร์ การตั้งค่าสภาพแวดล้อมและการติดตั้งแพ็คเกจต่าง ๆ เช่น NumPy SciPy Pandas Scikit-Learn และ TensorFlow ใน Python

บทที่ 2 : การเรียนรู้ของเครื่องจากตัวอย่างการใช้งานจริงบน Python ในการสำรวจชุดข้อมูลกลุ่มข่าวด้วยเทคนิคการวิเคราะห์ข้อความ

บทที่ 3 : การทำเหมืองข้อมูลสำหรับชุดข้อมูลกลุ่มข่าวด้วยอัลกอริทึมการจัดกลุ่มและการสร้างแบบจำลองการประยุกต์ใช้การจัดกลุ่มข้อมูลข่าวสารโดยใช้ k-mean clustering การนำ k-mean ไปใช้กับ scikit-learn การเลือกค่า k ที่เหมาะสม และการนำเสนอข้อมูลด้วย t-SNE

บทที่ 4 : การตรวจจับอีเมลขยะด้วย Naïve Bayes ซึ่งเป็นการเริ่มต้นเรียนรู้เกี่ยวกับแบบจำลองของแมชชีนเลิร์นนิงแบบมีผู้สอน (supervised learning) ที่เน้นในเรื่องการใช้ตัวแบบการจำแนกด้วยเทคนิค Naïve Bayes ซึ่งครอบคลุมเนื้อหาเกี่ยวกับทฤษฎีบทของ Bayes กลไกการทำงานของ Naïve Bayes การใช้งาน Naïve Bayes ตั้งแต่ขั้นเริ่มต้น การนำ Naïve Bayes ไปใช้กับ scikit-learn การประเมินประสิทธิภาพการจำแนกด้วยการตรวจสอบแบบไขว้ (cross validation) และการปรับแบบจำลอง

บทที่ 5 : การจำแนกหัวข้อย่อยด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine -- SVM) การหาขอบเขตการแบ่งข้อมูลด้วย SVM การศึกษาวิธีการทำงานของ SVM ผ่านกรณีตัวอย่าง เช่น การระบุเพื่อแยกไฮเปอร์เพลน การพิจารณาไฮเปอร์เพลนที่เหมาะสมที่สุด การจัดการค่าผิดปกติ การนำ SVM มาใช้จัดการกับชุดข้อมูลที่มีมากกว่าสองคลาส ตัวอย่างเพิ่มเติมเกี่ยวกับการจำแนกสถานะของทารกในครรภ์ในการทำ cardiotocography ตัวอย่างการจำแนกมะเร็งเต้านมโดยใช้ SVM กับ TensorFlow

บทที่ 6 : การทำนายการคลิกโฆษณาออนไลน์ด้วยอัลกอริทึมแบบต้นไม้ เนื้อหาในบทนี้จะกล่าวถึงการแนะนำองค์ความรู้เกี่ยวกับต้นไม้การตัดสินใจ (decision tree) เทคนิค Random Forest การใช้งาน Random Forest โดยใช้ TensorFlow ซึ่งเป็นการเรียนรู้จากตัวอย่างการใช้งานจริงที่เกี่ยวกับปัญหาการคาดการณ์อัตราการคลิกโฆษณาออนไลน์

บทที่ 7 : การคาดการณ์การคลิกโฆษณาออนไลน์ด้วยการวิเคราะห์ถดถอยโลจิสติก เป็นการอธิบายการใช้การวิเคราะห์ถดถอยโลจิสติกในการแก้ปัญหาการคาดการณ์อัตราการคลิกโฆษณาออนไลน์ เนื้อหาในบทนี้จะเกี่ยวข้องกับการแปลงคุณสมบัติ

(feature⁶) เจริญพุดหมู่ การเข้ารหัสตัวแปรจัดประเภทการเลือก Feature ด้วยวิธี L/1L 2Regularization สำหรับแบบจำลองการถดถอยโลจิสติกเพื่อแก้ปัญหา overfitting⁷

บทที่ 8 : การขยายขนาดการคาดการณ์การคลิกโฆษณาออนไลน์ เมื่อมีการบันทึกตัวอย่างการคลิกเป็นจำนวนหลายล้านรายการ เนื้อหาในบทนี้จะนำเสนอเกี่ยวกับการใช้แมชชีนเลิร์นนิงในสเกลขนาดใหญ่ซึ่งมีข้อมูลในระดับเทราไบต์ ซึ่งนำการประยุกต์ใช้เครื่องมือสำหรับการประมวลผลแบบขนานที่มีประสิทธิภาพสูงมาช่วยจัดการปัญหา เช่น Apache Hadoop และ Apache Spark มีการอธิบายสาระสำคัญของ Spark ตั้งแต่ขั้นตอนการติดตั้ง หลักการโปรแกรมใน PySpark ส่วนประกอบของแมชชีนเลิร์นนิงในงานระดับสเกลขนาดใหญ่ และการฝึกปฏิบัติกับตัวอย่างข้อมูลหลายล้านรายการเพื่อใช้สร้างแบบจำลองการจำแนก รวมถึงการประเมินและการคัดเลือกแบบจำลอง

บทที่ 9 : การทำนายราคาหุ้นด้วยอัลกอริทึมการถดถอย การทำเหมืองข้อมูลราคาหุ้น การประมาณค่าด้วยการถดถอยเชิงเส้น การประมาณค่าการถดถอยด้วยต้นไม้การตัดสินใจ การประมาณค่าการถดถอยด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine--SVM) การประมาณค่าด้วยโครงข่ายประสาทและการประเมินประสิทธิภาพการทำนายราคาหุ้นจากตัวแบบการถดถอยทั้ง 4 แบบ

บทที่ 10 : กรณีศึกษาแนวปฏิบัติที่ดีของแมชชีนเลิร์นนิง : แนวทางแก้ปัญหาตามกระแสนของแมชชีนเลิร์นนิง ซึ่งแบ่งเป็น 4 ชั้น ได้แก่ 1) แนวปฏิบัติที่ดีในขั้นการออกแบบการไหลของกระแสน เช่น การเตรียมข้อมูล การระบุเป้าหมาย การจัดการกับข้อมูลที่ขาดหายไป การจัดเก็บข้อมูลขนาดใหญ่ 2) แนวปฏิบัติที่ดีในขั้นการสร้างชุดการฝึกแบบจำลอง เช่น การระบุคุณลักษณะเชิงหมวดหมู่ด้วยค่าตัวเลข การตัดสินใจเกี่ยวกับการเข้ารหัสคุณลักษณะที่เป็นหมวดหมู่ การแยกคุณสมบัติจากข้อมูลที่เป็นข้อความ 3) แนวปฏิบัติที่ดีในขั้นการฝึกแบบจำลอง การประเมิน และการคัดเลือกแบบจำลอง เช่น การเลือกอัลกอริทึมที่เหมาะสมขั้นเริ่มต้น Naïve Bayes, SVM, Random forest, ถดถอยโลจิสติก หรือ โครงข่ายประสาท การลดการ overfitting การวินิจฉัยปัญหา overfitting และ underfitting⁸ การติดตั้งเกินและไม่เหมาะสม และการสร้างแบบจำลองบนชุดข้อมูลขนาดใหญ่ และ 4) แนวปฏิบัติที่ดีในขั้นตอนการใช้และการติดตามตรวจสอบแบบจำลองแมชชีนเลิร์นนิง เช่น การบันทึกการไหลและการนำแบบจำลองกลับมาใช้ใหม่ การตรวจสอบประสิทธิภาพของแบบจำลอง การปรับปรุงแบบจำลองอย่างสม่ำเสมอ

จากเนื้อหาที่กล่าวมาทั้งหมดจะเห็นได้ว่า หนังสือเล่มนี้ครอบคลุมความรู้ที่จำเป็นเกี่ยวกับแมชชีนเลิร์นนิงตั้งแต่การติดตั้งซอฟต์แวร์ที่จำเป็นต้องใช้ หลักการพื้นฐานของแมชชีนเลิร์นนิง การเรียนรู้แมชชีนเลิร์นนิงผ่านตัวอย่างที่ใช้งานจริง และการเรียนรู้แมชชีนเลิร์นนิงผ่านการศึกษาที่เป็นแนวปฏิบัติที่ดี เมื่อผู้อ่านศึกษาหนังสือเล่มนี้ครบทั้งหมดจะช่วยให้ผู้อ่านมีความรู้เกี่ยวกับแมชชีนเลิร์นนิงครบถ้วน นอกจากนี้ผู้อ่านสามารถดาวน์โหลดซอร์สโค้ดที่อยู่ในตัวอย่างแต่ละบทได้ฟรีจากเว็บไซต์ <https://github.com/PacktPublishing/Python-Machine-Learning-By-Example-Second-Edition#python-machine-learning-by-example-second-edition> และเป็นความท้าทายของผู้อ่านในการนำแมชชีนเลิร์นนิง และตัวอย่างโค้ดในหนังสือไปประยุกต์ใช้กับงานวิจัยด้านสังคมศาสตร์ ด้านการวัดและประเมินผล ที่ต้องการบูรณาการองค์ความรู้ของแมชชีนเลิร์นนิงข้ามศาสตร์ หนังสือเล่มนี้จึงเหมาะกับผู้เริ่มต้นเรียนรู้ศาสตร์ด้านแมชชีนเลิร์นนิง และต้องการนำองค์ความรู้ไปใช้งานจริง

⁶ คุณสมบัติ หรือ Feature เป็นการเรียกตัวแปรอิสระในศาสตร์แมชชีนเลิร์นนิง แบ่งเป็น ตัวแปรจัดประเภทและตัวแปรต่อเนื่อง ซึ่งคล้ายกับการจัดประเภทของตัวแปรในการวิจัย

⁷ overfitting คือ เหตุการณ์ที่แมชชีนเลิร์นนิงเรียนรู้ข้อมูลได้ดีมากสำหรับชุดข้อมูลสอน แต่มีความสามารถต่ำในการทำนายหรือจำแนกข้อมูลที่ไม่เคยเจอมาก่อน

⁸ underfitting คือ เหตุการณ์ที่นำแมชชีนเลิร์นนิงไปใช้แล้ว ผลลัพธ์ที่ได้จากการทำนายหรือจำแนกไม่ค่อยถูกหรือไม่ใกล้เคียงกับคำตอบที่แท้จริง

บรรณานุกรม

Liu, Y. (2019). *Python machine learning by example: implement machine learning algorithms and techniques to build intelligent systems*, (2nd). Birmingham, UK : Packt Publishing.

