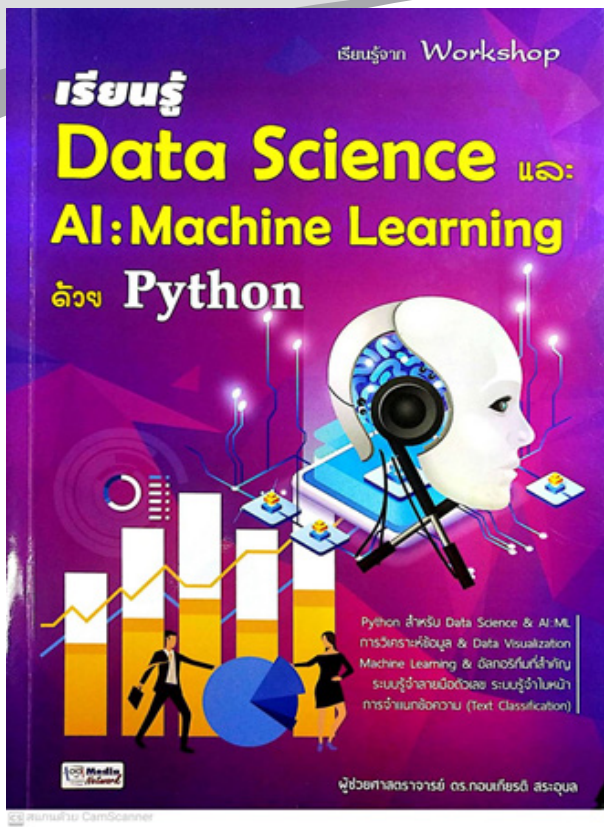
แนะนำหนังสือ
Book Review

โดย กุลวดี เถนว้อง
By Kulwadee Tanwong
หมวดวิชาศึกษาทั่วไป มหาวิทยาลัยอีสเทิร์นเอเชีย
General Education, Eastern Asia University



ชื่อเรื่อง: เรียนรู้ Data Science และ AI: Machine Learning ด้วย Python

ผู้แต่ง: ผู้ช่วยศาสตราจารย์ ดร.กอบเกียรติ สระอุบล

สำนักพิมพ์: หสม มีเดีย เนทเวิร์ค

ปีที่พิมพ์: 2563

จำนวนหน้า: 640 หน้า

เกริ่นนำ

หนังสือ “เรียนรู้ Data Science และ AI: Machine Learning ด้วย Python” เป็นหนังสือที่มีเนื้อหาครอบคลุม Data Science และ AI ในส่วนของ Machine Learning ซึ่งเป็นองค์ประกอบที่สำคัญอย่างหนึ่งของ AI เนื้อหาทุกบทจะมีการฝึกปฏิบัติ (Workshop) ซึ่งเป็นจุดเด่นของหนังสือเล่มนี้ มีการแทรกทฤษฎีที่สำคัญ ใช้ตัวอย่างที่เข้าใจง่าย รูปแบบการเขียนใช้ภาษาเข้าใจง่ายเหมาะสำหรับผู้ที่เริ่มต้นศึกษา Data Science

นอกจากนี้เนื้อหาในหนังสือ ผู้เขียนจัดเรียงเป็นลำดับจากประสบการณ์ที่สอนจริง แต่การเรียนรู้ Data Science และMachine Learning ด้วย Python จะต้องมีการเขียนโค้ดโปรแกรมเป็นหลัก ดังนั้นก่อนจะทดลองฝึกปฏิบัติการตามหนังสือเล่มนี้ จึงควรติดตั้งโปรแกรม Python ในเครื่องคอมพิวเตอร์ของท่านก่อน พร้อมทั้งศึกษาการเขียนโปรแกรม Python พื้นฐานก่อน เพราะหนังสือเล่มนี้ใช้ Python เป็นหลัก

จุดประสงค์

หนังสือเล่มนี้เน้น Workshop แสดงเป็นภาพพร้อมรายละเอียดอย่างชัดเจนทุกขั้นตอน เป็นการนำเอาข้อมูลมาใช้ให้เกิดประโยชน์ในงานด้านต่าง ๆ โดยการทำงานกับข้อมูลมีกระบวนการทางวิทยาศาสตร์อย่างเป็นลำดับ ตามศาสตร์ด้าน Data Science เก็บรวบรวมข้อมูลด้วยแบบสอบถาม Python pandas สถิติเบื้องต้น มีการอธิบายการได้มาซึ่ง Model การปรับจูน และประเมินผล

สาระสำคัญ

หนังสือเล่มนี้ประกอบด้วยเนื้อจำนวนมากถึง 24 บท ดังนี้ (1) บทนำ (2) การติดตั้งและเตรียมความพร้อม (3) Python สำหรับ Data Science (4) จัดการข้อมูลด้วย Pandas (5) สถิติเบื้องต้น (6) สรุปข้อมูลด้วย Pivot Table (7) Data Visualization 1 (8) Data Visualization 2 (9) การจัดการข้อมูลสูญหาย (10) การทำความสะอาดข้อมูล (11) ค่าผิดปกติและการกำจัด (12) ข้อมูล Time series (13) Machine Learning เบื้องต้น (14) Regression (15) Decision Tree (16) การประเมิน Model (17) อัลกอริทึมที่สำคัญ (18) ROC AUC และThreshold (19) ระบบรู้จำเลขลายมือเขียน (20) การตรวจจับและรู้จำใบหน้า (21) การลดมิติข้อมูล (22) การปรับจูน Model (23) การจำแนกหมวดหมู่ข้อความ และ (24) Recommender System

สะท้อนคุณค่า

1. บทที่ 1 2 และ 3 เป็นการกล่าวถึงพื้นฐานความรู้ทั่วไปที่จำเป็นต่อการทำงาน Data Science สามารถดาวน์โหลด Source Code ได้ที่ http://media.itpart.com และอธิบายการติดตั้งและเตรียมความพร้อมชุดโปรแกรมสำหรับ Python เพื่องานคำนวณประมวลผล Data Science, Machine Leaning ดาวน์โหลดและติดตั้งที่ http://www.anaconda.com/distribution สำหรับภาษาที่ใช้พัฒนาโปรแกรมคอมพิวเตอร์นี้เลือกใช้ Python เพราะมีความได้เปรียบที่เขียนโปรแกรมง่าย มีไลบรารีที่มีประสิทธิภาพสูง ไม่ต้องประกาศชนิดของตัวแปร สามารถกำหนดค่าแล้วเรียกใช้ได้เลย

2. บทที่ 4 จัดการข้อมูลด้วย Pandas เป็นไลบรารีแบบเปิด (open source) ที่มีประสิทธิภาพสูง สำหรับจัดการและวิเคราะห์ข้อมูลทั้งแบบโครงสร้างมิติเดียวและหลายมิติ โดยเนื้อหาจะเริ่มตั้งแต่การเตรียมข้อมูล การค้นหาความรู้หรือสิ่งใหม่ในข้อมูล การทำความสะอาดข้อมูล และมีการกล่าวถึงสถิติเบื้องต้น (mean standard deviation) ว่าควรจัดกระทำกับข้อมูลเบื้องต้นอย่างไร จึงทำให้บทนี้มีการฝึกปฏิบัติ (workshop) มากถึง 39 Workshop เพื่อสร้างความเข้าใจให้ผู้อ่านก่อนจะศึกษาขั้นตอนหลักของ Machine Learning

3. บทที่ 5-6 โดยบทที่ 5 เป็นส่วนขยายในบทที่ 4 แสดงรายละเอียดเฉพาะของสถิติเบื้องต้นที่จำเป็นต้องใช้ในการทำงานด้าน Data Science ในส่วนของ Machine Learning เช่น ค่าเฉลี่ย มัธยฐาน ฐานนิยม การแจกแจงความถี่ (frequency distribution) และ Pandas เมื่อตรวจสอบข้อมูลเบื้องต้นว่ามีความเหมาะสมในการนำไปวิเคราะห์ต่อได้ อีกขั้นตอนที่ควรศึกษาก่อนดำเนินงานคือ รูปแบบการสรุปข้อมูล สำหรับในบทที่ 6 เป็นการสรุปข้อมูลด้วยเครื่องมือชื่อ Pivot Table สามารถแสดงเป็นตาราง ค่าเฉลี่ย ผลรวม หรือกราฟได้ง่าย

4. บทที่ 7-8 Data Visualization เป็นการนำข้อมูลเชิงลึกที่เน้นความสัมพันธ์ของข้อมูลหลายตัวแปรจากช่องทางต่าง ๆ มาวิเคราะห์และแสดงผลในรูปแบบของแผนภูมิ (scatter pairplot heatmap) กราฟรูปแบบที่หลากหลาย วิดีโอที่แสดงให้เห็นถึงการเปลี่ยนแปลงของข้อมูลเชิงปริมาณ เนื้อหาจะเน้นศึกษาแนวโน้มและดูความสัมพันธ์ระหว่าตัวแปรเป็นหลัก

5. บทที่ 9-11 นำเสนอเนื้อหาเกี่ยวกับการจัดการข้อมูลสูญหาย การทำความสะอาดข้อมูล การตรวจสอบค่าผิดปกติและการกำจัด จากบทที่ 4 ที่เกริ่นนำการจัดการข้อมูลการเตรียมข้อมูลนั้น ในการทำงานกับข้อมูลจริงสิ่งที่มักจะเกิดขึ้น คือ ข้อมูลบางส่วนสูญหาย หรือได้มาไม่ครบ (missing data หรือ missing values) ความไม่สอดคล้องของข้อมูลในการบันทึก หรือระหว่างการจัดเก็บ และการพบจุดข้อมูลที่แตกต่างไปอย่างผิดปกติ สิ่งเหล่านี้ทำให้การสร้าง Model เกิดปัญหา ดังนั้นก่อนการนำข้อมูลไปใช้งานจึงต้องจัดการข้อมูลโดยการลบ และการแทนที่ข้อมูลที่เป็น Missing data ด้วยค่าเฉลี่ยหรือค่ากลางต่าง ๆ ทำความสะอาดข้อมูลโดยการแก้ไข ลบรายการที่ไม่ถูกต้อง ปรับแต่งโครงสร้างบางส่วน (ต้องไม่กระทบกับข้อมูลหลักที่นำไปใช้) ตรวจสอบค่าผิดปกติ Univariate Outliners (พิจารณาตัวแปร 1 ตัวหรือทีละตัว) และ Multivariate Outliers (พิจารณาตัวแปรมากกว่า 1 ตัวพร้อมกัน)

6. บทที่ 12 Time series เป็นการกล่าวถึงชุดข้อมูลที่อ้างอิงเวลา รวบรวมข้อมูลตามช่วงนั้น มาวิเคราะห์อดีต เพื่อทำนายอนาคต อาทิเช่น ฤดูกาล การจราจรรายวัน การเก็บข้อมูลประเภทนี้ต้องใช้ค่าอนุกรมเวลาเป็นตัวอ้างอิงหลัก และค่าวันเวลา เพื่อให้สามารถประมวลผลข้อมูล ณ ช่วงเวลาที่สนใจ ความแตกต่างของข้อมูลแบบ Time Series กับข้อมูลทั่วไป คือ Trend (เทรนด์) Seasonality (ความแปรผันตามฤดูกาล) และ Cycle (วัฏจักร)

7. บทที่ 13 Machine Learning เบื้องต้นในบทนี้เป็นการสะท้อนให้ผู้อ่านเข้าใจว่า Machine Learning มิใช่ระบบอัจฉริยะที่จะรอบรู้ไปทุกอย่าง มันจะทำงานได้ในเฉพาะส่วนที่ผู้พัฒนาโปรแกรมได้ออกแบบ โดยนำข้อมูล (data set) ต้องมากพอและมีคุณลักษณะเด่น (features) มาสอน (train) เพื่อทำการประมวลผลหรือทำนายได้อย่างอย่างแม่นยำเฉพาะเรื่องใดเรื่องหนึ่งเท่านั้น แสดงการเปรียบเทียบกระบวนการพัฒนาโปรแกรมทั่วไปกับ Machine Learning

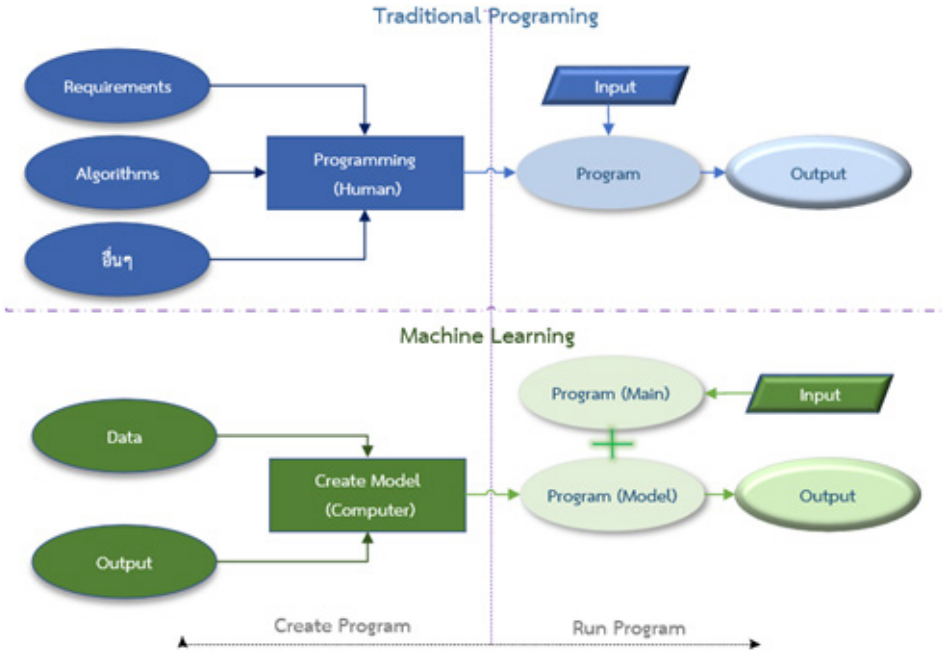
8. บทที่ 14-18 การได้มาของข้อมูลจะมาจากแบบสอบถาม แบบทดสอบ หรือแบบสัมภาษณ์ และนำ Machine Learning มาสร้าง model รูปแบบต่าง ๆ และอธิบายวิธีการประเมินที่เหมาะสมกับ Model นั้น เริ่มจากบทที่ 14 Regression เป็นการศึกษาความสัมพันธ์ของตัวแปรเพื่อสร้าง Model ทั้งในรูปแบบ Simple Linear Regression Multiple Linear Regression และ Polynomial Regression มีการประเมินหรือวัดด้วย Mean Absolute Error--MAE Mean Square Error--MSE Root Mean Square Error --RMSE และCoefficient of Determination บทที่ 15 Decision Tree หรือเรียกว่า ต้นไม้ตัดสินใจ จัดอยู่ในกลุ่ม Supervised Learning สร้าง Model ใช้กับการคัดแยกหรือจำแนก (classification) วิธีการประเมินเบื้องต้นจะแบ่งข้อมูลออกเป็น 2 ชุด คือ Training set และ Test set จะเห็นได้ว่าทั้งบทที่ 14 และ 15 เมื่อสร้าง Model แล้วก็จะมีการประเมินตามมา สำหรับในบทที่ 16 จึงเป็นการอธิบายการประเมิน Model เสริม เพื่อหาประสิทธิภาพความแม่นยำ ระยะคลาดเคลื่อน นอกจากนี้ในบทที่ 17 กล่าวถึงอัลกอริทึม (Algorithm) ที่ใช้ได้ทั้งการจำแนก Regression หรือจะคล้ายกับ Decision Tree เช่น Nai.ve Baye SVM kNN และ Logistic Regression โดยการจำแนกหรือแยกแยะ (classification) ต้องใช้ค่าหนึ่งค่าใดเป็นเป็นเกณฑ์ตัดสิน เรียกว่า Threshold เป็นตัวกำหนดระบบทำงานว่าควรจะแข็งหรืออ่อน ทำโดยการใช้กราฟ ROC curve และAUC (ค่าพื้นที่ใต้เส้น ROC curve) เป็นตัวชี้วัดประสิทธิภาพของ Model อธิบายเกณฑ์ไว้ในบทที่ 18

9. บทที่ 19-22 การได้มาของข้อมูลจะมาจากภาพนิ่ง หรือภาพเคลื่อนไหว โดยในบทที่ 19 จะกล่าวถึงระบบรู้จำเลขลายมือเขียน และบทที่ 20 จะกล่าวถึงการตรวจจับและรู้จำใบหน้า ทั้ง 2 บทต้องมีการแปลงข้อมูลภาพให้เป็นค่าตารางเล็ก ๆ เรียงต่อกันที่เรียกว่า Pixel มีการจัดเตรียมข้อมูลสำหรับ Train และ Test ใช้อัลกอริทึมแบบจำแนก (classification) และทำนาย จุดแตกต่างสำหรับเลขลายมือเขียน คือ แปลงข้อมูล Features ประกอบด้วยค่าความสว่าง โหมดเป็นโทนเทา ขาว หรือดำ และภาพนำเข้าเพียง 2 มิติ เท่านั้นก็พอ แต่การตรวจจับและรู้จำใบหน้านั้น ต้องประกอบด้วยชั้นข้อมูล RGB เพื่อแยกสี และภาพที่นำเข้าต้องมากกว่า 3 มิติ จึงจะสามารถทำนายได้ หากบางครั้ง Features มีจำนวนมาก หรือขนาดใหญ่ เกินความจำเป็น ทำให้ใช้ทรัพยากรเครื่องสูง ใช้เวลา Train นาน อาจเกิดปัญหา Over-fitting ก็ต้องลดมิติข้อมูลลง แต่ถ้าลดมากเกินไปก็จะทำให้ประสิทธิภาพของ Model แย่ ดังนั้นการปรับจูน Model ที่เหมาะสมต้องมีกฎว่าเมื่อกระทำไปแล้วผลลัพธ์ต้องไม่เกิด Trial and error ซึ่งในบทที่ 21 และ 22 ได้มีการอธิบายถึงลดมิติข้อมูล และการปรับจูน Model เป็น Workshop ไว้โดยละเอียด

10. บทที่ 23-24 การจำแนกหมวดหมู่ข้อความและ Recommender System 2 บทที่นี้เป็นเนื้อหาที่นำผลลัพธ์จากกระบวนการคิดวิเคราะห์ของ Data Science และ Machine Learning มาใช้งาน โดยที่การจำแนกหมวดหมู่ข้อความนั้นเมื่อผ่านการคำนวณหาค่าตัวแทนคุณลักษณะเด่น (features) เช่น การแยกข้อความที่เข้ามาในเมล์ จำแนกไปเป็นเมล์ดี หรือเมล์ขยะ หรือจะเป็น Recommender System เป็นระบบที่แนะนำสินค้าและบริการ โดยตรวจสอบจากพฤติกรรมบุคคลนั้น ๆ ระบบก็จะนำคุณสมบัติที่รวบรวมไว้ไปคำนวณหาดัชนีความคล้ายและแนะนำสินค้าใกล้เคียงให้กับลูกค้าได้

สรุป

หนังสือเล่มนี้เหมาะสำหรับผู้ที่ต้องการเน้นการฝึกปฏิบัติ (workshop) ที่ไม่ต้องการเนื้อหาบรรยายมาก เพราะเป็นการสรุปเฉพาะประเด็นสำคัญที่จำเป็นในการทำงานจริงเท่านั้น และเพื่อให้เกิดประโยชน์กับผู้อ่านสูงสุด ท่านควรติดตั้งโปรแกรม Python ในเครื่องคอมพิวเตอร์ของท่านก่อน พร้อมทั้งศึกษาการเขียนโปรแกรม Python พื้นฐาน จากนั้นค่อย ๆ ทดลองทำไปเป็นขั้นตอนตามหนังสือ ซึ่งท้ายที่สุดท่านจะเกิดทักษะในการประยุกต์ในการทำงานวิจัยของท่านต่อไปได้



ภาพ 1 Machine Learning

Reference

Saraubon, K. (2020). *Learn data science AI : Machine learning with Python*. Bangkok: Horsom Media Network (in Thai)