



สารบัญ

บทนำ	15
อิทธิพลของข้อมูล	15
วิทยาศาสตร์ข้อมูล (Data Science) คืออะไร?	15
สมมติฐานสร้างแรงบันดาลใจ: DataSciencecenter	16
ค้นหาตัวเชื่อมต่อที่สำคัญ	16
นักวิทยาศาสตร์ข้อมูลที่คุณอาจรู้จัก	19
รายได้และประสบการณ์	22
ประเมินการชำระหนี้	24
หัวข้อที่น่าสนใจ	25
ก้าวต่อไป	26
ภาษา Python ฉบับรวบรัด	29
หลักการของ Python	29
เริ่มใช้งาน Python	30
สภาพแวดล้อมเสมือนจริง	30
การจัดรูปแบบช่องว่าง	31
โมดูล	32
ฟังก์ชัน	32
ข้อความ	33
Exception	34
ลิสต์ (List)	34
Tuple	36
Dictionary	36
defaultdict	37
Counter	38
Set	39
ควบคุมการทำงาน	39
ค่าความจริง	40
การเรียงลำดับ	41
การแปลงลิสต์	42
การทดสอบและยืนยัน	42
การเขียนโปรแกรมเชิงวัตถุ (OOP)	43
การทำซ้ำและตัวสร้างค่า	44
การสุ่ม	46
Regular Expression	47
การเขียนโปรแกรมเชิงฟังก์ชัน	47



zip และการแยก Argument	47
args และ kwargs	48
การระบุชนิด	49
วิธีระบุชนิด	51
ยินดีต้อนรับสู่สังคม DataSciencester!	53
สำหรับการค้นคว้าเพิ่มเติม	53
แสดงข้อมูลด้วยภาพ	55
matplotlib	55
แผนภูมิแท่ง	57
แผนภูมิเส้น	60
แผนภูมิกระจาย	61
สำหรับการค้นคว้าเพิ่มเติม	63
พีชคณิตเชิงเส้น	65
เวกเตอร์	65
เมทริกซ์	70
สำหรับการค้นคว้าเพิ่มเติม	73
สถิติ	75
การอธิบายถึงชุดข้อมูล	75
แนวโน้มสู่ส่วนกลาง	77
การกระจายของข้อมูล	78
ค่าสหสัมพันธ์ (Correlation)	80
Paradox ของซิมป์สัน	82
เงื่อนไขอื่นๆ ของความสัมพันธ์	83
ความสัมพันธ์และสาเหตุ	84
สำหรับการค้นคว้าเพิ่มเติม	85
ความน่าจะเป็น	87
การขึ้นต่อกันและความเป็นอิสระ	87
ความน่าจะเป็นที่มีเงื่อนไข	88
ทฤษฎีบทของเบย์	89
ตัวแปรสุ่ม	90
การกระจายแบบต่อเนื่อง	90
การกระจายแบบปกติ	91
ทฤษฎีแนวโน้มเข้าสู่ศูนย์กลาง	94
สำหรับการค้นคว้าเพิ่มเติม	96



สมมติฐานและการอนุมาน	99
การทดสอบสมมติฐานทางสถิติ	99
ตัวอย่าง: การโยนเหรียญ	99
p-Value	102
ช่วงของความเชื่อมั่น	104
p-Hacking	104
ตัวอย่าง: ทำการทดสอบ A/B	105
การอนุมานแบบเบย์	106
สำหรับการค้นคว้าเพิ่มเติม	109
การเคลื่อนลงตามความชัน	111
แนวคิดของการไล่ระดับลง	111
การประมาณระดับความชัน	112
การใช้งาน	115
การเลือกระยะที่ถูกต้อง	116
ใช้ Gradient Descent เพื่อหาแบบจำลองที่ดี	116
การคำนวณแบบ Minibatch และ Stochastic	117
สำหรับการค้นคว้าเพิ่มเติม	119
การได้มาซึ่งข้อมูล	121
stdin และ stdout	121
การอ่านไฟล์	123
พื้นฐานของไฟล์ข้อความ	123
ไฟล์แบบมีตัวค้น	125
การเก็บข้อมูลจากเว็บ	126
HTML และการแยกวิเคราะห์	126
ตัวอย่าง: การเก็บข้อมูลของสภาองเกรส	128
การใช้ API	131
JSON และ XML	131
การใช้ API ที่ไม่มีการตรวจสอบผู้ใช้	132
การค้นหา API	133
ตัวอย่าง: การใช้ Twitter API	133
การลงทะเบียนกับ Twitter	133
สำหรับการค้นคว้าเพิ่มเติม	137
การทำงานกับข้อมูล	139
สำรวจข้อมูลของคุณ	139
สำรวจข้อมูลมิติเดียว	139
ข้อมูลสองมิติ	141



ข้อมูลหลายมิติ	143
การใช้ NamedTuple	145
Dataclass	146
ทำข้อมูลให้สะอาด	147
การจัดการข้อมูล	149
การปรับมาตราส่วนข้อมูล	152
แสดงความคืบหน้าด้วย tqdm	154
การลดมิติของข้อมูล	155
สำหรับการค้นคว้าเพิ่มเติม	161
Machine Learning	163
การสร้างแบบจำลอง	163
แมชีนเลิร์นนิงคืออะไร?	164
Overfitting และ Underfitting	164
ความถูกต้อง	167
จุดลงตัวของฟังก์ชันการสูญเสียและความแปรปรวน	169
การดึงและคัดเลือกคุณลักษณะ	170
สำหรับการค้นคว้าเพิ่มเติม	171
k-Nearest Neighbors	173
แบบจำลอง	173
ตัวอย่าง: ชุดข้อมูล Iris	175
คำสาปของมิติข้อมูล	178
สำหรับการค้นคว้าเพิ่มเติม	182
Naive Bayes	185
ตัวกรองสแปมแนสทีม	185
ตัวกรองสแปมที่ซับซ้อนขึ้น	186
สร้าง Classifier	187
ทดสอบแบบจำลอง	189
เริ่มใช้งานแบบจำลอง	190
สำหรับการค้นคว้าเพิ่มเติม	193
การถดถอยเชิงเส้นอย่างง่าย	195
แบบจำลอง	195
การเคลื่อนลงตามความชัน (Gradient Descent)	198
Maximum Likelihood Estimation	199
สำหรับการค้นคว้าเพิ่มเติม	200



การถดถอยพหุคูณ	203
แบบจำลอง	203
สมมติฐานเพิ่มเติมของแบบจำลองค่ายกกำลังสองที่น้อยที่สุด	204
ปรับแบบจำลองให้เหมาะสม	204
การแปลความแบบจำลอง	206
ความเหมาะสม	207
ขออนอกเรื่อง: Bootstrap	207
ค่าผิดพลาดมาตรฐานของค่าสัมประสิทธิ์การถดถอย	209
การทำให้โมเดลซับซ้อนน้อยลง	210
สำหรับการค้นคว้าเพิ่มเติม	213
การถดถอยโลจิสติก	215
ปัญหา	215
ฟังก์ชันโลจิสติก	217
การนำแบบจำลองไปใช้	220
ความเหมาะสม	221
Support Vector Machine	222
สำหรับการค้นคว้าเพิ่มเติม	225
Decision Tree	227
ต้นไม้สำหรับการตัดสินใจคืออะไร?	227
การวัดค่าความไม่แน่นอน	228
เอนโทรปีของพาร์ทิชัน	230
การสร้างดีชันทรี	231
รวมทั้งหมดเข้าด้วยกัน	234
โมเดล Random Forest	237
สำหรับการค้นคว้าเพิ่มเติม	238
โครงข่ายประสาทเทียม	241
Perceptron	241
โครงข่ายประสาทแบบ Feed-Forward	243
Backpropagation	246
ตัวอย่าง: Fizz Buzz	248
สำหรับการค้นคว้าเพิ่มเติม	251
Deep Learning	253
Tensor	253
โครงสร้างแบบเลเยอร์	255
เลเยอร์เชิงเส้น	257



โครงข่ายประสาทเทียมคืออนุกรมของเลเยอร์	260
การสูญเสีย (Loss) และการเพิ่มประสิทธิภาพ (Optimization)	261
ตัวอย่าง: XOR อีกครั้ง	264
ฟังก์ชันอื่นในการประมวลผล	265
ตัวอย่าง: FizzBuzz อีกครั้ง	267
Softmax และ Cross-Entropy	268
การละทิ้ง (Dropout)	271
ตัวอย่าง: MNIST	272
การเซฟและโหลดแบบจำลอง	276
สำหรับการค้นคว้าเพิ่มเติม	277
การจัดกลุ่ม	279
แนวคิด	279
แบบจำลอง	279
ตัวอย่าง: นัดจัดเลี้ยง	282
การเลือกค่า k	284
ตัวอย่าง: การจัดกลุ่มสี	285
การจัดกลุ่มลำดับชั้นจากล่างขึ้นบน	287
สำหรับการค้นคว้าเพิ่มเติม	293
การประมวลผลภาษาธรรมชาติ	295
กลุ่มคำ	295
โมเดลภาษาแบบ n-Gram	297
ไวยากรณ์	300
การสุ่มตัวอย่างแบบกิบส์	302
การสร้างแบบจำลองของหัวข้อ	303
เวกเตอร์คำ	309
โครงข่ายประสาทที่เกิดขึ้นซ้อนกัน	318
ตัวอย่าง: การใช้ RNN ในระดับตัวอักษร	320
สำหรับการค้นคว้าเพิ่มเติม	324
การวิเคราะห์เครือข่าย	327
ความเป็นศูนย์กลางระหว่างโหนด	327
หาค่าความเป็นศูนย์กลางด้วยเวกเตอร์ลักษณะเฉพาะ	332
การคูณเมทริกซ์	332
ค่าความเป็นศูนย์กลาง	333
กราฟระบุทิศทางและ PageRank	335
สำหรับการค้นคว้าเพิ่มเติม	337



ระบบให้คำแนะนำ	339
ถ้าจะทำได้ด้วยตัวเอง	340
การแนะนำสิ่งที่ได้รับความนิยม	340
การกรองโดยพิจารณาผู้ใช้	341
การกรองโดยพิจารณาหัวข้อ	344
การแยกตัวประกอบเมทริกซ์	346
สำหรับการค้นคว้าเพิ่มเติม	351
ฐานข้อมูล และ SQL	353
การสร้างตารางและเพิ่มข้อมูล	353
UPDATE	356
DELETE	357
SELECT	358
GROUP BY	360
ORDER BY	363
JOIN	364
การสอบถามย่อย (Subquery)	366
อินเด็กซ์	367
การเพิ่มประสิทธิภาพการสอบถาม	367
NoSQL	368
สำหรับการค้นคว้าเพิ่มเติม	369
MapReduce	371
ตัวอย่าง: นับจำนวนคำ	371
ทำไมต้องใช้ MapReduce	373
MapReduce ที่มีความทั่วไปมากขึ้น	373
ตัวอย่าง: วิเคราะห์การอัปเดตสเตตัส	375
ตัวอย่าง: การคูณเมทริกซ์	377
ตัวรวบรวม	379
สำหรับการค้นคว้าเพิ่มเติม	379
จริยธรรมด้านข้อมูล	381
จริยธรรมข้อมูลคืออะไร?	381
จริงๆ แล้วจริยธรรมข้อมูลคืออะไร?	381
เราควรใส่ใจเรื่องจริยธรรมในการใช้ข้อมูลหรือไม่?	382
การสร้างผลิตภัณฑ์ที่ไม่ดี	382
สมดุลของความแม่นยำและความเป็นส่วนตัว	382
การร่วมมือ	384
ความสามารถในการอธิบายการทำงาน	384



คำแนะนำ	384
ข้อมูลที่เอนเอียง	385
การปกป้องข้อมูล	385
สรุป	386
สำหรับการค้นคว้าเพิ่มเติม	386

ก้าวไปกับงานวิทยาศาสตร์ข้อมูล	389
IPython	389
คณิตศาสตร์	389
ไม่ได้มาแต่เริ่มแรก	389
NumPy	390
pandas	390
scikit-learn	390
การแสดงผลด้วยภาพ	390
R	390
Deep Learning	391
ค้นหาข้อมูล	391
โครงการที่สนใจ	391
Hacker News	391
รถดับเพลิง	391
เสื้อเชิ้ต	392
ทวีตบนโลก	392
แล้วคุณล่ะ?	392