Python Machine Learning Third Edition by Sebastian Raschka, Vahid Mirjalili
Copyright © Packt Publishing 2019.
First published in the English language under the title 'Python Machine Learning -Third Edition- (9781789955750)'
Korean Translation Copyright © 2021 by Gilbut Publishing co., Ltd.
The Korean edition was published by arrangement with Packt Publishing Ltd. through Agency-One, Seoul.

이 책의 한국어판 저작권은 에이전시 원을 통해 저작권자와의 독점 계약으로 길벗에 있습니다.
저작권법에 의해 한국 내에서 보호를 받는 저작물이므로 무단전재와 무단복제를 금합니다.

머신 러닝 교과서 with 파이썬, 사이킷런, 텐서플로 개정 3판

Python Machine Learning Third Edition

초판 발행 • 2021년 3월 31일

지은이 • 세바스찬 라시카, 바히드 미자리리
옮긴이 • 박해선
발행인 • 이종원
발행처 • (주)도서출판 길벗
출판사 등록일 • 1990년 12월 24일
주소 • 서울시 마포구 월드컵로 10길 56(서교동)
대표 전화 • 02)332-0931 | **팩스** • 02)323-0586
홈페이지 • www.gilbut.co.kr | **이메일** • gilbut@gilbut.co.kr

기획 및 책임편집 • 안윤경(yk78@gilbut.co.kr) | **디자인** • 박상희 | **제작** • 이준호, 손일순, 이진혁
영업마케팅 • 임태호, 전선하, 지운집, 박성용, 차명환 | **영업관리** • 김명자 | **독자지원** • 송혜란, 윤정아

교정교열 • 김윤지 | **전산편집** • 박진희 | **출력 · 인쇄** • 예림인쇄 | **제본** • 예림바인딩

▸ 잘못 만든 책은 구입한 서점에서 바꿔 드립니다.
▸ 이 책은 저작권법에 따라 보호받는 저작물이므로 무단전재와 무단복제를 금합니다. 이 책의 전부 또는 일부를 이용하려면 반드시 사전에 저작권자와 ㈜도서출판 길벗의 서면 동의를 받아야 합니다.

ISBN 979-11-6521-518-7 93000
(길벗 도서번호 080223)

정가 44,000원

독자의 1초를 아껴주는 정성 길벗출판사

길벗 | 길벗 IT실용서, IT/일반 수험서, IT전문서, 경제실용서, 취미실용서, 건강실용서, 자녀교육서
더퀘스트 | 인문교양서, 비즈니스서
길벗이지톡 | 어학단행본, 어학수험서
길벗스쿨 | 국어학습서, 수학학습서, 유아학습서, 어학학습서, 어린이교양서, 교과서

페이스북 • www.facebook.com/gbitbook
예제소스 • https://github.com/gilbutITbook/080223

1장 컴퓨터는 데이터에서 배운다 ····· 031

1.1 데이터를 지식으로 바꾸는 지능적인 시스템 구축 032

1.2 머신 러닝의 세 가지 종류 033

1.2.1 지도 학습으로 미래 예측 034

1.2.2 강화 학습으로 반응형 문제 해결 037

1.2.3 비지도 학습으로 숨겨진 구조 발견 038

1.3 기본 용어와 표기법 소개 040

1.3.1 이 책에서 사용하는 표기법과 규칙 041

1.3.2 머신 러닝 용어 043

1.4 머신 러닝 시스템 구축 로드맵 043

1.4.1 전처리: 데이터 형태 갖추기 044

1.4.2 예측 모델 훈련과 선택 045

1.4.3 모델을 평가하고 본 적 없는 샘플로 예측 046

1.5 머신 러닝을 위한 파이썬 046

1.5.1 파이썬과 PIP에서 패키지 설치 047

1.5.2 아나콘다 파이썬 배포판과 패키지 관리자 사용 048

1.5.3 과학 컴퓨팅, 데이터 과학, 머신 러닝을 위한 패키지 048

1.6 요약 049

2장 간단한 분류 알고리즘 훈련 ····· 051

2.1 인공 뉴런: 초기 머신 러닝의 간단한 역사 052

2.1.1 인공 뉴런의 수학적 정의 053

2.1.2 퍼셉트론 학습 규칙 055

2.2 파이썬으로 퍼셉트론 학습 알고리즘 구현 058

2.2.1 객체 지향 퍼셉트론 API 059

2.2.2 붓꽃 데이터셋에서 퍼셉트론 훈련 062

2.3 적응형 선형 뉴런과 학습의 수렴 068

2.3.1 경사 하강법으로 비용 함수 최소화 070

2.3.2 파이썬으로 아달린 구현 072

2.3.3 특성 스케일을 조정하여 경사 하강법 결과 향상 076
2.3.4 대규모 머신 러닝과 확률적 경사 하강법 078

2.4 요약 083

3장 사이킷런을 타고 떠나는 머신 러닝 분류 모델 투어 ····· 085

3.1 분류 알고리즘 선택 086

3.2 사이킷런 첫걸음: 퍼셉트론 훈련 087

3.3 로지스틱 회귀를 사용한 클래스 확률 모델링 093
3.3.1 로지스틱 회귀의 이해와 조건부 확률 094
3.3.2 다중 클래스를 위한 로지스틱 회귀 094
3.3.3 로지스틱 비용 함수의 가중치 학습 097
3.3.4 아달린 구현을 로지스틱 회귀 알고리즘으로 변경 100
3.3.5 사이킷런을 사용하여 로지스틱 회귀 모델 훈련 104
3.3.6 규제를 사용하여 과대적합 피하기 106

3.4 서포트 벡터 머신을 사용한 최대 마진 분류 110
3.4.1 최대 마진 110
3.4.2 슬랙 변수를 사용하여 비선형 분류 문제 다루기 112
3.4.3 사이킷런의 다른 구현 114

3.5 커널 SVM을 사용하여 비선형 문제 풀기 115
3.5.1 선형적으로 구분되지 않는 데이터를 위한 커널 방법 115
3.5.2 커널 기법을 사용하여 고차원 공간에서 분할 초평면 찾기 117

3.6 결정 트리 학습 121
3.6.1 정보 이득 최대화: 자원을 최대로 활용 122
3.6.2 결정 트리 만들기 127
3.6.3 랜덤 포레스트로 여러 개의 결정 트리 연결 133

3.7 k-최근접 이웃: 게으른 학습 알고리즘 136

3.8 요약 140

4장 좋은 훈련 데이터셋 만들기: 데이터 전처리 ····· 143

4.1 누락된 데이터 다루기 144
4.1.1 테이블 형태 데이터에서 누락된 값 식별 144
4.1.2 누락된 값이 있는 훈련 샘플이나 특성 제외 146
4.1.3 누락된 값 대체 147
4.1.4 사이킷런 추정기 API 익히기 151

4.2 범주형 데이터 다루기 153
4.2.1 순서가 있는 특성 매핑 154
4.2.2 클래스 레이블 인코딩 155
4.2.3 순서가 없는 특성에 원-핫 인코딩 적용 157

4.3 데이터셋을 훈련 데이터셋과 테스트 데이터셋으로 나누기 162

4.4 특성 스케일 맞추기 164

4.5 유용한 특성 선택 169
4.5.1 모델 복잡도 제한을 위한 L1 규제와 L2 규제 170
4.5.2 L2 규제의 기하학적 해석 171
4.5.3 L1 규제를 사용한 희소성 172
4.5.4 순차 특성 선택 알고리즘 176

4.6 랜덤 포레스트의 특성 중요도 사용 184

4.7 요약 188

5장 차원 축소를 사용한 데이터 압축 ····· 189

5.1 주성분 분석을 통한 비지도 차원 축소 190
5.1.1 주성분 분석의 주요 단계 191
5.1.2 주성분 추출 단계 192
5.1.3 총 분산과 설명된 분산 196
5.1.4 특성 변환 197
5.1.5 사이킷런의 주성분 분석 200

5.2 선형 판별 분석을 통한 지도 방식의 데이터 압축 205
5.2.1 주성분 분석 vs 선형 판별 분석 205
5.2.2 선형 판별 분석의 내부 동작 방식 207

5.2.3 산포 행렬 계산 207
5.2.4 새로운 특성 부분 공간을 위해 선형 판별 벡터 선택 210
5.2.5 새로운 특성 공간으로 샘플 투영 213
5.2.6 사이킷런의 LDA 214

5.3 커널 PCA를 사용하여 비선형 매핑 217
5.3.1 커널 함수와 커널 트릭 218
5.3.2 파이썬으로 커널 PCA 구현 223
5.3.3 새로운 데이터 포인트 투영 230
5.3.4 사이킷런의 커널 PCA 234

5.4 요약 239

6장 모델 평가와 하이퍼파라미터 튜닝의 모범 사례 ····· 241

6.1 파이프라인을 사용한 효율적인 워크플로 242
6.1.1 위스콘신 유방암 데이터셋 242
6.1.2 파이프라인으로 변환기와 추정기 연결 244

6.2 k-겹 교차 검증을 사용한 모델 성능 평가 247
6.2.1 홀드아웃 방법 247
6.2.2 k-겹 교차 검증 248

6.3 학습 곡선과 검증 곡선을 사용한 알고리즘 디버깅 254
6.3.1 학습 곡선으로 편향과 분산 문제 분석 254
6.3.2 검증 곡선으로 과대적합과 과소적합 조사 258

6.4 그리드 서치를 사용한 머신 러닝 모델 세부 튜닝 260
6.4.1 그리드 서치를 사용한 하이퍼파라미터 튜닝 260
6.4.2 중첩 교차 검증을 사용한 알고리즘 선택 267

6.5 여러 가지 성능 평가 지표 269
6.5.1 오차 행렬 269
6.5.2 분류 모델의 정밀도와 재현율 최적화 272
6.5.3 ROC 곡선 그리기 274
6.5.4 다중 분류의 성능 지표 280

6.6 불균형한 클래스 다루기 281

6.7 요약 284

7장 다양한 모델을 결합한 앙상블 학습 ····· 287

7.1 앙상블 학습 288

7.2 다수결 투표를 사용한 분류 앙상블 292

7.2.1 간단한 다수결 투표 분류기 구현 292

7.2.2 다수결 투표 방식을 사용하여 예측 만들기 299

7.2.3 앙상블 분류기의 평가와 튜닝 303

7.3 배깅: 부트스트랩 샘플링을 통한 분류 앙상블 310

7.3.1 배깅 알고리즘의 작동 방식 311

7.3.2 배깅으로 Wine 데이터셋의 샘플 분류 312

7.4 약한 학습기를 이용한 에이다부스트 317

7.4.1 부스팅 작동 원리 318

7.4.2 사이킷런에서 에이다부스트 사용 323

7.5 요약 330

8장 감성 분석에 머신 러닝 적용 ····· 331

8.1 텍스트 처리용 IMDb 영화 리뷰 데이터 준비 332

8.1.1 영화 리뷰 데이터셋 구하기 333

8.1.2 영화 리뷰 데이터셋을 더 간편한 형태로 전처리 333

8.2 BoW 모델 소개 336

8.2.1 단어를 특성 벡터로 변환 336

8.2.2 tf-idf를 사용하여 단어 적합성 평가 338

8.2.3 텍스트 데이터 정제 341

8.2.4 문서를 토큰으로 나누기 343

8.3 문서 분류를 위한 로지스틱 회귀 모델 훈련 345

8.4 대용량 데이터 처리: 온라인 알고리즘과 외부 메모리 학습 353

8.5 잠재 디리클레 할당을 사용한 토픽 모델링 357

8.5.1 LDA를 사용한 텍스트 문서 분해 358

8.5.2 사이킷런의 LDA 358

8.6 요약 362

9장 웹 애플리케이션에 머신 러닝 모델 내장 ····· 365

9.1 학습된 사이킷런 추정기 저장 366

9.2 데이터를 저장하기 위해 SQLite 데이터베이스 설정 370

9.3 플라스크 웹 애플리케이션 개발 372

9.3.1 첫 번째 플라스크 애플리케이션 373

9.3.2 폼 검증과 화면 출력 376

9.4 영화 리뷰 분류기를 웹 애플리케이션으로 만들기 384

9.4.1 파일과 폴더: 디렉터리 구조 살펴보기 385

9.4.2 메인 애플리케이션 app.py 구현 386

9.4.3 리뷰 폼 구성 389

9.4.4 결과 페이지 템플릿 만들기 390

9.5 공개 서버에 웹 애플리케이션 배포 393

9.5.1 PythonAnywhere 계정 만들기 393

9.5.2 영화 분류 애플리케이션 업로드 394

9.5.3 영화 분류기 업데이트 395

9.6 요약 399

10장 회귀 분석으로 연속적 타깃 변수 예측 ····· 401

10.1 선형 회귀 402

10.1.1 단순 선형 회귀 403

10.1.2 다중 선형 회귀 404

10.2 주택 데이터셋 탐색 405

10.2.1 데이터프레임으로 주택 데이터셋 읽기 405

10.2.2 데이터셋의 중요 특징 시각화 407

10.2.3 상관관계 행렬을 사용한 분석 409

10.3 최소 제곱 선형 회귀 모델 구현 411

10.3.1 경사 하강법으로 회귀 모델의 파라미터 구하기 412

10.3.2 사이킷런으로 회귀 모델의 가중치 추정 416

10.4 RANSAC을 사용하여 안정된 회귀 모델 훈련 419

10.5 선형 회귀 모델의 성능 평가 421

10.6 회귀에 규제 적용 425

10.7 선형 회귀 모델을 다항 회귀로 변환 427

10.7.1 사이킷런을 사용하여 다항식 항 추가 427

10.7.2 주택 데이터셋을 사용한 비선형 관계 모델링 429

10.8 랜덤 포레스트를 사용하여 비선형 관계 다루기 433

10.8.1 결정 트리 회귀 433

10.8.2 랜덤 포레스트 회귀 435

10.9 요약 438

11장 레이블되지 않은 데이터 다루기: 군집 분석 ····· 441

11.1 k-평균 알고리즘을 사용하여 유사한 객체 그룹핑 442

11.1.1 사이킷런을 사용한 k-평균 군집 443

11.1.2 k-평균++로 초기 클러스터 센트로이드를 똑똑하게 할당 447

11.1.3 직접 군집 vs 간접 군집 449

11.1.4 엘보우 방법을 사용하여 최적의 클러스터 개수 찾기 451

11.1.5 실루엣 그래프로 군집 품질을 정량화 452

11.2 계층적인 트리로 클러스터 조직화 457

11.2.1 상향식으로 클러스터 묶기 458

11.2.2 거리 행렬에서 계층 군집 수행 459

11.2.3 히트맵에 덴드로그램 연결 463

11.2.4 사이킷런에서 병합 군집 적용 465

11.3 DBSCAN을 사용하여 밀집도가 높은 지역 찾기 466

11.4 요약 472

12장 다층 인공 신경망을 밑바닥부터 구현 ····· 473

12.1 인공 신경망으로 복잡한 함수 모델링 474

12.1.1 단일층 신경망 요약 476

12.1.2 다층 신경망 구조 478

12.1.3 정방향 계산으로 신경망 활성화 출력 계산 480

12.2 손글씨 숫자 분류 483

12.2.1 MNIST 데이터셋 구하기 484

12.2.2 다층 퍼셉트론 구현 491

12.3 인공 신경망 훈련 503

12.3.1 로지스틱 비용 함수 계산 503

12.3.2 역전파 알고리즘 이해 506

12.3.3 역전파 알고리즘으로 신경망 훈련 508

12.4 신경망의 수렴 511

12.5 신경망 구현에 관한 몇 가지 첨언 512

12.6 요약 513

13장 텐서플로를 사용한 신경망 훈련 ····· 515

13.1 텐서플로와 훈련 성능 516

13.1.1 성능 문제 516

13.1.2 텐서플로란? 518

13.1.3 텐서플로 학습 순서 520

13.2 텐서플로 처음 시작하기 520

13.2.1 텐서플로 설치 520

13.2.2 텐서플로에서 텐서 만들기 522

13.2.3 텐서의 데이터 타입과 크기 조작 523

13.2.4 텐서에 수학 연산 적용 524

13.2.5 split( ), stack( ), concat( ) 함수 527

13.3 텐서플로 데이터셋 API(tf.data)를 사용하여 입력 파이프라인 구축 528

13.3.1 텐서에서 텐서플로 데이터셋 만들기 529

13.3.2 두 개의 텐서를 하나의 데이터셋으로 연결 530

13.3.3 shuffle( ), batch( ), repeat( ) 메서드 532
13.3.4 로컬 디스크에 있는 파일에서 데이터셋 만들기 535
13.3.5 tensorflow_datasets 라이브러리에서 데이터셋 로드 538

13.4 텐서플로로 신경망 모델 만들기 544
13.4.1 텐서플로 케라스 API(tf.keras) 545
13.4.2 선형 회귀 모델 만들기 545
13.4.3 .compile( )과 .fit( ) 메서드를 사용한 모델 훈련 550
13.4.4 붓꽃 데이터셋을 분류하는 다층 퍼셉트론 만들기 551
13.4.5 테스트 데이터셋에서 훈련된 모델 평가 555
13.4.6 훈련된 모델 저장하고 복원 556

13.5 다층 신경망의 활성화 함수 선택 560
13.5.1 로지스틱 함수 요약 560
13.5.2 소프트맥스 함수를 사용한 다중 클래스 확률 예측 562
13.5.3 하이퍼볼릭 탄젠트로 출력 범위 넓히기 565
13.5.4 렐루 활성화 함수 567

13.6 요약 569

14장 텐서플로 구조 자세히 알아보기 ····· 571

14.1 텐서플로의 주요 특징 573

14.2 텐서플로의 계산 그래프: 텐서플로 v2로 이전 574
14.2.1 계산 그래프 이해 574
14.2.2 텐서플로 v1.x에서 그래프 만들기 575
14.2.3 텐서플로 v2로 이전 576
14.2.4 입력 데이터를 모델에 주입: 텐서플로 v1.x 스타일 577
14.2.5 입력 데이터를 모델에 주입: 텐서플로 v2 스타일 577
14.2.6 함수 데코레이터로 계산 성능 높이기 578

14.3 모델 파라미터를 저장하고 업데이트하기 위한 텐서플로 변수 객체 581

14.4 자동 미분과 GradientTape로 그레이디언트 계산 584
14.4.1 훈련 가능한 변수에 대한 손실의 그레이디언트 계산 585
14.4.2 훈련하지 않는 변수에 대한 그레이디언트 계산 586
14.4.3 여러 개의 그레이디언트 계산 587

14.5 케라스 API를 사용하여 일반적인 구조 구현하기 588
14.5.1 XOR 분류 문제 풀어 보기 592
14.5.2 케라스 함수형 API로 유연성이 높은 모델 만들기 597
14.5.3 케라스의 Model 클래스 기반으로 모델 만들기 598
14.5.4 사용자 정의 케라스 층 만들기 601

14.6 텐서플로 추정기 604
14.6.1 특성 열 사용 605
14.6.2 사전에 준비된 추정기로 머신 러닝 수행 609
14.6.3 추정기를 사용하여 MNIST 손글씨 숫자 분류 613
14.6.4 케라스 모델에서 추정기 만들기 616

14.7 요약 618

15장 심층 합성곱 신경망으로 이미지 분류 ····· 619

15.1 합성곱 신경망의 구성 요소 620
15.1.1 CNN과 특성 계층 학습 621
15.1.2 이산 합성곱 수행 623
15.1.3 서브샘플링 632

15.2 기본 구성 요소를 사용하여 심층 합성곱 신경망 구성 633
15.2.1 여러 개의 입력 또는 컬러 채널 다루기 634
15.2.2 드롭아웃으로 신경망 규제 637
15.2.3 분류를 위한 손실 함수 640

15.3 텐서플로를 사용하여 심층 합성곱 신경망 구현 643
15.3.1 다층 CNN 구조 643
15.3.2 데이터 적재와 전처리 644
15.3.3 텐서플로 케라스 API를 사용하여 CNN 구현 645

15.4 합성곱 신경망을 사용하여 얼굴 이미지의 성별 분류 651
15.4.1 CelebA 데이터셋 로드 652
15.4.2 이미지 변환과 데이터 증식 653
15.4.3 CNN 성별 분류기 훈련 658

15.5 요약 664

16장 순환 신경망으로 순차 데이터 모델링 ····· 667

16.1 순차 데이터 소개 668
16.1.1 순차 데이터 모델링: 순서를 고려한다 669
16.1.2 시퀀스 표현 669
16.1.3 시퀀스 모델링의 종류 670

16.2 시퀀스 모델링을 위한 RNN 672
16.2.1 RNN 반복 구조 이해 672
16.2.2 RNN의 활성화 출력 계산 674
16.2.3 은닉 순환과 출력 순환 677
16.2.4 긴 시퀀스 학습의 어려움 680
16.2.5 LSTM 셀 681

16.3 텐서플로로 시퀀스 모델링을 위한 RNN 구현 683
16.3.1 첫 번째 프로젝트: IMDb 영화 리뷰의 감성 분석 684
16.3.2 두 번째 프로젝트: 텐서플로로 글자 단위 언어 모델 구현 699

16.4 트랜스포머 모델을 사용한 언어 이해 711
16.4.1 셀프 어텐션 메커니즘 이해 711
16.4.2 멀티-헤드 어텐션과 트랜스포머 블록 714

16.5 요약 715

17장 새로운 데이터 합성을 위한 생성적 적대 신경망 ····· 717

17.1 생성적 적대 신경망 소개 718
17.1.1 오토인코더 719
17.1.2 새로운 데이터 합성을 위한 생성 모델 721
17.1.3 GAN으로 새로운 샘플 생성 723
17.1.4 GAN의 생성자와 판별자 손실 함수 이해 724

17.2 밑바닥부터 GAN 모델 구현 726
17.2.1 구글 코랩에서 GAN 모델 훈련 726
17.2.2 생성자와 판별자 신경망 구현 731
17.2.3 훈련 데이터셋 정의 734
17.2.4 GAN 모델 훈련하기 736

17.3 합성곱 GAN과 바서슈타인 GAN으로 합성 이미지 품질 높이기 744
17.3.1 전치 합성곱 745
17.3.2 배치 정규화 747
17.3.3 생성자와 판별자 구현 749
17.3.4 두 분포 사이의 거리 측정 755
17.3.5 GAN에 EM 거리 사용 759
17.3.6 그레이디언트 페널티 759
17.3.7 WGAN-GP로 DCGAN 모델 훈련 760
17.3.8 모드 붕괴 765

17.4 다른 GAN 애플리케이션 766

17.5 요약 767

18장 강화 학습으로 복잡한 환경에서 의사 결정 ····· 769

18.1 경험에서 배운다 770
18.1.1 강화 학습 이해 771
18.1.2 강화 학습 시스템의 에이전트-환경 인터페이스 정의 772

18.2 강화 학습의 기초 이론 774
18.2.1 마르코프 결정 과정 774
18.2.2 마르코프 결정 과정의 수학 공식 775
18.2.3 강화 학습 용어: 대가, 정책, 가치 함수 778
18.2.4 벨먼 방정식을 사용한 동적 계획법 782

18.3 강화 학습 알고리즘 783
18.3.1 동적 계획법 783
18.3.2 몬테카를로를 사용한 강화 학습 786
18.3.3 시간 차 학습 788

18.4 첫 번째 강화 학습 알고리즘 구현 791
18.4.1 OpenAI 짐 툴킷 소개 791
18.4.2 Q-러닝으로 그리드 월드 문제 풀기 801
18.4.3 심층 Q-러닝 805

18.5 전체 요약 813

부록 A 윈도에 아나콘다, 사이킷런, 텐서플로 설치 ····· 817

A.1 아나콘다 설치 819

A.2 사이킷런, 텐서플로 설치 825

A.3 예제 노트북 실행 827

A.4 주피터 노트북 뷰어와 구글 코랩 사용 829

찾아보기 836