

화학 데이터 활용 머신러닝을 위한 파이썬 프로그래밍

화학데이터기반연구센터

장 승 훈

스터디 그룹 과제

과제명 : (Study Group) 화학데이터 활용 머신러닝을 위한
파이썬 프로그래밍 Study Group

사업구분 : 자체 사업

사업기간 : 2022-03-01 ~ 2022-12-31

참여연구원 : 내부(12), 내부2/학생(8)

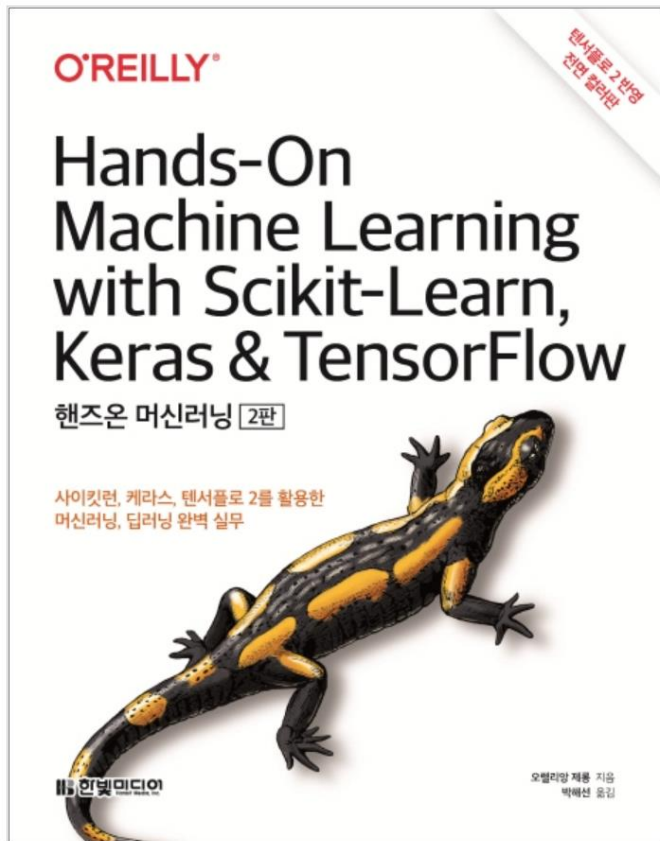
스터디 그룹 운영 목표

- 인공지능 알고리즘의 원리를 이해하고 실습.
- 파이썬 언어를 통한 자신의 연구데이터를 인공지능 알고리즘에 적용하고 결과를 분석.

스터디 시간 & 장소

- 스터디 시간
 - 매주 수요일 오후 3시 – 4시 반 (1시간 30분)
- 스터디 장소
 - 오프라인 : W6 연구동 3층 세미나실
 - 온라인 미팅 정보 : 스터디 시작 전, 메일 또는 카톡을 통해서 공유.

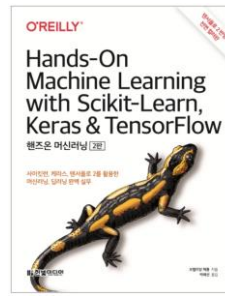
스터디 교재



Hands-On
Machine Learning
with Scikit-Learn,
Keras & TensorFlow
(한빛미디어, 오렐리앙 제롱 지음)

파이썬으로 작성된 오픈 소스 라이브러리
<https://scikit-learn.org/stable/>
<https://keras.io/>
<https://www.tensorflow.org/>

스터디 교재 목차



CHAPTER 1 한눈에 보는 머신러닝

- 1.1 머신러닝이란?
- 1.2 왜 머신러닝을 사용하는가?
- 1.3 애플리케이션 사례
- 1.4 머신러닝 시스템의 종류
- 1.5 머신러닝의 주요 도전 과제
- 1.6 테스트와 검증
- 1.7 연습문제

CHAPTER 2 머신러닝 프로젝트 처음부터 끝까지

- 2.1 실제 데이터로 작업하기
- 2.2 큰 그림 보기
- 2.3 데이터 가져오기
- 2.4 데이터 이해를 위한 탐색과 시각화
- 2.5 머신러닝 알고리즘을 위한 데이터 준비
- 2.6 모델 선택과 훈련
- 2.7 모델 세부 튜닝
- 2.8 론칭, 모니터링, 그리고 시스템 유지 보수
- 2.9 직접 해보세요!
- 2.10 연습문제

CHAPTER 3 분류

- 3.1 MNIST
- 3.2 이진 분류기 훈련
- 3.3 성능 측정
- 3.4 다중 분류
- 3.5 애러 분석
- 3.6 다중 레이블 분류
- 3.7 다중 출력 분류
- 3.8 연습문제

CHAPTER 4 모델 훈련

- 4.1 선형 회귀
- 4.2 경사 하강법
- 4.3 다항 회귀
- 4.4 학습 곡선
- 4.5 규제가 있는 선형 모델
- 4.6 로지스틱 회귀
- 4.7 연습문제

CHAPTER 5 서포트 벡터 머신

- 5.1 선형 SVM 분류
- 5.2 비선형 SVM 분류
- 5.3 SVM 회귀
- 5.4 SVM 이론
- 5.5 연습문제

CHAPTER 6 결정 트리

- 6.1 결정 트리 학습과 시각화
- 6.2 예측하기
- 6.3 클래스 확률 추정
- 6.4 CART 훈련 알고리즘
- 6.5 계산 복잡도
- 6.6 지니 불순도 또는 엔트로피?
- 6.7 규제 매개변수
- 6.8 회귀
- 6.9 불안정성
- 6.10 연습문제

CHAPTER 7 앙상블 학습과 랜덤 포레스트

- 7.1 투표 기반 분류기
- 7.2 배깅과 페이스팅
- 7.3 랜덤 패치와 랜덤 서브스페이스
- 7.4 랜덤 포레스트
- 7.5 부스팅
- 7.6 스택킹
- 7.7 연습문제

CHAPTER 8 차원 축소

- 8.1 차원의 저주
- 8.2 차원 축소를 위한 접근 방법
- 8.3 PCA
- 8.4 커널 PCA
- 8.5 LLE
- 8.6 다른 차원 축소 기법
- 8.7 연습문제

CHAPTER 9 비지도 학습

- 9.1 군집
- 9.2 가우시안 혼합
- 9.3 연습문제

[PART 2 신경망과 머신러닝]

CHAPTER 10 케라스를 사용한 인공 신경망 소개

- 10.1 생물학적 뉴런에서 인공 뉴런까지
- 10.2 케라스로 다층 퍼셉트론 구현하기
- 10.3 신경망 하이퍼파라미터 튜닝하기
- 10.4 연습문제

CHAPTER 11 심층 신경망 훈련하기

- 11.1 그레이디언트 소실과 폭주 문제
- 11.2 사전훈련된 층 재사용하기
- 11.3 고속 옵티마이저
- 11.4 규제를 사용해 과대적합 피하기
- 11.5 요약 및 실용적인 가이드라인
- 11.6 연습문제

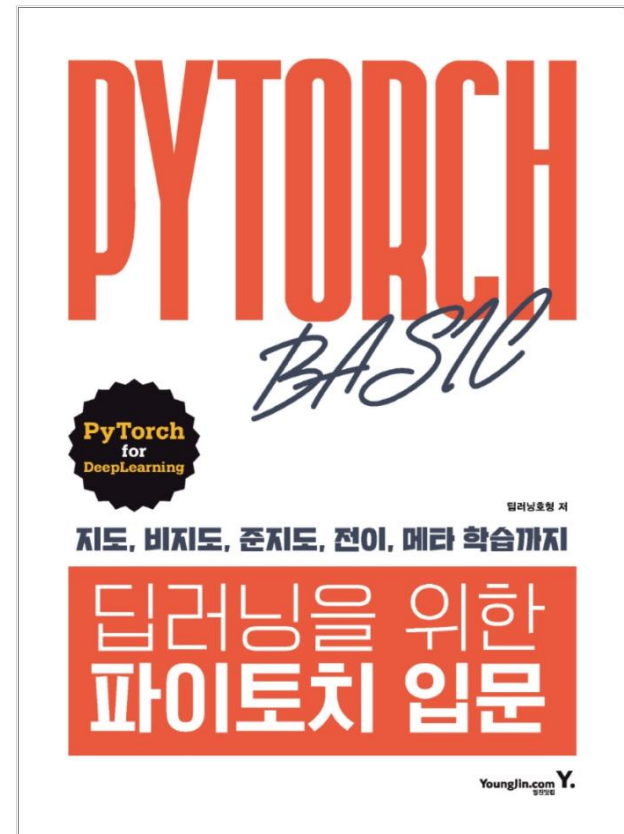
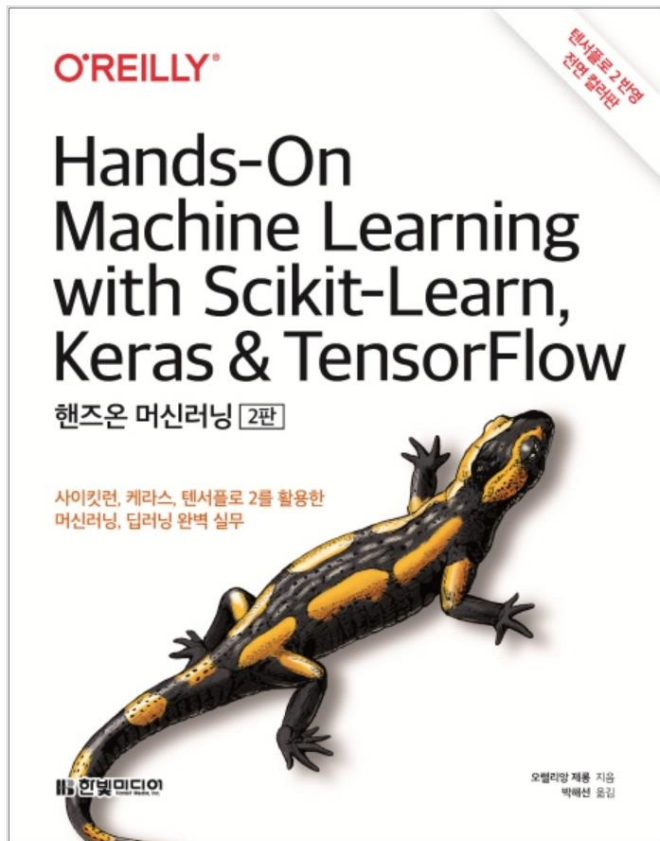
CHAPTER 12 텐서플로를 사용한 사용자 정의 모델과 훈련

- 12.1 텐서플로 훑어보기
- 12.2 넘파이처럼 텐서플로 사용하기
- 12.3 사용자 정의 모델과 훈련 알고리즘
- 12.4 텐서플로 함수와 그래프
- 12.5 연습문제

CHAPTER 13 텐서플로에서 데이터 적재와 전처리하기

- 13.1 데이터 API
- 13.2 TFRecord 포맷
- 13.3 입력 특성 전처리
- 13.4 TF 변환
- 13.5 텐서플로 데이터셋 (TFDS) 프로젝트
- 13.6 연습문제

스터디 교재



스터디 참여 연구원

번호	이름	소속	발표	참여방식
1	김동욱	화학데이터기반연구센터	예	오프
2	김민근	화학소재솔루션센터	예	온-오프
3	김영철	환경자원연구센터	예	온-오프
4	김윤희	고기능고분자연구센터	예	온-오프
5	김인	화학소재솔루션센터	아니오	온-오프
6	김진수	화학데이터기반연구센터	예	오프
7	나민주	화학안전연구센터	예	온-오프
8	문상진	에너지소재연구센터	아니오	온-오프
9	안현정	화학소재솔루션센터	예	온-오프
10	양진훈	화학데이터기반연구센터	예	온-오프
11	우미혜	에너지소재연구센터	아니오	온-오프
12	이수현	환경자원연구센터	아니오	온-오프
13	이주현	화학데이터기반연구센터	예	오프
14	장소민	화학데이터기반연구센터	예	오프
15	장승훈	화학데이터기반연구센터	예	오프
16	정지안	화학데이터기반연구센터	아니오	온-오프
17	조남정	화학소재솔루션센터	아니오	온-오프
18	최지원	화학안전연구센터	예	오프
19	한요셉	화학소재솔루션센터	예	온-오프

스터디 진행 일정 (매주 2-3명씩 발표)

순서	Chapter	스터디 내용	발표자 (호칭 생략)
1	0	머신러닝 스터디 소개 및 안내 사항 공유	장승훈
2	1	1.1 머신러닝이란? ~ 1.3 애플리케이션 사례 (29~35 p)	김동욱
3	1	1.4.1 지도학습과 비지도학습 (35~43 p)	김민근
4	1	1.4.2 배치/온라인 학습 ~ 1.4.3 사례기반/모델기반 학습 (43~53 p)	김영철
5	1	1.5 머신러닝의 주요도전과제 ~ 1.6 테스트와 검증 (53~64 p)	김윤호
6	2	2.1 실제 데이터로 작업하기 ~ 2.2 큰 그림 보기 (67~75 p)	김진수
7	2	2.3.1 작업환경 만들기 ~ 2.3.3 데이터 구조 훑어보기 (75~85 p)	나민주
8	2	2.3.4 테스트 세트 만들기 (85~90 p)	안현정
9	2	2.4 데이터 이해를 위한 탐색과 시각화 (91~98 p)	양진훈
10	2	2.5 머신러닝 알고리즘을 위한 데이터 준비 (99~110 p)	이주현
11	2	2.6 모델 선택과 훈련 ~ 2.8 론칭, 모니터링, 시스템 유지보수 (110~124 p)	장소민
12	3	3.1 MNIST ~ 3.3 성능측정 (127~145 p)	장승훈
13	3	3.4 다중분류 ~ 3.7 다중출력분류 (145~154 p)	최지원
14	4	4.1 선형회귀 ~ 4.2 경사하강법 (157~176 p)	한요셉

스터디 발표 준비 관련 사항

- 온라인 오프라인 자유롭게 참석 가능.
- 순서(이름순 정렬) 대로 발표 진행.
- 향후 발표 순서는 늦어도 발표 1달 전에는 미리 공지 예정.
- 30분 내외로 발표 내용 준비.
- 온라인 참석 발표자는 자신의 PC 환경(마이크,스피커,주변) 미리 체크.
- 별도의 PT 자료 없이 발표해도 전혀 무방.
- 발표 내용은 녹화될 예정, 녹화 파일은 유튜브 비공개 업로드 예정. (비번 공유)
- 메일 외에 공지 및 자료 공유를 위한 단톡방 개설? (의견수렴 후 결정)
- 스터디 진행 중간에 외부 연사 초청 예정. (실험1, 계산1, 인공지능1)
- 자유롭게 스터디에 대한 건의사항 제안.