

소셜데이터과학
INTRODUCTION TO DATA SCIENCE FOR SOCIOLOGISTS
(3-3-0)

2021년 2학기

Instructor	김현우, PhD (hxk271@psu.edu)
Office	106 Henderson Drive, State College, PA 16801
Classroom	충북대학교 사회과학대학(N15동) 302호
Class hours	10:00 AM–12:00 PM, 화요일; 10:00 AM–12:00 PM, 목요일
Course website	준비중

강의 개요

종종 4차 산업혁명과 초연결사회로 은유되는 오늘날, 통계분석과 데이터 과학(또는 기계학습)은 세부적인 전공 영역을 뛰어넘어 창의적 인재의 일반 소양으로 거듭나고 있습니다. 이 영역은 생산적인 학술 연구 뿐 아니라 정부 정책의 제안 및 비판, 데이터 저널리즘(data journalismism), 비영리단체/기업 운영관리 등 현실적 문제에서도 그 가능성을 폭넓게 증명해 왔습니다. 그러나 막상 현실적인 문제에 적용될 때 인문사회학적 소양을 함께 지닌 전문가의 부족으로 인해 학계와 산업체에서는 최근까지도 많은 아쉬움을 겪어왔습니다. 이에 따라 오늘날 사회학 교육 안에서도 데이터 과학과 코딩 교육, 그리고 사회통계에 관한 지식과 능력을 함양할 필요가 한층 더 요구됩니다.

사회학도로서 연관된 사회문제를 캐치하고 실제 데이터를 분석할 수 있는 능력은 다방면으로 유용합니다. 이 수업을 통해 향후 사회통계연습 등 다른 수업 이수에도 튼튼한 기초를 마련합니다. 하지만 이 수업이 추구하는 가장 궁극적인 목표는 여러분이 대학 졸업 이후에도 실무에서 활용할 수 있는 데이터 분석 능력을 함양하는 것입니다.

이 수업은 사회학을 전공하는 학부생을 대상으로 데이터 과학의 기초적인 아이디어와 논리, 그리고 알고리즘을 소개합니다. 이 수업에서 우리는 다음의 네 가지 토픽을 각 4주씩 차례로 학습합니다.

1. 데이터 과학 언어 파이썬(Python) 리뷰 (4주)
2. 파이썬 라이브러리를 사용한 데이터 전처리 및 시각화 (4주)
3. 지도학습(supervised learning) 머신러닝 알고리즘 소개 (4주)
4. 비지도학습(unsupervised learning) 알고리즘과 모델 평가 (4주)

구체적인 토픽별 내용은 아래 “토픽 개요”와 “세부 일정”을 참고하세요.

선수 과목

사회통계학 또는 그에 준하는 기초통계학 수업을 반드시 이수하여야 합니다.

강의 교재

- 송석리 · 이현아. 2019. *모두의 데이터 분석 with 파이썬*. 길벗. (필수)
- 박응용. 2019. *Do it! 점프 투 파이썬*. 이지스퍼블리싱 (참고)
(한글 원문 무료: <https://wikidocs.net/book/1>)
- 밴더플래스, 제이크. 2017. *파이썬 데이터 사이언스 핸드북*. 위키북스. (참고)
(영어 원서는 무료: <https://jakevdp.github.io/PythonDataScienceHandbook>)
- 프로보스트, 포스터 · 톰 포셋. 2014. *비즈니스를 위한 데이터 과학: 빅데이터를 바라보는 데이터 마이닝과 분석적 사고*. 한빛미디어. (참고)

학습 보조자료

- Python은 오픈소스이고 공식 튜토리얼은 무료입니다(<https://docs.python.org/3/tutorial>). 설명이 초보자에게는 다소 지나치게 세부적이지만 예제가 매우 파이썬답습니다("Pythonic"). 머신러닝 알고리즘을 구현하는데 필요한 라이브러리인 sklearn 역시 오픈소스로 마찬가지로 무료 공식 튜토리얼을 가지고 있습니다(<https://scikit-learn.org>). 설명을 온전히 이해하기 위해서는 약간의 수학 지식을 요구하지만 예제만으로도 충분히 훌륭합니다.
- 유튜브(<https://www.youtube.com>)나 Udacity(<https://www.udacity.com>) 등에서는 수많은 무료 Python 강의가 열려 있습니다. 영어를 겁내지 말고 (필요하다면) 캡션을 사용하세요.
- 구글(<https://www.google.com>)과 스택오버플로(<https://stackoverflow.com>)는 학부 1학년부터 업계 시니어에 이르기까지 모든 데이터 과학자들의 가장 든든한 우군입니다. 잘 모를 때는 질문을 영어로 옮겨 검색하세요. 초보자가 가질만한 거의 모든 질문은 이미 누군가가 던졌고 게다가 대답도 있을 가능성이 높습니다.

강의 구성

- 강의: 모든 학생은 반드시 수업에 참여해야 합니다. 수업을 통해 기초적인 개념과 알고리즘을 배우고 컴퓨터를 사용하여 실습합니다. 수업 내용은 진행될수록 누적되기 때문에 결석은 향후 이해에 큰 방해가 됩니다.

- 실습: 모든 학생은 반드시 실습에 참여해야 합니다. 전산실에서 Python을 사용하여 실제 코딩을 진행합니다.
- 퀴즈: 각각의 토픽을 마칠때마다 take-home quiz가 주어집니다. 4개의 토픽에 맞추어 퀴즈도 4개입니다. 각 퀴즈는 20점 만점으로 전체 학점의 80%에 해당합니다. 테드라인을 넘기면 감점됩니다.

학점 안내사항

모든 과제와 중간 및 기말발표 점수는 정보시스템에 업로드됩니다. 최종 학점은 다음 기준에 따라 산출됩니다.

- 출석 및 수업 참여 (20%)
- 퀴즈 4개 (80%)

수강생 유의사항

- 모든 수업은 별도의 안내가 없는 이상 원칙적으로 대면으로 진행됩니다.
- 수업 전일에 해당 주차 강의안과 퀴즈가 다음 GitHub 레파지토리에 업로드됩니다:
<https://github.com/hxk271/SocDataSci>
- 공결은 증빙서류를 모두 갖추어 담당교수에게 직접 제출된 경우만 인정하며, 개별적인 공결 행정처리는 일체 무효임에 주의하십시오.
- 3회 지각은 1회 결석으로 처리합니다. (첫 수업 및 공결을 포함하여) 4회 이상 결석한 경우 무조건 F이며 예외는 없습니다. 불가피한 사정으로 공결하였을 경우 그 외 일체 지각이나 결석을 하지 않아야 하겠습니다.
- 시험관련 부정행위자로 판명되었을 때는 학칙 또는 내규에 의거 해당 교과목의 성적을 취소합니다.
- 조기 취업자는 담당교수와 상담 후 출석과 과제, 중간/기말시험 등을 적절한 과제로 모두 대체할 수 있습니다. 이 경우 점수 상한은 B+로 합니다.

장애학생 수업안내

장애학생은 본 수업과 관련하여 본인 희망시 다음과 같은 지원이 가능합니다. 담당교수 및 장애학생지원센터와 상담 바랍니다.

- 공통: 도우미 지원(수업, 이동), 대체평가, 별도 발표/시험장소 제공, 선수강 지원, 노트북 사용
- 시각장애: 점자/확대/녹음 교재 및 시험지 제공, 발표/시험시간 연장, 강의자료 텍스트제공
- 청각장애: 지정좌석제, 동영상 자막지원
- 지체장애: 강의실 변경, 지정좌석제, 발표/시험시간 연장

토픽 개요

1주차 데이터 과학 소개

토픽1: 파이썬 리뷰	2주차 Hello World in Python
	3주차 파이썬에서 조건문과 반복문
	4주차 파이썬에서 함수와 클래스

목표: 기초 파이썬을 복습하고 좀 더 파이썬다운(Pythonic) 코딩을 연습한다.

5주차 numpy를 사용한 데이터 전처리

토픽2: 데이터 전처리와 시각화	6주차 pandas를 사용한 데이터 전처리
	7주차 matplotlib를 사용한 시각화
	8주차 Midterm

목표: 데이터 전처리를 위한 파이썬 라이브러리에 익숙해진다.

9주차 머신러닝과 인공지능 소개

토픽3: 머신러닝 지도학습	10주차 첫번째 머신러닝: 레이블이 두 개일때(Naïve Bayes)
	11주차 두번째 머신러닝: 레이블이 여러 개일때(Decision Tree)
	12주차 세번째 머신러닝: 레이블이 숫자일때(Linear Regression)

목표: 지도학습 알고리즘의 논리와 쓰임새를 파악한다.

13주차 네번째 머신러닝: 레이블이 없을때(k -Means)

토픽4: 머신러닝 비지도학습 외	14주차 머신러닝 퍼포먼스의 비교평가
	15주차 리뷰와 전망
	16주차 Final

목표: 비지도학습 알고리즘의 논리와 쓰임새를 파악하고, 여러 알고리즘을 비교한다.

세부 일정

1주차

데이터 과학의 소개

TOPICS 데이터 과학의 구성요소; 산업혁명들과 4차 산업혁명; 데이터 분석가, 데이터 엔지니어, 그리고 데이터 과학자; 스크립트 언어 및 통계 프로그램 개요; Anaconda 설치

GOAL 사회학 전공에 있어 데이터 과학의 쓸모를 탐색한다.

READINGS 프로보스트·포셋 1장

2주차

Hello World in Python

TOPICS 스크립트 언어로서 Python; IDE로서 Jupyter Notebook; 문자열과 슬라이싱(Slicing); 자료 타입(Lists, Tuples, Sets, Dicts); 파일 입출력(I/O)

GOAL 문자열과 여러 타입으로 데이터를 다룰 수 있고 보존할 수 있다.

READINGS 밴더플래스 1장; 박응용 1-2장

3주차

파이썬에서 조건문과 반복문

TOPICS 연산자(Operators); 조건문(if, elif, else); 반복문(for, range, while); 계속과 이탈(continue, pass, break); 동영상 파일의 재생 원리(OpenCV)

GOAL 단순대입 알고리즘(brute-force search)을 구현할 수 있다.

READINGS 박응용 3장; 밴더플래스 1장

4주차

파이썬에서 함수와 클래스

TOPICS 내장함수(Built-In Functions); 함수 작성과 호출; 클래스(Classes); 에러와 예외(Errors and Exceptions); 외부 라이브러리 이용; 이를바 객체지향 프로그래밍(Object-Oriented Programming)

GOAL 재귀 알고리즘(recursive algorithm)을 구현할 수 있다.

READINGS 박응용 4-5장; 밴더플래스 1장

세부 일정 (계속)

5주차 numpy를 사용한 데이터 전처리

TOPICS 사회과학도를 위한 벡터와 행렬; 기술통계; 재배열(Reshaping); 정렬(Sort)과 연산

GOAL numpy로 데이터를 요약할 수 있다.

READINGS 뱐더플래스 2장

6주차 pandas를 사용한 데이터 전처리

TOPICS DataFrame; numpy—pandas 변환; 인덱싱(Indexing); 피쳐 엔지니어링(Feature Engineering); 데이터 결합(merge, concat, join); 집계(Aggregation)와 groupby

GOAL pandas로 DataFrame을 원하는 형태로 변환할 수 있다.

READINGS 뱐더플래스 3장

7주차 matplotlib를 사용한 시각화

TOPICS 시각화의 힘; Boxplot, Histogram; Scatterplot; Wordcloud; Pie Chart; 지도 그리기(geopandas); 네트워크 그리기(NetworkX); 그림 파일의 조작 원리; 멋진 쇼케이스와 나쁜 예제들; Front-End Web Development 소개

GOAL matplotlib으로 간단한 그래프를 그릴 수 있다.

READINGS 뱐더플래스 4장

8주차 Midterm

세부 일정 (계속)

9주차

머신러닝과 인공지능 소개

TOPICS 전문가 시스템(Expert System)과 규칙기반 의사결정(Rule-Based Decision-Making); 기계에 의한 학습; 레이블(Label); 분류(Classification), 회귀(Regression), 그리고 군집(Clustering); scikit-learn

GOAL 인공지능의 역사와 머신러닝의 종류와 쓰임새를 간략히 설명할 수 있다.

READINGS 프로보스트 · 포셋 2장

10주차

첫번째 머신러닝: 레이블이 두 개일때(Naïve Bayes)

TOPICS 사회과학도를 위한 확률론; 조건부확률(Conditional Probability); 베이즈 공리 (Bayes Theorem); 조건부 독립성(Conditional Independence) 가정; 두 종류의 레이블; 연습문제(Spam Mail Classifier)

GOAL sklearn으로 나이브 베이즈를 구현할 수 있다.

READINGS 밴더플래스 5장; 프로보스트 · 포셋 9장

11주차

두번째 머신러닝: 레이블이 여러 개일때(Decision Tree)

TOPICS 사회과학도를 위한 정보이론; 정보의 불확실성과 엔트로피(Entropy); 스무고개 놀이; 분할정복(Divide-and-Conquer) 알고리즘; 가지치기(Pruning and Splitting); 몇 종류의 레이블; 랜덤 포레스트(random forest); 연습문제(Tennis Play and Weather)

GOAL sklearn으로 의사결정나무를 구현할 수 있다.

READINGS 밴더플래스 5장; 프로보스트 · 포셋 3장

12주차

세번째 머신러닝: 레이블이 숫자일때(Linear Regression)

TOPICS 사회과학도를 위한 미적분; 오차(Errors); 적합선(Fitting Line); 비선형 최적화 (Nonlinear Optimization); 숫자 레이블; 연습문제(Boston House Prices)

GOAL sklearn으로 선형회귀를 구현할 수 있다.

READINGS 밴더플래스 5장

세부 일정 (계속)

13주차

네번째 머신러닝: 레이블이 없을때(k -Means)

TOPICS 사회과학도를 위한 선형대수학; 유클리드 거리(Euclidean Distance); argmax 또는 argmin 문제; 국지적 최적화(Local Optima); 레이블이 존재하지 않음; 연습문제(Opening Delivery Stores)

GOAL sklearn으로 k -평균 클러스터링을 구현할 수 있다.

READINGS 밴더플래스 5장

14주차

머신러닝 퍼포먼스의 비교평가

TOPICS 과학의 퍼포먼스로서 설명, 예측, 통제; 왜 정확도(accuracy rate)는 정확하지 않은가; Precision–Recall Trade-Off; Confusion Matrix; F1 Score; Receiver Operating Characteristic (ROC) 곡선; 과적합(Overfitting); k -Fold Cross-Validation

GOAL sklearn으로 공짜점심은 없음(No Free Lunch Theorem)을 확인할 수 있다.

READINGS 밴더플래스 5장; 프로보스트 · 포셋 5, 7-8장

15주차

리뷰와 전망

TOPICS 여러 알고리즘의 공통점과 차이점; 이른바 빅데이터(Big Data); 서버, 슈퍼컴퓨터, 그리고 클라우드 컴퓨팅; Hadoop과 MapReduce 소개; GitHub과 L^AT_EX 소개; 오픈소스와 가치창출

GOAL 한 학기 동안의 수업 내용을 복습하고 새로운 배움을 준비한다.

READINGS 밴더플래스 5장; 프로보스트 · 포셋 13-14장

16주차

Final

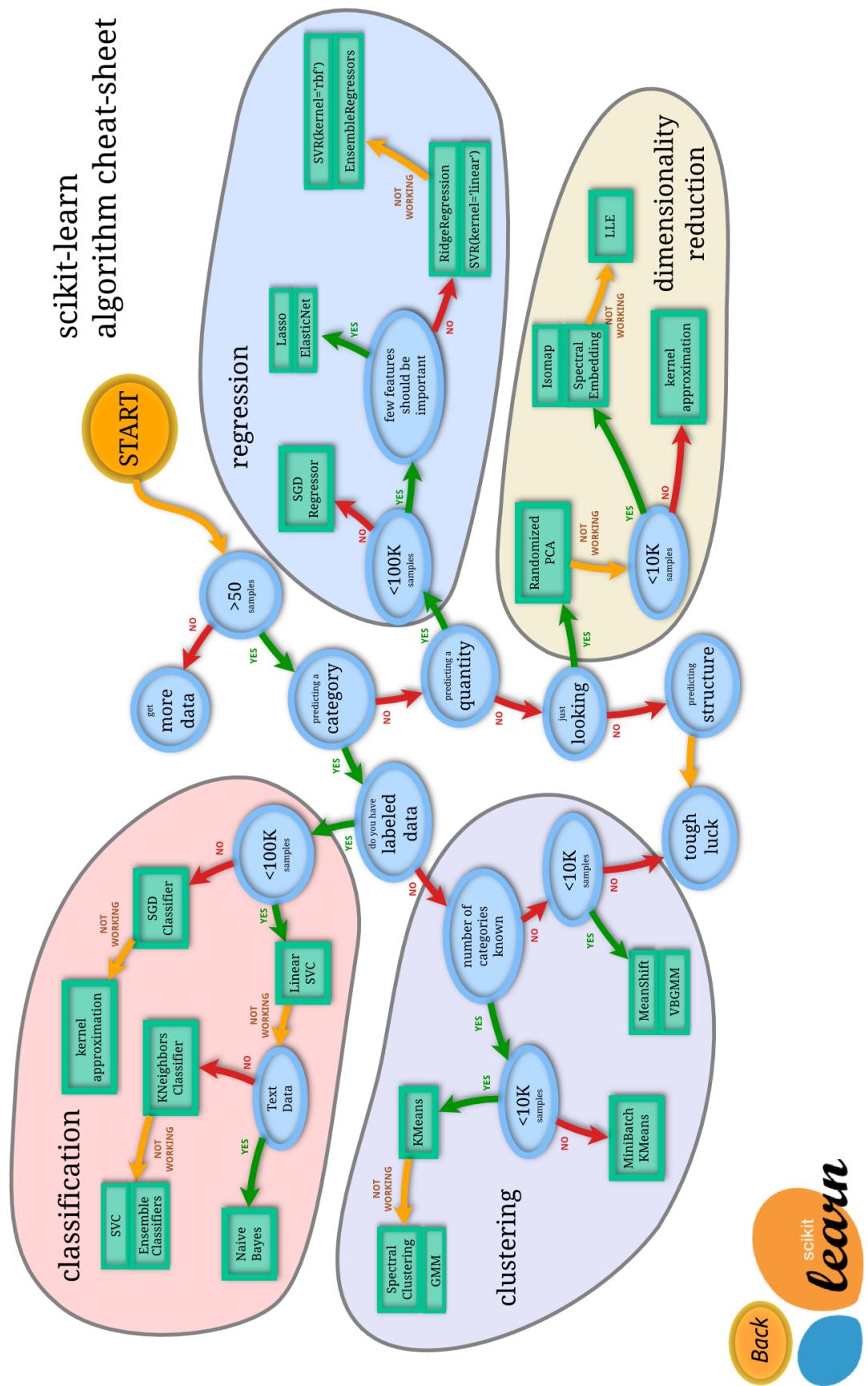


Figure 1: sklearn을 통해 구현할 수 있는 알고리즘