

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

2022년 2학기

Instructor	김현우, PhD (hxk271@cbnu.ac.kr)
Office	N15동 414호
Classroom	N15동 사회조사실습실
Class hours	03:00 PM – 5:50 PM, 목요일
Course website	https://github.com/hxk271/SocDataSci

강의 개요

종종 4차 산업혁명과 초연결사회로 온유되는 오늘날 통계분석과 데이터 사이언스는 세부적인 전공 영역을 뛰어넘어 창의적 인재의 일반 소양으로 거듭나고 있습니다. 이 영역은 학술 연구 뿐 아니라 정부 정책의 제안 및 비판, 데이터 저널리즘(data journalism), 헬스케어(health care), 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 사회학도에게도 데이터 사이언스와 코딩 교육, 그리고 사회통계에 관한 지식과 능력이 요구되고 있습니다.

데이터 사이언티스트에게는 코딩(coding), 수학(mathematics), 그리고 전공 지식(domain knowledge)이라는 세 가지 스킬셋이 요구됩니다. 이 수업은 특히 수학적 기초와 코딩 경험이 튼튼하지 않은 사회과학도를 위해 설계되었습니다. 수학은 알고리즘에 관한 직관적 이해를 돋는 정도로만 사용되고, 코딩은 알고리즘을 손쉽게 구현하는 예제를 중심으로 짜여집니다. 그러므로 수학이나 컴퓨터 코딩이 두려워 이 수업을 기피할 필요는 없습니다.

이 수업은 사회학을 전공하는 학부생을 대상으로 데이터 사이언스의 기초적인 아이디어와 논리, 그리고 알고리즘을 소개합니다. 이 수업에서 우리는 다음의 세 가지 토픽을 차례로 학습합니다.

1. 데이터 사이언스의 소개와 파이썬(Python) 입문
2. 데이터 전처리, API 사용법 및 웹 스크래핑
3. 통계 분석과 머신러닝(machine learning) 알고리즘

사회학도로서 연관된 사회문제를 캐치하고 실제 데이터를 분석할 수 있는 능력은 다방면으로 유용합니다. 이 수업을 통해 궁극적으로는 여러분이 대학 졸업 이후에도 실무에서 활용할 수 있는 데이터 분석 능력의 함양을 목표로 합니다.

선수 과목

사회통계, 사회통계연습, 사회조사방법론을 반드시 먼저 이수하여야 합니다. 데이터 사이언스 분야의 타전공 학생은 최종 학점에서 일정한 패널티를 받게 됩니다.

강의 교재

아래 박해선(2023)은 복습과 과제 수행을 위해 필수적입니다. 송석리·이현아 (2019)은 우리 수업 전반부에서 계속 사용하지만 구입이 필수는 아닙니다.

- 박해선. 2023. 혼자 공부하는 데이터 분석 *with* 파이썬. 한빛. (필수)
- 송석리 · 이현아. 2019. 모두의 데이터 분석 *with* 파이썬. 길벗. (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 사이언티스트의 가장 든든한 우군입니다. 잘 모를 때는 질문을 영어로 읊겨 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.

- 과제: 중간시험과 기말시험을 대신하여 과제가 두 개 주어집니다. 과제는 전체 학점의 70%에 해당합니다. 데드라인을 넘기면 감점됩니다. 하나라도 미제출시 무조건 수업에서 낙제합니다.

학점 안내사항

모든 과제의 점수는 충북대학교 개신누리에 업로드됩니다. 최종 학점은 다음 기준에 따라 산출됩니다.

- 출석 및 수업 참여 (30%)
- 과제 2개 (70%)

수강생 유의사항

- 모든 수업은 별도의 안내가 없는 이상 원칙적으로 **대면으로** 진행됩니다.
- 수업 전일에 해당 주차 강의안과 자료가 다음 GitHub 레파지토리에 업로드됩니다:
<https://github.com/hxk271/SocDataSci>
- 공결은 증빙서류를 모두 갖추어 담당교수에게 **직접** 제출된 경우만 인정하며, 개별적인 공결 행정처리는 일체 무효임에 주의하십시오.
- 2회 지각은 1회 결석으로 처리합니다. (첫 수업 및 공결을 포함하여) 3회 이상 결석한 경우 무조건 F이며 예외는 없습니다. 불가피한 사정으로 공결하였을 경우 그 외 일체 지각이나 결석을 하지 않아야 하겠습니다.
- 과제를 1회 이상 제출하지 않으면 무조건 F이며 예외는 없습니다. 과제는 선택사항이 아닙니다.
- 시험관련 부정행위자로 판명되었을 때는 학칙 또는 내규에 의거 해당 교과목의 성적을 취소합니다.
- 이 수업은 조기 취업자에게 적절하지 않습니다. 다만 파이썬과 기계학습에 이미 익숙한 경우에 한하여 담당교수와 상담 후 적절한 과제로 대체할 수 있습니다. 이 경우 점수 상한은 B+로 합니다.

장애인 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

파이썬 입문

-
- | | |
|-----|-----------------------|
| 1주차 | 데이터 사이언스 소개 |
| 2주차 | Hello World in Python |
| 3주차 | 조건문과 반복문 |
| 4주차 | matplotlib를 이용한 시각화 |
| 5주차 | 공공데이터 분석실습 1 |
| 6주차 | 공공데이터 분석실습 2 |
-

목표: 파이썬을 이용하여 간단한 공공데이터를 분석하고 시각화를 실습한다.

데이터 전처리 API 사용법 및 웹 스크래핑

- | | |
|------|------------------|
| 7주차 | 수업없음 |
| 8주차 | numpy와 pandas 입문 |
| 9주차 | API 사용하기 |
| 10주차 | 데이터베이스와 SQL |
| 11주차 | 자료의 전처리 |
-

목표: 데이터 수집 및 관리를 위해 유용한 도구를 익힌다

통계 분석과 머신러닝 알고리즘

- | | |
|------|---------|
| 12주차 | 통계적 추론 |
| 13주차 | 머신러닝 |
| 14주차 | 소셜 네트워크 |
| 15주차 | 수업없음 |
-

목표: 통계 추론과 지도/비지도학습 알고리즘의 논리와 쓰임새를 파악한다.

세부 일정

1주차

데이터 사이언스의 소개

TOPICS 데이터 사이언스의 구성요소; 산업혁명들과 4차 산업혁명; 데이터 분석가, 데이터 엔지니어, 그리고 데이터 사이언티스트; 구글 콜랩(Colab); 스크립트 언어로서 Python; IDE로서 Jupyter Notebook

READINGS 박해선 1장 1절, 1장 2절

2주차

Hello World in Python

TOPICS 문자열(string)과 슬라이싱(slicing); 자료 타입(lists, tuples, sets, dicts); 파일과 자료의 입출력(I/O); 내장함수(built-in functions)

READINGS 송석리·이현아 부록

3주차

조건문과 반복문

TOPICS 연산자(Operators); 조건문(if, elif, else); 반복문(for, range, while); 계속과 이탈(continue, pass, break); 단순대입 알고리즘(brute-force search)

READINGS 송석리·이현아 파트1

4주차

matplotlib를 사용한 시각화

TOPICS 시각화의 힘; 히스토그램(histogram); 상자-수염 그림(Box-Whisker Plot); 산점도(scatterplot); matplotlib

READINGS 송석리·이현아 파트2

5주차

공공데이터 분석실습 1

TOPICS 반복문과 조건문 활용

READINGS 송석리·이현아 파트3

세부 일정 (계속)

6주차 공공데이터 분석실습 2

TOPICS 시계열 그래프(time-series line chart)

READINGS 송석리·이현아 파트4

7주차 수업없음

8주차 numpy와 pandas 입문

TOPICS DataFrame; numpy—pandas 변환; 정렬(sort)

READINGS 박해선 1장 3절; 송석리·이현아 파트5

9주차 API 사용하기

TOPICS Application Programming Interface (API); JSON과 XML 데이터 다루기

READINGS 박해선 2장 1절, 2장 2절(선택)

10주차 데이터베이스와 SQL

TOPICS 데이터 웨어하우스(data warehouse); SQLite; MySQL

READINGS 박해선 부록A

11주차 자료의 전처리

TOPICS 인덱싱(indexing); 피쳐 엔지니어링(feature engineering); 데이터 결합(merge, concat, join); 집계(aggregation)와 groupby; 기술통계; 재배열(reshaping)

READINGS 박해선 3장, 5장

세부 일정 (계속)

12주차 통계적 추론

TOPICS 확률분포(probability distribution); t 검정; 분산분석(ANOVA); sklearn

READINGS 박해선 4장, 6장, 7장 1절

13주차 머신러닝 지도학습: 회귀분석

TOPICS 회귀(Regression); 오차(Errors); 적합선(Fitting Line); Precision–Recall Trade-Off; Confusion Matrix; F1 Score; Receiver Operating Characteristic (ROC) 곡선; 연습문제(Boston House Prices, Breast Cancer Diagnosis)

READINGS 박해선 7장 2절

14주차 머신러닝 비지도학습: 소셜 네트워크

TOPICS 노드(nodes)와 엣지(edge); 중심성(centrality); 군집(cluster)과 커뮤니티(community); NetworkX

READINGS 미정

15주차 수업없음

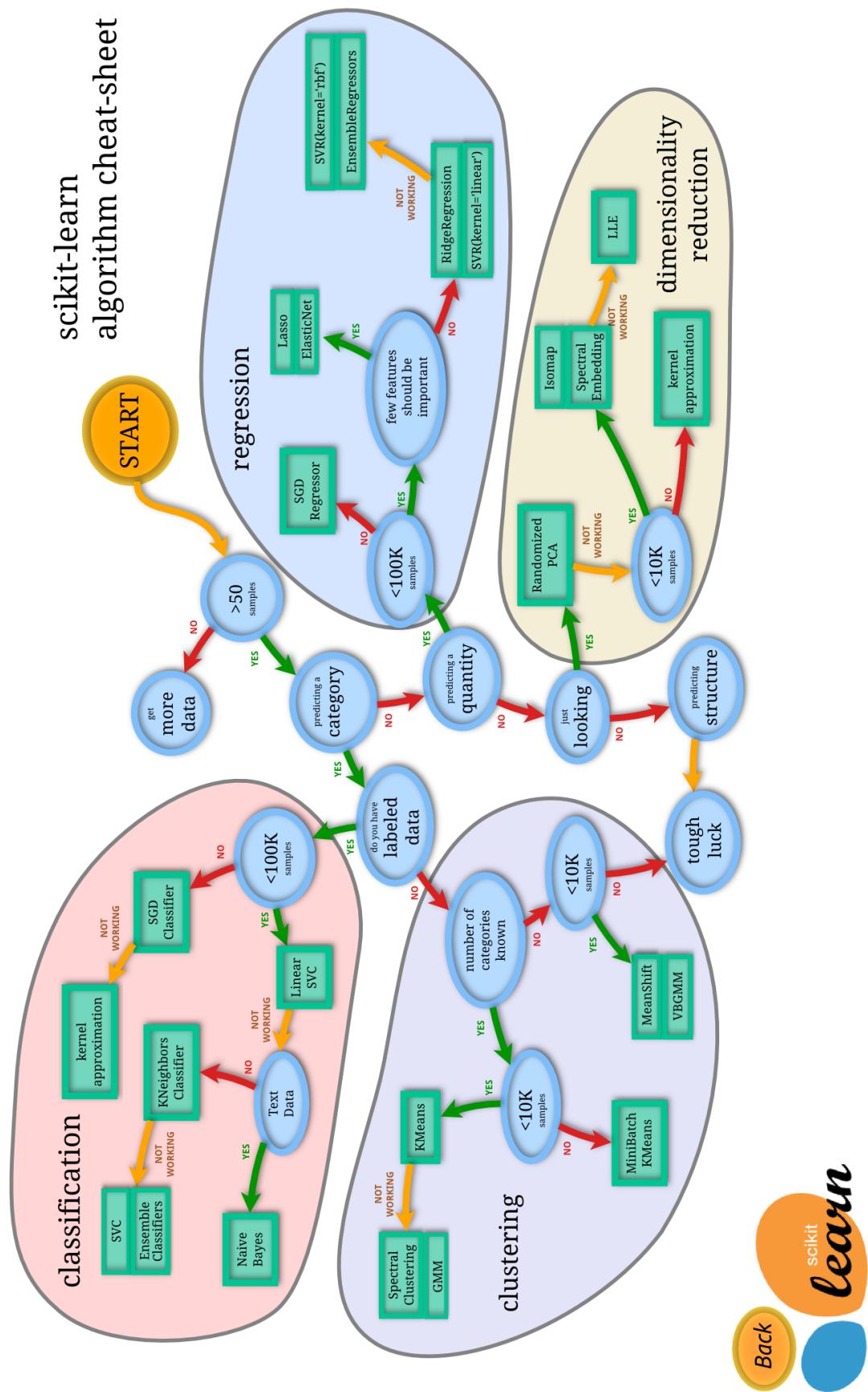


Figure 1: sklearn을 통해 구현할 수 있는 알고리즘