금융인을 위한

통계와 데이터 분석 입문

금융인을 위한 데이터분석

한국금융연수원
KOREA BANKING INSTITUTE

학습 내용

1. 데이터분석에서 통계지식의 필요성
2. 데이터분석 절차
3. 데이터분석 도구
4. 파이썬 작업환경 준비

빅데이터 시대의 도래

- 스마트폰과 같은 디지털 기기의 발전, 사물인터넷 등의 보급 등으로 규모를 가늠할 수 없을 정도로 많은 정보와 데이터가 생산되고 있음
- 데이터의 형태도 다양해지고 있음 : 정형/비정형 데이터

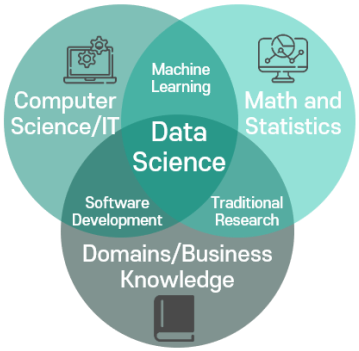
빅데이터 시대의 도래

- 데이터의 홍수 속에서 데이터로부터 의미있는 통찰을 얻기 위해서 데이터를 수집, 정제, 정리, 요약, 분석하고 이를 통해 의사결정을 하는 학문인 통계학(statistics) 기초지식 습득이 필요함

빅데이터 시대의 도래

데이터 과학(data Science)

컴퓨터 도구를 효율적으로 이용하고 적절한 방법을
사용하여 실제적인 문제에 **데이터를 활용**하여
합리적인 답을 도출하려는 활동



- 데이터 사이언티스트에 요구되는 역량
  - 컴퓨터 활용 능력, 모델에 대한 이해(통계적 지식),
    실무 지식 뿐만 아니라 협업 능력 및 태도도 중요함

데이터분석을 통해 알아내고자 하는 질문들

Q. 부동산 가격을 예측할 수 있을까?

Q. 개인/기업의 회생 및 파산을 예측할 수 있을까?

Q. 한 기업에 입사하는데 남녀의 성별 차이가 있을까?

Q. 온라인 광고에서 클릭 여부를 예측할 수 있을까?

Q. 새로운 마케팅이 고객의 만족도를 높이는데 효과가 있을까?

Q. 의심이 되는 신용카드 거래를 어떻게 감지해낼 수 있을까?

Q. 스팸 메일을 어떻게 분류해낼 수 있을까?

Q. 환자 자료(예: 의료 영상 이미지)를 활용하여 특정 암을 예측할 수 있을까?

데이터분석 절차

1. 문제 정의(problem definition)
2. 데이터 정의(data definition)
3. 데이터 취득(data acquisition)
4. 데이터 가공(data processing)
5. 탐색적 분석과 데이터 시각화
   (exploratory data analysis, data visualization)
6. 모형화(modeling)
7. 분석 결과 정리(reporting)

데이터분석 절차

- 현실에서는 데이터분석 절차가 순서대로 진행되지 못하고 중간에 다시 **이전 단계로 돌아가야 하는 경우**들이 생김

예

데이터 수집에서 문제가 생기면
필요한 데이터 또는 문제를 수정해야 할 수 있음

예

탐색적 자료분석 단계에서 수집된 데이터의 문제가 발견되는 경우 새로 데이터를 수집하거나
문제 가설을 바꿔야 할 수도 있음

데이터분석 절차

- 현실에서는 데이터분석 절차가 순서대로 진행되지 못하고 중간에 다시 **이전 단계로 돌아가야 하는 경우**들이 생김

예

모형화 결과 의미있는 결과가 나오지 않은 경우
이유를 알아내기 위해 탐색적 분석 과정을
다시 해야 할 수도 있음

예

분석 결과 정리 및 공유 후
새로운 문제와 데이터 정의로 이어질 수도 있음

데이터분석 도구

- 파이썬(Python), R, SAS, MATLAB 등
  다양한 분석 언어가 존재함
- 파이썬과 R은 무료로 사용 가능하다는 장점이 있음

데이터분석 도구

파이썬(python)

데이터분석 이외에도 다양한 분야에서
개발 언어로 활용함

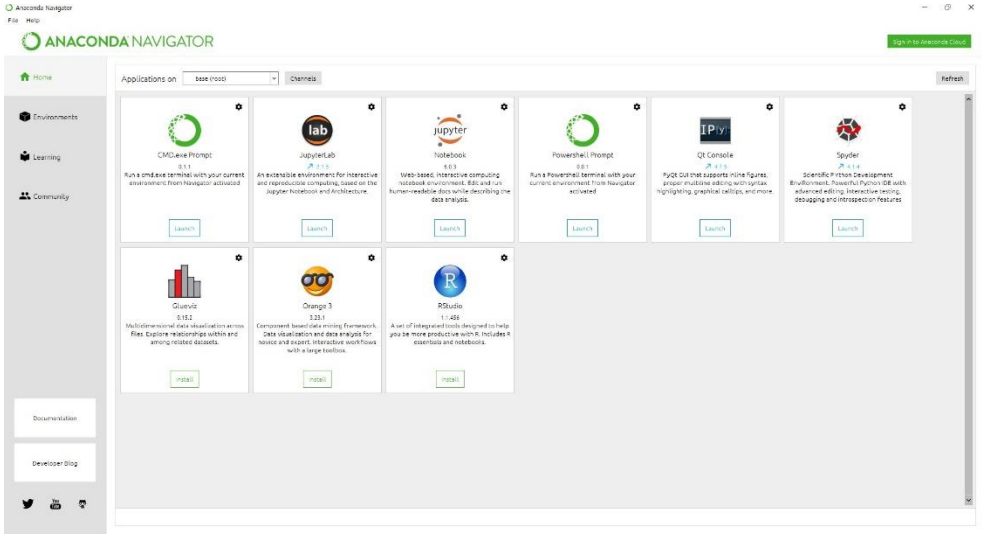
→ 데이터분석 결과를 다른 웹 애플리케이션에 접목하거나 통계적인 코드를 데이터베이스에 포함시켜야 할 때 범용적으로 더 많은 기능을 적용할 수 있는 장점이 있음

Python 사용을 위한 환경 설정 : Anaconda

- 파이썬을 직접 설치(python.org)할 수 있지만 Anaconda를 설치하고 수업을 진행함
- Why Anaconda?
  - 다양한 라이브러리와 기능이 통합되어 있음
    : 데이터 과학에 필요한 대부분의 패키지들 (scikit-learn, numpy, pandas, matplotlib 등), 주피터 노트북, 스파이더 등을 모두 포함함

Python 사용을 위한 환경 설정 : Anaconda

- Anaconda 설치
  - https://www.anaconda.com/download/



주피터 노트북(Jupyter Notebook)

- 브라우저 기반의 interactive한 파이썬 개발도구
- 마크다운(markdown) 문서화, 수학식 표현이 가능함

학습정리

- 데이터의 홍수 속에서 데이터로부터
  의미있는 통찰을 얻기 위해서는
  통계학(statistics)의 기초지식 습득이 필요함

- 데이터분석 절차를 알아보았고 현실에서는 데이터 결과
  보고 전에 다시 **이전 단계로 돌아가야 하는 경우**들이 생김

- 데이터분석에 활용할 수 있는 **분석 언어**로
  파이썬(python), R, SAS, MATLAB 등이 존재함

- **무료**로 사용할 수 있고 **범용성**이 있는 파이썬

- **파이썬 작업환경** 준비
  - **Anaconda** 설치, **Jupyter Notebook** 기본 사용법