가톨릭대학교
THE CATHOLIC UNIVERSITY OF KOREA

기계학습

-머신러닝 소개 및 실습

미디어기술콘텐츠학과 / 의료인공지능학과
강호철

왜 머신러닝인가?

- 초창기 지능형 애플리케이션
  - 조건문 사용
    - 예) 스팸 메일 필터링
    - 결정 규칙을 사람이 직접 모델링

- 결정 규칙을 직접 만들 때의 단점
  - 결정에 필요한 로직은 한 분야나 작업에 국한됨
    - 작업이 조금만 변경 되더라도 전체 시스템을 다시 개발해야 함
    - 숫자 8인식 → 2인식?
  - 규칙을 설계하려면 그 분야 전문가들이 내리는 결정 방식에 대해 잘 알아야 함
    - 얼굴 인식 문제

왜 머신러닝인가?

- 결정 규칙을 직접 만들 때의 단점
  - 얼굴 인식 문제
    - 컴퓨터의 인식 방식 vs. 사람의 인식하는 방식
    - 얼굴의 다양성. 해결책은?

왜 머신러닝인가?

- 머신러닝으로 풀 수 있는 문제
  - 지도 학습
    - 주어진 입력에 대한 출력 예측
    - 학습에 필요한 입, 출력 데이터는
  - 지도 학습의 예
    - 편지 봉투에 손으로 쓴 우편번호 숫자 판별
      - 입력, 출력?
      - 학습에 필요한 데이터?
    - 의료 영상 이미지에 기반한 종양 판단
      - 입력, 출력?
      - 학습에 필요한 데이터?
    - 의심되는 신용 카드 거래 감지
      - 입력, 출력?
      - 학습에 필요한 데이터?

왜 머신러닝인가?

- 머신러닝으로 풀 수 있는 문제
  - 비지도 학습
    - 주어진 입력에 대한 출력이 제공되지 않음
    - 학습을 이해하거나 평가하는 일이 쉽지 않음
  - 비지도 학습의 예
    - 블로그 글의 주제 구분
      - 텍스트 데이터 요약 및 핵심주제 찾기
    - 고개들을 취향이 비슷한 그룹으로 묶기
      - 어떤 고객들의 취향이 비슷한지 파악
      - 비슷한 취향의 고객 그룹화
    - 비 정상적인 웹사이트 접근 탐지
      - 웹 트래픽을 이용한 탐지

왜 머신러닝인가?

- 컴퓨터에 머신러닝 적용하려면
  - 컴퓨터가 인식할 수 있는 데이터 제공
    - 샘플 (데이터 포인트), 특성
    - 좋은 특성 추출 필요
      - 예) 성씨로 성별 구분?

왜 머신러닝인가?

- 문제와 데이터 이해하기
  - 머신러닝 프로세스에서 가장 중요한 과정은 사용할 데이터를 이해하고 그 데이터가 해결해야 할 문제와 어떤 관련이 있는지를 이해하는 것
    - 어떤 질문에 대한 답을 원하는가? 데이터가 답을 줄 수 있나?
    - 내 질문을 잘 기술할 수 있는 머신러닝 방법은?
    - 문제를 풀기 위한 충분한 데이터를 모았는가?
    - 내가 추출한 데이터의 특성으로 좋은 결과를 만들 수 있나?
    - 머신러닝 성능 측정 방법?
    - 다른 연구나 제품과의 협력은?

왜 파이썬인가?

- 파이썬은 데이터 과학 분야를 위한 표준 프로그래밍 언어
  - 파이썬은 범용 프로그래밍 언어의 장점은 물론 매트랩MATLAB과 R 같은 특정 분야를 위한 스크립팅 언어의 편리함을 함께 갖춤
  - 다양한 도구: 데이터 적재, 시각화, 통계, 자연어 처리, 이미지 처리 등에 필요한 라이브러리 존재
  - 터미널이나 주피터 노트북(Jupyter Notebook) 같은 도구로 대화하듯 프로그래밍할 수 있음
  - 머신러닝과 데이터 분석은 데이터 주도 분석이라는 점에서 근본적으로 반복 작업, 따라서 반복 작업을 빠르게 처리하고 손쉽게 조작할 수 있는 도구가 필수
  - 범용 프로그래밍 언어로서 파이썬은 복잡한 그래픽 사용자 인터페이스(GUI)나 웹 서비스도 만들 수 있으며 기존 시스템과 통합하기도 좋음

Scikit-Learn

- 오픈소스로 자유롭게 사용하거나 배포 가능
  - 잘 알려진 머신러닝 알고리즘들은 물론 알고리즘을 설명한 풍부한 문서도 제공
    - http://scikit-learn.org/stable/documentation
  - 사이킷런은 매우 인기가 높고 독보적인 파이썬 머신러닝 라이브러리임
  - 산업 현장이나 학계에도 널리 사용되고 많은 튜토리얼과 예제 코드를 온라인에서 쉽게 찾을 수 있음
  - 사이킷런은 다른 파이썬의 과학 패키지들과도 잘 연동됨

Scikit-Learn

- 설치
  - Scikit-learn은 두 개의 다른 파이썬 패키지인 넘파이(NumPy)와 사이파이(SciPy)를 사용
  - 그래프를 그리려면 맷플롯립(matplotlib)을, 대화식으로 개발하려면 아이파이썬(Ipython)과 주피터 노트북도 설치해야 함
  - 필요한 패키지들을 모아 놓은 파이썬 배포판을 설치하는 방법을 권장
    - **Anaconda: 대용량 데이터 처리, 예측 분석, 과학 계산용 파이썬 배포판**
    - Enthought Canopy: 과학 계산용 파이썬 배포판
    - Python(x,y): 윈도우 환경을 위한 과학 계산용 무료 파이썬 배포판

Scikit-Learn

- 설치
  - 주피터 노트북
    - 주피터 노트북은 프로그램 코드를 브라우저에서 실행해주는 대화식 환경을 제공
  - NumPy
    - 파이썬으로 과학 계산을 하려면 꼭 필요한 패키지임. 다차원 배열을 위한 기능과 선형 대수 연산과 푸리에 변환 같은 고수준 수학 함수와 유사(pseudo) 난수 생성기를 포함
  - SciPy
    - 과학 계산용 함수를 모아놓은 파이썬 패키지임. SciPy는 고성능 선형 대수, 함수 최적화, 신호 처리, 특수한 수학 함수와 통계 분포 등을 포함한 많은 기능을 제공

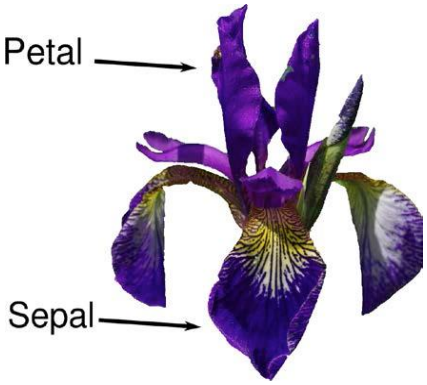
Scikit-Learn

- 설치
  - matplotlib
    - 파이썬의 대표적인 과학 계산용 그래프 라이브러리임. 선 그래프, 히스토그램, 산점도 등을 지원하며 출판에 쓸 수 있을 만큼의 고품질 그래프를 그려줌
  - pandas
    - 데이터 처리와 분석을 위한 파이썬 라이브러리임
  - mglearn
    - Helper functions for the book 'Introduction to machine learning with Python'
    - pip install mglearn

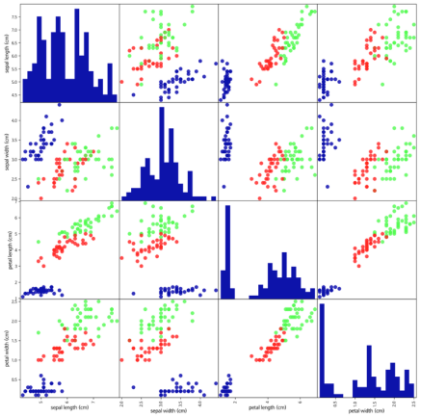
Review) Python 테스트

실습 – 붓꽃의 품종 분류

- 어떤 품종인지 구분해 놓은 측정 데이터를 이용, 새로 채집한 붓꽃의 품종을 예측하는 머신러닝 모델 구현
  - 데이터 적재
  - 성과 측정: 훈련 데이터와 테스트 데이터
  - 가장 먼저 할일: 데이터 살펴보기
  - 첫 번째 머신러닝 모델: K- 최근접 이웃 알고리즘
  - 예측 및 모델 평가하기



▲그림 1-2 붓꽃의 부위



▲그림 1-3 클래스 레이블을 색으로 구분한 Iris 데이터셋의 산점도 행렬

참고자료

- Introduction to Machine Leanring with Python

  (파이썬 라이브러리를 활용한 머신러닝)

  - 안드레아스 뮐러, 세라 가이도 지음 / 박해선 옮김
  - 한빛미디어, 2019