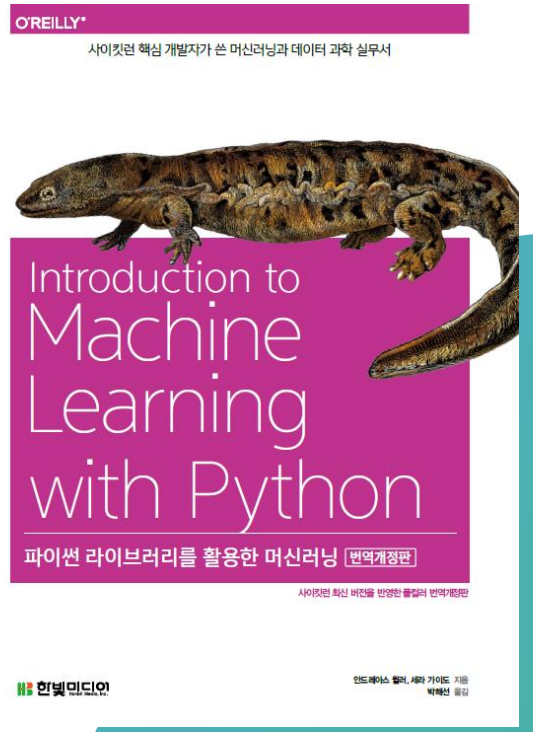
Chapter 01 : 소개

파이썬라이브러리를 활용한머신러닝 (개정판)



시작하기전에

- **책에서 사용하는 소프트웨어 버전**
  - Python 버전: 3.7.2 (default, Dec 29 2018, 06:19:36)
  - [GCC 7.3.0]
  - pandas 버전: 0.23.4
  - matplotlib 버전: 3.0.2
  - NumPy 버전: 1.15.4
  - SciPy 버전: 1.1.0
  - IPython 버전: 7.2.0
  - scikit-learn 버전: 0.20.2
- **예제 다운로드 링크**
  - https://github.com/rickiepark/introduction_to_ml_with_python
  - https://nbviewer.jupyter.org/github/rickiepark/introduction_to_ml_with_python/tree/master/

이 책의 학습 목표

- 1장: 머신러닝과 머신러닝 애플리케이션의 기초 개념을 소개 및 사용 환경
- 2장: 지도 학습 알고리즘
- 3장: 비지도 학습 알고리즘
- 4장: 머신러닝에서 데이터를 표현하는 방법
- 5장: 모델 평가와 매개변수 튜닝을 위한 교차 검증과 그리드 서치
- 6장: 모델을 연결하고 워크플로를 캡슐화하는 파이프라인 개념
- 7장: 텍스트 데이터에 적용하는 방법과 텍스트에 특화된 처리 기법
- 8장: 개괄적인 정리와 어려운 주제에 대한 참고 자료 안내

Contents

CHAPTER 01 소개

머신러닝을 사용한 문제 해결 예시와 머신러닝 모델의 중요 개념 학습

SECTION 1.1 왜 머신러닝인가? (1/2)

결정 규칙을 직접 만들 때의 단점

- 결정에 필요한 로직은 한 분야나 작업에 국한됨.
  따라서, 작업이 조금만 변경되더라도 전체 시스템을 다시 개발해야함
- 규칙을 설계하려면 그 분야 전문가들이 내리는 결정 방식에 대해 잘 알아야 함

머신러닝으로 풀 수 있는 문제

- 입력과 출력(정답, 레이블)으로 부터 학습하는 머신러닝 알고림 지도 학습 알고리즘
- 학습된 알고리즘은 사람의 도움 없이도 새로운 입력이 주어지면 적절한 출력 가능
- 데이터셋 (입력, 출력) , 분석하기 좋고, 성능 측정 쉬움
- 지도학습의 예
  - 편지 봉투에 손으로 쓴 우편번호 숫자 판별
  - 의료 영상 이미지에 기반한 종양 판단
  - 의심되는 신용카드 거래 감지

SECTION 1.1 왜 머신러닝인가? (2/2)

- 머신러닝으로 풀 수 있는 문제
  - 입력은 주어지지만, 출력(정답, 레이블) 없는 데이터를 비슷한 특징끼리 군집화 하여 새로운 데이터에 대한 결과를 예측하는 비지도 학습 알고리즘
  - 성공 사례는 많지만 비지도 학습을 이해하거나 평가하는 일은 쉬지 않음
  - 비지도 학습의 예
    - 블로그 글의 주제 구분
    - 고객들을 취향이 비슷한 그룹으로 묶기
    - 비정상적인 웹사이트 접근 탐지
- 문제와 데이터 이해하기
  - 지도 학습과 비지도 학습 모두 컴퓨터가 인식할 수 있는 형태러 입력 데이터를 준비하는 것이 매우중요
  - 머신러닝 프로세스에서 가장 중요한 과정은 사용할 데이터를 이해하고 그 데이터가 해결해야할 문제와 어떤 관련이 있는지를 이해하는 것임

SECTION 1.2 왜 파이썬인가?

- 파이썬(Python)은 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

SECTION 1.3 scikit-learn

- 오픈 소스인 사이킷런(scikit-learn)은 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
- 사이킷런 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

SECTION 1.4 필수 라이브러리와 도구들

- 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
- 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

SECTION 1.5 파이썬 2 vs. 파이썬 3

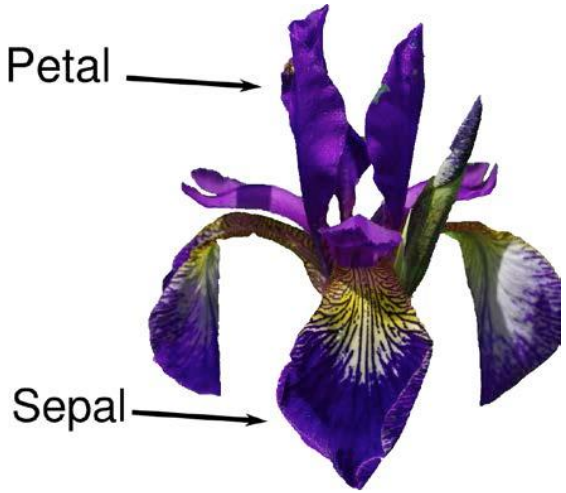
- 파이썬 2는 더 이상 큰 개선은 진행되지 않으며 파이썬 3에서 변경 사항이 많아 파이썬 2로 작성한 코드는 파이썬 3에서 실행되지 않는 경우가 많음
- 파이썬 처음 사용자나 새로운 프로젝트는 파이썬 3의 최신 버전 사용이 바람직함
- 대부분의 경우 새로운 코드가 파이썬 2와 3에서모두 실행되도록 작성하는 것은 가능함
- 이 책의 모든 코드는 두 버전에서 모두 작동하나 파이썬 2에서는 출력 모양이 조금 다를 수 있음
- 또한 matplotlib, numpy, scikitlearn과 같은 패키지들은 더 이상 파이썬 2.7에 맞추어 새로운 기능을 릴리스하지 않을 것임
- 새로운 버전에 포함된 기능을 사용하려면 파이썬 3.7로 업그레이드 해야함

SECTION 1.6 이 책에서 사용하는 소프트웨어 버전

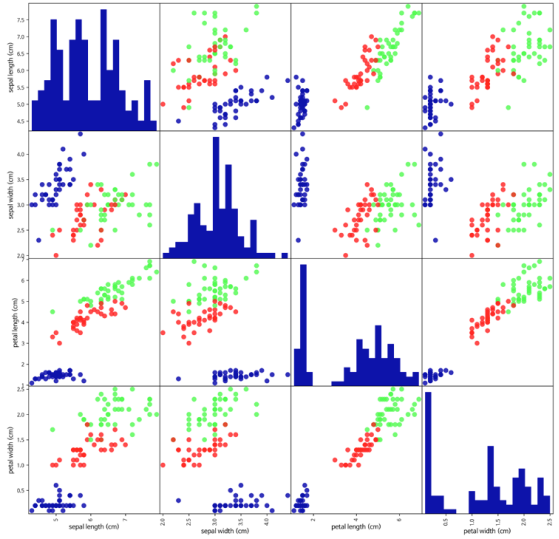
- Python 버전: 3.7.2 (default, Dec 29 2018, 06:19:36)
- [GCC 7.3.0]
- pandas 버전: 0.23.4
- matplotlib 버전: 3.0.2
- NumPy 버전: 1.15.4
- SciPy 버전: 1.1.0
- IPython 버전: 7.2.0
- scikit-learn 버전: 0.20.2
  - 버전이 정확히 같아야 하는 것은 아니지만 scikit-learn은 가능한 한 최신 버전이어야 함

SECTION 1.7 첫 번째 애플리케이션: 붓꽃의 품종 분류

- 어떤 품종인지 구분 해놓은 측정 데이터를 이용해 새로 채집한 붓꽃의 품종을 예측하는 머신러닝 모델을 만들어 보기
  - 데이터 적재
  - 성과 측정: 훈련 데이터와 테스트 데이터
  - 가장 먼저 할 일: 데이터 살펴보기
  - 첫 번째 머신러닝 모델: k-최근접 이웃 알고리즘
  - 예측하기
  - 모델 평가하기



▲그림 1-2 붓꽃의 부위



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

SECTION 1.8 요약 및 정리

- 머신러닝과 머신러닝 애플리케이션에 대한 간략한 소개
- 지도 학습과 비지도 학습의 차이
- 책에서 사용할 도구 개요와 소개
- 실측한 자료를 사용하여 붓꽃의 품종이 무엇인지 예측하는 작업
  - 모델을 구축하기 위해 전문가가 정확한 품종으로 구분 해놓은 데이터셋을 사용했으므로 지도 학습에 해당하는 문제를 학습함
  - 또한 품종이 세 개(setosa, versicolor, virginica)이므로 세 개의 클래스를 분류하는 문제도 학습함