1.2 머신러닝용 파이썬 패키지

이 절에서는 머신러닝에서 많이 사용되는 파이썬 패키지를 소개한다. 이 패키지들은 교재의 예제 코드에서 사용되므로 미리 실치하기를 권장한다. pip 또는 conda 패키지 관리자로 설치할 수 있다.

기본 패키지

다음에 소개하는 패키지는 머신러닝 뿐 아니라 모든 종류의 데이터 분석 업무에 공통적으로 사용된다. 본 교재의 코드는 항상 이 패키지들을 임포트하고 있다고 가정한다. 자주 사용하는 패키지이므로 짧은 별명(alias)으로 임포트하여 사용하는 경우가 많다.

numpy 패키지

numpy("넘파이"라고 읽는다) 패키지는 파이썬에서 수치 해석, 특히 선형대수 계산 기능을 제공한다. 자료형이 고정된 다차원 배열 클래스(n-dimensional array)와 벡터화 연산(vectorized operation)을 지원하며 수학 연산에서 가장 기본적이고 중요한 패키지다

다음 명령으로 설치한다.

```
pip install numpy
```

임포트할 때는 보통 `np` 라는 별명으로 임포트한다.

```
import numpy as np
```

scipy 패키지

scipy("사이파이"라고 읽는다) 패키지는 고급 수학 함수, 수치적 미적분, 미분 방정식 계산, 최적화, 신호 처리 등에 사용하는 다양한 과학 기술 계산 기능을 제공한다.

다음 명령으로 설치한다.

```
pip install scipy
```

임포트할 때는 보통 `sp` 라는 별명으로 임포트한다.

```
import scipy as sp
```

pandas 패키지

pandas("판다스" 또는 "팬더스"라고 읽는다) 패키지는 테이블 형태의 데이터를 다루는 데이터프레임(DataFrame) 자료형을 제공한다. 자료 탐색과 정리에 아주 유용하여 데이터 분석에 빠질 수 없는 필수 패키지다. 원래는 R 언어에서 제공하는 데이터프레임 자료형을 파이썬에서 제공할 목적이었으나 더 다양한 기능이 추가되었다.

다음 명령으로 설치한다.

```
pip install pandas
```

임포트할 때는 보통 `pd` 라는 별명으로 임포트한다.

```
import pandas as pd
```

matplotlib 패키지

matplotlib("맷플롯립"이라고 읽는다) 패키지는 파이썬에서 각종 그래프나 챠트 등을 그리는 시각화 기능을 제공한다. Tkinter, wxPython, Qt, GTK+ 등의 다양한 그래픽 엔진을 사용할 수 있다. 또한 MATLAB의 그래프 기능을 거의 동일하게 사용할 수 있는 pylab이라는 서브패키지를 제공하므로 MATLAB에 익숙한 사람들은 바로 맷플롯립을 사용할 수 있다.

다음 명령으로 설치한다.

```
pip install matplotlib
```

임포트할 때는 보통 pylab 서브패키지를 `plt` 라는 별명으로 임포트한다.

```
import matplotlib.pylab as plt
```

seaborn 패키지

seaborn("시본"이라고 읽는다) 패키지는 맷플롯립 패키지에서 지원하지 않는 고급 통계 차트를 그리는 통계용 시각화 기능을 제공한다.

다음 명령으로 설치한다.

```
pip install seaborn
```

임포트할 때는 보통 `sns` 라는 별명으로 임포트한다.

```
import seaborn as sns
```

머신러닝 패키지

statsmodels 패키지

statsmodels("스탯츠모델즈"라고 읽는다) 패키지는 추정 및 검정, 회귀분석, 시계열분석 등의 기능을 제공하는 파이썬 패키지다. 기존에 R에서 가능했던 다양한 회귀분석과 시계열분석 방법론을 그대로 파이썬에서 이용할 수 있다. 다음은 statsmodels 패키지가 제공하는 기능의 일부다.

- 예제 데이터셋
- 검정 및 모수추정
- 회귀분석
    - 선형회귀

    - 강건회귀
    - 일반화 선형모형
    - 혼합효과모형
    - 이산종속변수
- 시계열 분석
    - SARIMAX 모형
    - 상태공간 모형
    - 벡터 AR 모형
- 생존분석
- 요인분석

교재에서는 선형회귀분석, 로지스틱회귀분석, 시계열분석에서 statsmodels 패키지를 사용한다.

다음 명령으로 설치한다.

```
pip install statsmodels
```

임포트할 때는 보통 api 서브패키지를 `sm` 이라는 별명으로 임포트한다.

```
import statsmodels.api as sm
```

scikit-learn 패키지

scikit-learn("사이킷런"이라고 읽는다) 패키지는 머신러닝 교육을 위한 최고의 파이썬 패키지다. scikit-learn 패키지의 장점은 다양한 머신러닝 모형을 하나의 패키지에서 모두 제공하고 있다는 점이다. 다음은 scikit-learn 패키지에서 제공하는 머신러닝 모형의 목록의 일부다.

- 데이터셋
    - 회귀분석, 분류, 클러스터링용 가상 데이터셋 생성
    - 각종 벤치마크 데이터셋
- 전처리
    - 스케일링
    - 누락데이터 처리
    - 텍스트 토큰화
- 지도학습
    - 회귀분석
    - LDA/QDA
    - 서포트벡터머신
    - 퍼셉트론, SGD
    - KNN
    - 가우스프로세스
    - 나이브베이즈
    - 의사결정나무
    - 랜덤포레스트, 부스팅
- 비지도학습
    - 가우스 혼합모형
    - 클러스터링
    - PCA
- 성능 최적화
    - 교차검증
    - 특징선택

- 하이퍼파라미터 최적화

설치와 임포트에 사용하는 이름은 sklearn이다. 다음 명령으로 설치한다.

```
pip install sklearn
```

임포트할 때는 보통 `sk` 이라는 별명으로 임포트한다.

```
import sklearn as sk
```

데이터 전처리용 패키지

missingno 패키지

pandas 데이터프레임 데이터에서 누락된 데이터를 찾고 시각화하는 기능을 제공한다.

다음 명령으로 설치한다.

```
pip install missingno
```

patsy 패키지

pandas 데이터프레임 데이터에서 선택, 변형하는 기능을 제공한다.

statsmodels가 의존하는 패키지이므로 statsmodels 패키지를 설치하면 별도로 설치할 필요가 없다.

텍스트 전처리용 패키지

nltk 패키지

spacy 패키지

konlpy 패키지

soynlp 패키지

gensim 패키지

이미지 전처리용 패키지

opencv 패키지

사운드 전처리용 패키지

librosa 패키지

지리정보 전처리용 패키지

geopandas 패키지