Machine Learning & Deep Learning

Lecture 20

Keras를 이용한 딥러닝 코드 분석

(피마 인디언의 당뇨병 예측)

김희숙 (H.S.Kim)

당뇨병 예측 예제

- 데이터의 양보다 훨씬 중요한 것은, **'필요한'** 데이터가 **얼마나 많은가**임
- 준비된 데이터가 우리가 사용하려는 **머신러닝**과 **딥러닝**에 얼마나 효율적으로 사용될 수 있도록 가공되었는지가 역시 중요함
- 머신러닝 프로젝트의 성공과 실패는 **얼마나 좋은 데이터**를 가지고 시작하느냐에 영향을 많이 받음
- 여기서 좋은 데이터란 **내가 알아내고자 하는 정보를 잘 담고 있는 데이터**를 말함
- 한쪽으로 치우치지 않고, 불필요한 정보를 가지고 있지 않으며, 왜곡되지 않은 데이터이어야 함
- 머신러닝, 딥러닝 개발자들은 **데이터**를 들여다 보고 **분석**할 수 있어야 함
- 내가 이루고 싶은 목적에 맞춰 가능한 한 많은 정보를 모았다면 이를 **머신러닝**과 **딥러닝**에서 사용할 수 있게 잘 **정제된 데이터 형식**으로 바꿔야 함
- 이 작업은 모든 머신러닝 프로젝트의 첫 단추이자 **가장 중요한 작업**

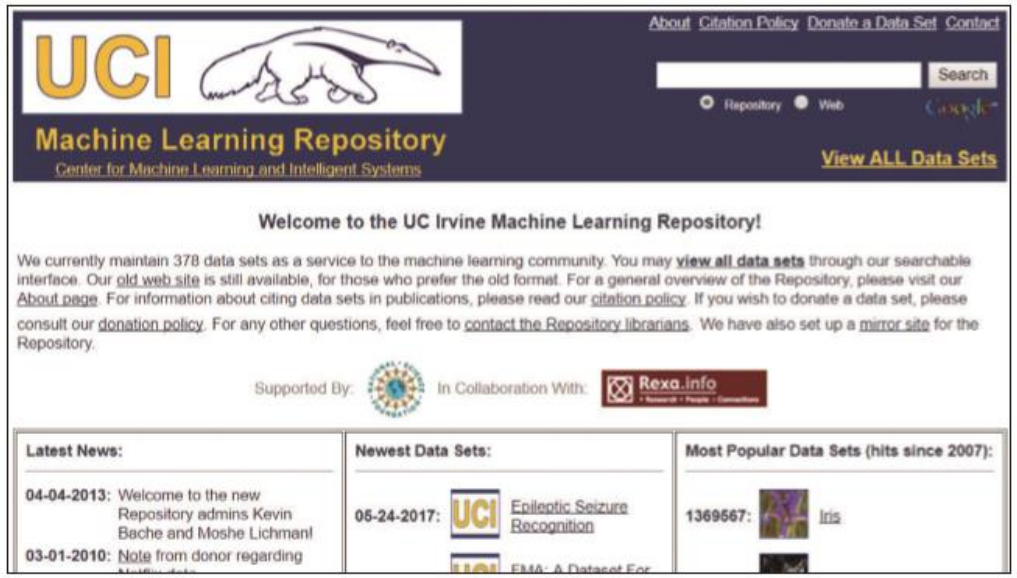
- **비만은 유전일까? 아니면 식습관 조절에 실패한 자신의 탓일까?**
  - ✓ **비만**이 **유전** 및 **환경** **모두가 원인**이라는 것을 증명하는 좋은 사례가 바로 미국 남서부에 살고 있는 **피마 인디언의 사례**
  - ✓ 피마 인디언은 **1950년대까지만 해도 비만인 사람이 단 한 명도 없는 민족**이었음
  - ✓ 그런데 지금은 전체 부족의 **60%가 당뇨, 80%가 비만**으로 고통받고 있음
  - ✓ 이는 생존하기 위해 **영양분을 체내에 저장하는 뛰어난 능력을 물려받은 인디언**들이 미국의 **기름진 패스트푸드 문화**를 만나면서 벌어진 일

피마 인디언 옛 모습

- **UCI 머신 러닝 저장소(UCI Machine Learning Repository)**

  → 머신러닝 공부에 필요한 각종 데이터를 모아 놓은 사이트 (UC Irvine에서 운영)

University of California, Irvine (캘리포니아 대학교 어바인)



http://archive.ics.uci.edu

피마 인디언 당뇨병 예측을 위한 데이터 **pima-indians-diabetes.csv**

- 데이터를 열어 보면 모두 **768명의 인디언**으로부터 **8개의 속성**과 **1개의 클래스**를 추출한 데이터임을 알 수 있음

속성 / 클래스 / 샘플

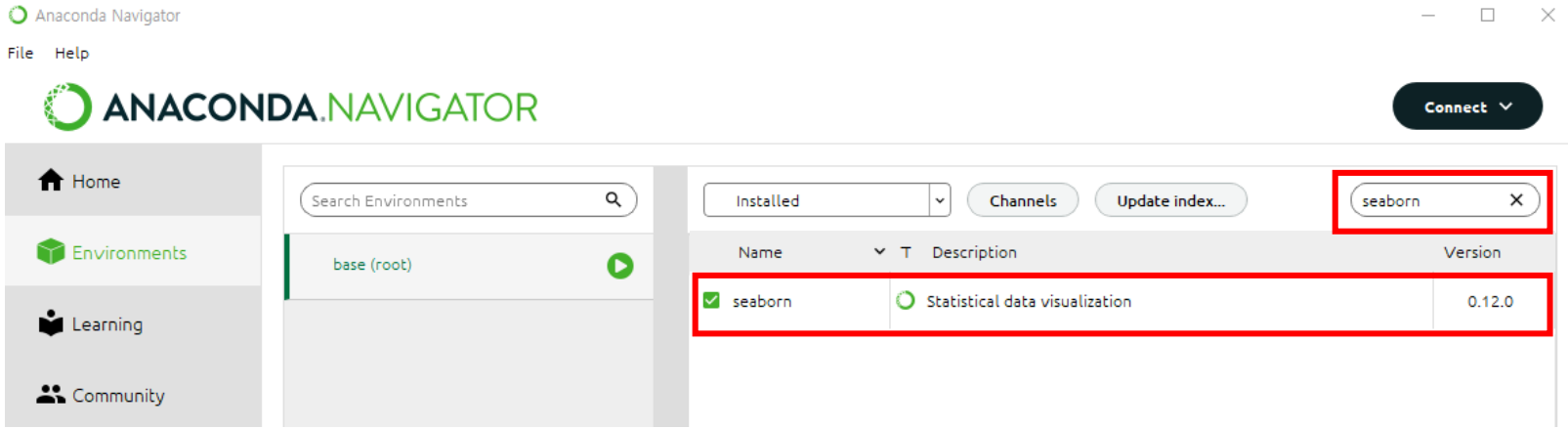
pregnant	plasma	pressure	thickness	insulin	BMI	pedigree	age	class
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1

- 샘플 수: 768
- 속성: 8
  - 정보 1 (pregnant): 과거 임신 횟수
  - 정보 2 (plasma): 포도당 부하 검사 2시간 후 공복 혈당 농도(mm Hg)
  - 정보 3 (pressure): 확장기 혈압(mm Hg)
  - 정보 4 (thickness): 삼두근 피부 주름 두께(mm)
  - 정보 5 (insulin): 혈청 인슐린(2-hour, mu U/ml)
  - 정보 6 (BMI): 체질량 지수(BMI, weight in kg/(height in $m)^2$)
  - 정보 7 (pedigree): 당뇨병 가족력
  - 정보 8 (age): 나이
- 클래스: 당뇨(1), 당뇨 아님(0)

- 데이터의 각 정보가 의미하는 의학, 생리학 배경 지식을 모두 알 필요는 없지만, 딥러닝을 구동하려면 반드시 **속성**과 **클래스**를 먼저 구분해야 함
- 모델의 정확도를 향상시키기 위해서는 **데이터의 추가 및 재가공이 필요**할 수도 있으므로 딥러닝의 구동에 앞서 **데이터의 내용과 구조를 잘 파악**하는 것이 중요

- **데이터를 잘 파악하는 것**이 **딥러닝을 다루는 기술의 1단계**
  - ✓ 데이터의 크기가 커지고 정보량이 많아지면 데이터를 불러오고 **내용을 파악**할 수 있는 **효과적인 방법**이 필요함
  - ✓ 이때 가장 유용한 방법이 **데이터를 시각화**해서 눈으로 직접 **확인**해 보는 것
  - ✓ 데이터를 다룰 때에는 데이터를 다루기 위해 만들어진 **라이브러리를 사용**하는 방법을 권장
  - ✓ **파이썬 데이터 관련 라이브러리 중 pandas**와 **matplotlib** 및 **seaborn**을 사용하는 것도 좋음

- **seaborn** 라이브러리 설치가 안되어있으면 설치하기



- **cmd**창에서 **pip install seaborn** 으로 설치해도 됨

- 딥러닝 설계 및 구현

pima-indians-diabetes.csv

20_(8 page) 강의용_ Diabetes.ipynb

pandas를 활용한 분석 예시

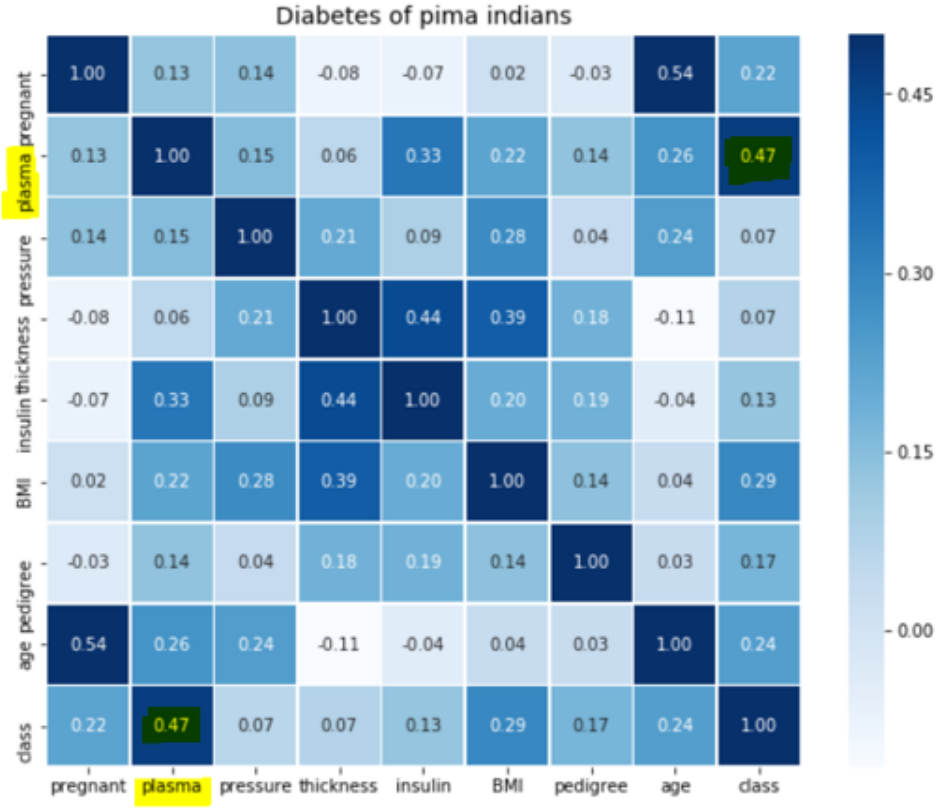
많은 데이터를 단순히 나열하는 것은 한눈에 들어오지 않으므로 큰 의미가 없다 데이터를 잘 다루려면 데이터를 한 번 더 가공해야 함
< 데이터를 가공할 때의 주의점 >
우리가 무엇을 위해 작업을 하는지 그 목적을 잊어서는 안 됨
이 프로젝트의 목적 : 당뇨병 발병을 예측하는 것
그렇다면 모든 정보는 당뇨병 발병과 어떤 관계가 있는지를 중점에 놓아야 함

```
1 print(df[['pregnant','class']].groupby(['pregnant'], as_index=False).mean().sort_values(by='pregnant', ascending=True))
```

```
    pregnant     class
0          0  0.342342
1          1  0.214815
2          2  0.184466
3          3  0.360000
4          4  0.338235
5          5  0.368421
6          6  0.320000
7          7  0.555556
8          8  0.578947
9          9  0.642857
10        10  0.416667
11        11  0.636364
12        12  0.444444
13        13  0.500000
14        14  1.000000
15        15  1.000000
16        17  1.000000
```

- pandas에서 제공하는 groupby() 함수를 사용해 pregnant 정보를 기준으로 하는 새 그룹을 만듦
- as_index=False는 pregnant 정보 좌측에 0, 1, 2 … 와 같은 새로운 인덱스(index)를 만듦
- mean() 함수를 사용해 평균을 구하고, sort_values() 함수를 써서 pregnant 컬럼을 오름차순(ascending)으로 정리하도록 설정하여 임신 횟수 당 당뇨병 발병 확률을 출력함

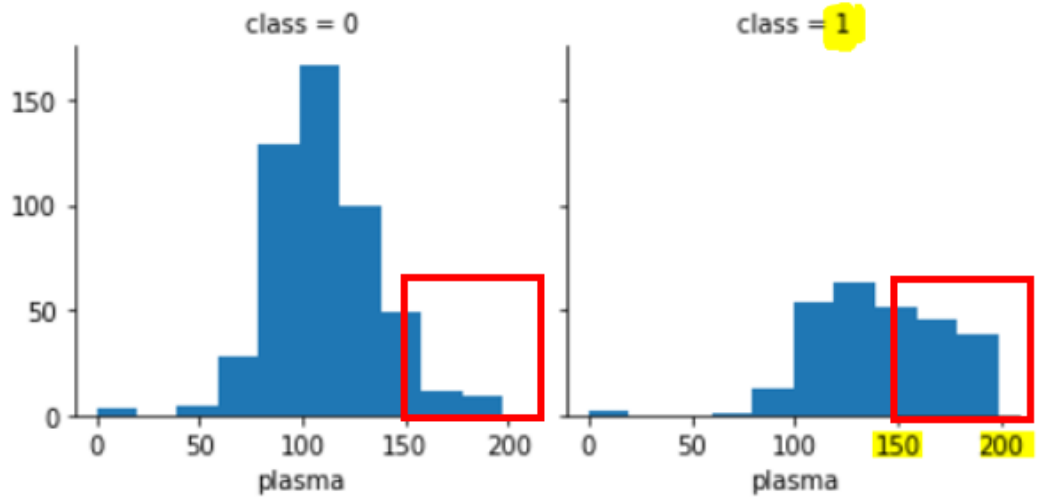
matplotlib를 활용한 분석 예시



- 데이터를 그래프로 표현해서 그 성격을 파악 하는 것도 중요함
- matplotlib는 파이썬에서 그래프를 그릴 때 가장 많이 사용되는 라이브러리
- matplotlib 라이브러리와 이를 기반으로 좀 더 정교한 그래프를 그릴 수 있도록 도와주는 seaborn 라이브러리를 사용해 각 정보끼리 어떤 상관관계가 있는지를 시각화
- 그래프를 통해 plasma 항목(공복 혈당 농도)이 class 항목과 상관관계가 높다는 것을 알 수 있음
- 즉, 이 항목이 딥러닝 결과를 만드는 데 가장 중요한 역할을 한다는 것을 예측할 수 있음

matplotlib를 활용한 분석 예시

- plasma와 class 항목만 따로 떼어 두 항목 간의 관계를 그래프로 다시 한번 확인함
- 당뇨병 환자의 경우 : class가 1에 해당하며 plasma 수치가 150이상인 경우가 많다는 것을 확인
- 이와 같이 딥러닝을 수행하기 전에 딥러닝 결과에 미치는 영향이 큰 항목을 발견하는 것이 필요하며 이를 '데이터 전처리 과정'이라고 함



- 딥러닝 모델 활용

Diabetes.h5

20_(12 page) 강의용_AFTER_Diabetes.ipynb