PyTorch로 딥러닝 제대로 배우기

- 기초편 -

Part4. 데이터

강사: 김 동 희

목차

I. 데이터

1. 데이터

❑ 데이터

- 데이터베이스, 데이터 마이닝, 데이터 기반 의사결정, 데이터 사이언스, ...

- 좋은 모델을 만들기 위해서는 좋은 데이터를 확보하는 것이 필수적!
  - 데이터의 **양** (크기가 크면 좋다)
  - 데이터의 **완결성** (비어있는 값이 없으면 좋다)
  - 데이터의 **신뢰도** (현실을 잘 계측한 데이터가 좋다)
  - 데이터의 **시기절적함** (timeliness, 필요할 때 수집하고 사용할 수 있어야 좋다)

1. 데이터

❑ **데이터를 기술하는데 사용하는 단어**

- 데이터의 **특성**: 정형 데이터와 비정형 데이터
- 데이터의 **크기**: 데이터의 행 수, 차원 수, 크기
- 데이터의 **타입**: 범주형 변수와 연속형 변수
- 데이터의 **용도**: 학습 데이터, 검증 데이터, 평가 데이터

2. 정형 데이터

❑ 정형데이터

- **데이터의 행 수**: 몇 개의 개체가 포함되어 있는가? 5개
- **데이터의 열 수**: 각 개체는 몇개의 특성을 가지는가? 3개 (이름, 수학 점수, 영어 점수)
- **데이터의 크기**: 데이터를 저장하기 위해 필요한 공간은 얼마인가? 1KB

학생 이름	수학 점수	영어 점수
A	50	80
B	70	65
C	24	30
D	47	97
E	25	43

두번째 행

[표1]

2. 정형 데이터

❑ 정형데이터

- 정형 데이터의 열은 차원(dimension), 특성(attribute 또는 feature), 측정값(measuer)

X
150
160
173
175

1차원 데이터

X	Y
150	50
160	30
173	40
175	70

2차원 데이터

X	Y	Z
150	50	1.3
160	30	2.5
173	40	1.2
175	70	2.4

3차원 데이터

X	Y	Z	V	W
150	50	1.3	..	..
160	30	2.5	..	..
173	40	1.2	..	..
175	70	2.4	...	..

다차원 데이터

[표2]

3. 비정형 데이터

❑ 비정형데이터

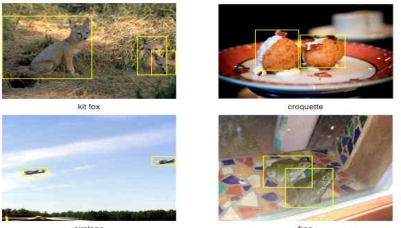
- 일정한 구조를 가지지 않는 데이터 (unstructured data)
  - 현대 잉공지능 분야의 대다수의 연구들이 분석하고자 하는 유형

- 텍스트 데이터 (자연어 처리)
- 이미지 데이터 (컴퓨터 비전)
- 비디오 데이터
- 음성(발화) 데이터
- 시퀀스 데이터 (예: 이동경로)
- ...

3. 비정형 데이터

❑ ImageNet

- 약 천 오백만 장의 이미지 구성
- 다양한 Task를 수행 가능하도록 구성
  - 고해상도 이미지
  - Object의 속성(spotted white, wet, etc…)
  - Bounding box
- 다운로드에 저작권 동의가 필요 (연구목적으로만 활용 가능)
- 공식사이트: https://image-net.org/



3. 비정형 데이터

❑ WikiText-103

- **Wikipedia 페이지에서 추출한 약 1억개의 토큰이 있는 데이터 셋**

- **예시 문제: 문장의 앞부분이 주어졌을 때 다음에 올 단어 맞히기**
  - 2022년 기준 Perplexity 10.81

3. 비정형 데이터

❑ coco

- 마이크로소프트에서 만든 약 33만개의 이미지로 구성되어있는 데이터
- 이미지에 있는 사물/동물의 이름, 위치, 이미지의 설명(caption)이 존재

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

4. 데이터 특성의 종류

- **범주형 변수(categoriacl variable)**
  - 이산적인(discrete) 값을 가지는 특성
  - 나라 이름, 합격여부, 계급, 나이(나이대로 구분된), ...

- **연속형 변수(conriacl variable)**
  - 연속적인 값을 가지는 특성
  - GDP, 점수, 나이, …

5. 범주형 변수

- **순서가 없는 범주형 변수(nominal variable)**
  - 이름과 같이 순서가 없는 값들
  - 유럽의 국가 이름: 독일, 이탈리아, 프랑스, ...
  - 우리나라의 행정 구역: 강원도, 경기도, 충청도, ...
  - 편의를 위해 특정 순서(예:가나다순)로 정렬하지만 실제 값에는 내재적인 순서가 없다.

- **순서가 있는 범주형 변수(ordinal variable)**
  - 계급: 소위, 중위, 대위, ...
  - 옷의 크기: S, M, L, XL, XXL, ...

6. 데이터 확보 전략

- **양질의 데이터를 획득하는 것이 중요하므로, 수집 목적과 데이터 종류에 대해 구체적인 계획을 세우는 것이 좋다.**
  - 어떤 곳에서, 어떤 방식으로, 어떤 주기로, 어떤 식으로 저장하고, 어떤 형태의 데이터인지

- 수집된 날 것의 데이터(raw data)는 분석을 위해 수정이 필요한 경우가 많다. 이러한 수정 과정을 **데이터 정제**(data cleaning 또는 data wrangling)라고 한다.
  - 불필요한 값의 제거
  - 결손치 보정
  - 표현의 일관화("2022년 1월 1일" vs "20220101")

7. 데이터 활용 전략

- **딥 러닝에서는 일반적으로 큰 데이터가 선호된다.**

- **데이터의 행의 개수는 최대한 많으면 좋다.**
  - 테이블에서 행의 개수, 이미지 개수 등

- **테이블 데이터에서 특성의 개수는 너무 많을 경우 문제가 된다.**
  - Overfitting의 위험성 존재
  - 모델이 복잡해짐
  - 중요한 특성만 남기자: 특징 추출(feature selection)
  - 특성을 합쳐 수를 줄이자: 차원 축소(dimensionality reduction)

8. 공개 데이터

- **AI Hub:** https://aihub.or.kr/
- **공공데이터포털:** https://www.data.go.kr/
- **Kaggle Datasets:** https://www.kaggle.com.datasets/
- **Google Dataset Search:** https://datasetsearch.research.google.com
- **Amazon Datasets:** https://registry.opendata.aws/
- **Datasets in Paper with Code:** https://paperswithcode.com/datasets

감사합니다.