⌨️

데이터 전처리, 특성 공학, 특성 학습

신경망에 입력 데이터와 타깃 데이터를 주입하기 전에 어떻게 준비해야 할까?

많은 데이터 전처리와 특성 공학 기법은 특정 도메인에 종속적이다. 여기서는 모든 종류의 데이터에 공통되는 기본 사항을 살펴볼 것이다.

신경망을 위한 데이터 전처리

데이터 전처리 목적은 주어진 원본 데이터를 신경망에 적용하기 쉽도록 만드는 것이다. 벡터화, 정규화, 누락된 값 다루기, 특성 추출 등이 포함된다.

벡터화

신경망에서 모든 입력과 타깃은 부동 소수 데이터로 이루어진 텐서여야 한다. 사운드, 이미지, 텍스트 등 처리해야 할 것이 무엇이든지 먼저 텐서로 변환해야 한다. 이 단계를 **데이터 벡터화**라고 한다.

예를 들어 이전에 나온 2개의 텍스트 분류 예에서 텍스트를 정수 리스트로 변환했다. 그 다음 원-핫 인코딩을 사용하여 float32 타입의 데이터로 이루어진 텐서로 바꾸었다.

값 정규화

숫자 이미지 분류 예에서 이미지 데이터를 그레이스케일 인코딩이나 0~255 사이의 정수로 인코딩 했다. 이 데이터를 네트워크에 주입하기 전에 float타입으로 변경하고 255로 나누어서 최종적으로 0~1 사이의 부동 소수 값으로 만들었다. 주택 가격을 예측할 때는 특성들의 범위가 제각각이었다. 이 데이터들을 네트워크에 주입하기 전에 각 특성을 독립적으로 정규화하여 평균이 0이고 표준 편차가 1이 되도록 만들었다. 일반적으로 비교적 큰 값, 균일하지 않은 데이터를 신경망에 주입하는 것은 위험하다. 이렇게 하면 업데이트할 그래디언트가 커져 네트워크가 수렴하는 것을 방해한다. 네트워크를 쉽게 학습시키려면 데이터가 다음 특징을 따라야한다.

- 작은 값을 취한다. 일반적으로 대부분의 값이 0~1 사이여야 한다.
- 균일해야 한다. 즉 모든 특성이 대체로 비슷한 범위를 가져야 한다.

- (추천) 각 특성별로 평균이 0이 되도록 정규화
- (추천) 각 특성별로 표준 편차가 1이 되도록 정규화한다.

넘파이 배열에서 하는 방법은 간단하다.

```
x -= x.mean(axis=0) # x가 크기인 2D 행렬이라고 가정
x /= x.std(axis=0)
```

누락된 값 다루기

이따금 데이터에 값이 누락된 경우가 있다. 일반적으로 신경망에서 0이 사전에 정의된 의미 있는 값이 아니라면 누락된 값을 0으로 입력해도 괜찮다. 네트워크가 0이 누락된 데이터를 의미한다는 것을 학습하면 이 값을 무시하기 시작할 것이다.

테스트 데이터에 누락된 값이 포함될 가능성이 있다고 가정하자. 하지만 네트워크가 누락된 값이 없는 데이터에서 훈련되었다면 이 네트워크는 누락된 값을 무시하는 법을 알지 못한다. 이런 경우에는 누락된 값이 있는 훈련 샘플을 고의적으로 만들어야 한다. 훈련 샘플의 일부를 여러번 복사해서 테스트 데이터에서 빠질 것 가은 특성을 제거한다.

특성 공학

모델이 수월하게 작업할 수 있는 어떤 방식으로 데이터가 표현될 필요가 있다.

예를 들어 시계의 시간을 읽기 위한 특성공학이라면, 여러 방법으로 문제를 쉽게 표현할 수 있다.

특성공학의 핵심은 특성을 더 간단한 방식으로 표현하여 문제를 쉽게 만드는 것이다.

최근 딥러닝엔느 대부분 특성 공학이 필요하지 않다. 신경망이 자동으로 원본 데이터에서 유용한 특성을 추출할 수 있기 때문이다. 심층 신경망을 사용할 때는 특성 공학에 대해 신경 쓰지 않아도 될까? 아니다.

- 좋은 특성은 적은 자원을 사용하여 문제를 더 멋지게 풀어낼 수 있다. 예를 들어 시계 바늘을 읽는 문제에 합성곱 신경망을 사용하는 것은 어울리지 않는다.
- 좋은 특성은 더 적은 데이터로 문제를 풀 수 있다. 딥러닝 모델이 스스로 특성을 학습하는 능력은 가용한 훈련 데이터가 많을 때 발휘된다. 샘플의 개수가 적다면 특성에 있는 정보가 매우 중요해진다.