영화 관객수 예측

20163146 이진주

날짜

1.데이터 불러오기

Pandas 모듈을 이용해 csv 파일을 불러온다.

2.데이터 전처리(필드 수정, 제거)

- **필드 수정 및 제거(A)**
  로드한 영화 데이터 중 상영시점, 개봉 4 주 전부터 1 주 전 트위터 개수와 개봉 1 주 전부터 개봉 전까지의 트위터 개수, 관객수 필드를 제외한 나머지 필드를 제거한다.

- **2 개 이상의 필드 조합 후 필드 이름 변경 (B)**
  개봉 4 주 전부터 1 주 전 트위터 개수와 개봉 1 주 전부터 개봉 전까지의 트위터 개수 필드를 합친 뒤 필드의 이름을 tweet_count 로 변경해주었다.

- **입력 데이터 수치화 및 표준화 처리(C,D)**
  개봉시점(screen time), tweet_count 필드의 정수형 데이터를 minimum 값과 maximum 값을 구하여 0~1 사이의 값으로 표준화 처리를 해주었다

3.모델 생성

학습 모델은 keras 를 이용하여 input 차원은 합친 트위터의 개수와 개봉시점 값을 이용하였고 활성함수는 relu 를 사용하여 총 2 개의 layers 를 통해 학습시켰다. Loss 는 mse(mean squared error), 가중치 최적화를 위한 최적화 함수로는 adam 을 사용했다.

입력 X 는 개봉영화에 대한 트위터 개수를 입력으로 주고 결과값은 관객수이며 보다 정확한 예측을 위해 정확도를 확인하는 test 데이터를 분류하기 위해서 test 데이터는 총 데이터 수의 20%를 test 데이터로 분류하였다. 한 step 당 12 개의 데이터에 대해서 총 반복 수 100 으로 학습한다.

4.학습 결과

```
Epoch 1/100
96/96 [==============================] - 1s
5ms/step - loss: 0.0375
...
Epoch 100/100
96/96 [==============================] - 0s
114us/step - loss: 0.0053
```

Loss 가 점차 줄어드는 것을 확인할 수 있다. (step 1 과 step 100 에 대한 결과)

Matplolib 을 이용해 Loss 에 대한 그래프를 출력한 결과는 다음과 같다.