인공지능 과제

딥러닝 오픈소스 기반 협력추천

컴퓨터전공

2013011457 송지수

1. 평점 예측 모델

주어진 데이터 중 오직 User ID와 Movie ID, 그리고 rating 만을 Input으로 사용하였다. 장르 등 영화의 Metadata를 이용할 경우, 해당 데이터의 Content가 오히려 유저의 취향을 온전히 분석하는 데 방해가 될 가능성이 있다. 따라서 그런 가능성을 배제하고 순수한 Collaborate Filtering을 구현하고자 추천에 꼭 필요한 데이터만을 Input으로 사용하기로 결정했다.

Backpropagation Neural Network 를 사용하여 모델을 구현하였다. Neural Network 를 구축하는 데 중요하게 사용된 파라미터들은 다음과 같다.

Batchsize = 50

nEpoch = 2

Learning Rate = 0.0001

Layer : 7층 (Input, Output, Hidden Layer 포함)

Node : Input Layer 2개, Output Layer 5개, Hidden Layer 최대 40개

nEpoch와 Layer의 층수, 개수 등을 달리하여 모델을 학습시키고, 그 결과를 확인하여 해당 파라미터 값을 결정하였다. nEpoch 가 일정 수준 이상으로 높아질 경우 예측된 평점의 범위가 좁아지는 등의 Overfitting 문제가 발생했다. (nEpoch = 30 시 평점 예측 범위 3.8~4.7, nEpoch = 10 시 평점 예측 범위 3.4~4.8) 따라서 적당한 수준으로 nEpoch의 값을 결정했다.

Backpropagation Neural Network를 학습시킬 때에 Hidden Layer 층수와 Node 개수를 어떻게 하는 것이 제일 적합한지는 알 수 없지만, 일반적으로 적어도 각각 Training Data의 Input으로 주어진 변수의 개수들보다는 많아야 학습이 제대로 진행된다는 것이 정설로 알려져 있다. DL4J의 예제로 주어진 MLPLinearClassifier 의 경우 Output Layer를 포함해 Layer가 총 2개였고, 이것을 그대로 Epinions Data에 적용했더니 학습이 제대로 되지 않는 것을 볼 수 있었다(학습 데이터가 달라도 똑같은 결과를 출력). 따라서 Output Layer를 포함한 전체 Layer 층수를 최소한 5개로 잡고, Layer 층수와 각 Hidden Layer의 층에 포함된 Node의 개수를 달리하며 모델을 학습시켰다.

Hidden Layer의 층수와 각 Layer에 포함된 Node의 개수가 이상으로 너무 많아질 경우 또한Overfitting이 되는 것으로 추정되어, 적절하게 학습이 되는 값으로 파라미터를 결정하였다.

1. 모델의 결과 출력

모델은 최종 결과로 Input Data (User ID, Movie ID) 에 따른 5개의 Double형 소수가 담긴 리스트를 리턴한다. 이는 각각 1점, 2점, 3점, 4점, 5점으로 예측될 확률이며, 따라서 최종 Rating은 각 Double형 소수에, 각 데이터에 할당된 점수 가중치를 곱하여 계산하였다.

(예시: [0.03, 0.05, 0.15, 0.30, 0.47] 이 출력되었을 경우 최종 예상 Rating은 1\*0.03 + 2\*0.05 + 3\*0.15 + 4\*0.30 + 5\*0.47 = 4.13)

1. 추가 구현한 함수, 사용한 오픈소스

createNewCSVFile 이라는 함수를 추가로 구현하였다. 사용하지 않는 텍스트 데이터를 제거하기 위해, CSVReader로 파일을 읽어 필요한 부분만 빼어 새로 파일을 작성하고 그 파일을 이용하여 Training과 Testing을 수행한다. 또한 이 함수를 구현하기 위해서 openCSV라는 오픈소스 CSV 파싱 라이브러리를 사용하였다. Pom.xml 파일에 정보를 추가해 뒀으며 Maven에 의해 자동으로 Import될 것이다.