**김형원\_계획**

**1. 이미지와 text정보들을 이용한 쇼핑카테고리 분류AI (네이버 쇼핑 검색 플랫폼 개발) 요약**

**구현 과정**

1. **텍스트 데이터 처리: Word2Vec스크린샷, 모니터, 시계, 앉아있는이(가) 표시된 사진

   자동 생성된 설명스크린샷, 사진, 앉아있는, 모니터이(가) 표시된 사진

   자동 생성된 설명**
2. **카테고리 분류 모델**

스크린샷이(가) 표시된 사진

자동 생성된 설명

1. **다중 신경망 모델 간단한 구조 코드**

**스크린샷이(가) 표시된 사진

자동 생성된 설명**

1. **TEXT 학습 모델로 TEXT CNN을 사용**

**스크린샷이(가) 표시된 사진

자동 생성된 설명**

1. **이미지 처리 신경망은 MobileNETV2 사용**

**스크린샷이(가) 표시된 사진

자동 생성된 설명**

**(이하 생략)**

**\*추가사항**

**스크린샷이(가) 표시된 사진

자동 생성된 설명**

출처: 이미지와 text정보들을 이용한 쇼핑카테고리 분류AI (네이버 쇼핑 검색 플랫폼 개발)

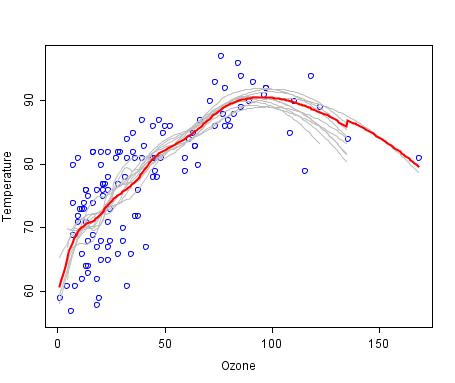
**ML에서의 부트스트랩 (Bootstrap in machine learning)**

이 “bootstrapping”은 ML에서도 사용되며, 이는 랜덤 샘플링을 통해 training data를 늘리는 방법이다.

“bootstrapping”이 사용되는 경우로 **데이터 셋(training set) 내의 데이터 분포가 고르지 않은 경우**를 말할 수 있다. 사과와 오렌지를 구분하는 classifier를 트레이닝한다고 하자. 만약 training set에 사과 이미지 1만장과 오렌지 이미지 100장이 포함되어 있다면, 항상 사과만 찍는 멍청한 classifier도 99%의 트레이닝 정확도를 보일 것이다. 이렇게 균형이 맞지 않은 상황에서는 데이터가 적은 클래스의 error는 무시되는 방향으로 트레이닝되기 쉽다. 이를 해결하기 위한 방법에는 세 가지가 있는데, 첫 번째는 weight를 줄 수 있는 알고리즘을 사용하는 것이고, 두 번째는 “bootstrapping”을 이용해서 오렌지의 데이터 수를 늘리는 것이며, 마지막은 사과 데이터의 수를 역으로 줄이는 것이다.

“bootstrapping”은 **over-fitting을 줄이는 데**에도 도움이 되며, 이를**“bagging”**이라고 한다. Over-fitting을 줄이기 위해서는 데이터가 많은 게 제일 좋지만 그게 마땅치 않을 때가 많다. 그렇다고 model의 complexity를 줄이는 건 그건 역으로 정확도에 악영향을 줄 수 있다. 만약 계산 시간이 크게 문제가 되지 않는 경우라면, 여러 개의 모델을 써서 model ensemble(여러 모델을 통해 결과를 뽑고, 그 값들을 평균하거나 voting하여 최종 값을 산출하는 방법)을 통해 모델의 안정성을 크게 높일 수 있다. 각각의 모델이 over-fitting 되어있더라도, 그들을 평균내면 서로서로 상쇄되어 더 general한 모델이 되기 때문이다.

“bootstrapping”을 이용하면 주어진 데이터가 충분하지 않아도 model ensemble을 만들 수 있다. 똑같은 알고리즘을 통해 m 번 학습시킨다고 하자. 이 때 매번 training data를 random sampling하면, 서로 다른 m개의 모델이 만들어진다. 각각의 모델은 학습 과정에서 사용된 데이터에 over-fitting 되어 있겠지만, m개 전체를 사용하여 결정을 내리면 over-fitting 걱정이 크게 줄어들 것이다. 이에 대한 추가적인 내용은 [Wikipedia:Bootstrap\_aggregating](https://en.wikipedia.org/wiki/Bootstrap_aggregating)을 참조할 것.

회색으로 표시된 각각의 model은 overfit되어 있지만, 그들의 평균으로 이루어진 빨간선은 더 안정적이다. (출처: Wikipedia/Bootstrap\_aggregating)

이 외에도, “bootstrapping”을 단일 모델의 성능을 높이기 위한 학습 방법으로도 사용할 수 있다. 모델을 정확도를 높이기 위해서는 같은 데이터를 반복해서 학습시키는 것이 필요한데, 단순히 전체 데이터 셋을 트레이닝하는 것은 시간이 오래 걸리고 비효율적이다. 충분히 구분히 잘 되는 샘플을 쓸데없이 여러번 학습시키는 데에 많은 시간이 소요되기 때문이다. 이럴 때, “bootstrapping”을 이용하여 각 iteration에 사용되는 학습 데이터의 갯수를 줄이고, 어려운 샘플에 큰 비중을 둬서 학습하도록 조절할 수 있다. “Bootstrapping”을 통하여 학습시키는 방법은 아래와 같다.

1. 전체 training sample 중에서 n개를 추출하여 모델을 학습시킨다.  
   (이 때 n은 원래 학습 데이터의 크기보다 작아도 된다)
2. 학습된 모델을 이용하여, training sample을 분류(classify)한다.
3. 잘못 분류된 학습 데이터가 선택될 probability를 높이고, 제대로 분류된 데이터가 선택될 probability를 낮춘다. 이렇게 하면 다음 번 학습 시 “어려운” 샘플의 비율이 커진다.
4. 다시 1번으로 돌아가 학습하기를 반복한다.

출처: <https://learningcarrot.wordpress.com/2015/11/12/%EB%B6%80%ED%8A%B8%EC%8A%A4%ED%8A%B8%EB%9E%A9%EC%97%90-%EB%8C%80%ED%95%98%EC%97%AC-bootstrapping/>