5주차_머신러닝 어플리케이션 설정하기

강의 딥러닝 2단계

링크 https://velog.io/@pehye89/Euron-5주차-머신러닝-어플리케이션-설정하기

Add a property

Train/Dev/Test Set

훈련, 개발, 테스트 세트를 어떻게 설정하냐에 따라 네트워크의 성능이 달라진다

> 💡 신경망을 훈련시킬 때 설정해야하는 것
>
> - 신경망 층의 개수
> - 은닉층의 개수
> - 학습률
> - 활성화 함수

처음에는 어떤 값이 좋을지 추정하기 어렵기 때문에, 테스트를 여러 번 반복하고 아이디어를 수정해서 특정 데이터에 맞는 코드를 찾는다.

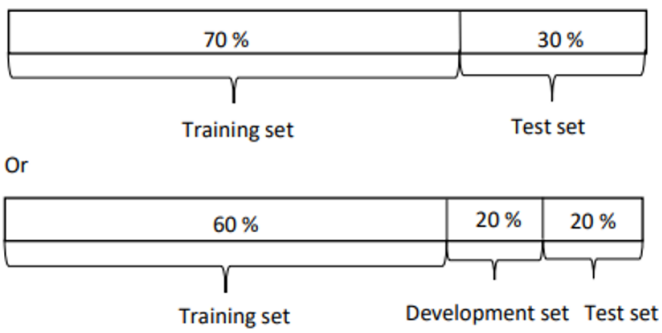
현재 딥러닝은 NLP, 컴퓨터 비전, 음성인식, 그리고 다른 광고, 검색, 보안이나 물건 배송과 같은 구조화된 데이터에 적용된 다양한 어플리케이션 등 여러 분야에 적용되고 있다. 같은 딥러닝이어도, 다른 데이터에서 같은 하이퍼 파라미터에 대한 직관이 공유되지 않는다. 그렇기 때문에 아무리 경험이 많은 사람이라도 첫 모델부터 좋은 하이퍼파라미터를 찾는 것은 쉽지 않다.

그래서 빠른 진전을 이루기 위해 중요한 것은, (1) 이 사이클을 얼마나 효율적으로 돌 수 있는지와 (2) 얼마나 데이터 세트를 잘 설정하는지다.

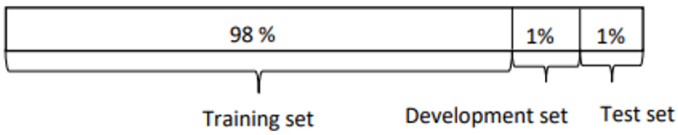
훈련-개발-테스트 데이터의 비율

데이터가 있으면, 이 데이터의 대부분을 훈련 세트, 일부분을 개발과 테스트 세트로 설정한다.

현재보다 상대적으로 적은 데이터, 약 1000~10000개의 데이터를 다뤘던 머신러닝 이전의 시대에는 훈련-개발 세트를 70-30의 비율, 또는 훈련-개발-테스트 데이터를 60-20-20의 비율로 나누곤 했다.



그러나 100만개의 데이터가 넘어가는 현재의 머신러닝에서는 개발과 테스트 세트의 비율을 훈련 세트보다 훨씬 적은 비율로 나누고 있다.



왜냐하면 개발 세트의 목적은 단순히 모델을 테스트해서 수정하기 위한 것이고, 테스트 데이터의 목적은 모델의 성능을 수정하기 위한 것이기 때문에, 만약 100만개의 데이터가 있으면 각각 1만의 데이터만 사용해도 모델을 테스트하는 데 큰 문제는 없을 것이다. (해당 예시의 비율을 살펴보면 98-1-1의 비율이다) 이미지 출처

일치하지 않은 훈련/테스트 분포

또한 현재 딥러닝의 트렌드는 더 많은 사람이 일치하지 않은 훈련-테스트 분포에서 훈련한다는 것이다.

사용자가 사진을 업로드하는 앱을 개발한다고 한다. 훈련 세트를 온라인에 업로드된 사진으로 사용하고, 개발/테스트를 앱으로 찍은 사진을 사용한다고 한다. 이런 경우, 이 둘의 데이터가 같은 분포에서 와야 좋은 성능을 갖는 모델이 될 것이다.

하지만 딥러닝의 경우 데이터가 많으면 많을수록 좋기 때문에, 같은 분포를 따르지 않더라도 온라인에서 크롤링과 같은 창의적인 방법을 활용해서 더 많은 데이터로 훈련하기도 한다.

물론 같은 분포를 가진 데이터라면 더 효율적일 것이다.

테스트 세트의 유무

> 💡 테스트 세트의 목적은 네트워크의 성능에 대한 비편향적인 추정을 제공하기 위함이다.

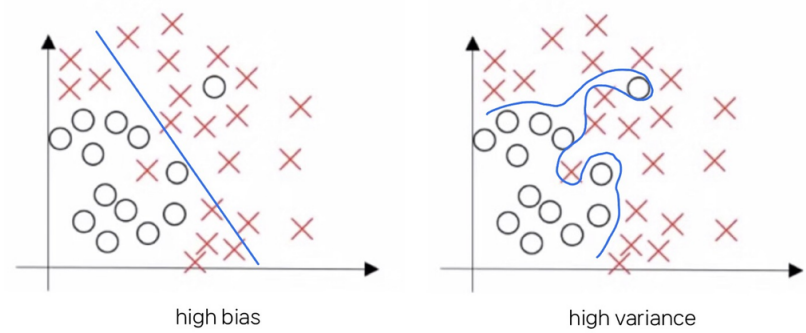
그렇기 때문에 만약 성능에 대한 비편향적인 추정이 필요 없을 때는, 테스트 세트가 없어도 괜찮다. 테스트 세트 대신, 개발 세트에서 모델을 수정한다. (모델을 개발 데이터에서 수정하기 때문에 비편향적인 추정은 불가능하다)

편향과 분산

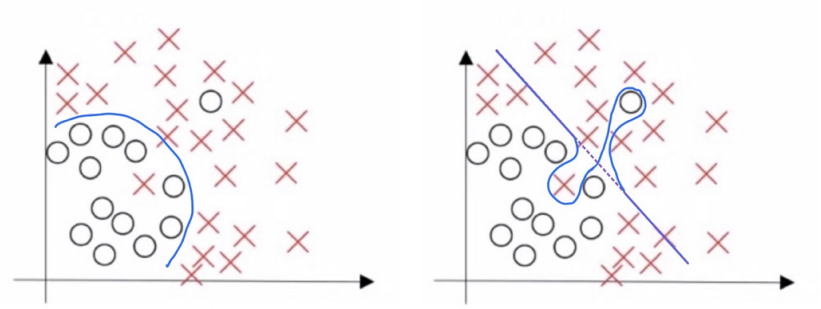
훈련 세트와 개발 세트의 알고리즘 오차를 살펴봄으로써, 높은 편향을 갖는 경우, 높은 분산을 갖는 경우 또는 둘 다 높거나 낮은 경우를 진단할 수 있다.

고양이와 개 분류 모델

인간 수준의 오차가 0%라는 가정을 했을 때,



1. **높은 분산을 갖는 경우**
    - 훈련 세트 오차가 1%, 개발 세트 오차가 11%
    - 훈련 세트에서는 잘 분류했지만, 훈련 세트에서 과대적합되어 개발 세트에서 일반화하지 못한 경우라고 볼 수 있다.
2. **높은 편향을 갖는 경우**
    - 훈련 세트 오차가 15%, 개발 세트 오차가 16%
    - 인간은 대략 0%의 오차를 낸다고 가정했을 때, 이 알고리즘은 데이터에서 과소적합되어, 훈련 세트와 개발 세트에서 모두 좋은 성능을 갖지 않는다는 것을 알 수 있다,
    - 하지만 훈련 세트와 개발 세트를 비교했을 때는 차이가 거의 나지 않기 때문에 합리적인 수준의 개발세트에서 일반화되고 있다.



low bias and low variance　　high bias and high variance

3. **높은 분산과 편향을 갖는 경우** (🌟최악의 경우)
    - 훈련 세트 오차가 15%, 개발 세트 오차가 30%
4. **낮은 분산과 편향을 갖는 경우** (🌟최선의 경우)
    - 훈련 세트 오차가 0.5%, 개발 세트 오차가 1%

이 분석은 인간 수준의 분류 오차가 0%에 가깝다는 가정을 하고 있다. 더 일반적으로는 베이 오차가 0%에 가깝다는 가정을 하고 있다. 하지만 만약 이 베이즈 오차가 15%일 경우, 저 2번째 경우는 낮은 분산과 편향을 갖고 있는 모델이 좋은 모델임을 알 수 있다.

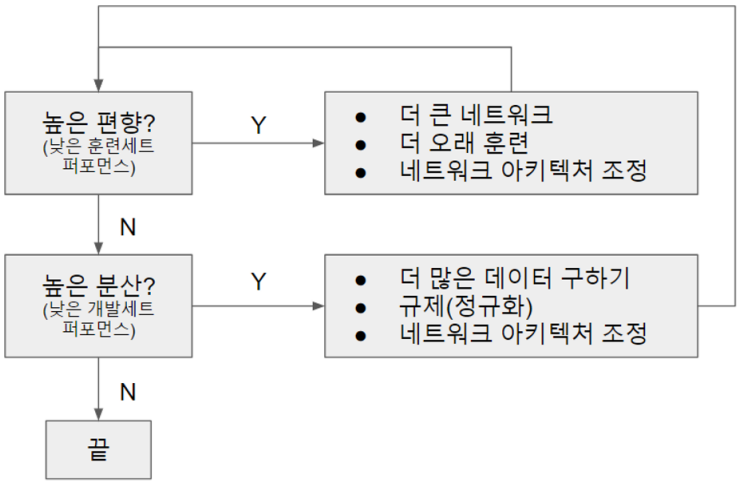
> 📌 **베이즈 에러**는 모든 기계 학습에서 가능한 이론적 최소 오차로, 어떤 알고리즘이나 모델이 베이즈 에러에 비해 얼마나 오차가 큰지를 비교하는 경우가 많다. — 출처

머신러닝을 위한 기본 레시피

신경망 모델을 구축하고 나서, 더 좋은 모델을 위해 아래의 질문들을 해보며 모델을 업데이트 시키는 것을 반복할 수 있다.

1. 높은 편향을 갖고 있나?
    - 높은 편향을 확인하기 위해 훈련 데이터에서의 성능을 확인해 본다.
    - 더 큰 네트워크, 즉 더 많은 은닉층이나 은닉 유닛을 더하기 (항상 도움이 된다)
    - 더 오랜 시간 훈련시키기 (항상 도움이 된다고 할 수는 없지만, 해가 되지는 않는다)
    - 다른 발전된 최적화 알고리즘을 사용하기
    - 작동하지 않을 수 있지만, 더 다른 신경망 아키텍처를 사용하는 것

2. 높은 편차를 갖고 있나?
    - 높은 편차를 확인하기 위해 개발 데이터에서의 성능을 확인해 본다.
    - 더 많은 데이터를 얻기
    - 과대적합을 줄이기 위해 정규화하기
    - 작동하지 않을 수 있지만, 더 다른 신경망 아키텍처를 사용하는 것



신경망에 어떤 문제가 있느냐에 따라서 특정 방법이 효율적이지 않을 수 있다.

예를 들면 만약 모델이 높은 편향을 갖고 있다면, 더 많은 데이터를 더하는 것이 모델의 성능을 높이는 데 큰 도움이 되지 않을 수 있다는 것이다.

그렇기 때문에 높은 편향을 갖고 있는지, 또는 높은 편차를 가졌는지 진단하여 가장 유용한 시도를 선택해 집중할 수 있게 하는 것이 중요하다.

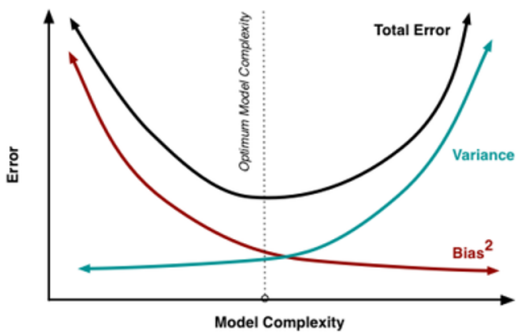
편향-분산 트레이드오프

딥러닝 초기에는 편향과 분산 중 하나만을 낮출 수 있는 방법이 없었다. 편향을 낮추면 분산이 올라가기 때문에 "트레이드오프"를 해야 하는 상황이 많았던 것이다.

하지만 빅데이터의 시대인 현재에는, 네트워크의 크기와 데이터의 크기가 충분히 크기 때문에 이 문제가 많이 줄어들었다. 만약 네트워크와 데이터의 크기가 충분히 크면, 아무리 더 많은 은닉층을 더하거나 데이터를 추가해도 트레이드오프를 할 필요 없이 편향/분산 중 하나만 줄어든다. 또한 정규화를 시킬 경우 편향이 높아질 수 있지만, 충분히 큰 네트워크를 사용한다면 크게 문제가 되지 않는 경우가 대부분이다.

편향과 분산이란?

우리가 무언가를 학습시킨 뒤 예측할 때 그로 인한 오차가 발생하기 마련인데 이때 발생하는 세 가지 두 가지 오차가 바로 bias와 variance이다. 쉽게 말해 그냥 오차의 유형이다. 그리고 이 둘은 trade-off 관계가 있어서 시소처럼 한쪽이 올라가면 한쪽이 내려가는 관계다.



- Bias (편향) 에러가 높아지는 것은 과소적합한 상황으로, 많은 데이터를 고려하지 않아(=모델이 너무 단순해) 정확한 예측을 하지 못하는 경우를 말한다.
- Variance (분산) 에러는 과대적합한 상황으로, 노이즈까지 전부 학습해(=모델이 너무 복잡해) 약간의 input에도 예측 Y 값이 크게 흔들리는 것을 말한다.

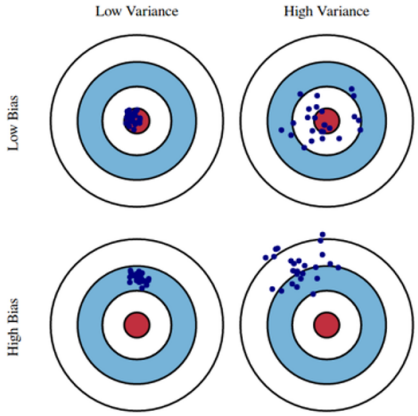


Fig. 1 Graphical illustration of bias and variance.

출처 — https://datacookbook.kr/48