

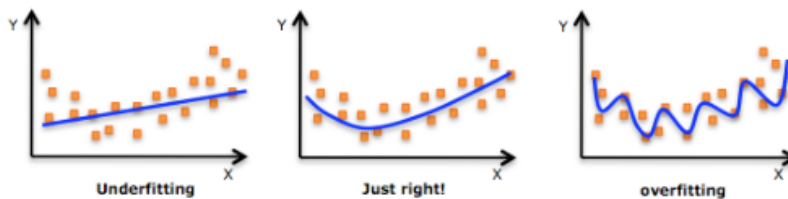
Overfitting

즉, 함수 모델이 주어진 데이터를 너무 자세히 본떠서 만들어진 현상

Overfitting이란 문자 그대로 너무 과도하게 데이터에 대해 모델을 learning을 한 경우를 의미한다. 현재에 대해 잘 설명하는 것만으로 충분하지 않을까라고 생각할 수 있지만, 우리가 사실 원하는 정보는 (기존에 알고 있는 데이터에 대한 것들이 아니라) 새롭게 우리가 알게되는 데이터에 대한 것들을 알고 싶은 것인데, 정작 새로운 데이터에 대해서는 하나도 못맞추고, 즉 제대로 설명할 수 없는 경우라면 그 시스템은 그야말로 무용지물이라고 할 수 있을 것이다. 이전에 [적은 포스트](#)에서도 간략하게 다루고 있다.

Regularization

Overfitting을 어떻게 해결할까 고민하기에 앞서 먼저 Overfitting이 일어나는 이유는 무엇인가에 대해서 한 번 생각해보자. 먼저 overfitting의 가장 간단한 예시를 하나 생각해보자.



위 그림에서도 알 수 있 듯, 만약 우리가 주어진 데이터에 비해서 높은 complexity를 가지는 model을 learning하게 된다면 overfitting이 일어날 확률이 높다. 그렇다면 한 가지 가설을 세울 수 있는데, 'complexity가 높을 수록 별로 좋은 모델이 아니다.' 라는 가설이다. 이는 Occam's razor, 오컴의 면도날이라 하여 문제의 solution은 간단하면 간단할수록 좋다는 가설과 일맥상통하는 내용이다. 하지만 그렇다고해서 너무 complexity가 낮은 model을 사용한다면 역시 부정확한 결과를 얻게 될 것은 거의 자명해 보인다. 그렇기 때문에 우리는 원래 cost function에 complexity와 관련된 penalty term을 추가하여, 어느 정도 '적당한' complexity를 찾을 수 있다. 이를 regularization이라 한다. 이 이외에도 다양한 설명이 있을 수 있기에 [위키 링크](#)를 첨부한다.