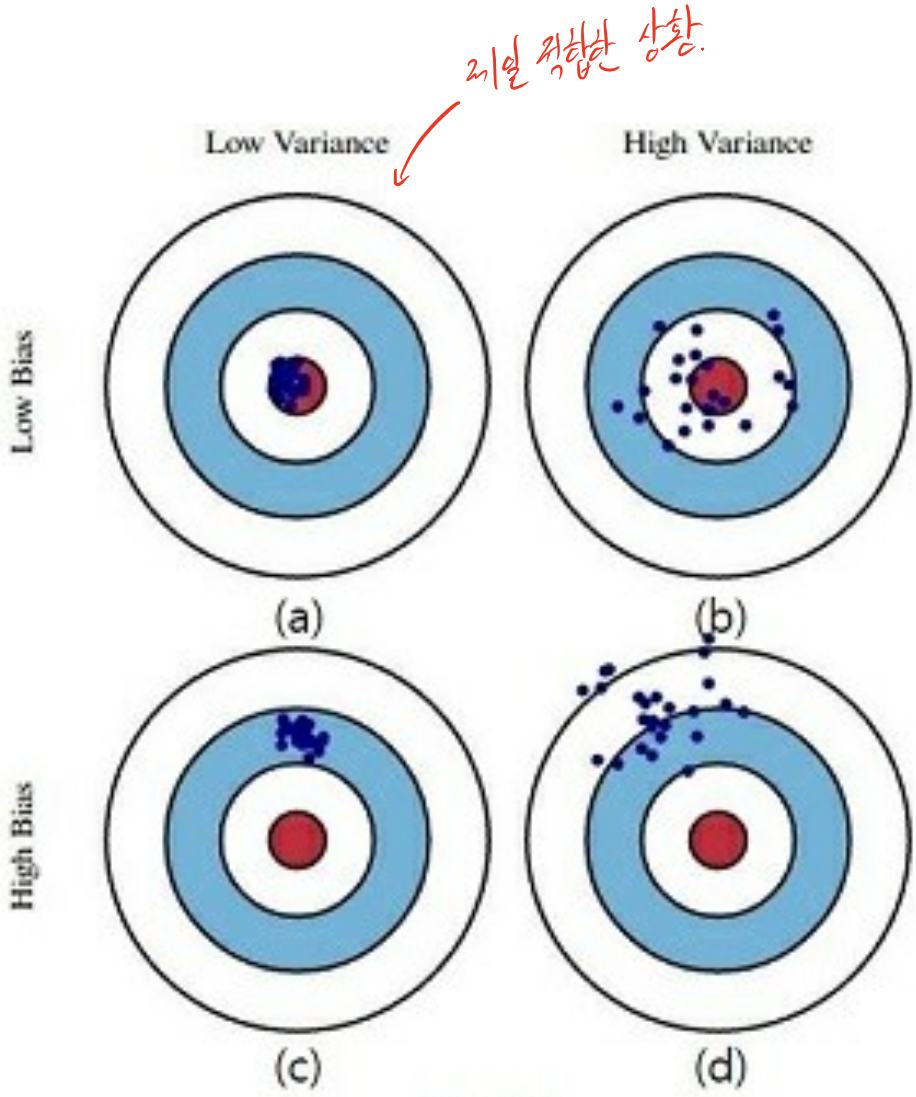
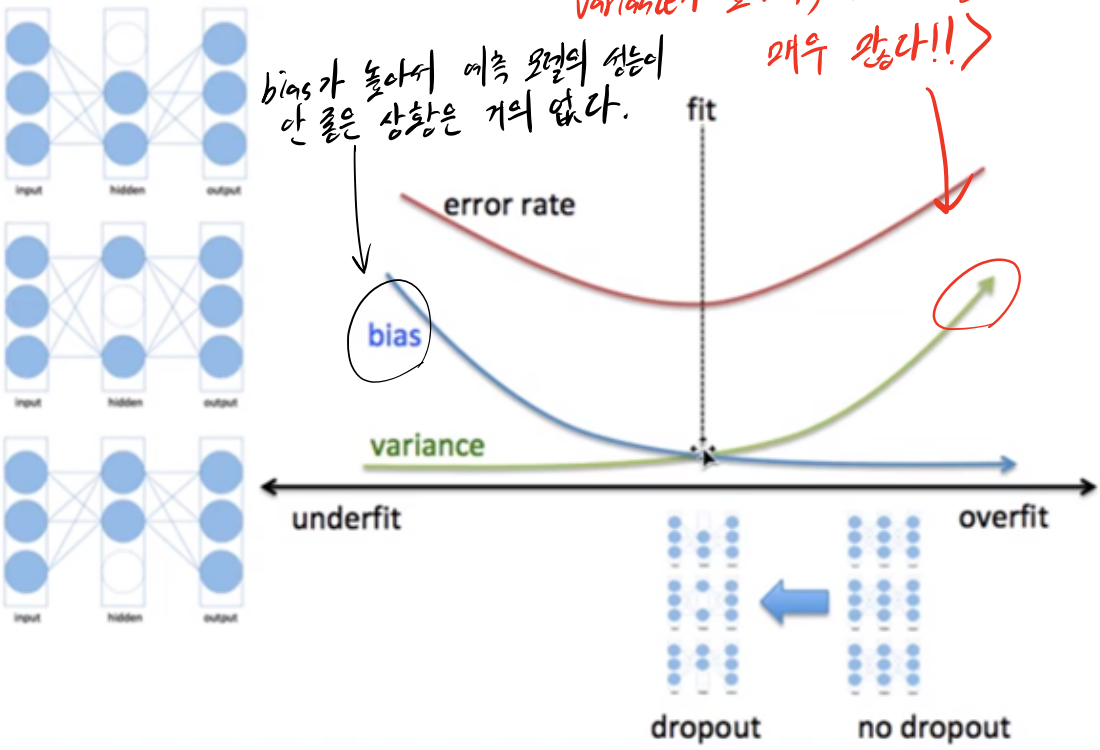


그림 1

Overcome Overfitting!



Bias

여기서 말하는 편향은 회귀모델의 상수항이 아닙니다.

편향이란, y_pred의 평균과 y_true와의 관계입니다. 다르게 말하면 외부적으로 얼마나 영향을 받는지를 뜻합니다. 외부란 정답이라고 생각하면 쉽겠죠?

다르게 말해서, y_pred의 값들과 y_true의 값들이 떨어져 있는 정도가 클 경우를 '편향이 높다' 라고 표현합니다.

편향이 클 경우엔 정답값들과의 거리가 멀테니 이를 과소적합이라고 표현할 수 있습니다 - underfitting

$$Bias = (E[f^{pred}(x)] - f(x))^2$$

Variance

분산이란, 예측값들 간의 관계입니다. 즉, 밑의 식에 따라 예측값과 예측값들의 평균의 차이에 대한 평균입니다.

(=예측값 편차들의 평균)

즉, 예측값들끼리 얼마나 떨어져 있는가 입니다.

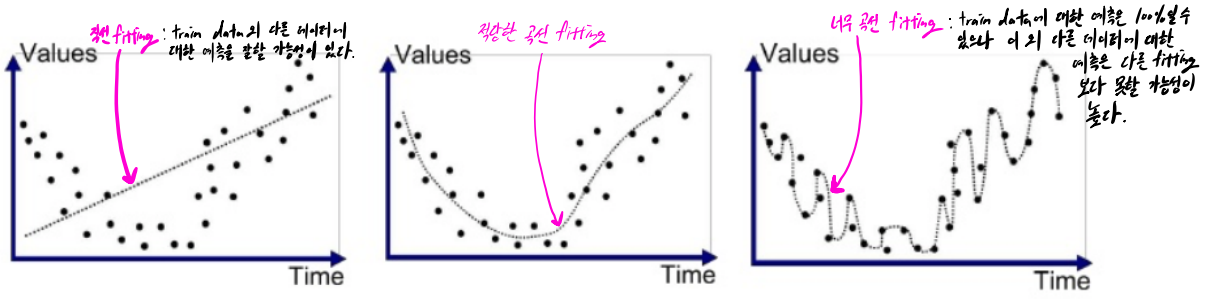
예측값들이 자기들끼리 떨어져 있는 정도가 클 경우를 '분산이 높다' 라고 표현합니다.

높은 분산을 가질 경우 과대적합이라고 표현할 수 있습니다. - overfitting

그리는 선의 종류가 구불구불하게 복잡해져 새로운 데이터를 예측하기가 쉽지 않기 때문이죠.

$$Variance = E[[f^{pred}(x) - E[f^{pred}(x)]]^2$$

↑ 예측값의 편차



high bias
low variance

medium bias
medium variance

low bias
high variance

실제로 첫번째 그림은 편향이 높고 분산이 낮습니다.

예측값들은 한 직선 위에 있으니 분산이 낮고, 데이터들이 모델과 멀어져 있으니 편향이 높습니다.

세번째 그림은 구불구불한 직선 위에 있으니 분산이 높고, 모델과 거리가 가까우니 편향이 낮게 됩니다.

실제 데이터를 다룰때는, 데이터가 충분히 많다는 가정하에 bias가 높아 underfitting으로 고생하는 연구자들은 잘 보지 못했지만, overfitting때문에 고생하는 분들은 상당수 보았습니다.

overfitting은 트레이닝시에 training error(acc)와 validation error(acc)가 함께 같은방향으로 진행되다가 갑자기 벌어지는 경우입니다.

X · 예측모델의 함수식이 복잡하게 되어있다 => 높은차수의 함수'를 의미함.
· 예측모델의 함수식이 복잡하게 되어있으면, 예측값들의 분산이 크다!