정답 하나를 맞추기 위해 컴퓨터는 여러 번의 예측값 내놓기를 시도하는데,
컴퓨터가 내놓은 예측값의 동태를 묘사하는 표현이 '편향' 과 '분산' 입니다.

① 예측값들과 정답이 대체로 멀리 떨어져 있으면 결과의 편향(bias)이 높다고 말하고,
② 예측값들이 자기들끼리 대체로 멀리 흩어져있으면 결과의 분산(variance)이 높다고 말합니다.

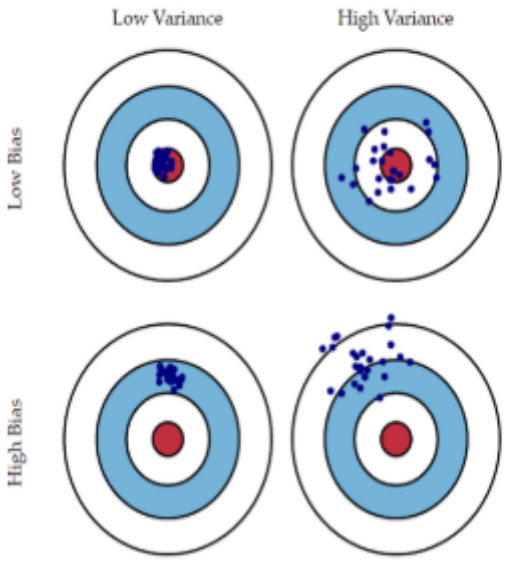
활쏘기로 비유를 들어 봅니다.
아래 그림을 보시면, 가운데 빨간 점이 사람이 정해준 '정답'이고
파랗게 여러 번 찍힌 점이 컴퓨터가 예측한 값들 입니다.

1. 왼쪽 상단 과녁은
   예측값들이 대체로 정답 근방에서 왔다갔다 합니다.->편향이 낮습니다.
   예측값들끼리 서로 몰려 있습니다.->분산이 낮습니다.

2. 오른쪽 상단 과녁은
   예측값들이 대체로 정답 근방에서 왔다갔다 합니다.->편향이 낮습니다.
   예측값들끼리 서로 흩어져 있습니다.->분산이 높습니다.

3. 왼쪽 하단 과녁은
   예측값들이 대체로 정답으로부터 멀어져 있습니다.->편향이 높습니다.
   예측값들끼리 서로 몰려 있습니다.->분산이 낮습니다.

4. 오른쪽 하단 과녁은
   예측값들끼리 대체로 정답으로부터 멀어져 있습니다.->편향이 높습니다.
   예측값들끼리 서로 흩어져 있습니다.->분산이 높습니다.



편향과 분산은, 머신러닝 모델이 '복잡하게 생긴 정도'와 큰 관련이 있습니다.

아래 세 그림은, 세 가지 서로다른 머신러닝 모델로 같은 데이터를 설명하는 모습입니다.
회귀(Regression) 모델인데요.
정답들은 '점'으로 찍혀 있고
모델이 내놓을 예측값은 직선 혹은 구불구불한 곡선으로 표현되어 있습니다.
이 문제에서는 여러 점들의 경향을 표현하는 모델을 찾는 것이 목적입니다.

첫 번째 그림을 보시면, 데이터들이 모델과 멀어져 있으므로 편향(bias)이 높고,
"모델이 내놓는 값"들 끼리는 별로 떨어져 있지 않게 되므로(왜냐면 같은 직선위의 점들이니까) 분산(variance)은 낮습니다. (=예측값)

세 번째 그림을 보시면, 정답들이 모델과 아주 붙어 있으므로 편향이 낮고,
모델이 내놓는 값들 끼리는 매우 흩어져 있게 되므로(왜냐면 구불구불한 선 위의 점들이니까) 분산이 높습니다.

두 번째 그림 정도가 어지간히 적당하다 고 볼 수 있죠.