

- 표본내 성능과 표본외 성능

- 보통회귀분석 모델의 성능 척도는 결정계수 (Coefficient of determine) 을 이용한다.
이런 성능은 표본내 성능검증 (In-sample testing) 이라 한다.

- 훈련모델 소집단에 학습데이터 외에 얻어진 학습데이터의 값을 얼마나 잘 예측할 수 있는지를 평가하는 것이 표본외 검증 (out-of sampling testing) 및 교차검증 (Cross Validation) 이라 한다.

- 과적합화 (over fitting)

- 일반적으로 표본내 성능과 표본외 성능은 비슷한 수준로 보이지만, 표본내 성능보다 표본외 성능이 현저하게 낮아질 수 있다. 이를 과적합화라 한다.
- 과적합화가 발생하면 표본내 데이터는 잘 예측하지만 새로운 표본데이터에 대한 예측력이 현저하게 떨어진다.

$y = ax + b$ 회귀모델이 있을 때,



데이터에 대한 예측이 잘된다면,



새로운 표본에 대한 예측이 잘된다면,

- 이를 방지하기 위해서 정규화 (regularization) 을 이용하며, 이를 달성하기 위해서는 교차검증을 진행해야 한다.

• 여기서 교차검증은 train 데이터와 test 데이터로 나누어 검증하는 것을 말한다.

- K 폴드 교차검증

- 데이터의 수(표본)가 적은 경우에는 검증데이터의 수를 적절하게 증가시키는 것이 필요하다.
이런때 K 폴드 교차검증은 효과적이다.

- 교차검증 방법 -