Overview

기계학습은 많은 최적 방법을 포함함.
이 강좌에서는 기계학습과 데이터 과학의 공통영역

기계학습은 사전에 결정된 특정 규칙을 따르는 것이 아니라, 데이터에 기반하여 의사결정과 행동하는 모델을 만드는 것임.

기계는 함수로 표현 가능 → $Y = f(X) + \epsilon$

Y의 유무에 따라 학습의 방법이 달라짐.
$f()$는 사전 정의되지 않고, X에 따라 찾아내야 함
X는 input 데이터. ϵ은 줄일 수 없는 노이즈임.

새로운 X에 대한 Y를 알 수 있을 때 $f()$를 먼저 추정해야 함.

지도학습에서는 주로 Y를 통해 $f()$를 모델링할 수 있음. $f(x)$로 결정된 수가 Y와 차이가 적도록 모델링 해야 함.

우리의 예측 함수를 최적화 하기 위한 식의 전개

$$E(Y-\hat{Y})^2 = E(Y - \hat{f}(X))^2$$
$$= E[f(X) + \epsilon - \hat{f}(X)]^2$$
$$= E[f(X) - \hat{f}(X)]^2 + Var(\epsilon)$$

→ $E[\sim]^2$는 reducible error.
$Var(\epsilon)$는 irreducible error
→ 줄일 수 있는 에러를 줄여야 함!

회귀와 분류의 에러 MSE

회귀 MSE $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$
분류 MSE $= \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i))$

Y와 관련된 X를 찾는 Inference

Select State : X중에 시 몇 종류를 중요한
Statistical significance : 통계 테스트로 유의성 검정

$f()$를 추정 (estimate)하는 방법.

모수적 (Parametric) : X와 Y의 특정 관계를 가정
→ 정확한 관계를 알 수 없으므로 근사값을 가짐.
ex) 시험점수 = $\beta_0 + \beta_1$ 난이도 + β_2 공부시간

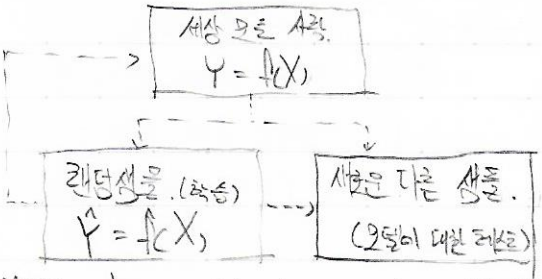
비모수적 (Non-Parametric) : X와 Y사이에 관계가 없음을 가정
→ 복잡한 유연성을 가진 모델을 만들어야 함.

유연성과 해석 가능성의 trade off 관계

유연성이 개념. 모델이 더 복잡하고 곡률성이 있다면 모델의 유연성이 높아짐. 해석 가능성은 떨어짐.
해석가능성이 높으면, 복잡한 관계를 모델링할 수 없음.

비지도 학습 (자율 학습)은 Y가 주어지지 않은 채로 $f()$를 추정해야 함. 오직 X로 $f()$를 추정함.
강화학습은 동작에 따라 보상을 받고, 보상을 최대화하는 다음 X를 찾는 것임. → $Y = f(a|x)$

새로운 모집단이 어떤 특성을 찾는 것으로 훈련에 사용된 X와 이번, 새로운 샘플을 통해 테스트 해야 함.



MSE도 train set과 test set에 대해 구할 수 있고, 최종적으로 test set에 대한 MSE를 줄여야 함.
이 수식을 통해 오버피팅, 언더피팅을 확인할 수 있음.