Supervised learning IV

모델 선택에 대한 문제. (모델 평가, 교차 검증, 부분 집합 선택, regularization)에 대해 선택.

예측을 목적으로 참 $f(x)$와 유사한 $\hat{f}(x)$를 찾는 것임. 모든 모집단에 대해 직접적으로 $\hat{f}(x)$을 찾을 수 없기 때문에 표본을 통해 $Y = f(X)$를 구하게 됨. 이 $\hat{f}(x)$가 전체 모집단의 $f(x)$와 유사한지 판별하기 위해서, Test set으로 검증을 함. 지도학습의 모델은 MSE를 줄이는 것을 목표로 함. Train set의 MSE만을 고려하면 다른 데이터에는 적합하지 못한, Overfitting된 $\hat{f}(x)$를 갖게 됨.
최상은 test set에 대해 최소 MSE를 갖는 모델 $\hat{f}(x)$를 찾아야 하는 것이고, 모델 적합시에는 train set만 사용해야 함.

모델 선택 기준 (변형하기 기준)

독립변수 d에 따라 값으로 이 기준값이 최소화 혹은 최대화 되도록 작성하려는 기준을 제공함

→ Mallow's C_p

$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$ → 

→ Bayesian Information Criteria

$BIC = \frac{1}{n}(RSS + \log n\, d\hat{\sigma}^2)$ → 

→ Adjusted R^2

$Adj\ R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$ → 

모델 선택 기준은 이론적 배경이 강하며, 전체 중에 관찰되지 않은 전체 집단의 MSE를 반영한다는 점이 큰 장점.
↳ 방정식으로 계산되어, 작은 표본으로도 계산 가능함.

Cross validation (CV) : 교차검증.

훈련세트 내에서 트레이닝 셋으로 시뮬레이션 함. 모델선택 기준에 비해 많은 데이터와 계산능력을 요구하며, 데이터에 대한 가정이 적음.
모델에는 교차검증을 많이 사용함.

LOOCV (Leave One Only Cross Validation)

n개의 샘플 중 1개를 제외한 모델로 학습하고, 제외한 1개의 샘플로 검증하는 방식.

K-fold CV

샘플을 k개의 그룹으로 나누고 1개의 그룹을 제외한 상태로 학습 및 검증하는 방식.

LOOCV vs K-fold CV.

LOOCV가 훨씬 더 많이 학습함.
트레이닝 세트에 의존적이며, 더 유연함 (편향을 낮추, 분산을 높임)
K-fold는 계산의 횟수가 훨씬 더 적으며, k를 n으로 놓을 경우, LOOCV와 동일함

변수 선택의 문제

전통적으로는 변수의 개수가 샘플에 비해 적었으나 현대에는 오히려 변수가 샘플보다 많아지고 있음. 이런 상황에서는 전통적 학습법이 실패하기 때문에, 유익한 변수를 찾고, 그리고 작업이 필요하다.

→ n개의 변수 중 k개의 변수만 선택하는 것으로 모든 조합을 만들어보고 test set에 대한 최소의 MSE를 갖는 모델을 고르면 되겠으나, 현실적으로 어렵다.