Statistical Learning

Wednesday, August 25, 2021 9:17 PM

- Motivation
  - We wish to determine the relationship between the inputs and outputs of a system.
    - Assume such a relationship exists!

$Y \sim$ output (response, dependent variable)

$X \sim$ input (predictor, independent variable)

$X \to (X_1 \cdots X_p)$ (features) dimensions

$$Y = f(X) + \varepsilon$$

f: systematic information of Y from X

ε: error term independent of X

  - SL attempts to define and estimate f, we then attempt to evaluate the quality of our estimation!

In-Sample vs Out of Sample

- Estimating f
  - $f(X)$ is "ideal" or "best" representation we can have

↳ $\hat{f}(X)$

*Why?
↳ Inference
↳ Prediction

- Prediction

$$\hat{Y} = \hat{f}(X)$$

↖ "black box"

- Accuracy: $\hat{Y}$ as a prediction of Y
  - Reducible Error → Improve estimation of $\hat{f}$
  - Irreducible Error → ε, missing X's

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

Reducible Irreducible

↑

- Inference
  - We estimate f in order to
    - Understand association $Y \to X_1 \dots X_p$
    - Which X_j are used, and significantly

*Linearity → Is Y a linear function of $X_1 \dots X_p$

- Parametric vs Non-Parametric Methods

* Training Data ⇒ Estimation of f using available sample data

- Parametric:
  1. Assumption on function family/form of f:

Linear: $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$

  2. Use training data to fit or train $\hat{f}(X)$

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

* OLS Estimator

* If f has a lot of βs ⇒ Flexible
↓
overfitting
(noise)

- Non-Parametric:
  - No assumptions on structure of f
  - Estimate f as direct local/smoothed function of training data

$f(X)$ is piecewise defined over domain of X

- Accuracy
  - Quality of fit/training : minimize error between our predicted values and actuals
    - training (in-sample)
    - test (out-of-sample)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

mean error squared

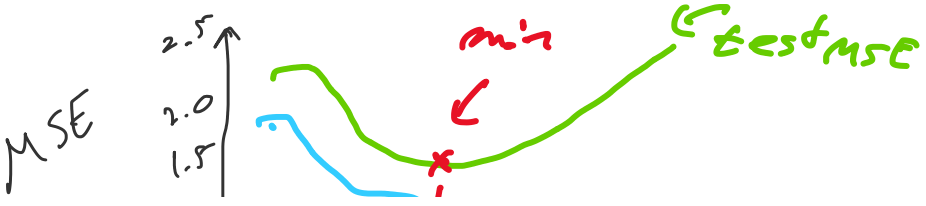
* sample x_i : iid

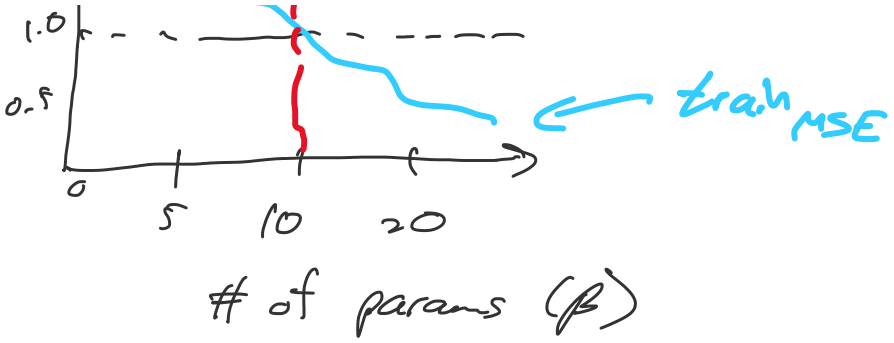
* We have minimized error on training observations $\{x_1 \dots x_n\}$, out-of-sample I get:

(x_0, y_0) as test data

$$Avg\left(y_0 - \hat{f}(x_0)\right)^2$$ ← in test!

"generalize in test with $\hat{f}(x)$"





- Bias - Variance
  - Expectations of Error
    - $E[\text{train}_{MSE}] = \text{train}_{MSE}$
    - $E[\text{test}_{MSE}] = ?$
      - Test Set separate from Sample
      - Test / Train split of Sample
      - k-Fold / etc Cross-Validation (Holdouts)

* for a given x_0, $E[\text{test}_{MSE}]$
  - variance of $\hat{f}(x_0)$
  - squared bias $\hat{f}(x_0)$
  - variance of error ε

$$E[(y_0 - \hat{f}(x_0))^2] = \text{Var}(\hat{f}(x_0)) + (\text{Bias}\,\hat{f}(x_0))^2 + \text{Var}(\varepsilon)$$

↑ Expected test_{MSE} for any test x_0

Variance → Change in $\hat{f}$'s params as x_0 changes
Bias → Error in prediction vs actual as x_0 changes

#Trade-off: More flexible we make f in terms of # of parameters, the lower the bias, the higher the variance.