Cross-Validation

Wednesday, September 22, 2021 6:07 PM

- Hold-out / Validation
  - Training Error vs. Test Error ← out of sample

$E[e_{oos}] = ?$

e_{Train}

$e_{Test-set}$ (Test/Train Split) [e.g., 80/20]

e_{CV}

- Test/Train Split
  - Holdout set: Randomly dividing available sample data into training and testing sets (Validation)

* Stratified Sampling → Equal ratios of classes/ranges

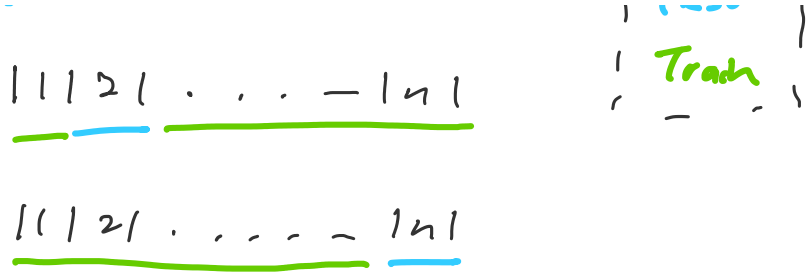
$E[e_{oos}] = e_{Test}$

* Issues:
  1. High variability in Test Set on random splits
  2. We end up with smaller Training set
     → Higher Error Rate in Training (Overestimate)

- Cross-Validation
  - LOOCV

| 1 | 2 | ..... | n |

Test



* Distribution estimate for e_{OOS} → $E[e_{OOS}] = \mu_{LOOCV}$
median
⋮

$$\therefore CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

$$MSE_i = (y_i - \hat{y}_i)^2$$

↳ Less bias ⇒ we fit on $n-1$ points!
↳ No randomness on test-train split since all n points are used as test sets

- K-Fold CV
  - k^{th} fold is used as test
  - $1..k-1$ fold is used as training

$$CV_{(K)} = \frac{1}{k}\sum_{i=1}^{K} MSE_i$$

* LOOCV is k-fold with $k=n$

* Bias-Variance Issues in K-Fold / LOOCV
  - We are taking mean error of k models to estimate e_{OOS}

all k models as $k \to n$ get more correlated!

- Mean of many correlated values has a higher variance than than sets of many uncorrelated values!