2/12/23
Day 11

5.1.2: Leave-One-Out Cross-Validation (LOOCV)

- A single observation is used as the validation set, (x_i, y_i). The MSE is then $(y_i - \hat{y}_i)^2$.

- This MSE is unbiased, but highly variable.
- What we do is repeat this procedure leaving out each observation, giving n squared errors. $\{MSE_i\}_{i=1}^n$.
- Then: the LOOCV test error estimate is
$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n MSE_i$$

Advantages:
1) Far less bias than the validation set approach.
2) LOOCV always yields the same result for a given dataset.

Disadvantages:
1) Can be slow to implement if n is large or fitting the model is slow.

Note that, for polynomial regression, the following ~~result~~ ~~holds.~~ fast formula holds:

$$\boxed{CV_{(n)} = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - h_i} \right)^2}$$

where h_i is the leverage of observation i and $\hat{y}_i$ comes from fitting the model on all n observations.

A very general method.

5.1.3: k-Fold Cross-Validation

1) Randomly break the observations into k ~~groups~~ folds
2) Fit the model on the last k-1 ~~groups~~ folds
3) ~~Estimate~~ Calculate the test MSE on the first ~~set~~ fold.
4) Iterate ~~of~~ over [k] to get $\{MSE_i\}_{i=1}^k$
5) Estimate the test MSE as
$$\boxed{CV_{(k)} = \frac{1}{k} \sum_{i=1}^k MSE_i}$$

Often k=5 or 10. Why?

1) Computational advantage for large n or complex fitting model.

2) If we only want to identify the correct level of flexibility, then k-fold CV does a good job of identifying the degrees of freedom giving the minimum test MSE.

12/13/23

Day 12

5.1.4: Bias-Variance Trade-off for k-Fold Cross-Validation

1) k-fold CV is more computationally efficient than LOOCV

☆ 2) k-fold CV gives more accurate estimates of the test error than LOOCV

LOOCV reduces bias the most, but has high variance
k-fold CV reduces bias. Why

LOOCV estimates the test MSE with a mean of ~~highly correlated observations, whereas~~ nearly identical datasets, which causes higher correlation in the test error estimates from iteration.

k-folds Operations are less correlated

5.1.5: Cross-Validation on Classification Problems

Instead of using MSE to quantify error, we use the misclassification rate. Thus, the LOOCV error rate is

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

And likewise for validation ~~of~~ & k-fold