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 observations are less correlated

5.1.5: Cross-Validation on Classification Problems

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

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

And likewise for validation set & k-fold

5.2: The Bootstrap

- Used to quantify the uncertainty associated with a given estimator or statistical learning method.

- example: Assess the variability associated with the regression components of a linear model.

We wish to invest a fixed sum of money into two financial assets that yield returns of X and Y, respectively.

We invest a fraction α into X, $1-\alpha$ into Y.

We want to minimize the total risk (the variance) $Var(\alpha X + (1-\alpha)Y)$. One can show that the minimizer is

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

We'll need to estimate these σ's.
How can we assess the accuracy of α?

- We resample with replacement from the data, calculate α for those & then average.
- Repeat B times, for large B, get $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \dots, \hat{\alpha}^{*B}$

Then

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2}$$

Ch.6: Linear Model Selection & Regularization

Alternative fitting procedures can yield better prediction accuracy and model interpretability.

- Constraining or shrinking coefficients can give better prediction accuracy if $p \geq n$.

- Feature/variable selection improves model interpretability by removing irrelevant variables from the model

- **Subset selection:** Find the subset of predictors which is relevant
- **Shrinkage (regularization)** reduces the variance
- **Dimension reduction:** ???

6.1: Subset Selection

6.1.1: Best Subset Selection

Fit a different model to each of the 2^p subsets of the predictors, then take the one with the best CV prediction error, C_p, AIC, BIC, or adjusted R^2

Deviance = $-2 \times \max(\log(L))$; RSS for a broader class of models; the smaller the better

Way too computationally expensive

6.1.2: Stepwise Selection

Forward Stepwise Selection
- Begin with empty model, then adds predictors until they're all in the model
- Specifically, the variable giving the greatest additional improvement to the fit is added to the model.
- Select the single best model

You only need to fit $1 + p(p+1)/2$ models here.

This works for $n < p$ too, but not well since you can only fit $n-1$ models.

Backward stepwise selection
Only needs to fit $1 + p(p+1)/2$ models
Needs $n > p$

Hybrid approaches exist too.

6.1.3: Choosing the Optimal Model

We need to estimate the test error.

1) Adjust the training error
2) Directly estimate test error.

C_p, AIC, BIC, Adjusted R^2

- These can be used to select among different number of variables.

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

where $\hat{\sigma}^2$ is an estimate of $Var(\epsilon)$ and d is the number of predictors
(Note ϵ is often estimated with the full model)

- AIC is defined for models fit by maximum likelihood

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS - 2d\hat{\sigma}^2)$$

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