Resampling Methods

Resampling Methods: tool in modern statistics, involving repeatedly drawing samples from a training set & refitting a model on each sample

Example:

We can draw samples from training & repeatedly fit a linear model & see how each differs

two most common resampling

1. cross validation
2. bootstrap

Model Assessment: the process of evaluating a model's performance

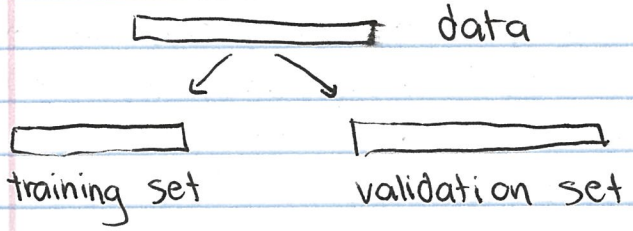
Model Selection: the process of selecting the proper level of flexibility for a model

Cross-Validation

Given a dataset  $\rightarrow f(x) \rightarrow y$

statistical learning method is only warranted if it results in low test error

Validation Set Approach



← provides an estimate for test set

↳ you can do multiple splits that are all different
- can be highly variable

Leave-One-Out Cross-Validation

- single observation is used for validation

$$CV_{(n)} = \frac{1}{n} \sum MSE$$

Benefits:

- no randomness in training/validation
- doesn't overestimate test error

k-Fold Cross Validation

involves randomly dividing the set of observations into k groups of approximately equal size

1st fold is treated as the validation set, & model is fit on k-1 folds

↑ repeat procedure k times

this procedure results in k estimates of test errors $mse_1, mse_2, \ldots mse_k$

$$CV_{(k)} = \frac{1}{k} \sum MSE$$

When we perform cross-validation our goal might be to determine how well a given statistical learning procedure can be expected to perform on independent data

↳ we want to eventually choose the model with the lowest test error

Bias Variance Tradeoff for k Fold Cross Validation

k Fold CV gives an accurate estimate of test error

when your test set is small you can have overestimates of test error rate

has bias

The Bootstrap

extremely powerful statistical tool that can be used to quantify uncertainty associated with a given estimator or statistical learning method

↓ gives you a standard error for your estimates

Example:

value that minimizes risk

$$\alpha = \frac{\sigma^2 y - \sigma xy}{\sigma^2 x + \sigma^2 y - 2\sigma xy}$$

↑ we get estimates of parameter from data

we estimate α on simulated data

we get 1000 estimates for α, $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_{1000}$

$$\bar{\alpha} = \frac{1}{1000} \sum^{1000} \hat{a}$$

- get μ & σ

from different samples we expect a to differ by 0.08

z_1, $n=3$

	x	y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8