The Validation Set Approach

Two potential drawbacks of the validation set approach are:

1. {The validation estimate of the test error rate can be highly variable, dependent on which observations are included in the training and test sets};

2. {Only a subset of observations (the training set) are used to train the model. Since statistical methods tend to perform worse when trained on fewer observations, this suggests that the validation set error rate may tend to overestimate the test error rate for the model fit on the entire data set}.

Back:

Leave-One-Out Cross-Validation

LOOCV involves leaving a single observations out of the training set, building the model using the remaining observations and calculating the MSE on the test set observation, MSE_j.
The LOOCV estimate for the test MSE is the {average of the n test error estimates,} {

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \text{MSE}_i.$$

}

Back:

Leave-One-Out Cross-Validation

With least squares linear or polynomial regression, the cost of LOOCV is the same as a single model fit. The CV formula is, {

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2,$$

} where {$\hat{y}_i$ is the i^{th} fitted value from the original least squares fit}, and {h_i is the leverage statistic}.

Back:

k-fold Cross-Validation

k-fold CV involves randomly dividing the set of observations into k groups. For each fold j, the model is built using the remaining folds, and MSE calculated on the j^{th} fold. The k-fold CV estimate is {computed by averaging the MSE values}, {

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \text{MSE}_i.$$

}

Back:

– Resampling

LOOCV vs. k-fold CV

LOOCV has {lower bias} then k-folds CV as {more observations are included in the training set}.
However, in LOOCV, each model is trained on {an almost identical set of observations, and therefore the outputs are highly (positively) correlated}. Since the {mean of many highly correlated quantities has higher variance then quantities that are less correlated}, the test error estimate from LOOCV {tends to have higher variance than that resulting from k-folds CV}.

Back:

Resampling–

CV on Classification Problems

When Y is qualitative, instead of using {MSE to quantify test error, we use the number of misclassified observations}. For example, the LOOCV classification error rate takes the form, {

$$\text{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \text{Err}_i,$$

} where $\{\text{Err}_i = I(y_i \neq \hat{y}_i)\}$.

Back:

Resampling–

The Bootstrap

Given B bootstrap data sets, $Z^{*1}, \ldots, Z^{*B}$, and B corresponding α estimates, $\hat{\alpha}^{*1}, \ldots, \hat{\alpha}^{*B}$, we can compute the standard error of these bootstrap estimates using the formula, {

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

} This serves as an estimate of {the standard error of $\hat{\alpha}$ estimated from the original data set}.

Back: