Model Selection by Predictive Accuracy

When predictive accuracy is more important than understanding the relationship between y and our explanatory variables, we use different metrics to compare models.

In particular, rather than goodness-of-fit metrics we use metrics that quantify the distance between observed response values and predicted response values. Optimality in this case corresponds to the minimization of prediction error. To quantify prediction accuracy effectively, we use cross-validation:

Training vs. Test data

- Here we randomly partition the observed data into a training set that is used to fit the model and a test set that is used to evaluate the fitted model's predictive capabilities.
  *usually an 80-20 split is used

- It is very important that the model is not trained/fit on the test. To get the fairest assessment of a model's predictive accuracy, it must be validated on data it has not seen before.

- Predictive accuracy can be summarized in a few ways:

Estimates of Prediction Error

- Predictive Mean Squared Error:

$$MSE_p = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

(where $y_i, i=1,2,\dots,m$ are the response observations in the test set).

- Predictive Root-Mean Squared Error:

$$RMSE_p = \sqrt{MSE_p}$$

- Predictive Mean Absolute Error:

$$MAE_p = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|$$

Although this approach is intuitive and simple it has two drawbacks:

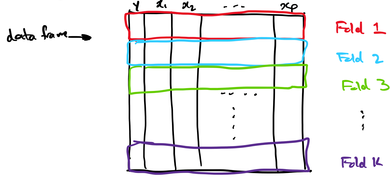
- The estimate of prediction error can be highly variable depending on the particular training vs. test split that we get.

- The models are never training on all of the data - just a subset of it.

To overcome both of the problems we use k-fold cross-validation.

K-fold Cross-validation

- Here we randomly partition the available data into k (roughly equal sized) sub-datasets (folds)



At each stage one fold is treated as the test set and all of the other folds comprise the training set. Using partition we calculate an estimate of the prediction error as we would in ordinary cross-validation.

Treating each fold as a test set exactly one time yields k estimates of prediction error

i.e. $MAE_1, MAE_2, \dots, MAE_k$

The k-fold cross validation estimate of prediction error is taken to be the average of the k fold-specific estimates.

i.e., $$MAE_{CV} = \frac{1}{k}\sum_{i=1}^{k} MAE_i$$

* Common choices of k are $k=5$ and $k=10$. Note that a special arises when $k=n$ and each time the "test set" is just a single observation. This n-fold cross validation is also referred to as "leave one out" cross validation. Although this can be computationally intensive, it provides a very accurate estimate of prediction error.

Thus, the general strategy is to perform k-fold cross validation on every model under consideration, and identify the optimal model as being the one that minizes the selected predictive accuracy metric.