

1 Resampling Methods

Sampling methods are a class of methods that serve a twofold purpose

- Provide a subset of data to be used for evaluating the test error rate
- Use different samples of data to assess the variability in the model parameters to choose the level of flexibility

1.1 Cross Validation

Validation set approach forms the bedrock for cross validation. We randomly split our training data into two sets, a training set and a validation set. We fit our models on the training set, and the validation set will serve as the unseen data. We can evaluate different models on the validation test to judge which of them performs the best.

A small problem with this approach is that the validation error will depend on the split of the data, i.e., we can expect slightly different validation errors based on which subset of data we train. Hence, it is better to do this sampling multiple times in order to confidently select models and report test errors.

Furthermore, by preparing a validation set, we are reducing the size of our training set. Larger training data will usually result in better models. Hence, we might be overestimating the test errors in this case. A simple way to avoid this problem is to choose small sizes of the validation set and do the tests multiple times.

1.1.1 Leave One Out CV

An extension of the validation approach, here we train the data on all but on example. This way, if the data has n examples, we build n models (each trained on $n - 1$ examples) and the error is

$$\text{test error} = \frac{1}{n} \sum_{i=1}^n \text{error}_i$$

where error_i is the error on i^{th} observation from the model trained on remaining $n - 1$ observations

Computing this can be extremely expensive when n is large. For linear regression, there exists a trick by which the time taken to get *test error* is exactly the same as the time to fit a single model on entire data set !

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$
$$\text{Hat matrix } H = X(X^T X)^{-1} X^T$$

Let X_i denote the matrix X but with the i^{th} row removed, and similarly Y_i . Let x_i^T denote the i^{th} row of X and h_i be the diagonal entry of H . Then we have the following

$$\begin{aligned} X_i^T X_i &= X^T X - x_i x_i^T \\ X_i^T Y &= X^T Y - x_i y_i \\ h_i &= x_i^T (X^T X)^{-1} x_i \\ \hat{\beta}_i &= (X_i^T X_i)^{-1} X_i^T Y_i \\ e_i &= y_i - x_i^T \hat{\beta}_i \end{aligned}$$

Also, we have the Sherman–Morrison formula for calculating the inverse of a perturbed matrix using the original matrix. Let A be the original invertible square matrix and u, v be column vectors. Then,

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1}u}$$

The formula can be verified by evaluating $LHS * RHS = RHS * LHS = I$.

Substituting $A = X^T X$ and $-u = v = x_i$,

$$\begin{aligned} (X^T X - x_i x_i^T)^{-1} &= (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i} \\ (X_i^T X_i)^{-1} &= (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i} \\ \hat{\beta}_i (X_i^T Y_i)^{-1} &= (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i} \\ \hat{\beta}_i &= [(X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i}] (X_i^T Y_i) \\ \hat{\beta}_i &= [(X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - h_i}] (X^T Y - x_i y_i) \\ \hat{\beta}_i &= \hat{\beta} - [\frac{(X^T X)^{-1} x_i}{1 - h_i}] (y_i (1 - h_i) - x_i^T \hat{\beta} + h_i y_i) \\ &= \hat{\beta} - [\frac{(X^T X)^{-1} x_i (y_i - x_i^T \hat{\beta})}{1 - h_i}] \end{aligned}$$

$$\begin{aligned} \text{Subsequently, } e_i &= y_i - x_i^T \hat{\beta}_i \\ &= y_i - x_i^T (\hat{\beta} - [\frac{(X^T X)^{-1} x_i (y_i - x_i^T \hat{\beta})}{1 - h_i}]) \\ &= (y_i - x_i^T \hat{\beta}) + \frac{h_i (y_i - x_i^T \hat{\beta})}{1 - h_i} \\ \text{or, } e_i &= \frac{y_i - x_i^T \hat{\beta}}{1 - h_i} \\ &= \frac{y_i - \hat{y}_i}{1 - h_i} \end{aligned}$$

Applying this formula for all the errors across the n models,

$$\begin{aligned} \text{test error} &= \frac{1}{n} \sum_{i=1}^n error_i \\ &= \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - h_i} \right)^2 \end{aligned}$$

which can be computed by simply building a single model on all the n data points.