Resampling Methods

William Brasic

The University of Arizona

Validation Set Approach

Definition 1: Validation Set Approach

1. Split data into training, validation, and testing sets.

Validation Set Approach

Definition 1: Validation Set Approach

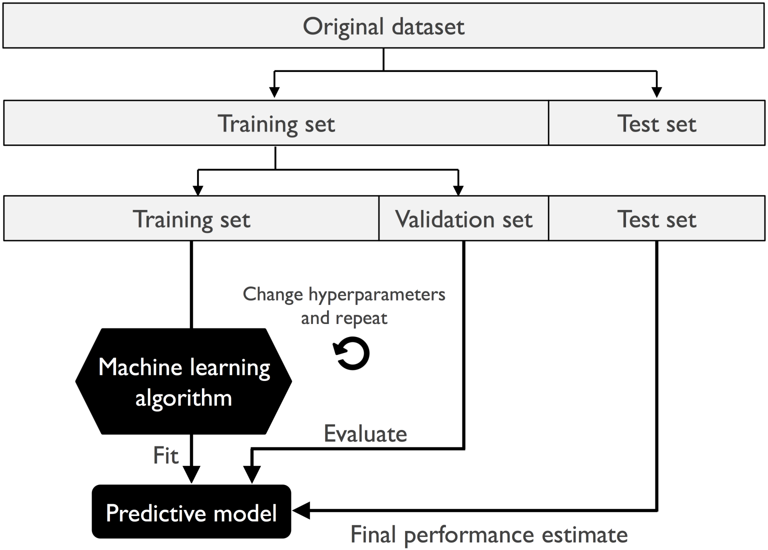
1. Split data into training, validation, and testing sets.
2. Fit model using training data.

Validation Set Approach

Definition 1: Validation Set Approach

1. Split data into training, validation, and testing sets.
2. Fit model using training data.
3. Estimate the test error rate by evaluating the model on the validation set.

Validation Set Approach



Validation Set Approach Issues

Property 1: Validation Set Approach Issues

- The validation set estimate of the test error can be highly variable.
  - Depends on which observations land in training or validation set.

Validation Set Approach Issues

Property 1: Validation Set Approach Issues

- The validation set estimate of the test error can be highly variable.
  - Depends on which observations land in training or validation set.
- Only a subset of the data is used to fit the model in training.
  - Perform worse than model fitted on all day implying test error overestimated.

- Cross-validation fixes these issues.

Resampling Methods

Definition 2: Resampling Methods

Resampling methods involve repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model.

- Allows us to get more estimates of how our model will perform out-of-sample.
- Often not too computationally expensive.

Leave-One-Out Cross-Validation (LOOCV)

Definition 3: Leave-One-Out Cross-Validation (LOOCV)

Leave-One-Out Cross Validation (LOOCV) involves the following process:

1. Split the data into training and testing sets.

Leave-One-Out Cross-Validation (LOOCV)

Definition 3: Leave-One-Out Cross-Validation (LOOCV)

Leave-One-Out Cross Validation (LOOCV) involves the following process:

1. Split the data into training and testing sets.
2. Fit the model on all training data besides a single observation.

Leave-One-Out Cross-Validation (LOOCV)

Definition 3: Leave-One-Out Cross-Validation (LOOCV)

Leave-One-Out Cross Validation (LOOCV) involves the following process:

1. Split the data into training and testing sets.
2. Fit the model on all training data besides a single observation.
3. Use the left out observation to estimate the test error rate.

Leave-One-Out Cross-Validation (LOOCV)

Definition 3: Leave-One-Out Cross-Validation (LOOCV)

Leave-One-Out Cross Validation (LOOCV) involves the following process:

1. Split the data into training and testing sets.
2. Fit the model on all training data besides a single observation.
3. Use the left out observation to estimate the test error rate.
4. Return to 2. now using a different single observation.

Leave-One-Out Cross-Validation (LOOCV)

Definition 3: Leave-One-Out Cross-Validation (LOOCV)

Leave-One-Out Cross Validation (LOOCV) involves the following process:

1. Split the data into training and testing sets.
2. Fit the model on all training data besides a single observation.
3. Use the left out observation to estimate the test error rate.
4. Return to 2. now using a different single observation.
5. Do this for all training observations.

Leave-One-Out Cross-Validation (LOOCV)

Definition 3: Leave-One-Out Cross-Validation (LOOCV)

Leave-One-Out Cross Validation (LOOCV) involves the following process:

1. Split the data into training and testing sets.
2. Fit the model on all training data besides a single observation.
3. Use the left out observation to estimate the test error rate.
4. Return to 2. now using a different single observation.
5. Do this for all training observations.
6. Average the errors to get a single estimate of the test error rate.

- Can be computationally expensive.

Leave-One-Out Cross-Validation (LOOCV)

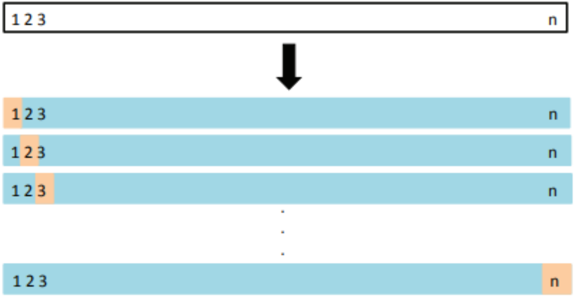


FIGURE 5.3. *A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSEs. The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.*

LOOCV test MSE and Error Rate Estimates

Definition 4: LOOCV test MSE and Error Rate Estimates

The LOOCV test MSE and error rate estimates, respectively, are given by

$$CV_{n,MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_{-i}(\boldsymbol{x}_i)\right)^2$$

$$CV_{n,Error} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(y_i \neq \widehat{f}_{-i}(\boldsymbol{x}_i)\right).$$

- $\widehat{f}_{-i}$ is trained using all, but the ith observation.

LOOCV Benefits

Property 2: LOOCV Benefits

- Zero variability in test error estimate because LOOCV always yields the same estimate.
- Use $n-1$ observations for training (nearly all dataset observations) so test error estimate is not as overestimated.

k-Fold Cross Validation (k-Fold CV)

Definition 5: k-Fold Cross Validation (k-Fold CV)

k-Fold Cross Validation (k-Fold CV) involves involves the following process:

1. Split the data into training and testing sets.

k-Fold Cross Validation (k-Fold CV)

Definition 5: k-Fold Cross Validation (k-Fold CV)

k-Fold Cross Validation (k-Fold CV) involves involves the following process:

1. Split the data into training and testing sets.
2. Split the training data into k folds.

k-Fold Cross Validation (k-Fold CV)

Definition 5: k-Fold Cross Validation (k-Fold CV)

k-Fold Cross Validation (k-Fold CV) involves involves the following process:

1. Split the data into training and testing sets.
2. Split the training data into k folds.
3. Train model on all data, but the k-th fold.

k-Fold Cross Validation (k-Fold CV)

Definition 5: k-Fold Cross Validation (k-Fold CV)

k-Fold Cross Validation (k-Fold CV) involves involves the following process:

1. Split the data into training and testing sets.
2. Split the training data into k folds.
3. Train model on all data, but the k-th fold.
4. Using the fitted model, evaluate its performance using the k-th fold as the validation set.

k-Fold Cross Validation (k-Fold CV)

Definition 5: k-Fold Cross Validation (k-Fold CV)

k-Fold Cross Validation (k-Fold CV) involves involves the following process:

1. Split the data into training and testing sets.
2. Split the training data into k folds.
3. Train model on all data, but the k-th fold.
4. Using the fitted model, evaluate its performance using the k-th fold as the validation set.
5. Go back to 3. and do this for all folds of the training data.

- Much less computationally expensive relative to LOOCV.
- Typical values of k are five or ten.

k-Fold Cross Validation (k-Fold CV)

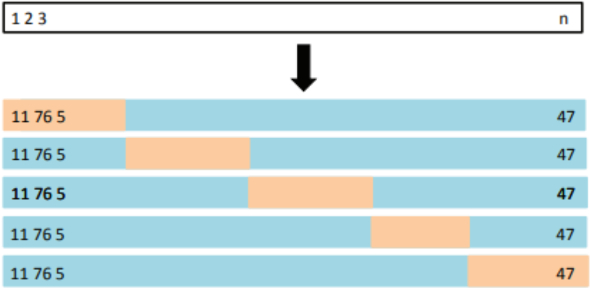
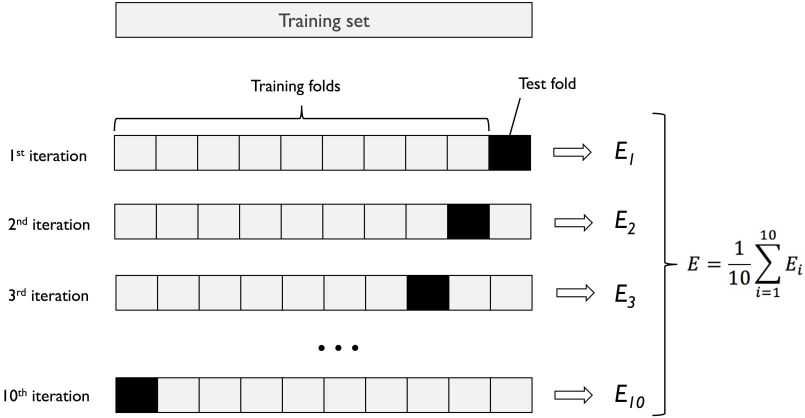


FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

k-Fold Cross Validation (k-Fold CV)



k-Fold CV

Property 3: k-Fold CV

- Typically used for model selection and hyperparameter tuning.
- After finding the “best” algorithm and its hyperparameters:
  1. Retrain model using all the training data.
  2. Evaluate its performance based on the test data.

- Preferred over LOOCV because:
  - More efficient
  - Much lower variance (although slightly higher bias)

k-Fold CV test MSE and Error Rate Estimates

Definition 6: k-Fold CV test MSE and Error Rate Estimates

The k-Fold CV test MSE and error rate estimates, respectively, are given by

$$CV_{k,MSE} = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{n_j}\sum_{i \in D_j}\left(y_i - \widehat{f}_{-j}(\boldsymbol{x}_i)\right)^2\right)$$

$$CV_{k,Error} = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{n_j}\sum_{i \in D_j}\mathbb{1}\left(y_i \neq \widehat{f}_{-j}(\boldsymbol{x}_i)\right)\right).$$

- D_j denotes the set of indices for observations contained in the jth fold.
- $\widehat{f}_{-j}$ is trained on all observations besides those in D_j.
- n_j is the number of observations in the jth fold.

k-Fold CV vs LOOCV

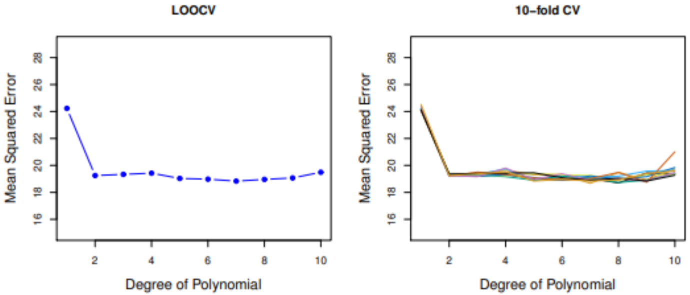


FIGURE 5.4. *Cross-validation was used on the* `Auto` *data set in order to estimate the test error that results from predicting* `mpg` *using polynomial functions of* `horsepower`. Left: *The LOOCV error curve.* Right: 10*-fold CV was run nine separate times, each with a different random split of the data into ten parts. The figure shows the nine slightly different CV error curves.*

Bootstrap

Definition 7: Bootstrap

Bootstrap is a resampling technique used to estimate statistics on a population by sampling a dataset *with replacement*.

- Repeatedly sample distinct data sets to evaluate models to test out-of-sample performance.
- Largely used to assess the variability of out-of-sample performance.
- **Widely used in econometrics to estimate variability of $\widehat{\beta}$, construct confidence intervals, and conduct hypothesis tests**.

Bootstrap Procedure for ML

Definition 8: Bootstrap Procedure for ML

To perform the bootstrap technique:

1. Randomly draw B training samples with replacement from the data to create multiple bootstrap samples.

Bootstrap Procedure for ML

Definition 8: Bootstrap Procedure for ML

To perform the bootstrap technique:

1. Randomly draw B training samples with replacement from the data to create multiple bootstrap samples.
2. Fit the model to each bootstrap sample obtaining $\widehat{f}_1, \ldots, \widehat{f}_B$.

Bootstrap Procedure for ML

Definition 8: Bootstrap Procedure for ML

To perform the bootstrap technique:

1. Randomly draw B training samples with replacement from the data to create multiple bootstrap samples.
2. Fit the model to each bootstrap sample obtaining $\widehat{f}_1, \ldots, \widehat{f}_B$.
3. Evaluate the model on the validation data.

Bootstrap Procedure for ML

Definition 8: Bootstrap Procedure for ML

To perform the bootstrap technique:

1. Randomly draw B training samples with replacement from the data to create multiple bootstrap samples.
2. Fit the model to each bootstrap sample obtaining $\widehat{f}_1, \ldots, \widehat{f}_B$.
3. Evaluate the model on the validation data.
4. Obtain the error.

Bootstrap Procedure for ML

Definition 8: Bootstrap Procedure for ML

To perform the bootstrap technique:

1. Randomly draw B training samples with replacement from the data to create multiple bootstrap samples.
2. Fit the model to each bootstrap sample obtaining $\widehat{f}_1, \ldots, \widehat{f}_B$.
3. Evaluate the model on the validation data.
4. Obtain the error.
5. Calculate the bootstrap estimate of the error as the average of the estimates from each bootstrap sample.

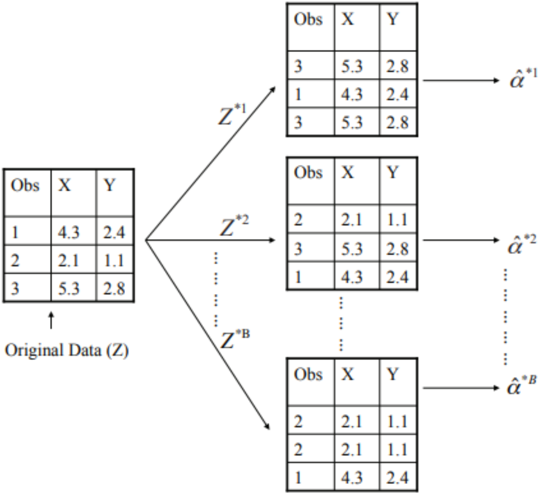
Bootstrap Procedure for ML

Definition 8: Bootstrap Procedure for ML

To perform the bootstrap technique:

1. Randomly draw B training samples with replacement from the data to create multiple bootstrap samples.
2. Fit the model to each bootstrap sample obtaining $\widehat{f}_1, \ldots, \widehat{f}_B$.
3. Evaluate the model on the validation data.
4. Obtain the error.
5. Calculate the bootstrap estimate of the error as the average of the estimates from each bootstrap sample.
6. Assess the variability of the estimate using the distribution of the bootstrap estimates.

Bootstrap Procedure



Single Bootstrapped Estimate of Empirical MSE

Definition 9: Single Bootstrapped Estimate of Empirical MSE

Denote the first bootstrapped sample as Z_1 which contains n indices corresponding to observations sampled from the original data *with replacement*. The first bootstrapped estimate of the empirical MSE is given by

$$MSE_{Z_1^*} = \sum_{i \in Z_1} \left(y_i - \widehat{f}_1(\boldsymbol{x}_i) \right)^2 .$$

Bootstrapped Estimate

Definition 10: Bootstrapped Estimate

A bootstrapped estimate is the aggregated estimate of a statistic calculated from all bootstrap samples.

- Typically involves averaging the bootstrap statistics .

Bootstrapped Estimate of Empirical MSE

Definition 11: Bootstrapped Estimate of Empirical MSE

Denote the B bootstrapped samples as $Z_1^*, \ldots, Z_B^*$ which each contain n observations sampled from the original data *with replacement*. The bootstrapped estimate of the empirical MSE is given by

$$\begin{aligned} MSE_B &= \frac{1}{B}\sum_{b=1}^{B} MSE_{Z_b^*} \\ &= \frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{n}\sum_{i \in Z_b}\left(y_i - \widehat{f}_b(\boldsymbol{x}_i)\right)^2\right). \end{aligned}$$

Bootstrap Distribution

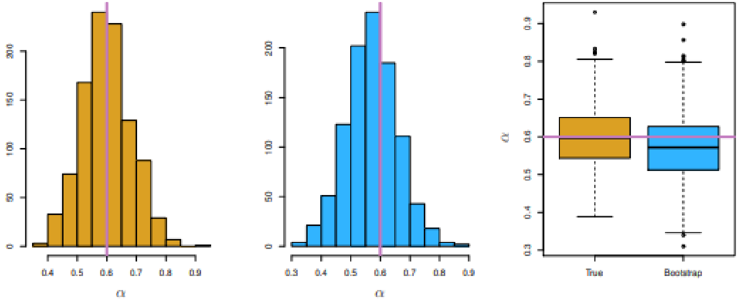


FIGURE 5.10. Left: *A histogram of the estimates of* α *obtained by generating 1,000 simulated data sets from the true population.* Center: *A histogram of the estimates of* α *obtained from 1,000 bootstrap samples from a single data set.* Right: *The estimates of* α *displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of* α*.*

Bootstrap vs Cross-Validation

Property 4: Bootstrap vs Cross-Validation

- **Cross-Validation**: Mainly used for model evaluation and selection by estimating prediction error.
- **Bootstrap**: Primarily used for estimating the distribution of a statistic and assessing the variability of an estimate.

Bootstrap for Econometrics

Property 5: Bootstrap for Econometrics

The bootstrap can also be used for econometric analysis to estimate the variability of our OLS estimator $\widehat{\beta}$. Via repeated bootstrapped sampling we can:

1. Construct the bootstrapped distribution of $\widehat{\beta}$.
2. Construct bootstrapped confidence intervals.
3. Conduct bootstrapped hypothesis tests.

Bootstrap for Econometrics

Property 5: Bootstrap for Econometrics

The bootstrap can also be used for econometric analysis to estimate the variability of our OLS estimator $\widehat{\beta}$. Via repeated bootstrapped sampling we can:

1. Construct the bootstrapped distribution of $\widehat{\beta}$.
2. Construct bootstrapped confidence intervals.
3. Conduct bootstrapped hypothesis tests.

- Particularly useful when:
  - We have small samples so asymptotic properties are unlikely to hold.
  - Complex models where closed-form solutions for standard errors are not straightforward.

Thank You!