Introduction to Machine Learning

Ranga Raju Vatsavai, Ph.D.
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)

Feb. 25-27, 2019

Challenges in Error Estimation

- For given test dataset, we can obtain error/accuracy, but how good are our measures?
  - Do the accuracy remain same for various training (and test datasets)?
- Getting a separate test data set is costly (though most desirable)
- Often training data set is used for validation of the model as well
  - Resampling

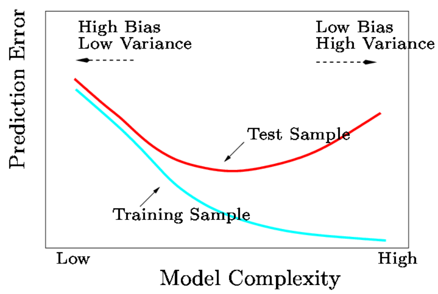
Resampling

- Repeated sampling of training dataset to fit multiple models to obtain additional information about the fitted models
- Most commonly used resampling methods are
  - Cross-validation
  - Bootstrap
- These methods can be used to
  - Estimate test error (model assessment)
  - Select appropriate level of flexibility (model selection)

Training Error versus Test Error

- Recall the distinction between the test error and the training error:
- The test error is the average error that results from using a statistical learning method to predict the response on a new observation, one that was not used in training the method.
- In contrast, the training error can be easily calculated by applying the statistical learning method to the observations used in its training.
- But the training error rate often is quite different from the test error rate, and in particular the former can dramatically underestimate the latter.

Training Error versus Test Error



General Idea of Using Training Data for Validation

Training Dataset

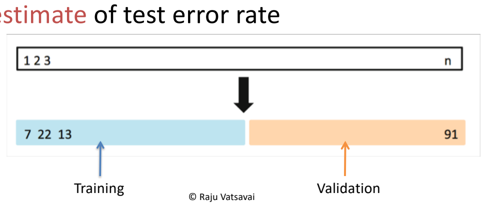
X	Y
150	5.5
155	5.6
125	5.3
130	5.4
135	5.6
160	5.5
165	5.2
160	5.6

Model Fit

Holdout data for validation

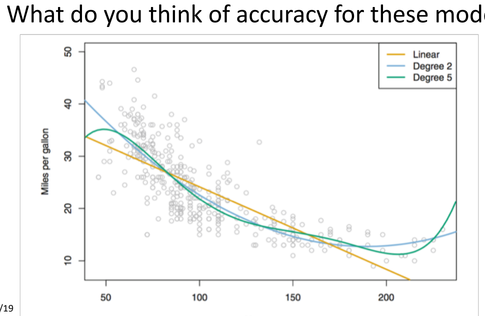
Validation Set Approach

- Simplest approach
- Randomly divide n available samples into a training set and validation or holdout set
- The resulting validation set error rate is an estimate of test error rate



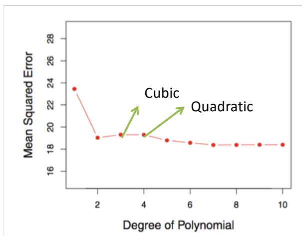
Accuracy on Different Models

- First, we can fit different models for same data
- What do you think of accuracy for these models?



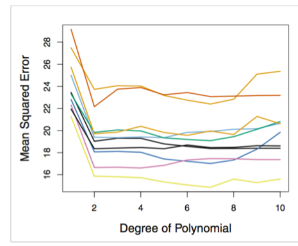
Accuracy on Different Models

- Does higher order terms (increasing complexity) improves accuracy?



Accuracy on Different Models

- How about accuracy on different random splits of training data



Observations

- Increasing complexity (e.g., higher order terms like cubic) may not lead to better prediction than less complex (e.g., quadratic) models
- Validation estimate of test error rate can be highly variable depending on which observations are included in the training set and which observations are included in the validation set (plot shows general trend)
- In the validation approach, only a subset of available data is included in fitting 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.

Cross-Validation

- A refinement over validation set approach that address the two issues: highly variable test error rates and overestimation of test error rates
- Leave-one-out cross-validation (LOOCV)
- k-fold cross-validation

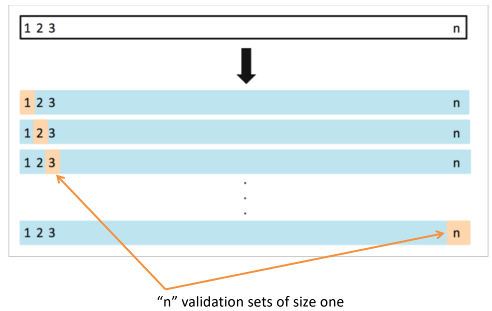
NC STATE UNIVERSITY

Leave-one-out cross-validation

- Like the validation set approach, LOOCV involves splitting the set of observations into two parts.
- However, instead of creating two subsets of comparable size, “n” sets (of training and test) are created, where a single observation (x_i, y_i) is used for the i^{th} validation set, and the remaining observations $\{(x_n, y_n) - (x_i, y_i)\}$ make up the i^{th} training set.

NC STATE UNIVERSITY

Leave-one-out cross-validation

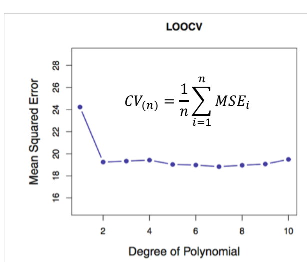


NC STATE UNIVERSITY

Leave-one-out cross-validation

- Could be expensive
- However, for least squares linear or polynomial regression, the following short-cut applies:

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

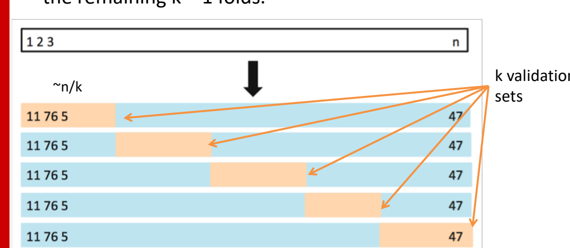


- Where leverage statistic h_i is given by $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$

NC STATE UNIVERSITY

k-Fold Cross-Validation

- k-fold CV involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining k − 1 folds.

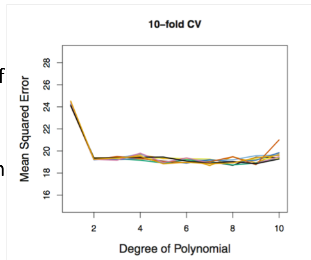


k-fold Cross-Validation

- The k-fold CV estimate is computed by averaging

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

- LOOCV is a special case of k-fold CV, where k = ?
- There is some variability, but this variability is typically much lower than the variability in the test error estimates that results from the validation set approach



Acknowledgements

- ISLR book; Chapter 5