NC STATE UNIVERSITY

CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

W13: 11/10/15-11/12/15

NC STATE UNIVERSITY

Review

- HW1: 4.85
- HW2: 4.75
- MT-1: 82.16
  - High (102), 30% in A-range.
- MT-2: If less than 80, extra work to catch-up
- Final: Cumulative (20% from topics covered after MT-2.

NC STATE UNIVERSITY

MT-2 Quick Review

- Q1 (Regression: 40%) – Similar to HW3-Q1
- Q2 (30%) (a) – W10-C2-Slide 5.
  - (b) principal components – similar to working example from W11-C1
- Q3 (30%) (a) – Sign test; (b) Wilcoxon rank sum test (similar to example from W11-C2)

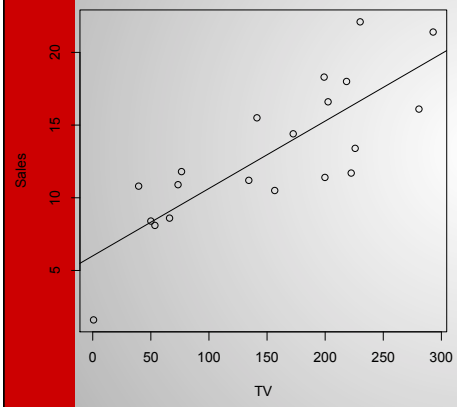
NC STATE UNIVERSITY

How do you find accuracy of a model

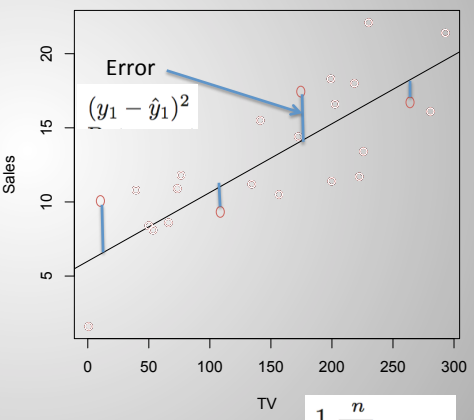
- General concepts
  - Training data (to fit a model)
  - Test data (to validate a model)

NC STATE UNIVERSITY

Regression



Training



Test Error $\frac{1}{n}\sum_{i=1}^{n} \text{MSE}_i$

NC STATE UNIVERSITY

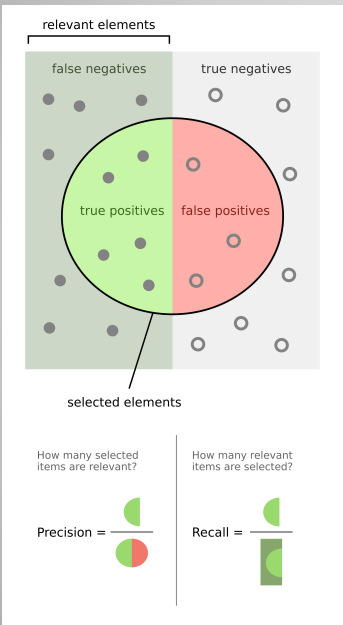
Classification

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

NC STATE UNIVERSITY

Classification Accuracy Measures



		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F - measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

NC STATE UNIVERSITY

Classification Accuracy Measures

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	10	0
	Class=No	10	980

$$\text{Precision (p)} = \frac{10}{10+10} = 0.5$$

$$\text{Recall (r)} = \frac{10}{10+0} = 1$$

$$\text{F - measure (F)} = \frac{2*1*0.5}{1+0.5} = 0.62$$

$$\text{Accuracy} = \frac{10}{1000} = 0.01$$

NC STATE UNIVERSITY

Classification Accuracy Measures

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	10	0
	Class=No	10	980

$$\text{Precision (p)} = \frac{10}{10+10} = 0.5$$

$$\text{Recall (r)} = \frac{10}{10+0} = 1$$

$$\text{F-measure (F)} = \frac{2*1*0.5}{1+0.5} = 0.62$$

$$\text{Accuracy} = \frac{990}{1000} = 0.99$$

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	1	9
	Class=No	0	990

$$\text{Precision (p)} = \frac{1}{1+0} = 1$$

$$\text{Recall (r)} = \frac{1}{1+9} = 0.1$$

$$\text{F-measure (F)} = \frac{2*0.1*1}{1+0.1} = 0.18$$

$$\text{Accuracy} = \frac{991}{1000} = 0.991$$

NC STATE UNIVERSITY

Challenges

- For given test dataset, we can obtain error/ accuracy, but how accurate (variable) 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

NC STATE UNIVERSITY

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)

NC STATE UNIVERSITY

Cross-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

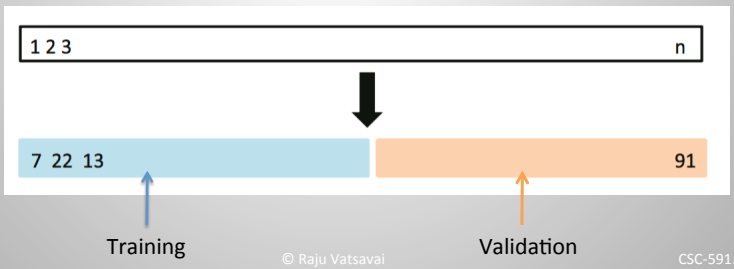
Model Fit

Holdout data for validation

NC STATE UNIVERSITY

Validation Set Approach

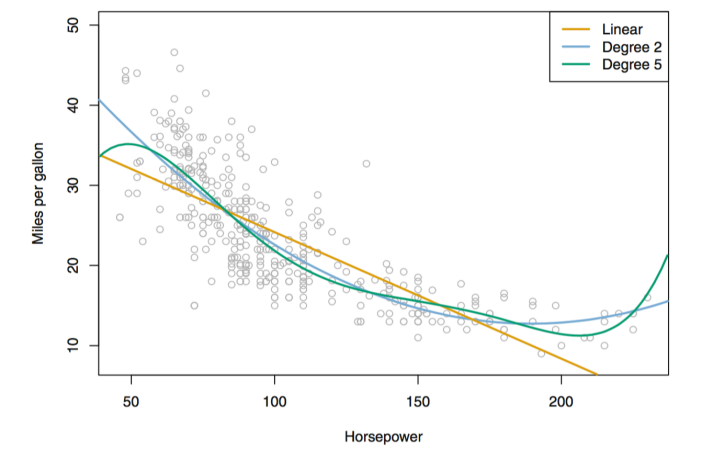
- Simplest approach
- Randomly divide n available samples into a training set and validation or holdout set
- The resulting validation set error rate is an estimate of test error rate



NC STATE UNIVERSITY

Accuracy on Different Models

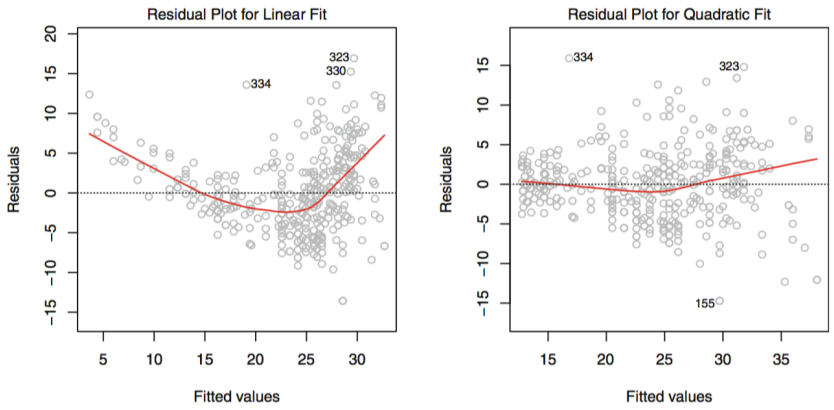
- First, we can fit different models for same data
- What do you think of accuracy for these models?



NC STATE UNIVERSITY

Accuracy on Different Models

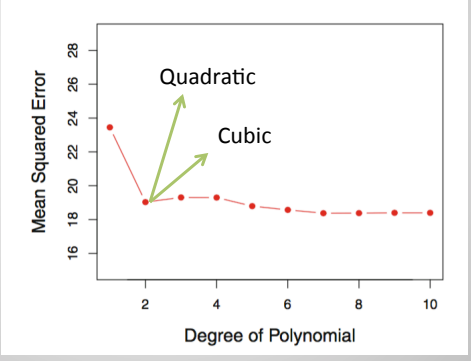
- Let us look at distribution of residuals



NC STATE UNIVERSITY

Accuracy on Different Models

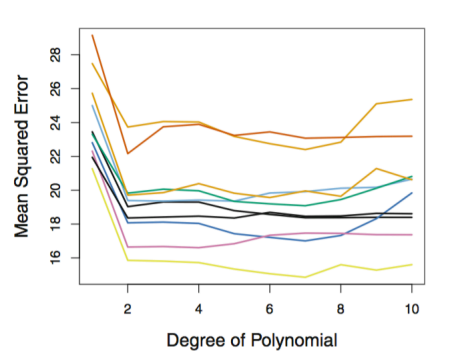
- Does higher order terms (increasing complexity) improves accuracy?



NC STATE UNIVERSITY

Accuracy on Different Models

- How about accuracy on different random splits of training data



NC STATE UNIVERSITY

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.

NC STATE UNIVERSITY

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

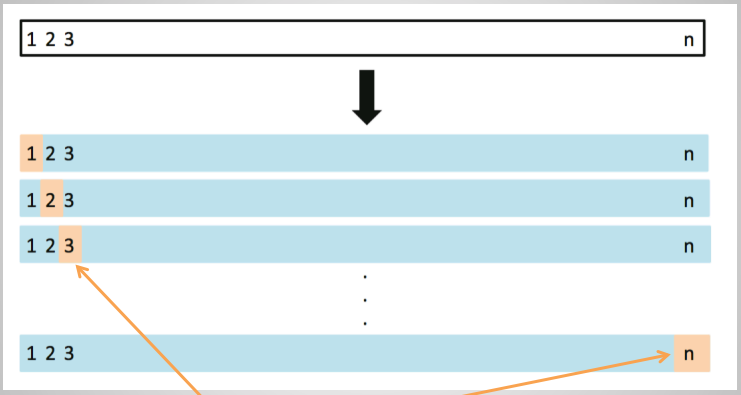
NC STATE UNIVERSITY

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



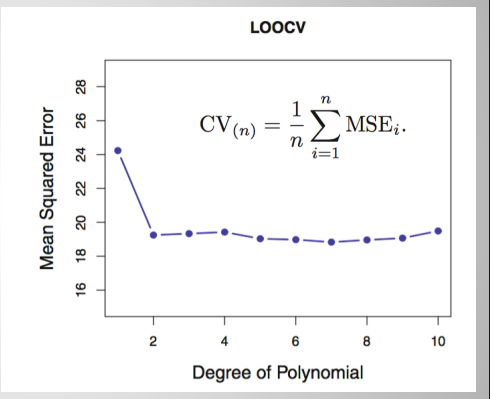
"n" validation sets of size one

NC STATE UNIVERSITY

Leave-one-out cross-validation

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

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



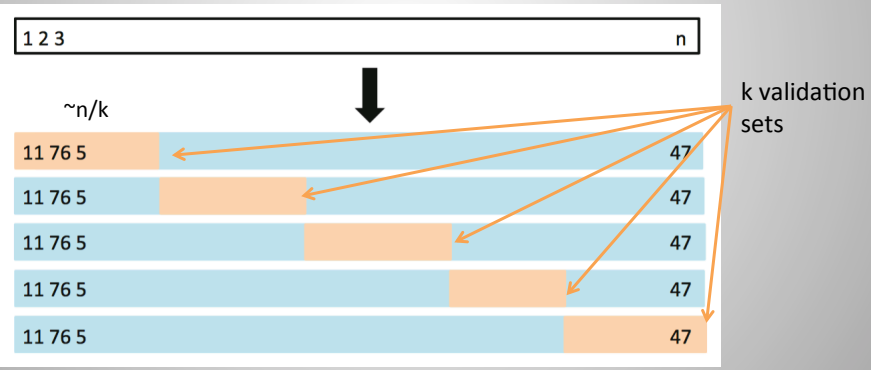
- Where leverage statistic h_i is given by

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}.$$

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.



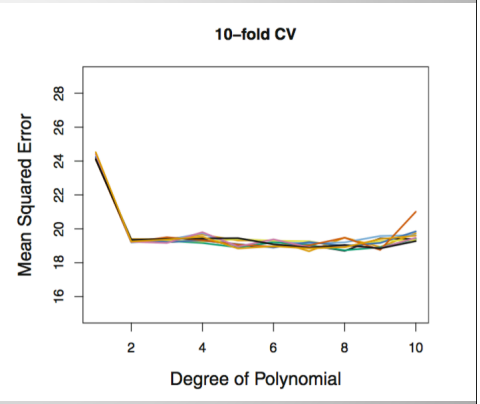
NC STATE UNIVERSITY

k-fold Cross-Validation

- The k-fold CV estimate is computed by averaging

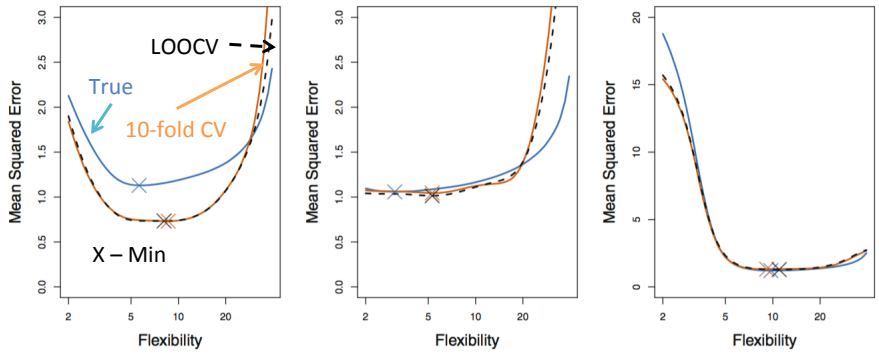
$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$$

- 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



NC STATE UNIVERSITY

True vs. Estimated Test MSE



True estimates are from simulated data

Goals of CV

- How well a statistical learning method performs on independent data
- We may be interested in location of minimum error to determine methods that result in lowest error

NC STATE UNIVERSITY

Acknowledgements

- Introduction to statistical learning with R (read all of chapter 5; except 5.1.5 which is optional).
- See 5.3 for R example on bootstrap
- Read chapter 6.1