Model Selection

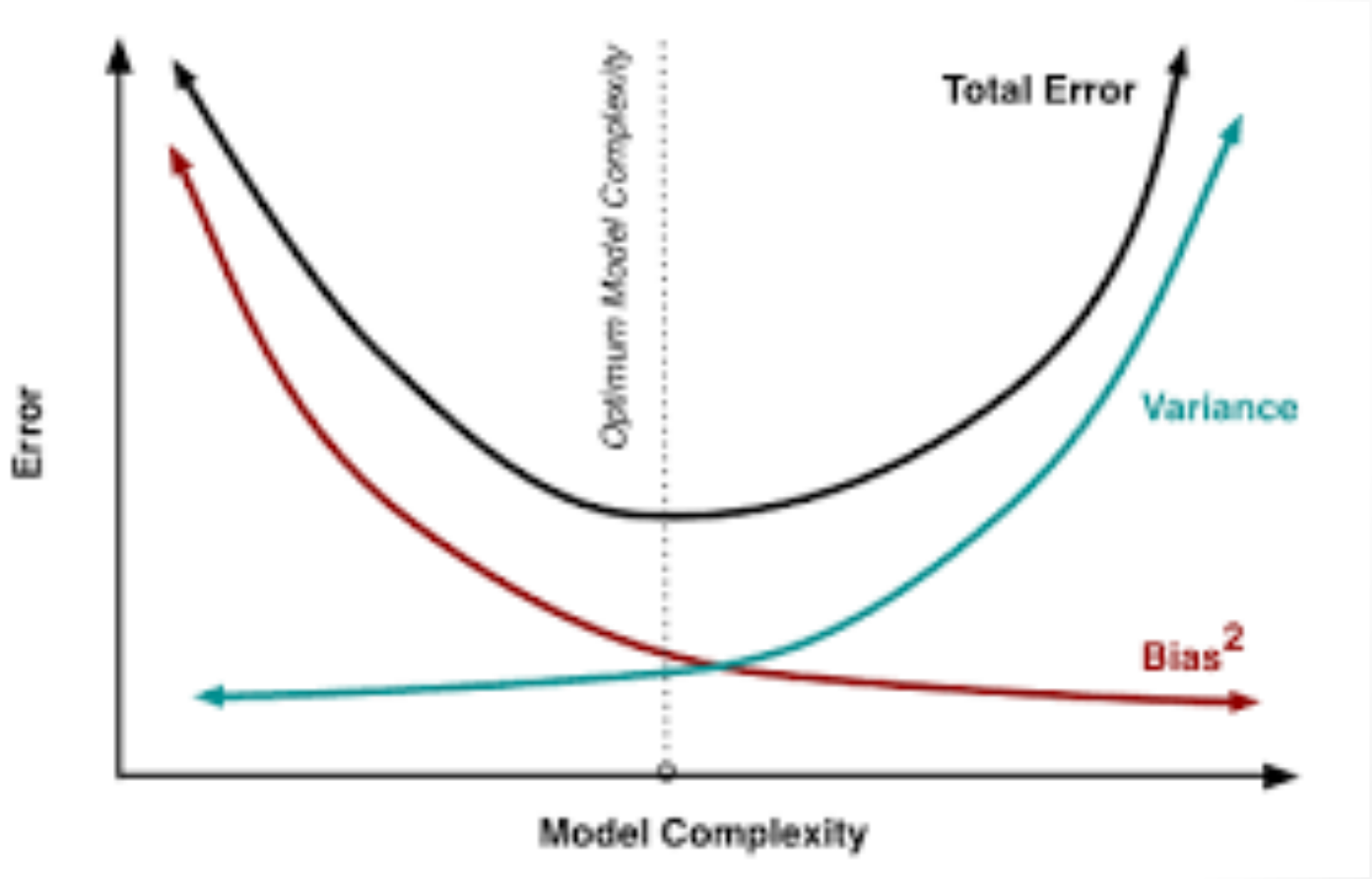
Instructor: Hamed Hasheminia

Lecture 5

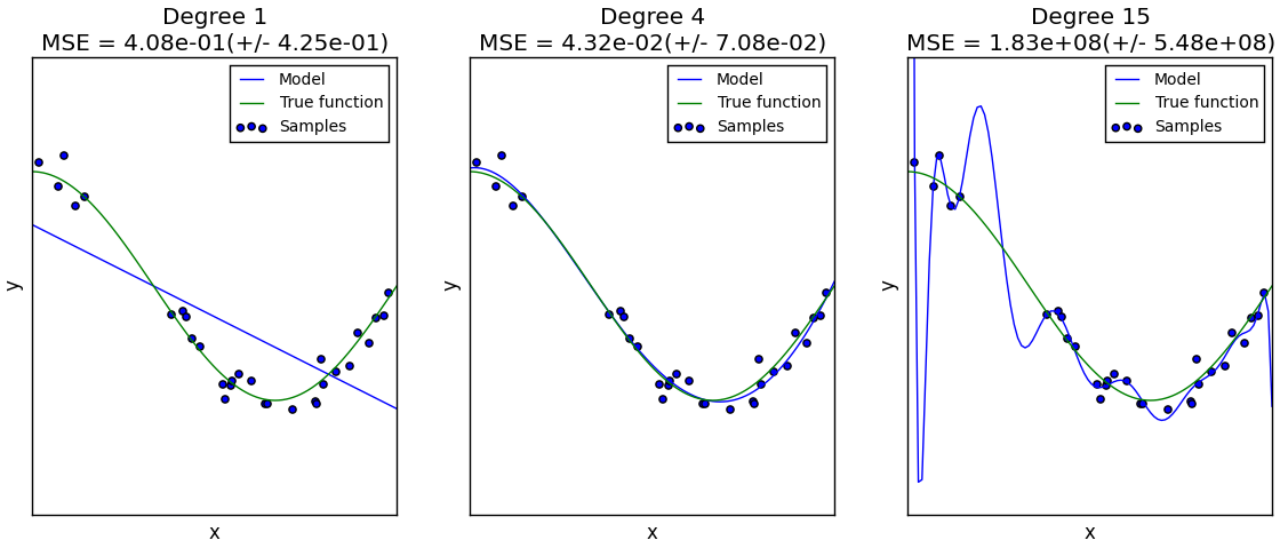
Agenda

- Bias-Variance Trade off
- Validation (Test vs Train set)
- Cross-Validation
- Ridge and Lasso Regression
- (Optional) Backward Selection, Forward Selection, All Subset Selection. (If you want to use these methods you need to use R)

Bias – Variance Trade-off



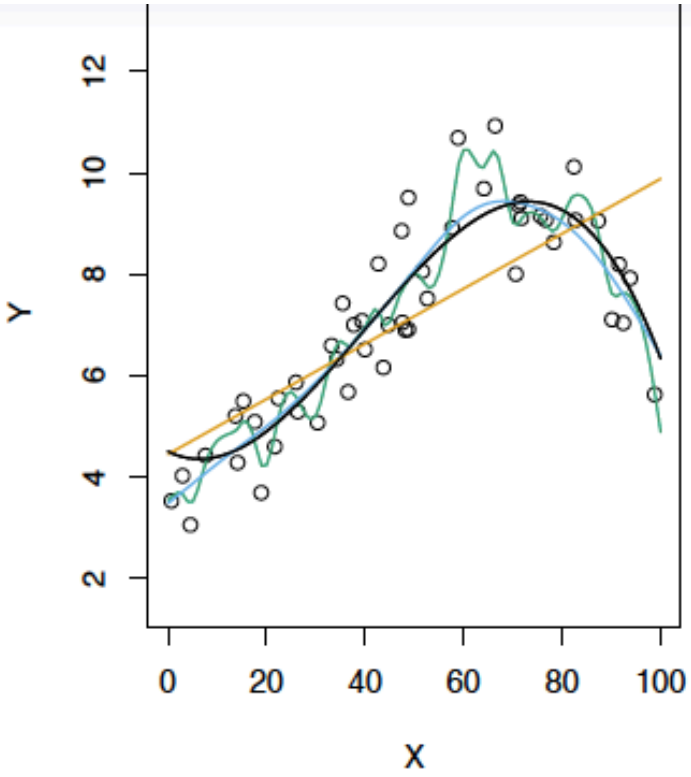
Bias – Variance Trade-off

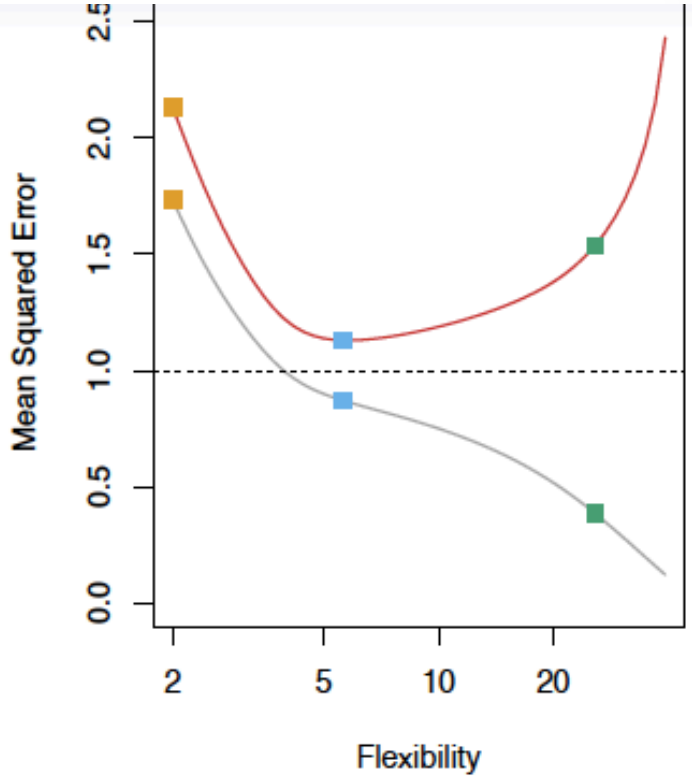


Validation

- We divide our dataset to test and training.
- We then train our algorithms with train data.
- We compute test error by using predictions of our models on our test set.

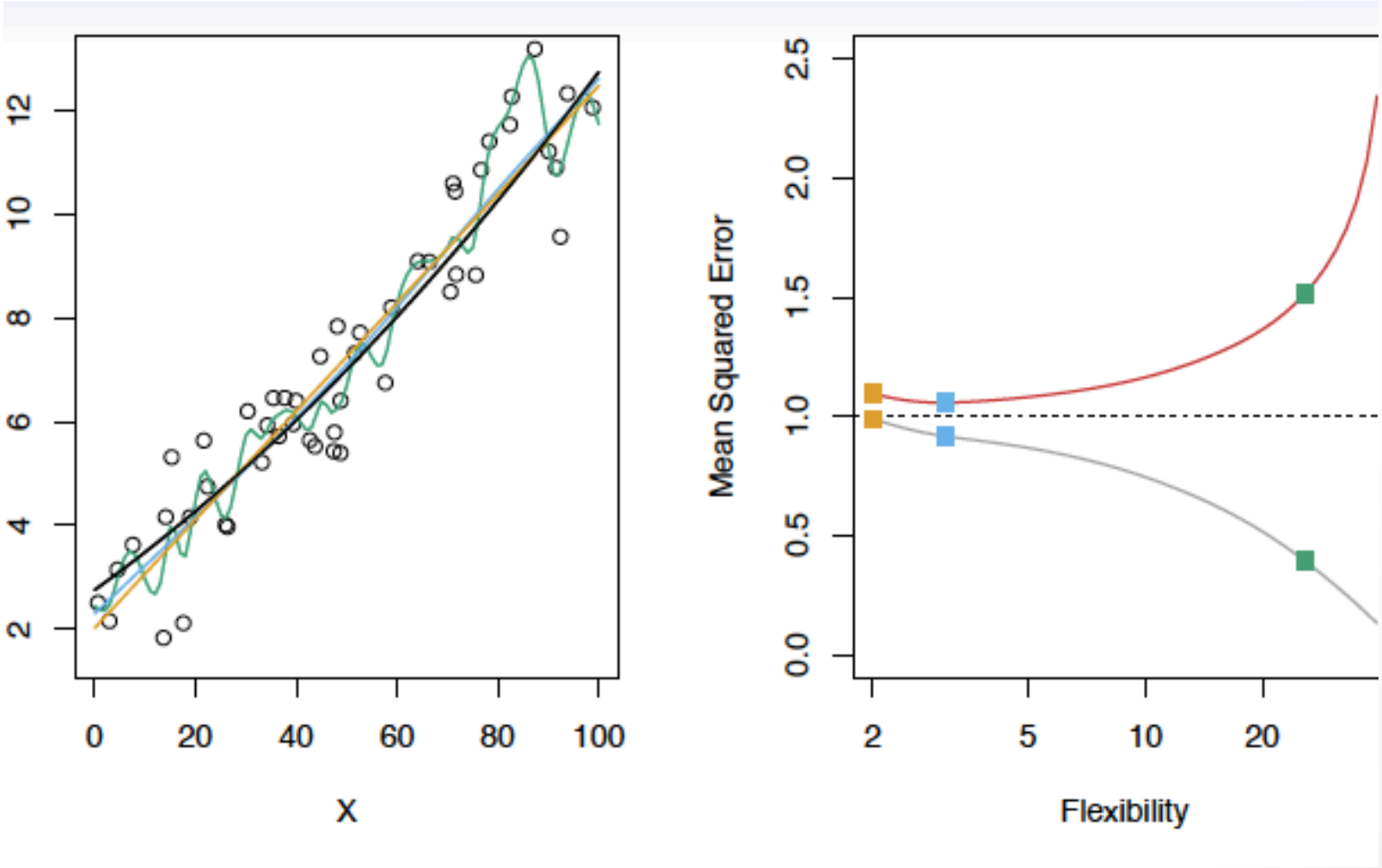
Validation example



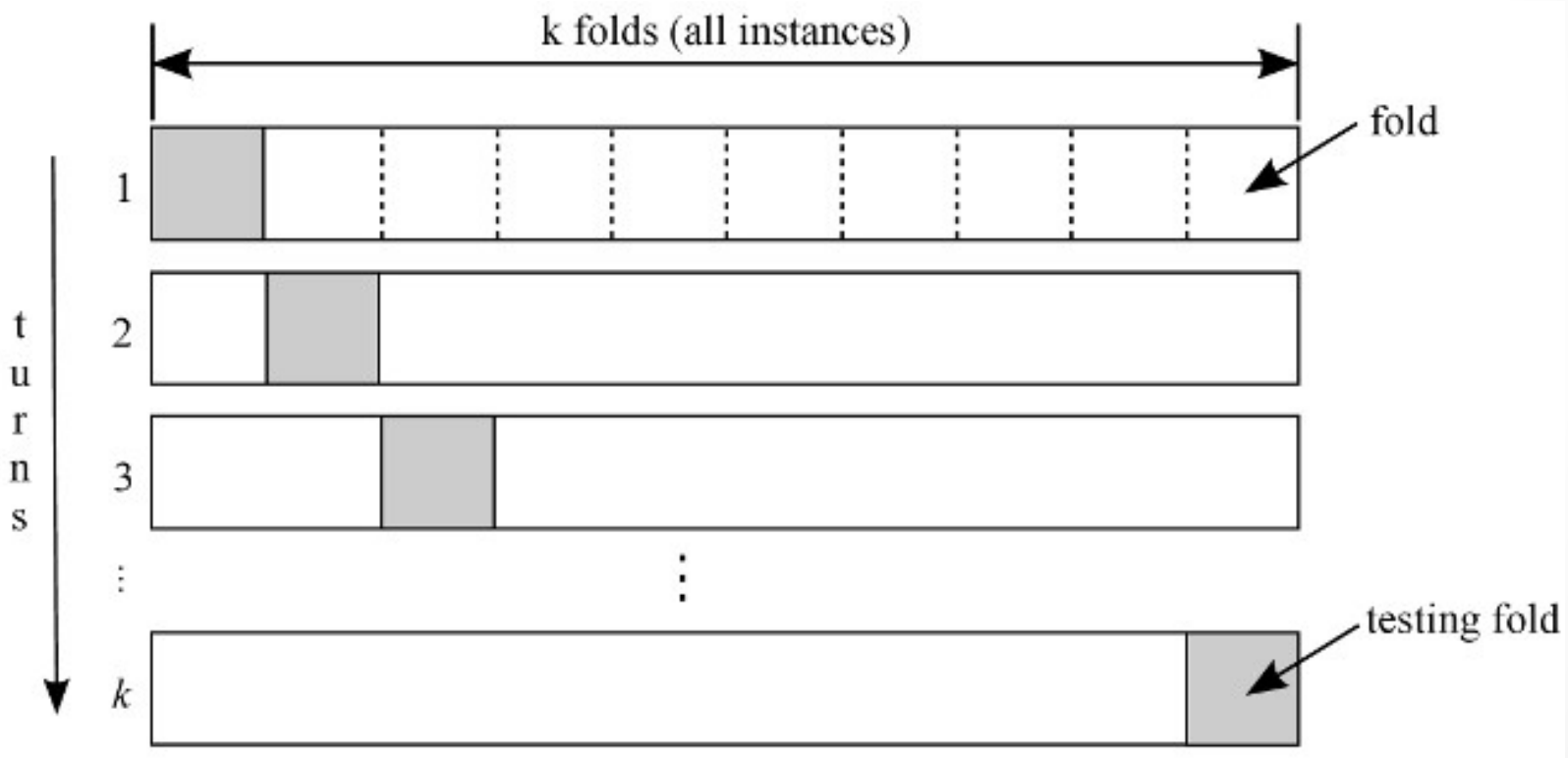


Black curve is truth. Red curve on right is MSE_{Te}, grey curve is MSE_{Tr}. Orange, blue and green curves/squares correspond to fits of different flexibility.

Validation Example



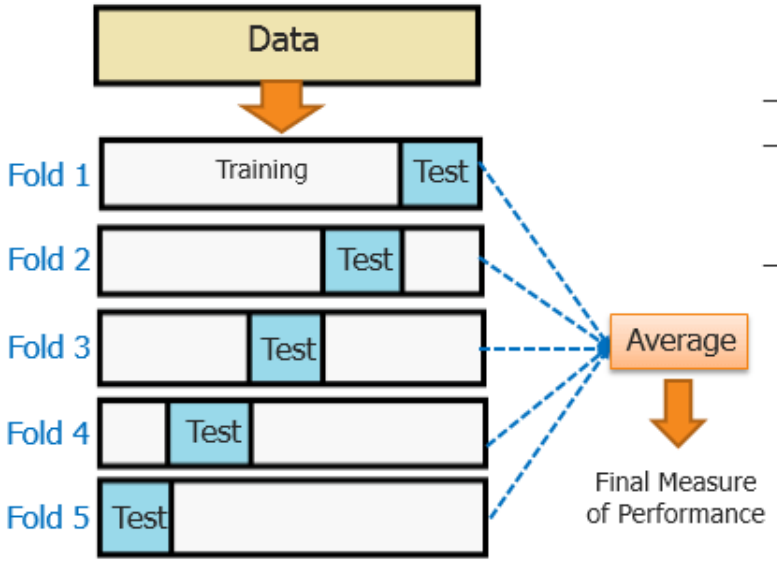
K-fold Cross-Validation



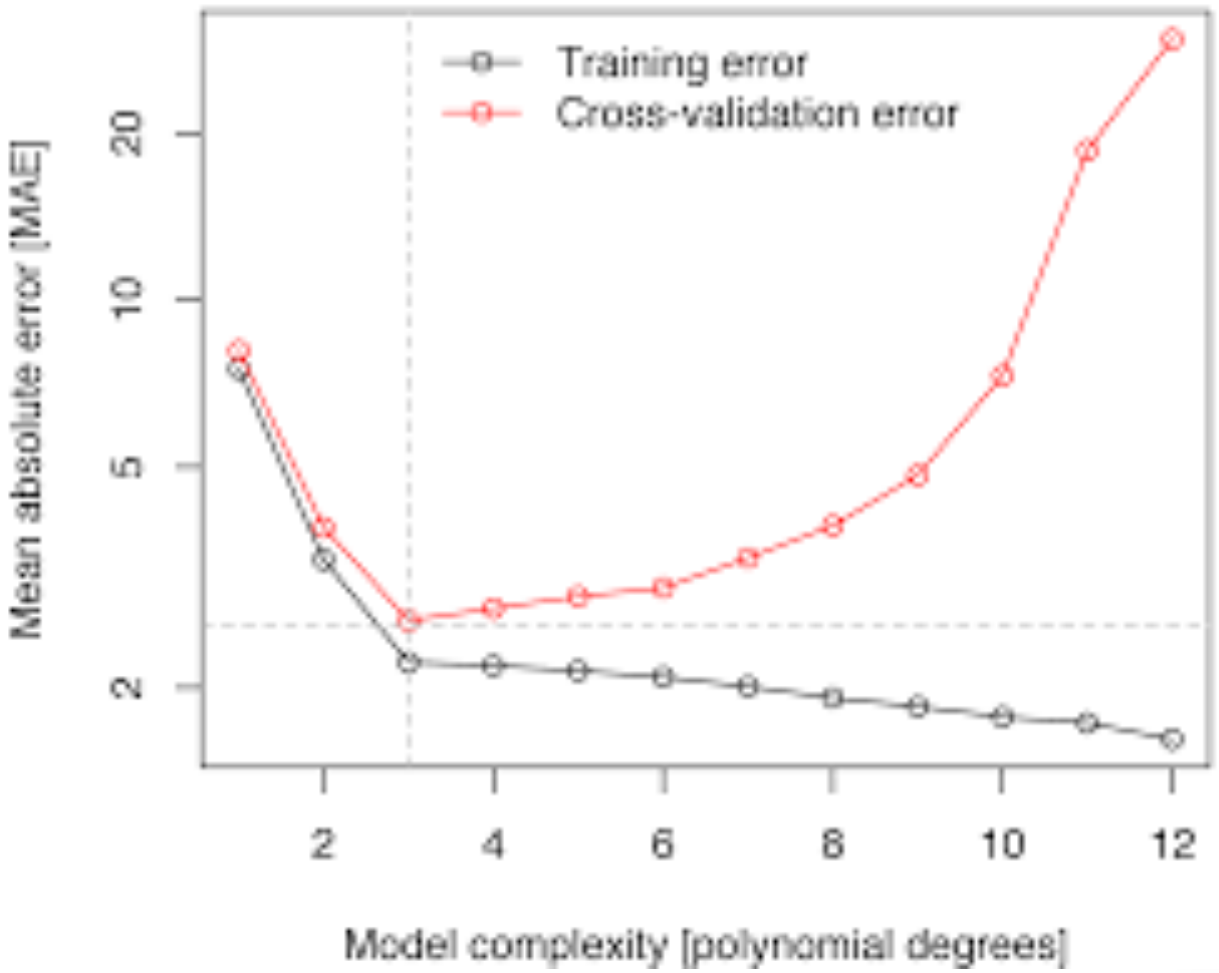
K-fold Cross-Validation

Cross-Validation (CV)

edureka!



- → Technique to validate models/classifiers
- → Method to estimate how accurately the model generalizes to unseen data i.e., how well it performs/predicts
- → K-fold CV
  - » Most popular
  - » k is typically set to 10
  - » Every sample/record is used both in training and test sets


Training error
Cross-validation error
Mean absolute error [MAE]
2
5
10
20
2
4
6
8
10
12
Model complexity [polynomial degrees]

Ridge Regression (Shrinkage method)

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

In ridge regression we penalize our model for adding more variables:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

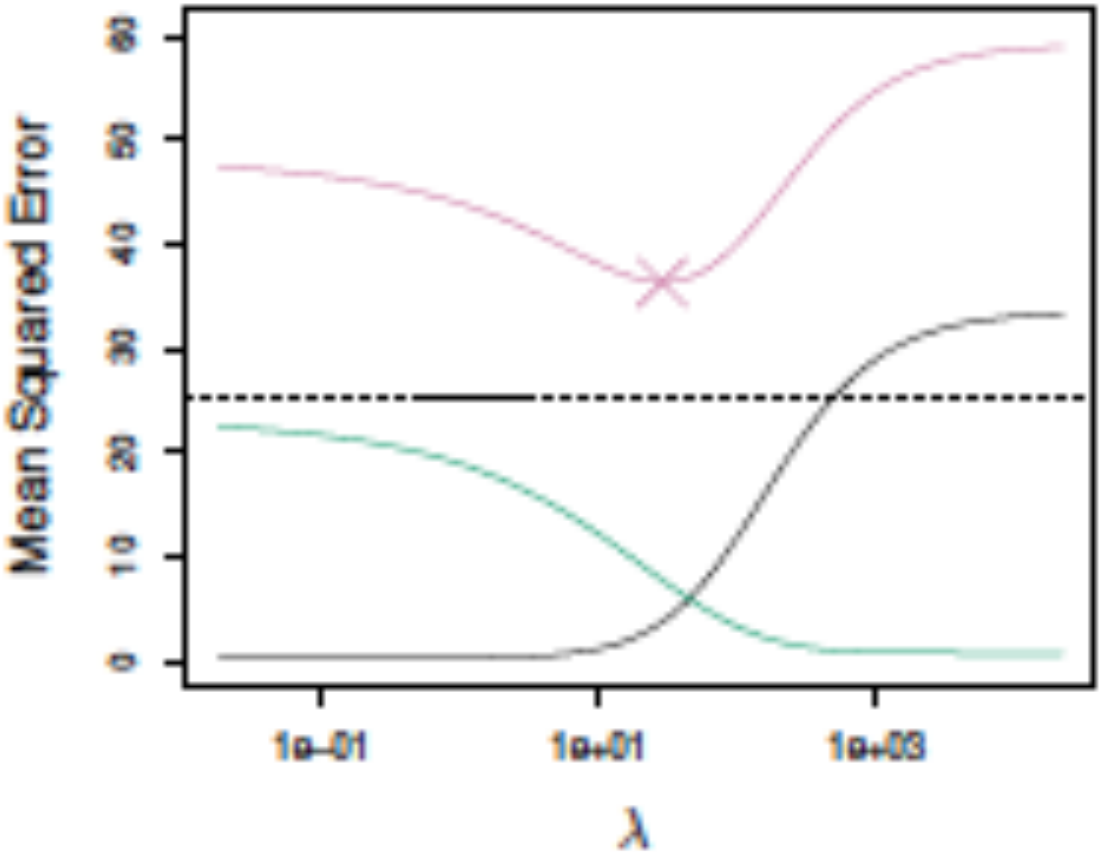
Lambda > 0, is a tuning parameter. We set it by cross-validation

Ridge Regression

- Coefficients in Ridge Regression are sensitive to scale of your data. It is highly recommended that you standardize your data before performing Ridge Regression. One way to standardize your data is dividing it by MAX. The other way, is chaining it to z-values.

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}}$$

Bias-Variance Trade-off in choosing the best lambda



Lasso Regression

Lasso is very similar to Ridge regression. The only difference is it penalizes objective function with Norm 1 – summation of absolute values - instead of Norm 2.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|$$

As in Ridge regression, selecting a good lambda for the lasso is critical; cross-validation is again the method of choice.

Lasso Regression

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero.
- However, in the case of the lasso, the Norm 1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter lambda is sufficiently large.

Other methods of model selection

- Forward Selection (Python is not that good for this method)
- Backward selection(Python is not that good for this method)
- All subset selection(Python is not that good for this method)
- Principle component methods. (Will cover that later in the course.)

Summary

- Bias-Variance Trade-off
- Validation
- Cross-Validation
- Ridge-Regression
- Lasso-Regression
- Backward, Forward, and all subset selection
- Resource (Chapter 6): http://www-bcf.usc.edu/~gareth/ISL/data.html
- Video Resource: http://dataminingclass.com/index.php/lectures/linear-model-selection-and-regularization/