Week 6: Model Validation

Kevin Lee

Department of Statistics
Western Michigan University

Model Validation

- In principle, model validation is very simple: after choosing a model and its hyperparameters, we can estimate how effective it is by applying it to some of the data and comparing the prediction to the known value.

The Validation Set Approach

- Advantages:
  - Simple and computationally less expensive.
- Disadvantages:
  - The validation error rate can be highly variable.
  - Only a subset of observations is used to fit the model (training data).

Visualization of Validation Set Approach

K-fold Cross Validation

- With k-fold Cross Validation, we divide the data set into k different parts (e.g. k = 5, or k = 10).
- We then remove the first part, fit the model on the remaining k-1 parts, and see how good the predictions are on the left out part (i.e. compute the error rate on the first part).
- We then repeat this k different times taking out a different part each time.
- By averaging the k different error rate's we get an estimated test error rate,

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

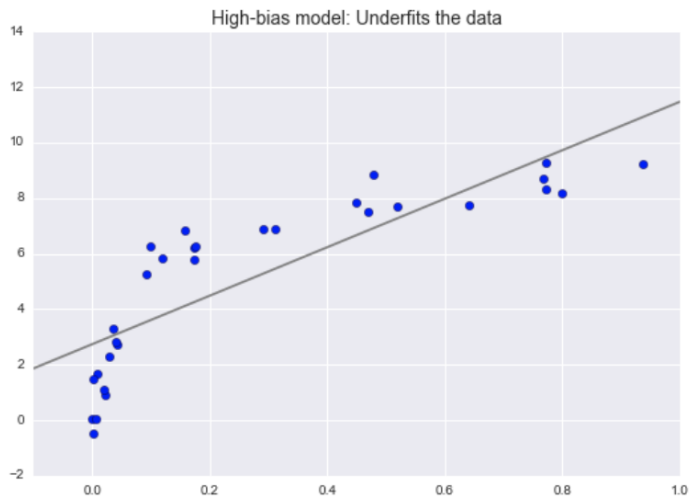
Visualization of 5-fold Cross Validation

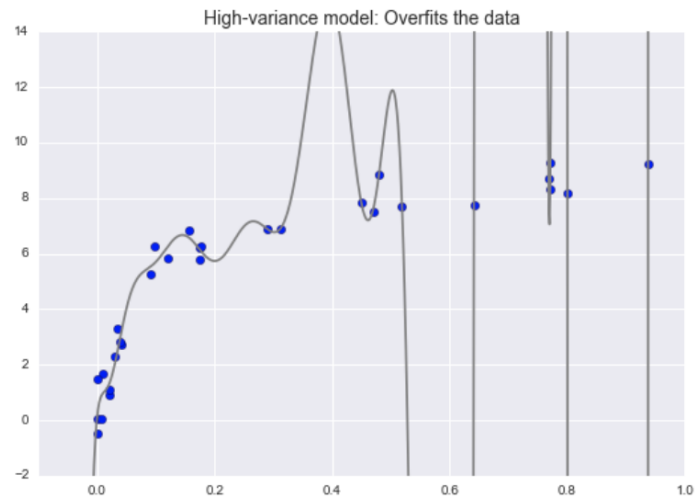
Selecting the Best Model

- If our model is underperforming, how should we move forward? There are several possible answers:
  - Use a more complicated / more flexible model
  - Use a less complicated / less flexible model
  - Gather more training samples
  - Gather more data to add features to each sample

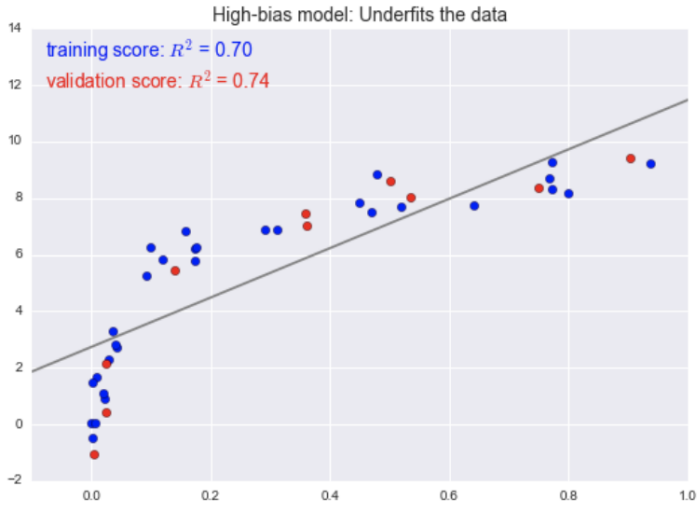
The Bias-Variance Trade-Off

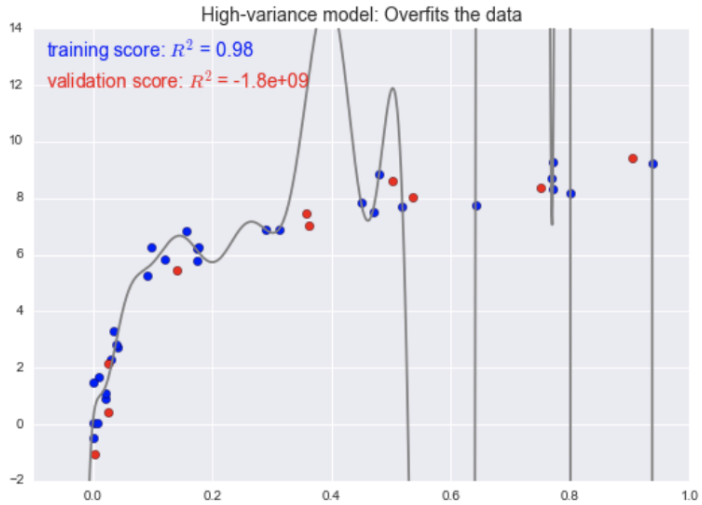
- In principle, the question of "the best model" is about finding a sweet spot in the trade-off between bias and variance.



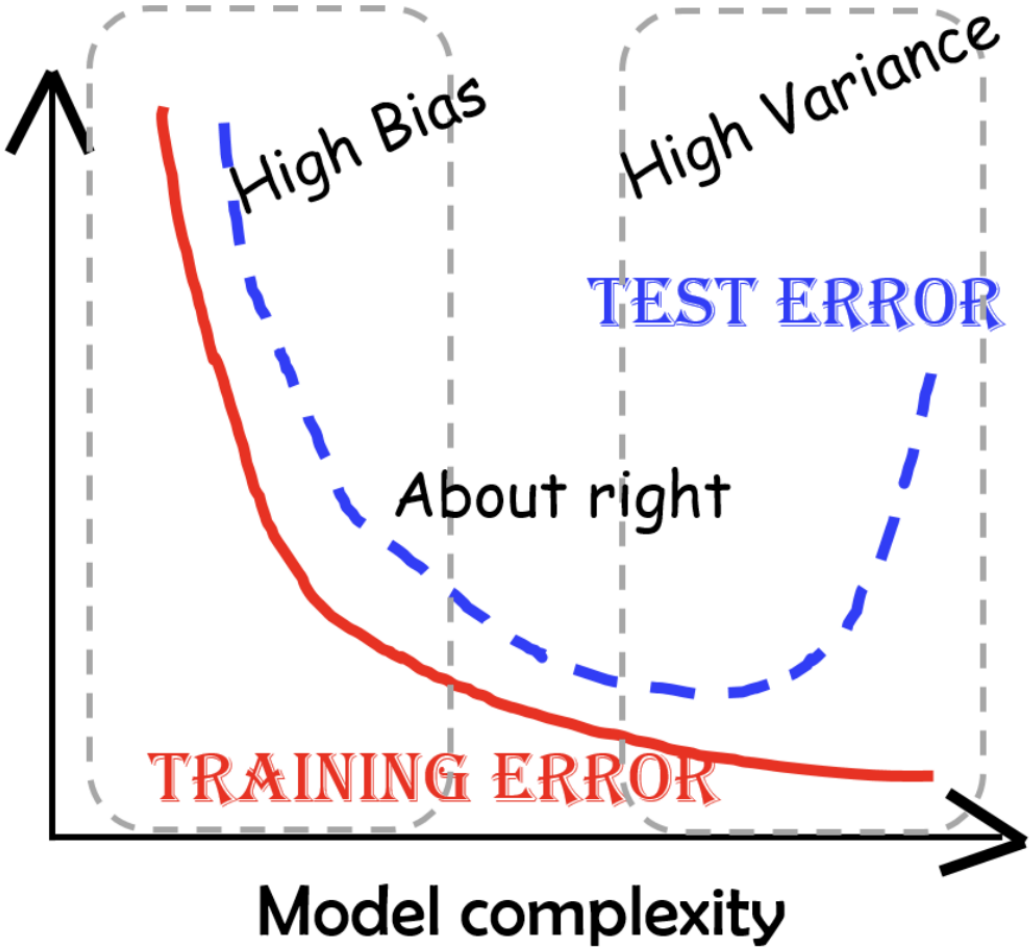


The Bias-Variance Trade-Off

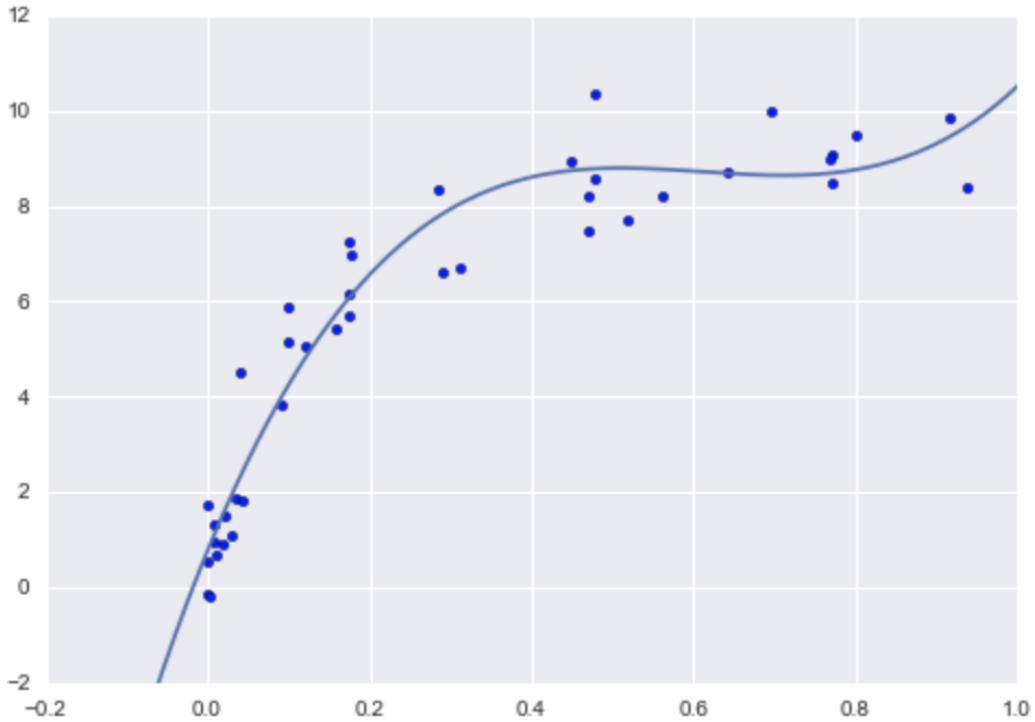




The Bias-Variance Trade-Off



The Bias-Variance Trade-Off



Feature Engineering

- In the real world, data rarely comes in a tidy format.
- Feature engineering is taking whatever information you have about your problem and turning it into numbers that you can use to build your feature matrix.