Resampling Methods for Uncertainty

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

Agenda

Motivation

- The real world is messy.
- Recognizing this mess will differentiate a sophisticated and useful analysis from one that is hopelessly naive.
- This is especially true for highly complicated models, where it becomes tempting to confuse signal with noise and hence "overfit."
- The ability to deal with this mess and noise is the most important skill you need.

Agenda

What are resampling methods?

- Tools that involves repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain more information about the fitted model
  - Parameter Assessment: estimate standard errors
  - Model Assessment: estimate test error rates
  - They are computationally expensive! But these days we have powerful computers

Agenda

The Bootstrap

- In general terms:
  - $Y_i \; i = 1, \ldots, n$
  - θ is the magnitude of interest
- To calculate it's variance
  1. Sample of size n with replacement (*bootstrap sample*)
  2. Compute $\hat{\theta}_j \; j = 1, \ldots, B$
  3. Repeat B times
  4. Calculate

$$\hat{V}(\hat{\theta})_B = \frac{1}{B}\sum_{j=1}^{B}(\hat{\theta}_j - \bar{\hat{\theta}})^2 \tag{1}$$

The Bootstrap

- There are two key properties of bootstrapping that make this seemingly crazy idea actually work.

  1. Each bootstrap sample must be of the same size (N) as the original sample
  2. Each bootstrap sample must be taken with replacement from the original sample

Example: Elasticity of Demand for Gasoline

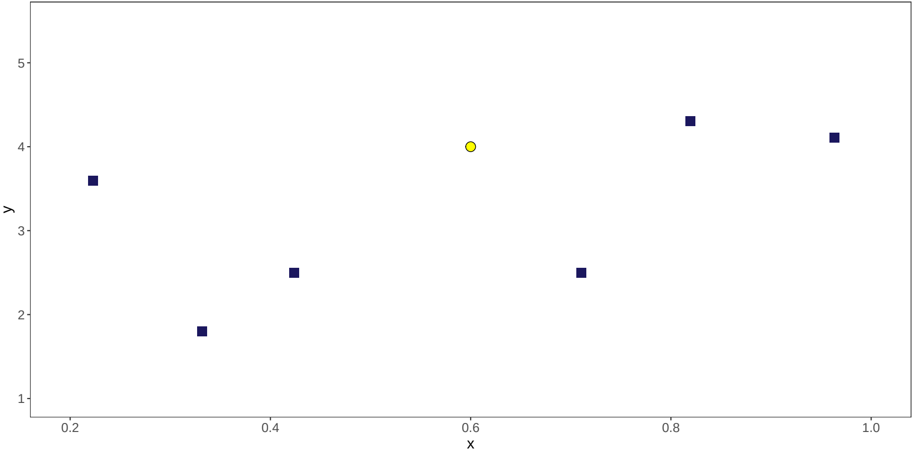
photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Agenda

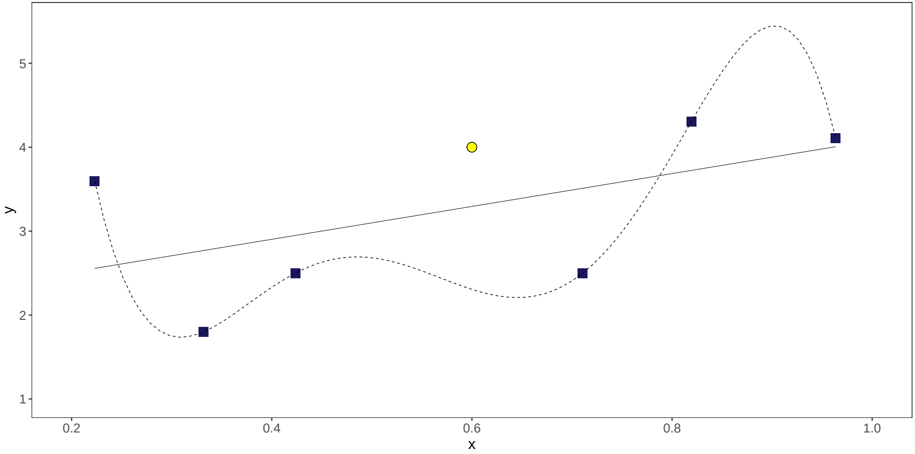
Resampling methods for Out of Sample Prediction

- The goal of machine learning is *out of sample* prediction i.e. the ability to predict on new data.
- Overfit: complex models predict well in sample but bad our of sample
- How to choose the optimal compelexity level?

Resampling methods for Out of Sample Prediction



Resampling methods for Out of Sample Prediction



Resampling methods for Out of Sample Prediction

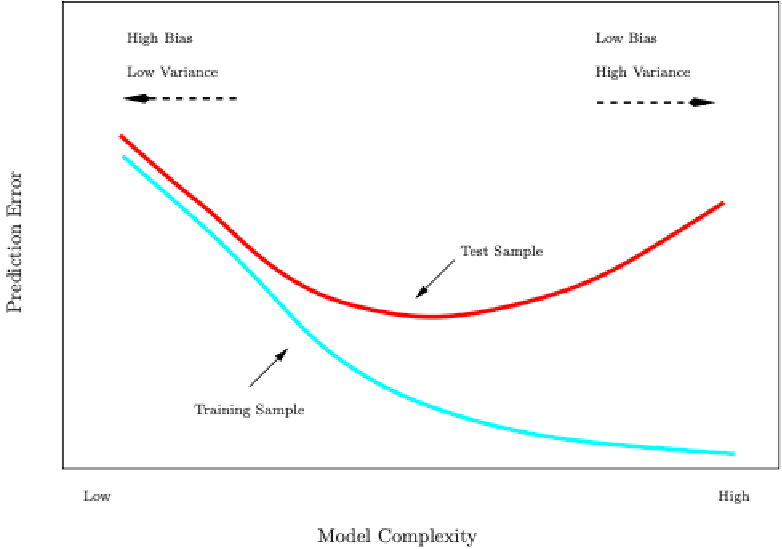
- Two concepts
  - *Test Prediction Error*: is the prediction error in a test sample

$$Err_{\mathcal{T}est} = E[L(Y, \hat{Y})|\mathcal{T}est] \tag{2}$$

  - *Training Prediction Error*: is the prediction error in the training sample

$$Err_{\mathcal{T}rain} = E[L(Y, \hat{Y})|\mathcal{T}rain] \tag{3}$$

Resampling methods for Out of Sample Prediction



Resampling methods for Out of Sample Prediction

- Two concepts
  - *Test Prediction Error*: is the prediction error in a test sample
$$Err_{\mathcal{T}est} = E[L(Y, \hat{Y})|\mathcal{T}est] \tag{2}$$
  - *Training Prediction Error*: is the prediction error in the training sample
$$Err_{\mathcal{T}rain} = E[L(Y, \hat{Y})|\mathcal{T}rain] \tag{3}$$
- Then how do we estimate the test prediction error?

Resampling methods for Out of Sample Prediction

- In the absence of a very large designated test set we can use some techinques:
    1. Validation Set
    2. Loocv
    3. K-fold Crossvalidation

Resampling methods for Out of Sample Prediction

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/