Generalization. Out of Sample Performance.

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

Agenda

1. Review
2. Generalization. Out-of-sample Performance
3. Out-of-Sample Error Estimation
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
4. Review

Predicting Well

$$y = f(X) + u \tag{1}$$

- Interest on predicting y

Linear Regression

$$y = f(X) + u \tag{2}$$
$$= \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u \tag{3}$$

- If $f(X) = X\beta$, obtaining $f(.)$ boils down to obtaining β
- We learn these βs by minimizing a loss function on the sample.

Linear Regression

- Quadratic loss $\rightarrow$ OLS

$$\mathcal{L}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f_\beta(X_i)\right)^2 \tag{4}$$

$$= \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{ji}\right)^2 \tag{5}$$

- Compute βs uisng sample data
  - QR
  - SVD
  - Gradient Descent

Agenda

Generalization Overview

- In ML we care in out-of-sample prediction
- Generalization refers to a model's performance on unseen data.
- The ultimate goal is **not** minimizing the in-sample loss, but achieving low error out-of-sample on unseen data.

Training and Test Loss

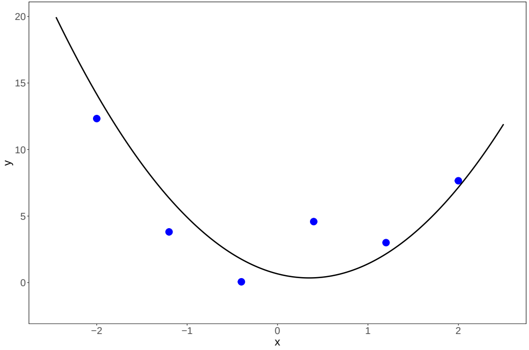
- Unseen data is typically refered as **test data**,
- While the sample data is called the **training data**.
- The expected loss over the test distribution is called the test loss.
- Test loss is defined as:

$$L(\theta) = \mathbb{E}_{(X,y)\sim F}[(y - f_\beta(X))^2]$$

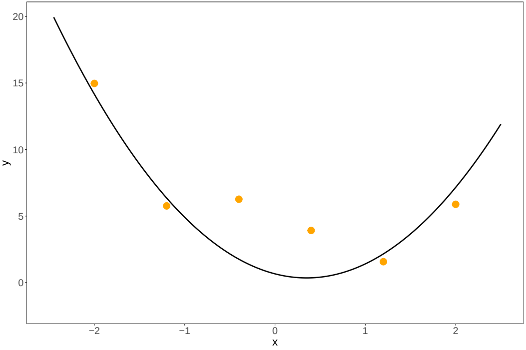
Overfitting and Underfitting

- Successfully minimizing training error does not always result in a small test error.
- A model is said to overfit if it predicts accurately on training data but poorly on test (unseen) data.
- A model underfits if its training error is relatively large, which usually means test error is also large.
- Understanding overfitting and underfitting helps in choosing appropriate model parameterizations.

Overfitting and Underfitting. Bias-Variance Tradeoff



(a) Training Data



(b) Testing Data

Overfitting and Underfitting. Bias-Variance Tradeoff

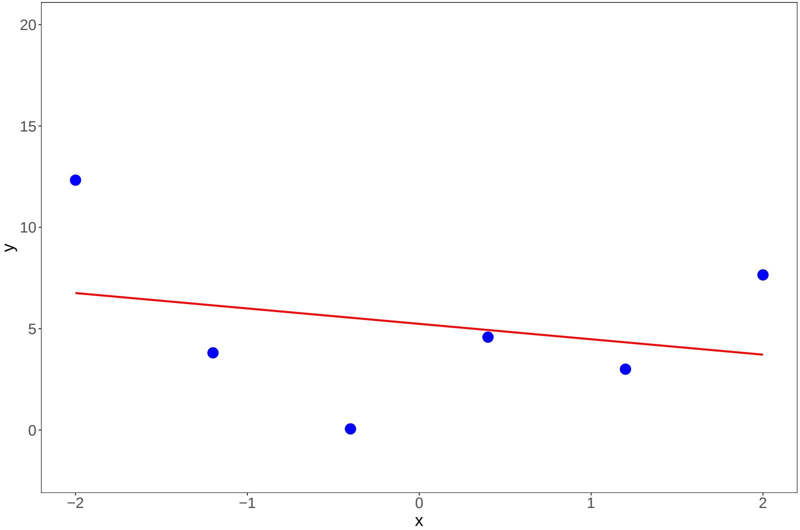
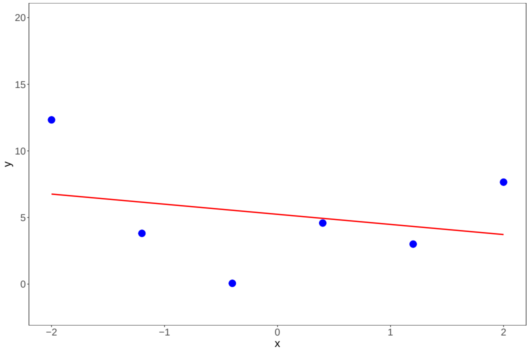


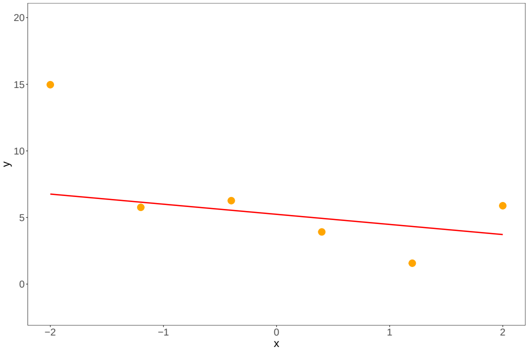
Figure 2: Training Data

Overfitting and Underfitting. Bias-Variance Tradeoff

Out-of-Sample Performance



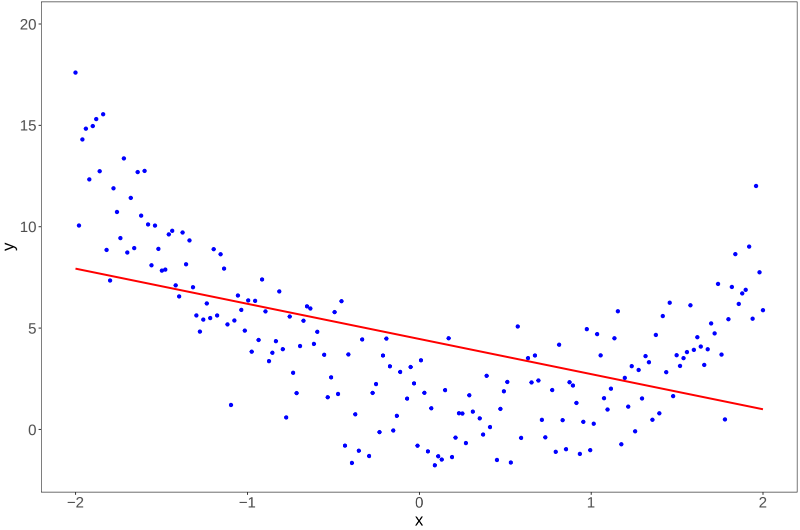
(a) Training Data



(b) Testing Data

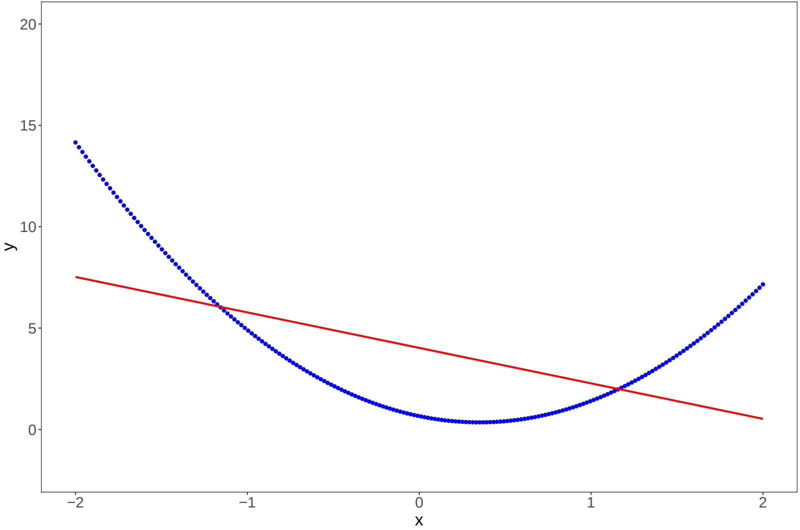
Overfitting and Underfitting. Bias-Variance Tradeoff

More data?



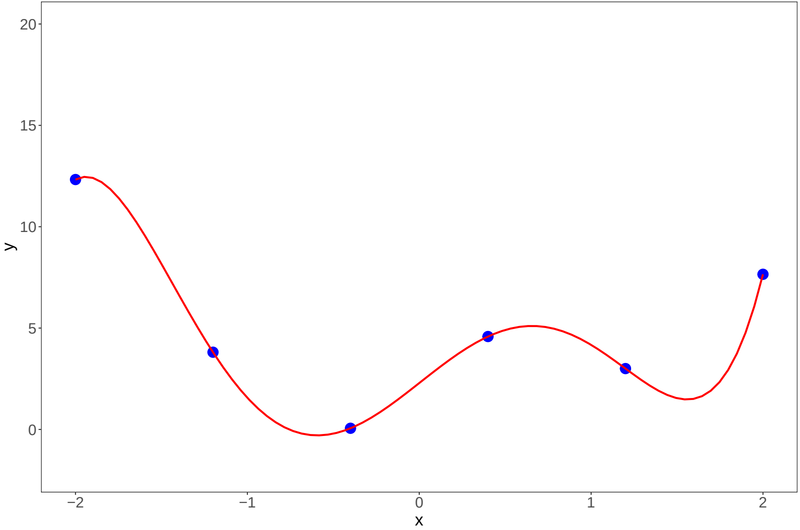
Overfitting and Underfitting. Bias-Variance Tradeoff

Noiseless data?



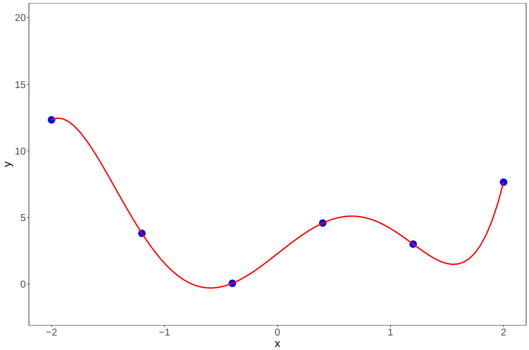
Overfitting and Underfitting. Bias-Variance Tradeoff

More Complex Model

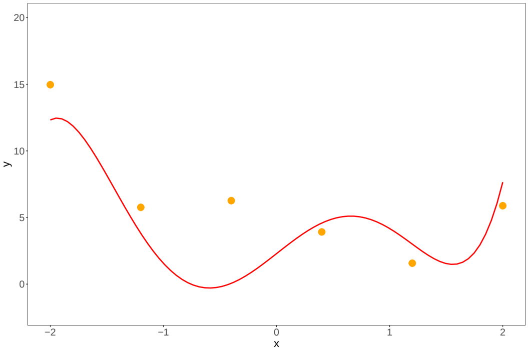


Overfitting and Underfitting. Bias-Variance Tradeoff

Out-of-Sample Performance



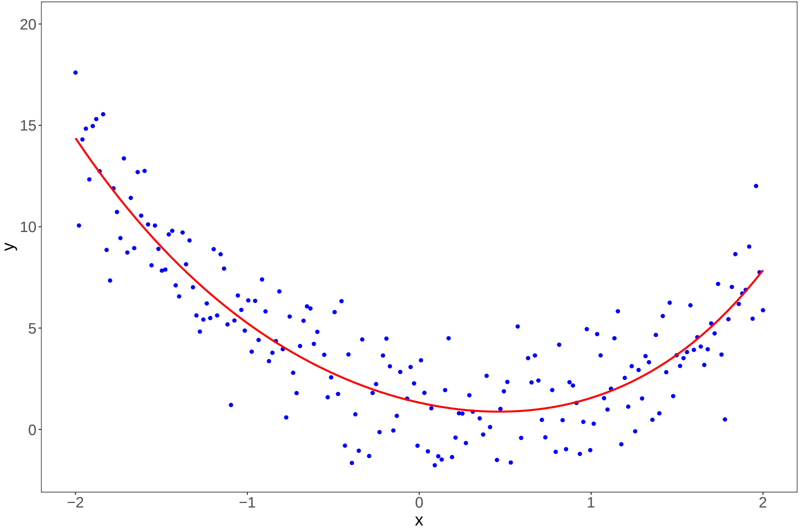
(a) Training Data



(b) Testing Data

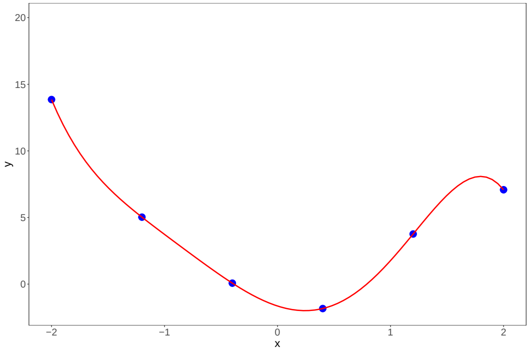
Overfitting and Underfitting. Bias-Variance Tradeoff

More Data

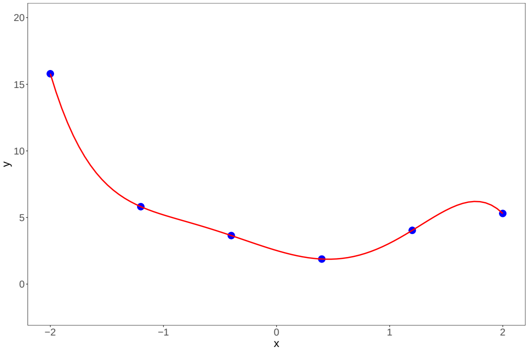


Overfitting and Underfitting. Bias-Variance Tradeoff

Variance



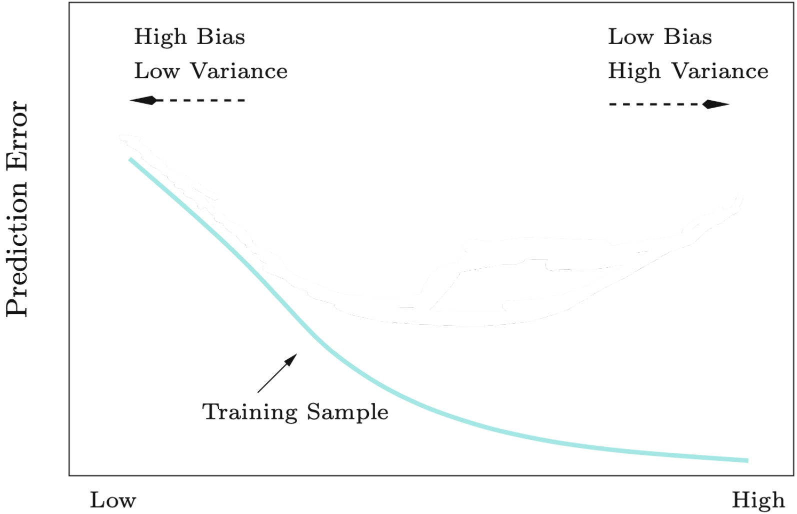
(a) Training Data 2



(b) Training Data 3

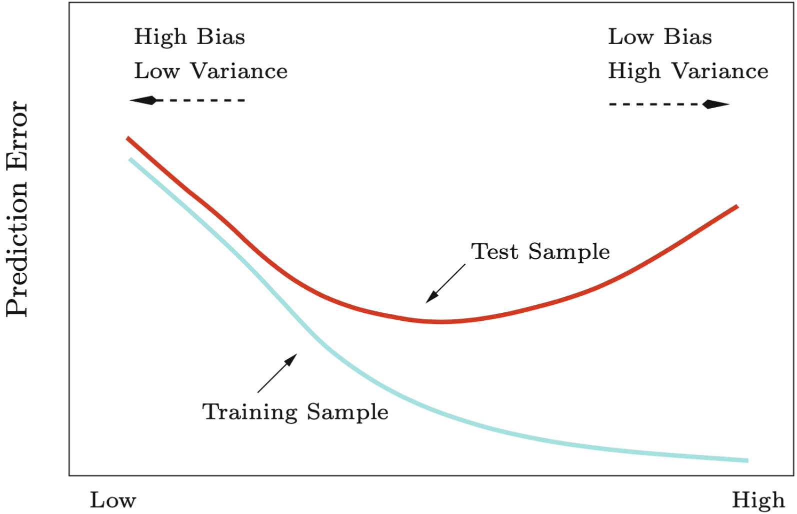
Overfitting and Underfitting. Bias-Variance Tradeoff

In-Sample Prediction and Overfit



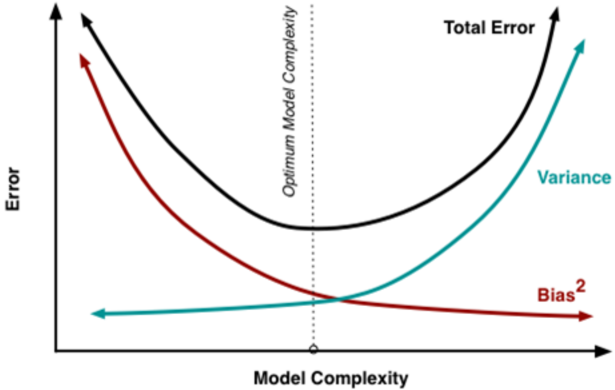
Overfitting and Underfitting. Bias-Variance Tradeoff

Out-of-Sample Prediction and Overfit



Mathematical Decomposition for Regression

Bias-Variance Tradeoff



Source: https://tinyurl.com/y4lvjxpc

- ML best kept secret: By tolerating some bias we can have significant gains in variance

Example polynomial revisited

- Suppose that the true model is $y = f(X) + u$
  - f is a polynomial of degree p^*
  - p^* is finite but unknown
  - $E(u) = 0$ and $V(u) = \sigma^2$
- We fit polynomials with increasing degrees $p = 1, 2, ...$
- What happens when we increase the degree of the polynomial?

Example polynomial revisited

- The expected prediction error of a regression fit $\hat{f}(X)$ at a point $X = x_0$, is

$$\begin{aligned} MSE(x_0) &= MSE(y - \hat{f}(x_0)|X = x_0) \\ &= Bias^2(f, \hat{f}(x_0)) + V(\hat{f}(x_0)) + Irreducible\,Error \end{aligned} \tag{6}$$

- The average expected prediction error

$$\frac{1}{n}\sum_{i=1}^{N} MSE(x_i) \tag{7}$$

Example polynomial revisited

- Bias ?

Example polynomial revisited

- Bias ?
- Variance?

Example polynomial revisited

- Bias ?
- Variance?

- Trace.
  - If $A_{m\times m}$ with typical element a_{ij}. The **trace** of A, $tr(A)$ is the sum of the elements of its diagonal: $tr(A) \equiv \sum_{s=1}^{m} a_{ii}$
  - Properties
    - For any square matrices A, B, and C: $tr(A+B) = tr(A) + tr(B)$
    - Cyclic property: $tr(ABC) = tr(BCA) = tr(CAB)$
    - If $m = 1$ tr(A)=A

Key Insights on Bias-Variance Tradeoff

- The bias term reflects the error introduced by the model's inability to approximate the true function f^*.
- The variance term reflects the sensitivity of the model to the specific training set.
- As dataset size increases, variance generally decreases.
- The noise term σ^2 is unavoidable and cannot be predicted.
- The decomposition for classification problems is less clear than for regression problems, but still present.

Agenda

1. Review
2. Generalization. Out-of-sample Performance
3. Out-of-Sample Error Estimation
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
4. Review

Train and Test Sets. In-Sample and Out-of-Sample Prediction.

- Como seleccionamos la parametrización que minimize el error de predicción fuera de muestra?
- Problema: solo contamos con una muestra

Test Error

- Para seleccionar la mejor parametrización con respecto al Test Error (error de prueba), es necesario estimarlo.
- Hay dos enfoques comunes:
  - Podemos estimar indirectamente el error de la prueba haciendo un ajuste al error de entrenamiento para tener en cuenta el sesgo debido al sobreajuste $\Rightarrow$ Penzalización ex post: AIC, BIC, etc.

Agenda

Test Error

AIC

- Akaike (1969) fue el primero en ofrecer un enfoque unificado al problema de la selección de modelos.
- Elegir el modelo j tal que se minimice:

$$AIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - p_j \tag{8}$$

Agenda

Test Error

SIC/BIC

- Schwarz (1978) mostró que el AIC es inconsistente, (cuando $n \rightarrow \infty$, tiende a elegir un modelo demasiado grande con probabilidad positiva)
- Schwarz (1978) propuso:

$$SIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - \frac{1}{2}p_j log(n) \tag{9}$$

Test Error

AIC vs BIC

$$AIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - p_j \tag{10}$$

$$SIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - p_j\frac{1}{2}log(n) \tag{11}$$

Test Error

- Para seleccionar la mejor parametrización con respecto al Test Error (error de prueba), es necesario estimarlo.
- Hay dos enfoques comunes:
  - Podemos estimar indirectamente el error de la prueba haciendo un ajuste al error de entrenamiento para tener en cuenta el sesgo debido al sobreajuste $\Rightarrow$ Penzalización ex post: AIC, BIC, etc.
  - Levantarnos de nuestros bootstraps (resampling methods) y estimar directamente el Test Error (error de prueba)

Agenda

Test Error

Cross-Validation

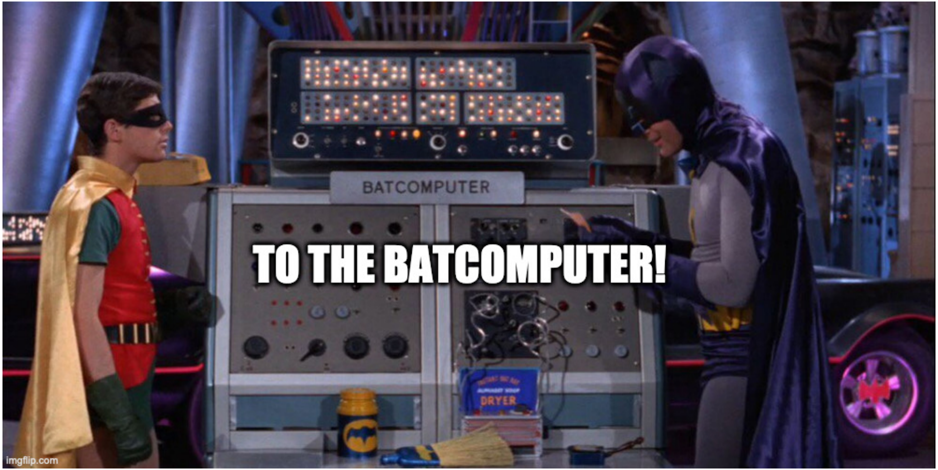


photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Agenda

Review

Hoy

- Bias-Variance Tradeoff (Dilema Sesgo/Varianza)
- Sobreajuste y Selección de modelos
  - AIC y BIC
  - Enfoque de Validación
  - LOOCV
  - K-fold Cross-Validation (Validación Cruzada)