The Predictive Paradigm

Machine Learning

Ignacio Sarmiento-Barbieri

Universidad de La Plata

¿Qué entendemos por Big Data y ML?

- ¿Que es Big Data?
  - Big n, es solo parte de la historia
  - Big también es big k, muchos covariates, a veces $n << k$
  - Vamos a entender Big también como datos que no surgen de fuentes tradicionales (cuentas nac., etc)
    - Datos de la Web, Geográficos, etc.
- Machine Learning
  - Cambio de paradigma de estimación a predicción

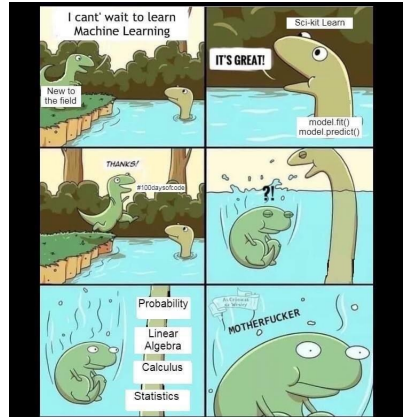
Agenda

1. About the Course
2. Machine learning is all about prediction
3. ML Tasks
4. Prediction vs Causality
5. Getting serious about prediction
   - The basic logic of prediction
   - Minimizing our losses
   - Best Predictor
6. Generalization. Out-of-sample Performance
7. Out-of-Sample Error Estimation
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
8. Review

Agenda

Lenguajes

- Estadística y Econometría
- Inglés
- Código
  - Elijan el que quieran:
    - `Python`, `R`, o cualquier otro
    - no hay restricción
    - yo me basare en `Python`
  - Github
- Aprender haciendo y mucha prueba y error!



Materiales

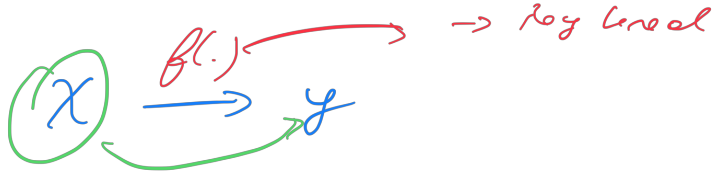
1. Página web: link
2. Statistical Learning (FREE!!! (as beer, not speech))
   https://www.gnu.org/philosophy/free-sw.en.html

   - James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (ISLP)
   - Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction
   - Békés, G., & Kézdi, G. (2021). Data analysis for business, economics, and policy. Cambridge University Press.

Agenda

Machine learning is all about prediction

- Machine learning is a branch of computer science and statistics, tasked with developing algorithms to predict outcomes Y from observable variables X.
- The learning part comes from the fact that we don't specify how exactly the computer should predict Y from X.
- This is left as an empirical problem that the computer can "learn".
- In general, this means that we abstract from the underlying model, the approach is pragmatic



Machine learning is all about prediction

- Machine learning is a branch of computer science and statistics, tasked with developing algorithms to predict outcomes Y from observable variables X.
- The learning part comes from the fact that we don't specify how exactly the computer should predict Y from X.
- This is left as an empirical problem that the computer can "learn".
- In general, this means that we abstract from the underlying model, the approach is pragmatic

"Whatever works, works...."

“Whatever works, works....”

“Whatever works, works....”????

- In many applications, ML techniques can be successfully applied by data scientists with little knowledge of the problem domain.
- For example, Kaggle competitions

“Whatever works, works....”????

- Much less attention has been paid to the limitations of pure prediction methods.
- When ML applications are used “off the shelf” without understanding the underlying assumptions or ensuring that conditions like stability are met, then the validity and usefulness of the conclusions can be compromised.

Agenda

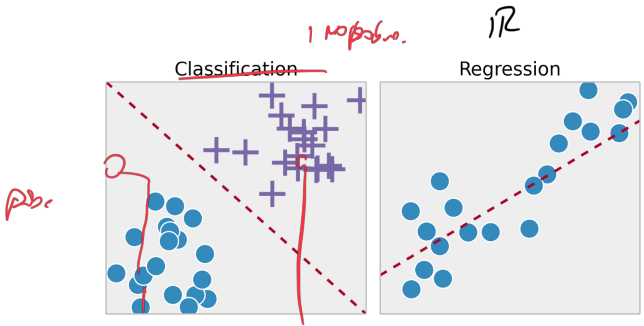
Agenda

ML branches

- ML tasks can (¿?) be divided into two main branches:
  1. Supervised Learning

ML branches

- Supervised Learning
  - for each predictor x_i a 'response' is observed y_i.
  - everything we have done in econometrics is supervised

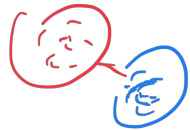


Source: shorturl.at/opqKT

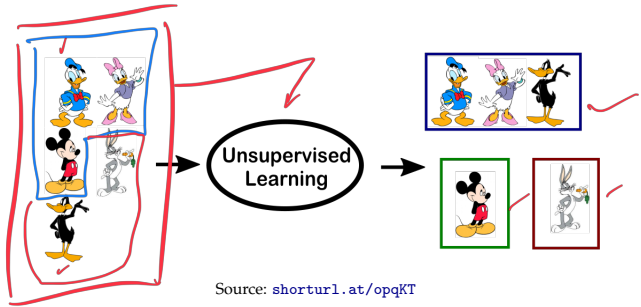
ML branches

- ML tasks can (¿?) be divided into two main branches:
  1. Supervised Learning
  2. Unsupervised Learning

ML branches



- Unsupervised Learning
  - observed x_i but no response.
  - example: cluster analysis



Source: shorturl.at/opqKT

Agenda

Policy Prediction Problems

- Empirical policy research often focuses on causal inference.
- Since policy choices seem to depend on understanding the counterfactual there's a tight link
- While this link holds in many cases, there are also many policy applications where causal inference is not central, or even necessary.

The Causal Paradigm

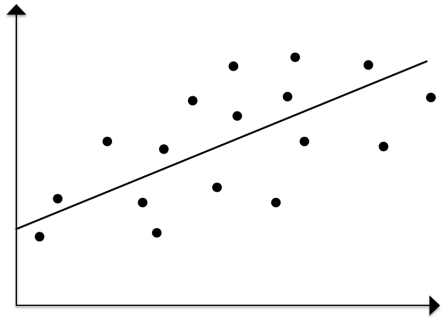
$$y = f(X) + u \tag{1}$$

- Interest lies on inference
- "Correct" $f()$ to understand how y is affected by X
- Model: Theory, experiment
- Hypothesis testing (std. err., tests)

Prediction vs. Causality: Target

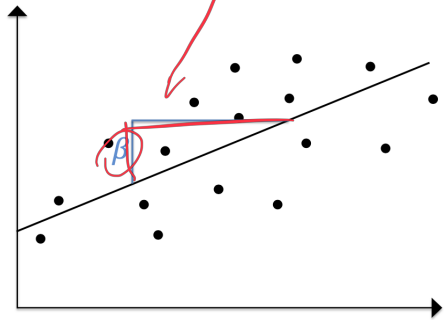
$$y = f(x) + \epsilon \tag{2}$$

$$y = \alpha + \beta x + \epsilon \tag{3}$$



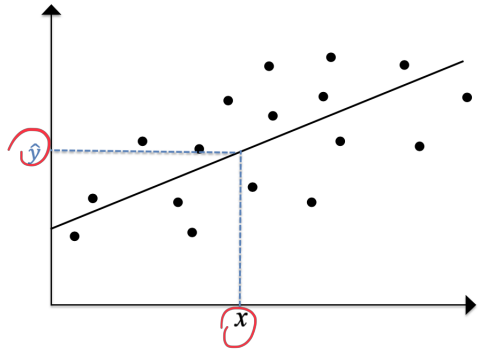
Prediction vs. Causality: Target

$$y = \alpha + \beta x + \epsilon \tag{4}$$

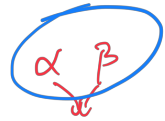


Prediction vs. Causality: Target

$$y = \underbrace{\alpha + \beta x}_{\hat{y}} + \epsilon \tag{5}$$



The Predictive Paradigm



$$y = f(X) + u \tag{6}$$

- Interest on predicting y
- "Correct" $f()$ to be able to predict (no inference!)
- Model? We treat $f()$ as a black box, and any approximation $\hat{f}()$ that yields a good prediction is good enough (*Whatever works, works.*).

Prediction vs. Causality: The garden of the parallel paths?

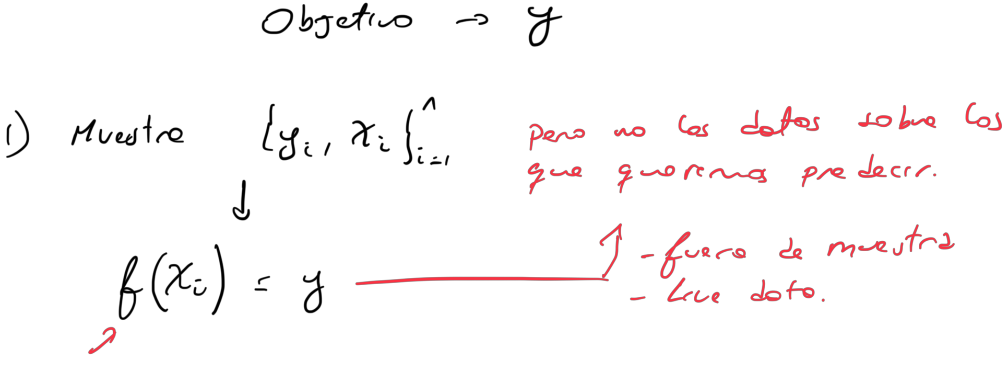
- We've seen that prediction and causality
  - Answer different questions
  - Serve different purposes
  - Seek different targets
- Different strokes for different folks, or complementary tools in an applied economist's toolkit?

Agenda

Agenda

1. About the Course
2. Machine learning is all about prediction
3. ML Tasks
4. Prediction vs Causality
5. Getting serious about prediction
   - The basic logic of prediction
   - Minimizing our losses
   - Best Predictor
6. Generalization. Out-of-sample Performance
7. Out-of-Sample Error Estimation
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
8. Review

The basic logic of prediction



Agenda

1. About the Course
2. Machine learning is all about prediction
3. ML Tasks
4. Prediction vs Causality
5. Getting serious about prediction
   - The basic logic of prediction
   - Minimizing our losses
   - Best Predictor
6. Generalization. Out-of-sample Performance
7. Out-of-Sample Error Estimation
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
8. Review

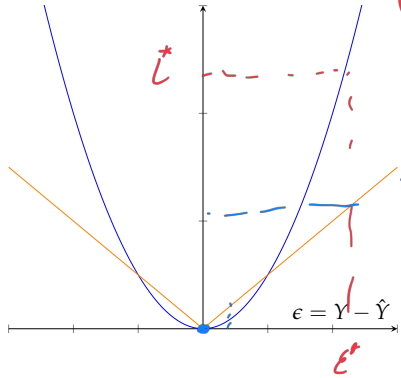
Prediction error

$$\epsilon = Y - \hat{Y} \tag{7}$$

$$L(Y, \hat{Y}) \tag{8}$$

Minimizing our losses

$$L(Y, \hat{Y}) \tag{9}$$

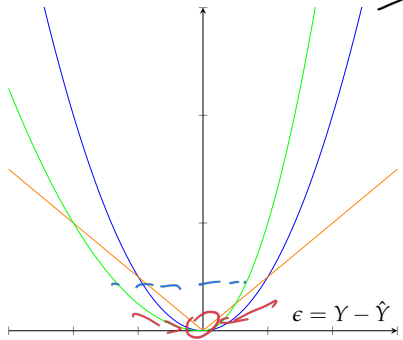


Minimizing our losses

$$L(Y, \hat{Y}) \tag{10}$$

$(y-\hat{y})^2$

MSE



Agenda

1. About the Course
2. Machine learning is all about prediction
3. ML Tasks
4. Prediction vs Causality
5. Getting serious about prediction
   - The basic logic of prediction
   - Minimizing our losses
   - Best Predictor
6. Generalization. Out-of-sample Performance
7. Out-of-Sample Error Estimation
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
8. Review

Best Predictor

The CEF solves the best prediction problem

$$\min_{m(W)} E[(Y - m(W))^2] = \int (Y - m(W))^2 Pr(dW, dY) \tag{11}$$

- conditioning on W we have that

$$E_W E_{Y|X}[(Y - m(W))^2 | W]$$

$$f(x) = E(y|x)$$

$$\mathcal{L} = \sum \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(\frac{x_i - \mu}{(12)\sigma^2}\right)$$

$$\mu = \frac{\sum x_i}{n}$$

$$\sigma^2 = \frac{1}{n}\sum (x_i - \bar{x})$$

Linear Regression

- If we assume

$$y = X\beta + \epsilon \tag{13}$$

- Using matrix algebra, the loss function:

$$\tilde{\epsilon}'\tilde{\epsilon} = (y - X\tilde{\beta})'(y - X\tilde{\beta}) \tag{14}$$

  - $SSR(\tilde{\beta})$ is the aggregation of squared errors if we choose $\tilde{\beta}$ as an estimator.

- The **least squares estimator** $\hat{\beta}$ will be

$$\hat{\beta} = \underset{\tilde{\beta}}{argmin}\, SSR(\tilde{\beta}) \tag{15}$$

Agenda

Generalization Overview

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \; SSR \quad (y - X\beta)^2$$

$$\hat{\beta} = (X'X)^{-1} X y$$

- 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:

Risk.

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

$\{y, X\}$ entrenamiento

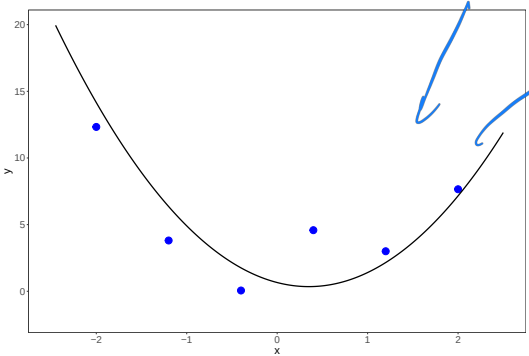
$\{y, X\}$ test → vienen del mismo GDP

Overfitting and Underfitting

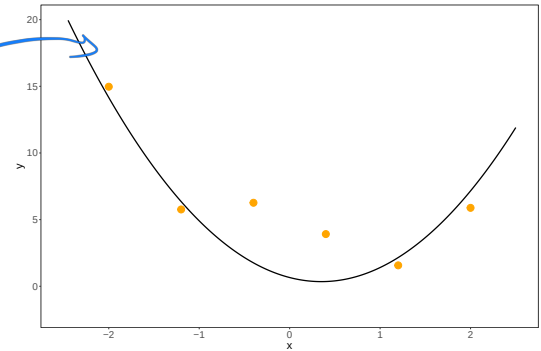
- Successfully minimizing training error does not always result in a small test error.
- A model is said to overfit if it predicts accurately on training data but poorly on test (unseen) data.
- A model underfits if its training error is relatively large, which usually means test error is also large.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

Overfitting and Underfitting. Bias-Variance Tradeoff

$y = f(x) + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$



(a) Training Data



(b) Testing Data

Overfitting and Underfitting. Bias-Variance Tradeoff

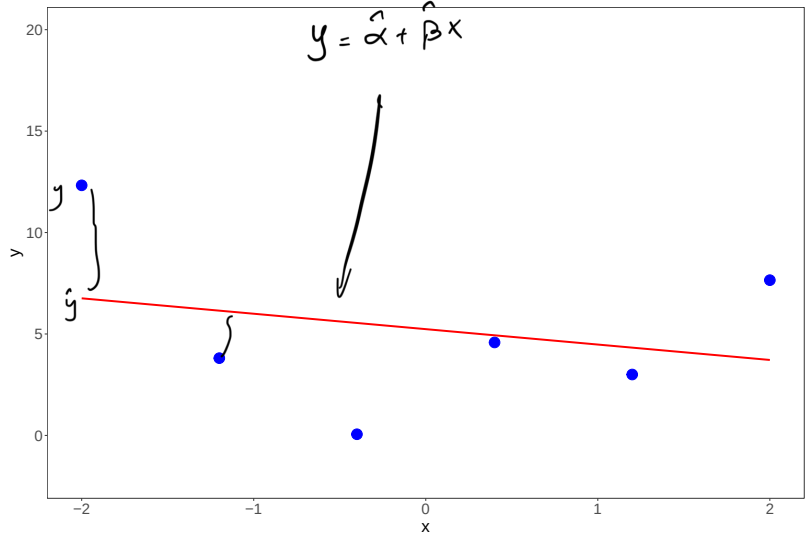
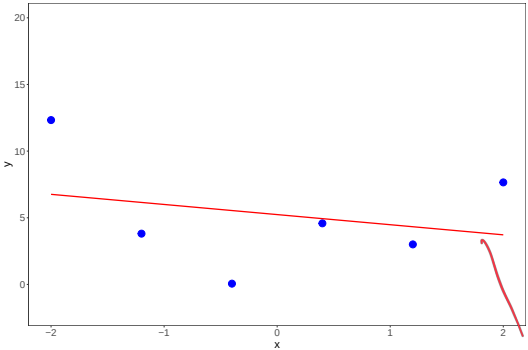


Figure 2: Training Data

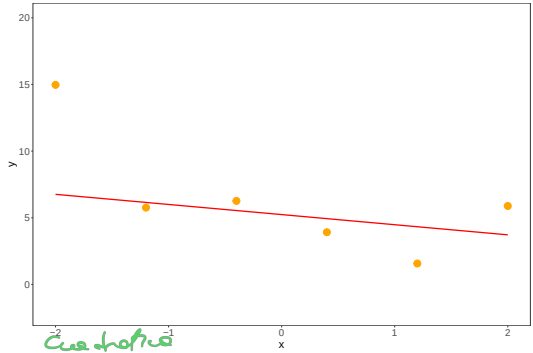
Overfitting and Underfitting. Bias-Variance Tradeoff

Out-of-Sample Performance

underfitting (sub ajuste) $E(\hat{\beta}) = \beta$



(a) Training Data



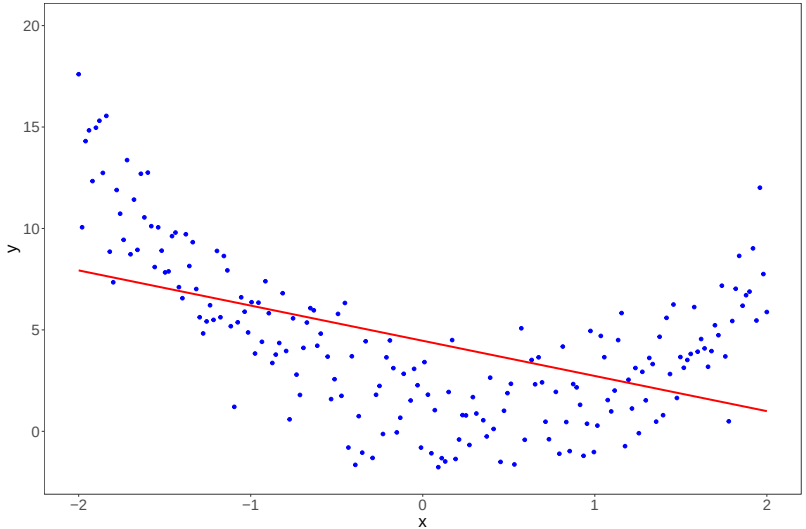
(b) Testing Data

lineal cuadrática

$E(\hat{f}) - f^* = 0$

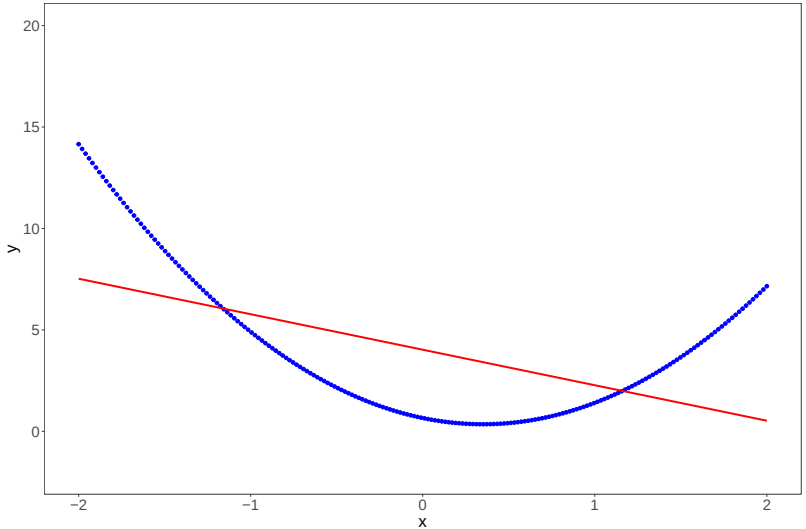
Overfitting and Underfitting. Bias-Variance Tradeoff

More data?



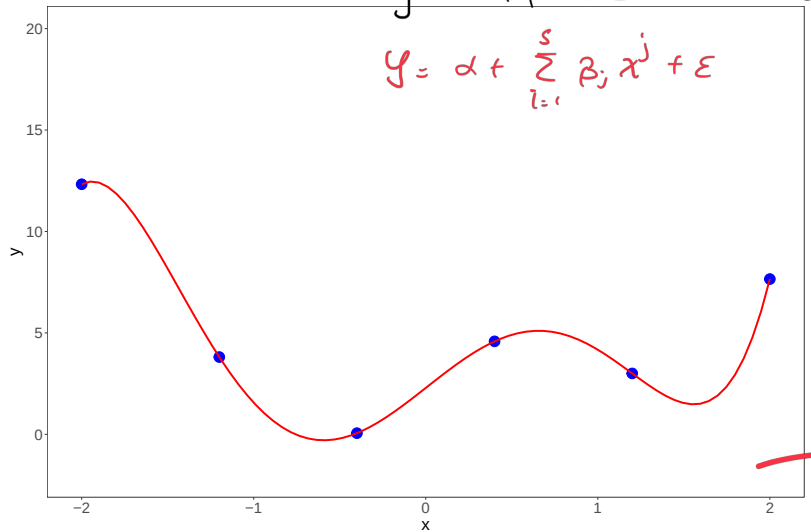
Overfitting and Underfitting. Bias-Variance Tradeoff

Noiseless data?



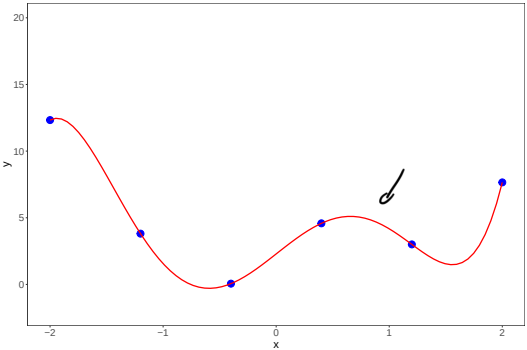
Overfitting and Underfitting. Bias-Variance Tradeoff

More Complex Model

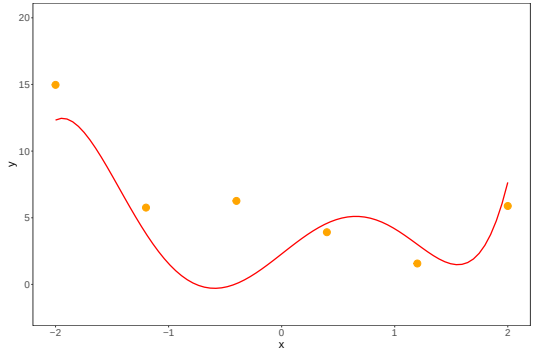


Overfitting and Underfitting. Bias-Variance Tradeoff

Out-of-Sample Performance



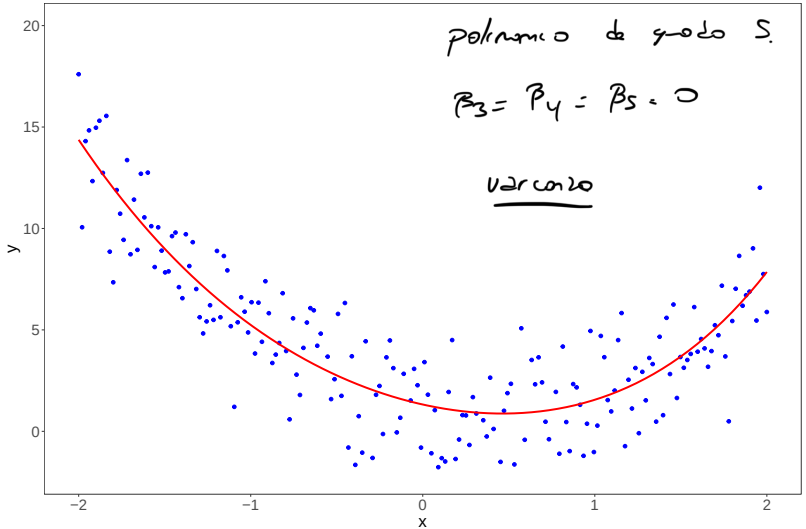
(a) Training Data



(b) Testing Data

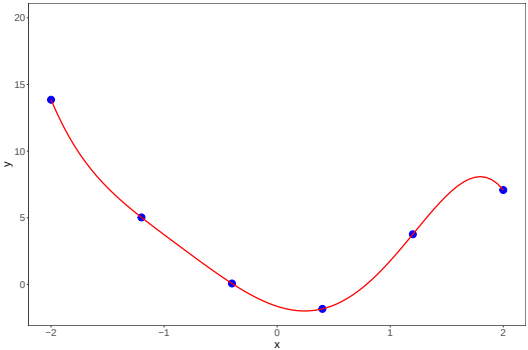
Overfitting and Underfitting. Bias-Variance Tradeoff

More Data

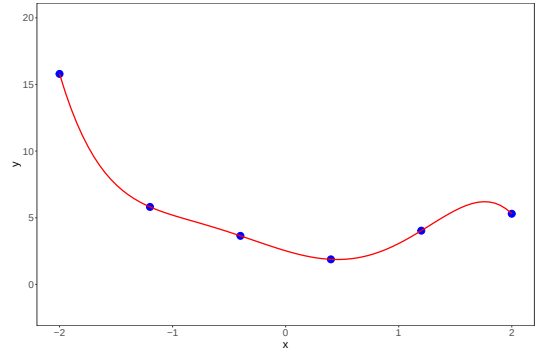


Overfitting and Underfitting. Bias-Variance Tradeoff

Variance



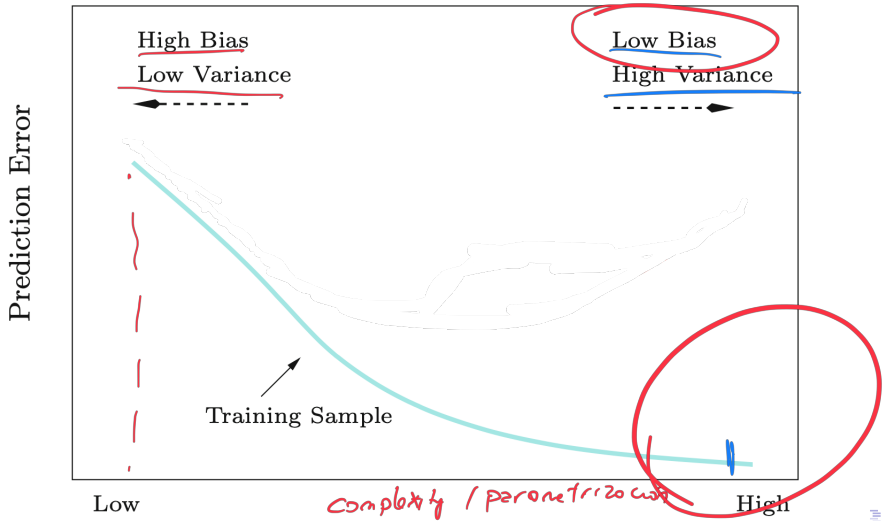
(a) Training Data 2



(b) Training Data 3

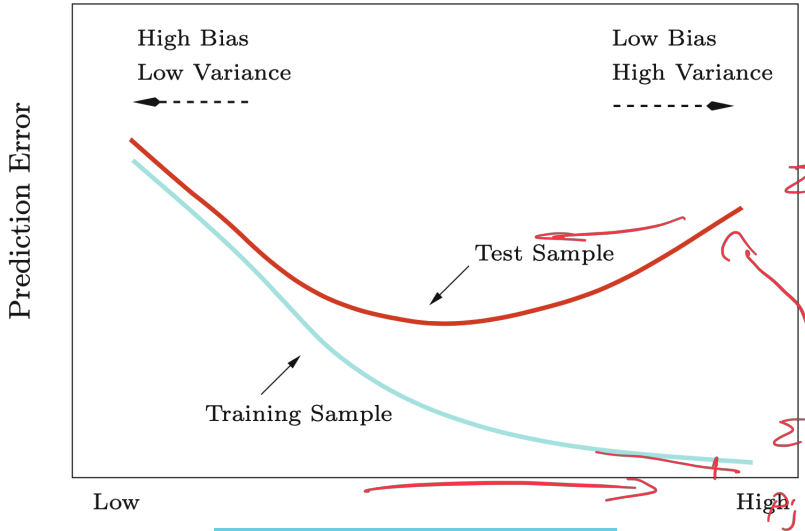
Overfitting and Underfitting. Bias-Variance Tradeoff

In-Sample Prediction and Overfit



Overfitting and Underfitting. Bias-Variance Tradeoff

Out-of-Sample Prediction and Overfit



Mathematical Decomposition for Regression

$$E(u^2) = \sigma^2$$

$$E(u) = 0$$

$$x \perp u = 0$$

$$y = f(x) + u$$

$$E\left[(y - \hat{f}(x))^2 \right]$$

$$E\left[\left(f(x) + u - \hat{f}(x) \right)^2 \right] = E\left[\left(f(x) - \hat{f}(x) \right)^2 \right] + \underbrace{E(u^2)}_{\sigma^2}$$

reducible

irreducible

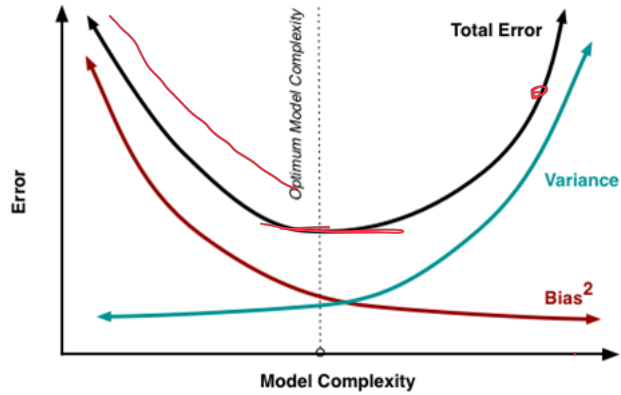
$$E\left[\left(f(x) - E(\hat{f}(x)) + E(\hat{f}(x)) - \hat{f}(x) \right)^2 \right]$$

$$\left[f(x) - E(\hat{f}(x)) \right]^2 + E\left[\left(\hat{f} - E(\hat{f}(x)) \right)^2 \right]$$

Bias^2

varianza

Bias-Variance Tradeoff



Source: https://tinyurl.com/y4lvjxpc

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

Agenda

1. About the Course
2. Machine learning is all about prediction
3. ML Tasks
4. Prediction vs Causality
5. Getting serious about prediction
   - The basic logic of prediction
   - Minimizing our losses
   - Best Predictor
6. Generalization. Out-of-sample Performance
7. Out-of-Sample Error Estimation
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
8. 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

1. About the Course
2. Machine learning is all about prediction
3. ML Tasks
4. Prediction vs Causality
5. Getting serious about prediction
   - The basic logic of prediction
   - Minimizing our losses
   - Best Predictor
6. Generalization. Out-of-sample Performance
7. Out-of-Sample Error Estimation
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
8. Review

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:

$$AIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) + p_j \tag{16}$$

Agenda

1. About the Course
2. Machine learning is all about prediction
3. ML Tasks
4. Prediction vs Causality
5. Getting serious about prediction
   - The basic logic of prediction
   - Minimizing our losses
   - Best Predictor
6. Generalization. Out-of-sample Performance
7. Out-of-Sample Error Estimation
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
8. Review

Test Error

SIC/BIC

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

$$SIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - \frac{1}{2}p_j log(n) \tag{17}$$

Test Error

AIC vs BIC

$$AIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - p_j \tag{18}$$

$$SIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - p_j\frac{1}{2}log(n) \tag{19}$$

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

1. About the Course
2. Machine learning is all about prediction
3. ML Tasks
4. Prediction vs Causality
5. Getting serious about prediction
   - The basic logic of prediction
   - Minimizing our losses
   - Best Predictor
6. Generalization. Out-of-sample Performance
7. Out-of-Sample Error Estimation
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Cross-Validation
8. Review

Test Error

Cross-Validation

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

Agenda

Review

- This Week:
  - Machine Learning is all about prediction
  - ML targets something different than causal inference, they can complement each other
  - Bias Variance trade-off: tolerating some bias is possible to reduce $V(\hat{f}(X))$ and lower MSE (ML best kept secret)
  - Overfit and Model Selecction
    - AIC y BIC
    - Validation Approach
    - LOOCV
    - K-fold Cross-Validation