

Lecture 6: Machine Learning Paradigma Predictivo

Big Data and Machine Learning en el Mercado Inmobiliario
Educación Continua

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

- 1 Recap
- 2 Predicción: estadística clásica vs la máquina de aprender
 - Tipos de Aprendizaje
 - Prediction vs Estimation
 - Sobreajuste y Predicción fuera de Muestra
- 3 Selección de Modelo
 - Enfoque Clásico
 - Métodos de remuestreo
 - Enfoque de conjunto de validación
 - LOOCV
 - Validación cruzada en K-partes
- 4 Para seguir leyendo
- 5 Break

Modelo Hedónico

- ▶ Podemos entonces estimar la función hedónica utilizando precio de las propiedades y las características de la misma

$$P = f(atrib_1, atrib_2, \dots, atrib_n) \quad (1)$$

- ▶ Sin embargo, la teoría no nos dice poco sobre
 - ▶ Cuáles son o cómo medir estos atributos que influyen sobre el precio
 - ▶ Como estas características influyen en el precio (forma funcional)

- 1 Recap
- 2 Predicción: estadística clásica vs la máquina de aprender
 - Tipos de Aprendizaje
 - Prediction vs Estimation
 - Sobreajuste y Predicción fuera de Muestra
- 3 Selección de Modelo
 - Enfoque Clásico
 - Métodos de remuestreo
 - Enfoque de conjunto de validación
 - LOOCV
 - Validación cruzada en K-partes
- 4 Para seguir leyendo
- 5 Break

Estadística clásica vs la máquina de aprender

$$y = f(X) + u \quad (2)$$

- ▶ Estadística Clásica
 - ▶ Inferencia
 - ▶ $f()$ "correcta" el interes es en entender como y afecta X
 - ▶ modelos surge de la teoria/experimentos
 - ▶ Interés es en test de hipótesis (std. err., ci's)
- ▶ Maquina de Aprender
 - ▶ Interés es predecir y
 - ▶ El $f()$ correcto es el que predice mejor
 - ▶ Modelo?

¿Qué es la máquina de aprender?

- ▶ El aprendizaje de máquinas es una rama de la informática y la estadística, encargada de desarrollar algoritmos para predecir los resultados y a partir de las variables observables X .
- ▶ La parte de aprendizaje proviene del hecho de que no especificamos cómo exactamente la computadora debe predecir y a partir de X .
- ▶ Esto queda como un problema empírico que la computadora puede "aprender".
- ▶ En general, esto significa que nos abstraemos del modelo subyacente, el enfoque es muy pragmático

¿Qué es la máquina de aprender?

- ▶ El aprendizaje de máquinas es una rama de la informática y la estadística, encargada de desarrollar algoritmos para predecir los resultados y a partir de las variables observables X .
- ▶ La parte de aprendizaje proviene del hecho de que no especificamos cómo exactamente la computadora debe predecir y a partir de X .
- ▶ Esto queda como un problema empírico que la computadora puede “aprender”.
- ▶ En general, esto significa que nos abstraemos del modelo subyacente, el enfoque es muy pragmático

“Lo que sea que funciona, funciona...”

“Lo que sea que funciona, funciona...”



Tipos de Aprendizaje

► ML se divide en dos ramas principales:

1 Aprendizaje supervisado: Tenemos datos tanto sobre un resultado y como sobre las variables explicativas X .

- Esto es lo más cercano al análisis de regresión que conocemos.
- Si y es discreto, también podemos ver esto como un problema de clasificación.
- Es el enfoque de este curso.

2 Aprendizaje no supervisado: No tenemos datos sobre y , solo sobre X .

- Queremos agrupar estos datos (sin especificar qué agrupar).
- Permite reducir la dimensionalidad y explorar datos
- Algunos algoritmos destacados: PCA, y K-medias

Predicción y Error Predictivo

- ▶ El objetivo es predecir y dadas otras variables X . Ej: precio vivienda dadas las características
- ▶ Asumimos que el link entre y and X esta dado por el modelo:

$$y = f(X) + u \quad (3)$$

- ▶ donde $f(X)$ es cualquier función,
- ▶ u una variable aleatoria no observable $E(u) = 0$ and $V(u) = \sigma^2$

Como medimos: “Lo que sea que funciona, funciona...”

- ▶ En la práctica no conocemos $f(X)$
- ▶ Es necesario “aprenderla” (estimarla) $\hat{y} = \hat{f}(X)$
- ▶ La medida de cuán bien funciona nuestro modelo es

$$MSE(\hat{y}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (4)$$

Como medimos: “Lo que sea que funciona, funciona...”

- Podemos descomponer el MSE en dos partes

$$MSE(\hat{y}) = MSE(\hat{f}) + \sigma^2 \quad (5)$$

- el error de estimar f con \hat{f} . (*reducible*)
- el error de no observar u . (*irreducible*)

Predicción y Error Predictivo

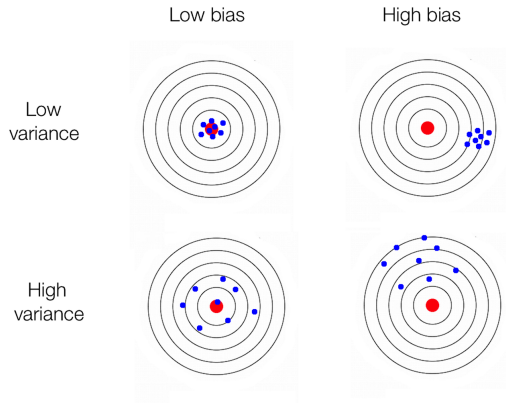
- ▶ Descomponiendo un poco más:

$$Err(Y) = MSE(\hat{f}) + \sigma^2 \quad (6)$$

$$= Bias^2(\hat{f}) + V(\hat{f}) + Error\ Irreducible \quad (7)$$

- ▶ Este resultado es muy importante,
 - ▶ Aparece el dilema entre sesgo y varianza

Dilema sesgo/varianza



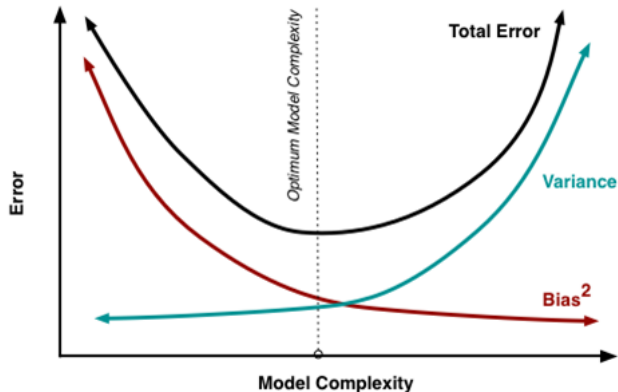
Source: <https://tinyurl.com/y4lvjxpc>

Dilema sesgo/varianza

- El secreto de ML: admitiendo un poco de sesgo podemos tener ganancias importantes en varianza

Dilema sesgo/varianza

- El secreto de ML: admitiendo un poco de sesgo podemos tener ganancias importantes en varianza



Source: <https://tinyurl.com/y4lvjxpc>

Predicción y regresión lineal

- El problema es:

$$y = f(X) + u \quad (8)$$

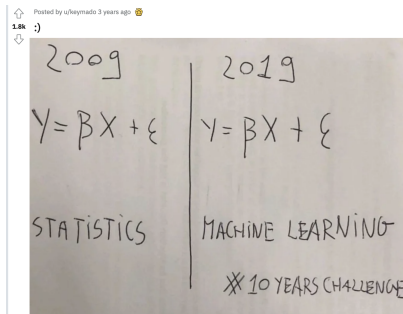
Predicción y regresión lineal

- El problema es:

$$y = f(X) + u \quad (8)$$

- proponemos :

$$f(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p \quad (9)$$



Fuente: <https://www.reddit.com/r/datascience/comments/ah0n69/>

Predicción y regresión lineal

- ▶ Y el dilema sesgo varianza?

Predicción y regresión lineal

- ▶ Y el dilema sesgo varianza?
- ▶ Bajo los supuestos clásicos (Gauss-Markov) el estimador de OLS es insesgado:

$$E(X\hat{\beta}) = E(\hat{\beta}_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p) \quad (10)$$

$$= E(\hat{\beta}_1) + E(\hat{\beta}_2) X_2 + \cdots + E(\hat{\beta}_p) X_p \quad (11)$$

$$= X\beta \quad (12)$$

- ▶ $MSE(\hat{y})$ se reduce a $V(\hat{\beta})$

Complejidad y compensación de varianza/sesgo

- ▶ En la econometría clásica, la elección de modelos se resume a elegir entre modelos más pequeños y más grandes.
- ▶ Considere los siguientes modelos para estimar y :

$$y = \beta_1 X_1 + u_1$$

$$y = \beta_1 X_1 + \beta_2 X_2 + u_2$$

- ▶ $\hat{\beta}_1^{(1)}$ el estimador de OLS y on X_1
- ▶ La predicción es:
- ▶ $\hat{\beta}_1^{(2)}$ y $\hat{\beta}_2^{(2)}$ con β_1 y β_2 los el estimador de OLS de y en X_1 y X_2 .
- ▶ La predicción es:

$$\hat{y}^{(1)} = \hat{\beta}_1^{(1)} X_1$$

$$\hat{y}^{(2)} = \hat{\beta}_1^{(2)} X_1 + \hat{\beta}_2^{(2)} X_2$$

Complejidad y compensación de varianza/sesgo

- ▶ Una discusión importante en la econometría clásica es la de la omisión de variables relevantes frente a la inclusión de variables irrelevantes.
 - ▶ Si el modelo (1) es verdadero entonces estimar el modelo más grande (2) conduce a estimadores ineficientes aunque no sesgados debido a que incluyen innecesariamente X_2 .

Complejidad y compensación de varianza/sesgo

Ejemplo

#Load Packages

```
require("tidyverse")  
require("fabricatr")  
require("stargazer")
```

#for reproducibility

```
set.seed(101010)
```

```
db1 <- fabricate(  
  N = 10000,  
  ability=rnorm(N,mean=.5,sd=2),  
  schooling = round(runif(N, 2, 14)),  
  logwage =rnorm(N, mean=7+.15*schooling, sd=2)  
)
```

Complejidad y compensación de varianza/sesgo

Ejemplo

```
reg1<-lm(logwage~schooling,db1)
reg2<-lm(logwage~schooling+ability,db1)
stargazer(reg1,reg2,type="text")
```

```
##
## =====
##                               Dependent variable:
##                               -----
##                               logwage
##                               (1)           (2)
## -----
## schooling                0.145***        0.145***
##                          (0.006)          (0.006)
##
## ability                      0.007
##                          (0.010)
##
## Constant                7.050***        7.046***
##                          (0.051)        (0.051)
## -----
## Observations              10,000         10,000
## R2                        0.059          0.059
## =====
## Note:                      *p<0.1; **p<0.05; ***p<0.01
```


Complejidad y compensación de varianza/sesgo

Ejemplo

```
db1<- db1 %>% mutate(yhat_reg1=predict(reg1),  
                     yhat_reg2=predict(reg2))
```

```
var(db1$yhat_reg1)
```

```
## [1] 0.2522197
```

```
var(db1$yhat_reg2)
```

```
## [1] 0.2524032
```

Complejidad y compensación de varianza/sesgo

- ▶ Una discusión importante en la econometría clásica es la de la omisión de variables relevantes frente a la inclusión de variables irrelevantes.
 - ▶ Si el modelo (1) es verdadero entonces estimar el modelo más grande (2) conduce a estimadores ineficientes aunque no sesgados debido a que incluyen innecesariamente X_2 .
 - ▶ Si el modelo (2) es verdadero, estimar el modelo más pequeño (1) conduce a una estimación de menor varianza pero sesgada si X_1 también se correlaciona con el regresor omitido X_2 .

Complejidad y compensación de varianza/sesgo

Ejemplo

```
db2 <- fabricate(  
  N = 10000,  
  ability=rnorm(N,mean=.5,sd=2),  
  schooling = round(runif(N, 2, 14)),  
  schooling = round(ceiling(schooling+1*ability)),  
  logwage =rnorm(N, mean=7+.15*schooling+.25*ability, sd=2)  
)
```

Complejidad y compensación de varianza/sesgo

Ejemplo

```
reg3<-lm(logwage~schooling,db2)
reg4<-lm(logwage~schooling+ability,db2)
stargazer(reg3,reg4,type="text")
```

```
##
## =====
##                               Dependent variable:
##                               -----
##                               logwage
##                               (1)             (2)
## -----
## schooling                0.216***          0.153***
##                          (0.005)           (0.006)
##
## ability                    0.254***
##                          (0.011)
##
## Constant                  6.563***          7.007***
##                          (0.051)           (0.053)
## -----
## Observations              10,000            10,000
## R2                        0.152             0.192
## =====
## Note:                      *p<0.1; **p<0.05; ***p<0.01
```

Complejidad y compensación de varianza/sesgo

Ejemplo

```
db2$yhat_reg3<-predict(reg3)  
db2$yhat_reg4<-predict(reg4)
```

```
var(db2$yhat_reg3)
```

```
## [1] 0.755213
```

```
var(db2$yhat_reg4)
```

```
## [1] 0.9538193
```

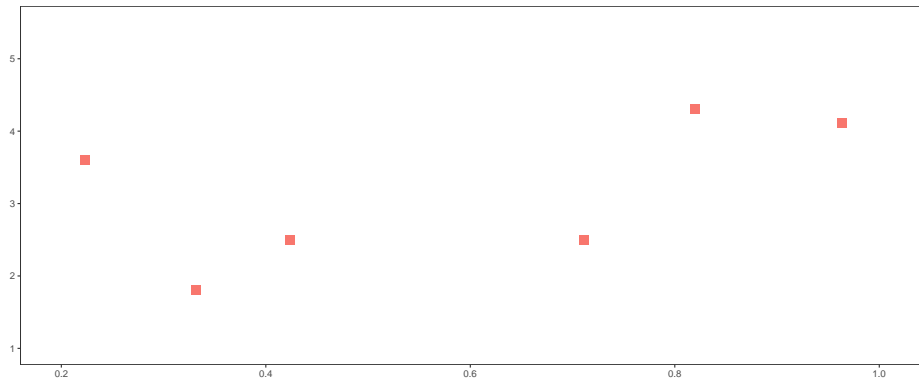
Complejidad y compensación de varianza/sesgo

- ▶ Una discusión importante en la econometría clásica es la de la omisión de variables relevantes frente a la inclusión de variables irrelevantes.
 - ▶ Si el modelo (1) es verdadero entonces estimar el modelo más grande (2) conduce a estimadores ineficientes aunque no sesgados debido a que incluyen innecesariamente X_2 .
 - ▶ Si el modelo (2) es verdadero, estimar el modelo más pequeño (1) conduce a una estimación de menor varianza pero sesgada si X_1 también se correlaciona con el regresor omitido X_2 .
- ▶ Esta discusión de pequeño vs grande siempre es con respecto a un modelo que se supone es verdadero.
- ▶ Pero en la práctica el modelo verdadero es desconocido!!!

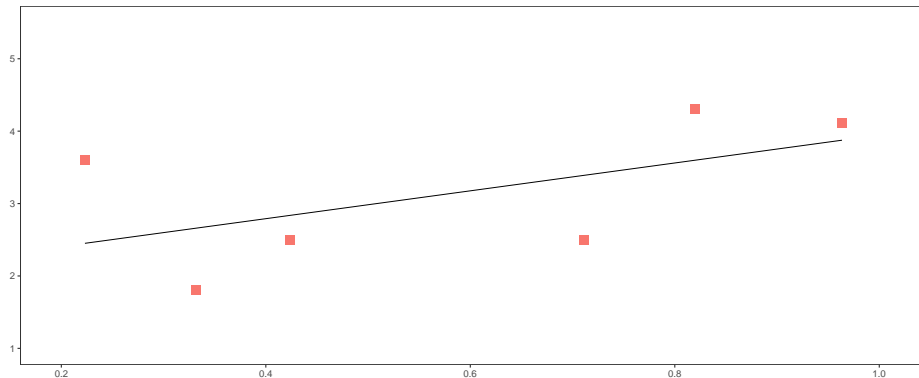
Complejidad y compensación de varianza/sesgo

- ▶ Elegir entre modelos implica un dilema *sesgo/varianza*
- ▶ La econometría clásica tiende a resolver este dilema abruptamente,
 - ▶ requiriendo una estimación no sesgada y, por lo tanto, favoreciendo modelos más grandes para evitar sesgos
- ▶ En esta configuración simple, los modelos más grandes son "más complejos", por lo que los modelos más complejos están menos sesgados pero son más ineficientes.
- ▶ Por lo tanto, en este marco muy simple, la complejidad se mide por el número de variables explicativas.
- ▶ Una idea central en el aprendizaje automático es generalizar la idea de complejidad,
 - ▶ Nivel óptimo de complejidad, es decir, modelos cuyo sesgo y varianza conducen al menor MSE.

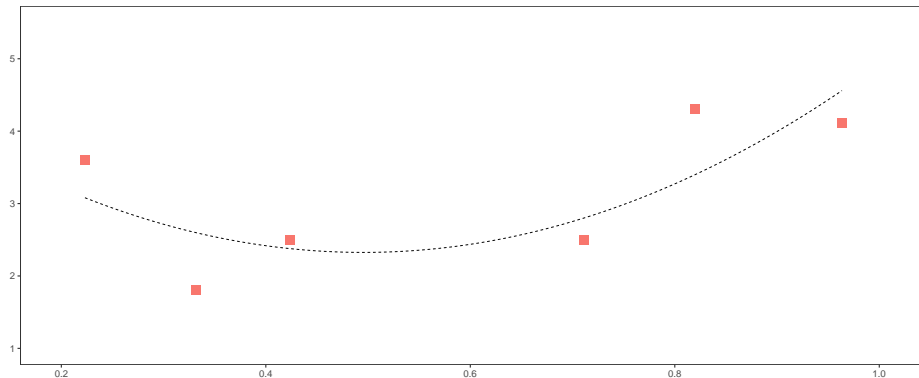
Sobreajuste



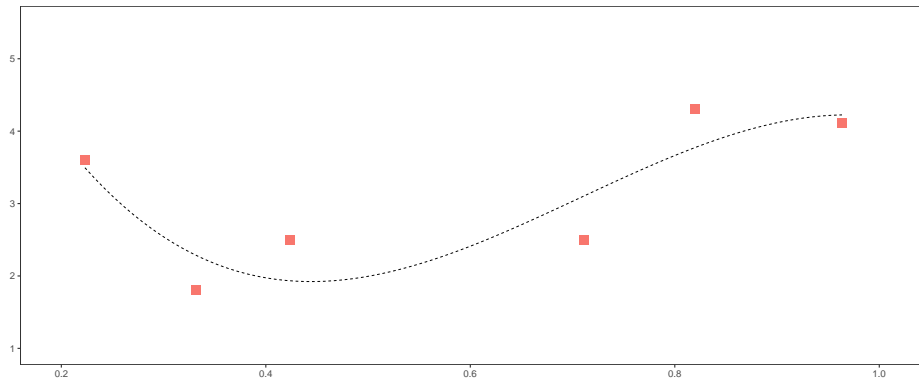
Sobreajuste



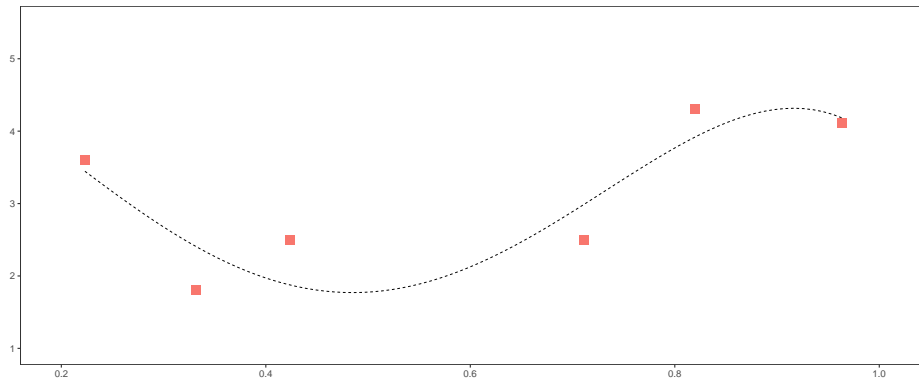
Sobreajuste



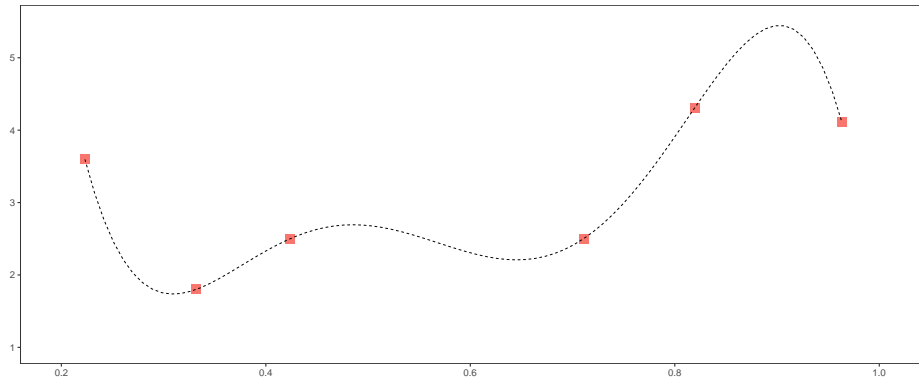
Sobreajuste



Sobreajuste



Sobreajuste



Sobreajuste

- Notemos que esto no es otra cosa que la suma de los residuales al cuadrado

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(X))^2 \quad (13)$$

$$= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y})^2 \quad (14)$$

$$= \frac{1}{n} \sum_{i=1}^n (e)^2 \quad (15)$$

$$= RSS \quad (16)$$

- Esta medida nos da una idea de *lack of fit* que tan mal ajusta el modelo a los datos

Sobreajuste

- ▶ Un problema del RSS es que nos da una medida absoluta de ajuste de los datos, y por lo tanto no está claro que constituye un buen RSS.
- ▶ Una alternativa muy usada en economía es el R^2
- ▶ Este es una proporción (la proporción de varianza explicada),
 - ▶ toma valores entre 0 y 1,
 - ▶ es independiente de la escala (o unidades) de y

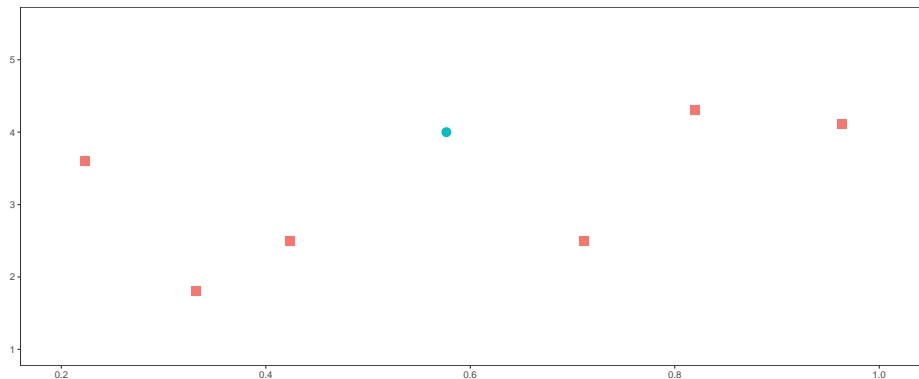
$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (17)$$

$$= 1 - \frac{RSS}{TSS} \quad (18)$$

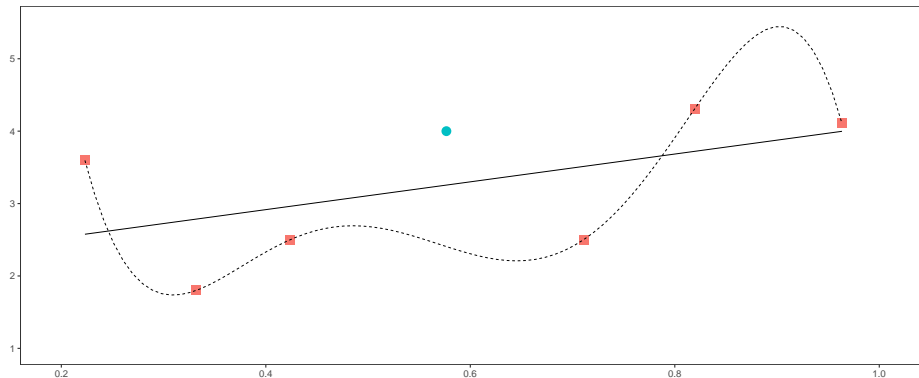
Sobreajuste y Predicción fuera de Muestra

- ▶ ML nos interesa la predicción fuera de muestra
- ▶ Overfit: modelos complejos predicen muy bien dentro de muestra, pero tienden a hacer un trabajo fuera de muestra
- ▶ Hay que elegir el nivel adecuado de complejidad
- ▶ Como medimos el error de predicción fuera de muestra?
- ▶ R^2 no funciona: se concentra en la muestra y es no decreciente en complejidad

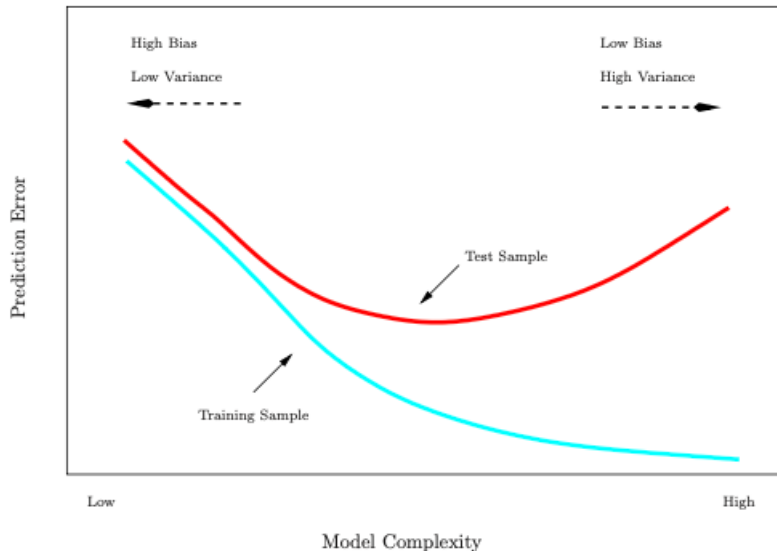
Sobreajuste y Predicción fuera de Muestra



Sobreajuste y Predicción fuera de Muestra



Sobreajuste y Predicción fuera de Muestra



Selección de Modelo AIC (Akaike Information Criterion)

- ▶ Akaike (1969) fue el primero en ofrecer un enfoque unificado al problema de la selección de modelos.
- ▶ Su punto de vista fue elegir un modelo del conjunto f_i que funcionó bien cuando se evaluó sobre la base del rendimiento de la previsión.
- ▶ Su criterio, que ha llegado a llamarse criterio de información de Akaike, es

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

Selección de Modelos: SIC/BIC (Schwarz/Bayesian Information Criterion)

- ▶ Schwarz (1978) mostró que, si bien el enfoque *AIC* puede ser bastante satisfactorio para seleccionar un modelo de pronóstico
- ▶ Sin embargo, tiene la desafortunada propiedad de que es inconsistente, (cuando $n \rightarrow \infty$, tiende a elegir un modelo demasiado grande con probabilidad positiva)
- ▶ Schwarz (1978) formalizó el problema de selección de modelos desde un punto de vista bayesiano:

$$SIC(j) = \log \left(\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y})^2 \right) - \frac{1}{2} p_j \log(n) \quad (20)$$

$$AIC(j) = \log \left(\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y})^2 \right) - p_j \quad (21)$$

$$SIC(j) = \log \left(\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y})^2 \right) - p_j \frac{1}{2} \log(n) \quad (22)$$

► Note que

$$\frac{1}{2} \log(n) > 1 \text{ for } n > 8 \quad (23)$$

- La penalidad de SIC es mayor que la penalidad de AIC,
- SIC tiende a elegir modelos más pequeños.
- En efecto, al dejar que la penalización tienda al infinito lentamente con n , eliminamos la tendencia de AIC a elegir un modelo demasiado grande.

Error de Prueba y de Entrenamiento

- ▶ Dos conceptos importantes

- ▶ *Test Error*: es el error de predicción en la muestra de prueba (test)

$$Err_{\mathcal{T}_{est}} = MSE[(y, \hat{y}) | \mathcal{T}_{est}] \quad (24)$$

- ▶ *Training error*: es el error de predicción en la muestra de entrenamiento (training)

$$Err_{\mathcal{T}_{rain}} = MSE[(y, \hat{y}) | \mathcal{T}_{rain}] \quad (25)$$

Train and test samples

- Cómo elegimos \mathcal{T}_{est} ?

Train and test samples

- ▶ Cómo elegimos \mathcal{T}_{est} ?
- ▶ Una alternativa simple sería dividir los datos en dos:
 - ▶ Training sample: para construir/estimar/entrenar el modelo
 - ▶ Test sample: para evaluar el desempeño
- ▶ Desde una perspectiva estrictamente clásica
 - ▶ Tiene sentido si los datos de entrenamiento son iid de la población, incluso funciona si es iid condicional en X
 - ▶ Dos problemas con esta idea:
 - ▶ El primero es que, dado un conjunto de datos original, si parte de él se deja de lado para probar el modelo, quedan menos datos para la estimación (lo que lleva a una menor eficiencia).
 - ▶ Un segundo problema es cómo decidir qué datos se usarán para entrenar el modelo y cuáles probarlo.

Enfoque de conjunto de validación

- Podemos entonces aproximar esta idea partiendo la muestra en 2

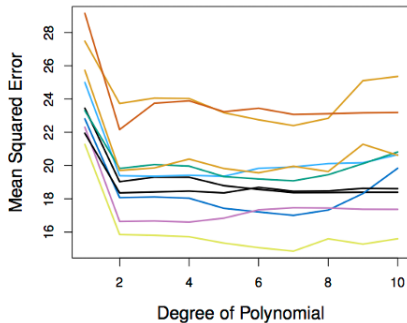
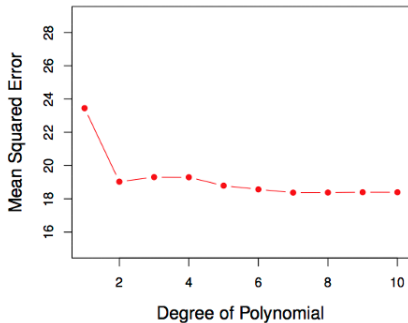


Training Data

Testing Data

Enfoque de conjunto de validación

- Modelo $y = f(x) + u$ donde f es un polinomio de grado p^* .
- Izquierda: error de predicción en la muestra de prueba para una sola partición
- Derecha: error de predicción en la muestra de prueba para varias particiones
- Hay un montón de variabilidad. (Necesitamos algo más estable)



Enfoque de conjunto de validación

- ▶ Ventajas:

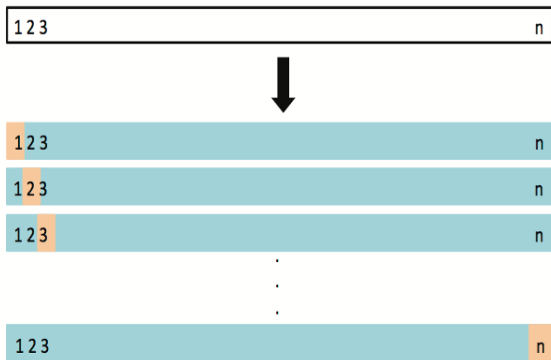
- ▶ Simple
- ▶ Fácil de implementar

- ▶ Desventajas:

- ▶ El MSE de validación (prueba) puede ser altamente variable
- ▶ Solo se utiliza un subconjunto de observaciones para ajustar el especificationo (datos de entrenamiento). Los métodos estadísticos tienden a funcionar peor cuando se entrenan con pocas observaciones.

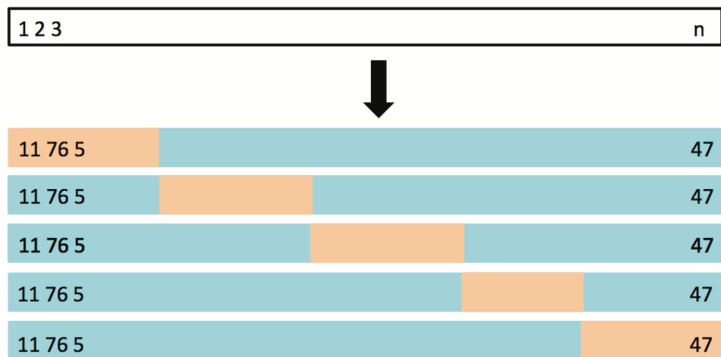
Leave-One-Out Cross Validation (LOOCV)

- Este método es similar al enfoque de validación, pero trata de abordar las desventajas de este último.



Validación cruzada en K-partes

- ▶ LOOCV es computacionalmente intensivo, por lo que podemos ejecutar k-fold Cross Validation



Validación cruzada en K-partes

- ▶ Dividir los datos en K partes ($N = \sum_{j=1}^K n_j$)
- ▶ Ajustar el modelo dejando afuera una de las partes (folds) $\rightarrow f_{-k}(x)$
- ▶ Calcular el error de predicción en la parte (fold) que dejamos afuera

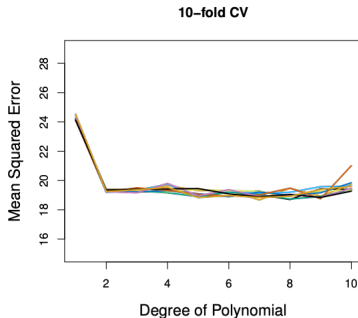
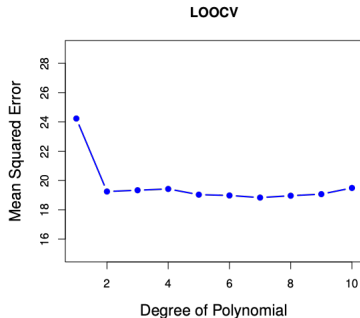
$$MSE_j = \frac{1}{n_j} \sum (y_j^k - \hat{y}_{-j})^2 \quad (26)$$

- ▶ Promediar

$$CV_{(k)} = \frac{1}{k} \sum_{j=1}^k MSE_j \quad (27)$$

Validación cruzada en K-partes

- ▶ Izquierda: LOOCV error
- ▶ Derecha: 10-fold CV
- ▶ LOOCV es caso especial de k-fold, donde $k = n$
- ▶ Ambos son estables, pero LOOCV (generalmente) es mas intensivo computacionalmente!



Trade-off Sesgo-Varianza para validación cruzada en K-partes

► Sesgo:

- El enfoque del conjunto de validación tiende a sobreestimar el error de predicción en la muestra de prueba (menos datos, peor ajuste)
- LOOCV, agrega más datos → menos sesgo
- K-fold un estado intermedio

► Varianza:

- LOOCV promediamos los resultados de n modelos ajustados, cada uno está entrenado en un conjunto casi idéntico de observaciones → altamente correlacionado
- K partes esta correlación es menor, estamos promediando la salida de k modelo ajustado que están algo menos correlacionados

► Por lo tanto, existe un trade-off

- Tendemos a usar k-fold CV con ($K = 5$ y $K = 10$)
- Se ha demostrado empíricamente que producen estimaciones del error de predicción que no sufren ni de un sesgo excesivamente alto ni de una varianza muy alta Kohavi (1995)

Validation and Cross-validation en la práctica

248

QUARTERLY JOURNAL OF ECONOMICS

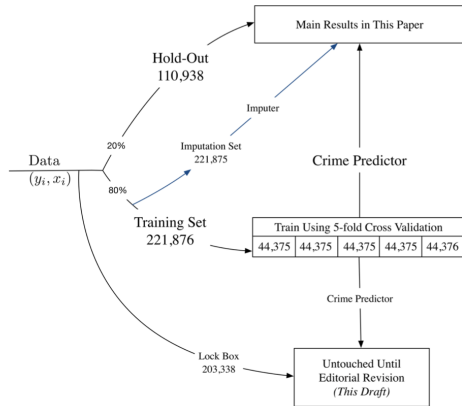


FIGURE I

Partition of New York City Data (2008–13) into Data Sets Used for Prediction and Evaluation

Source: Kleinberg et al (2018)

- 1 Recap
- 2 Predicción: estadística clásica vs la máquina de aprender
 - Tipos de Aprendizaje
 - Prediction vs Estimation
 - Sobreajuste y Predicción fuera de Muestra
- 3 Selección de Modelo
 - Enfoque Clásico
 - Métodos de remuestreo
 - Enfoque de conjunto de validación
 - LOOCV
 - Validación cruzada en K-partes
- 4 Para seguir leyendo
- 5 Break

Para seguir leyendo

- ▶ Davidson, R., & MacKinnon, J. G. (2004). Econometric theory and methods (Vol. 5). New York: Oxford University Press.
- ▶ James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- ▶ Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
- ▶ Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.
- ▶ Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. Journal of political economy, 82(1), 34-55.

Break