Lecture 7: Machine Learning
Overfit & Cross Validation

Big Data and Machine Learning en el Mercado Inmobiliario
Educación Continua

Ignacio Sarmiento-Barbieri

Universidad de los Andes

September 8, 2021

Agenda

1. Recap
2. Overfit
   - Overfit y Predicción fuera de Muestra
3. Métodos de Remuestreo
   - Enfoque de conjunto de validación
   - LOOCV
   - Validación cruzada en K-partes
4. Further Readings
5. Break

Predicción y Error Predictivo

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

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

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

Predicción y Error Predictivo

- En la práctica no conocemos $f(X)$
- Es necesario estimarla $\hat{Y} = \hat{f}(X)$

Entonces

$$Err(Y) = MSE(\hat{f}) + \sigma^2 \tag{2}$$
$$= Bias^2(\hat{f}) + V(\hat{f}) + Irreducible\ Error \tag{3}$$

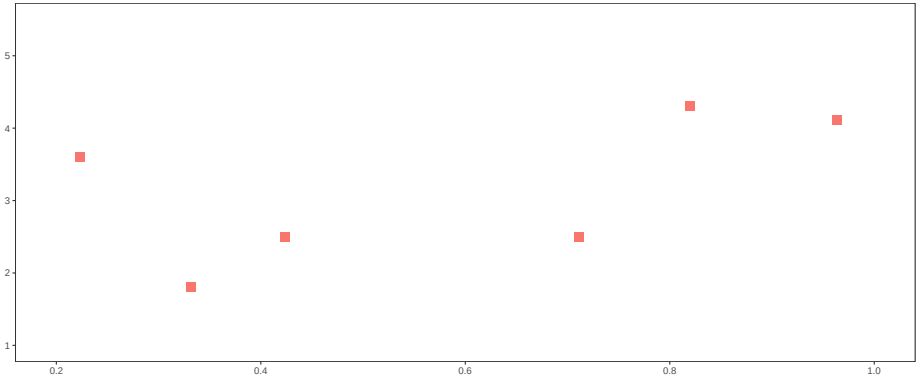
Dos partes

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

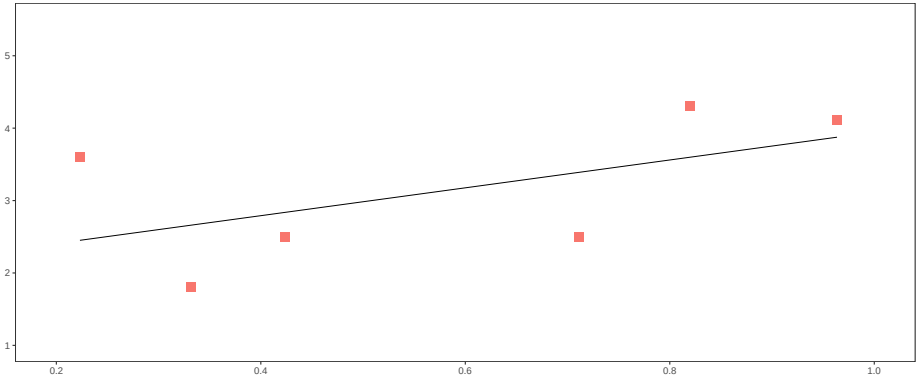
Este resultado es muy importante,

- predecir Y implica predecir bien f.
- existe un dilema entre sesgo y varianza

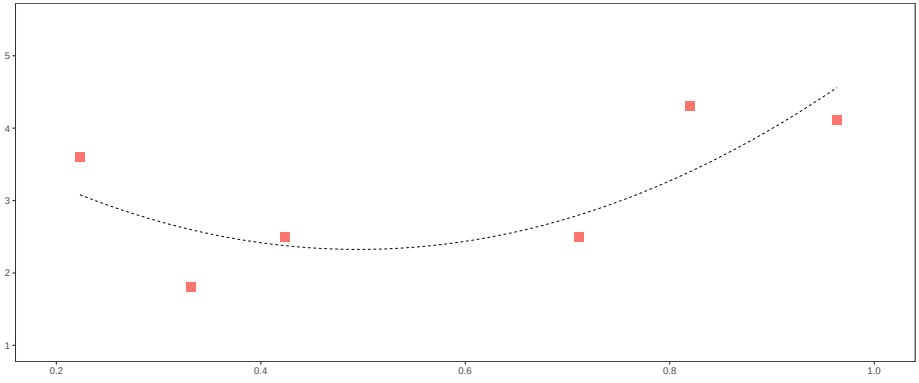
Overfit



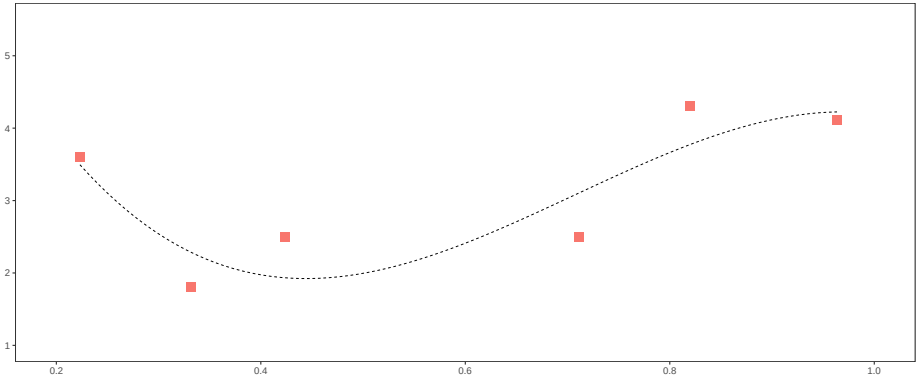
Overfit



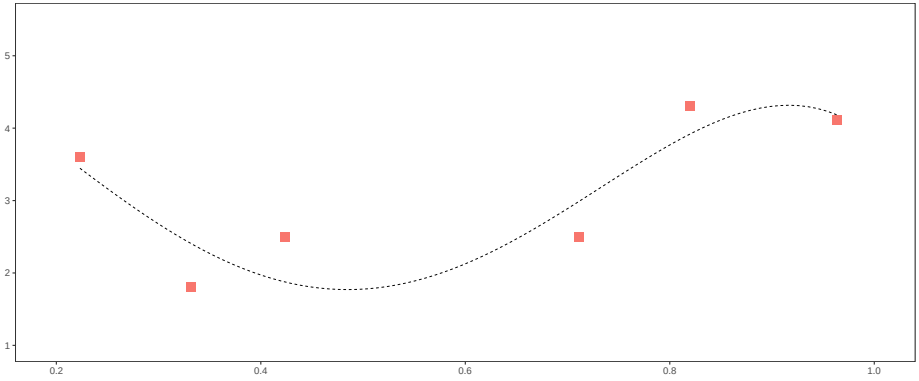
Overfit



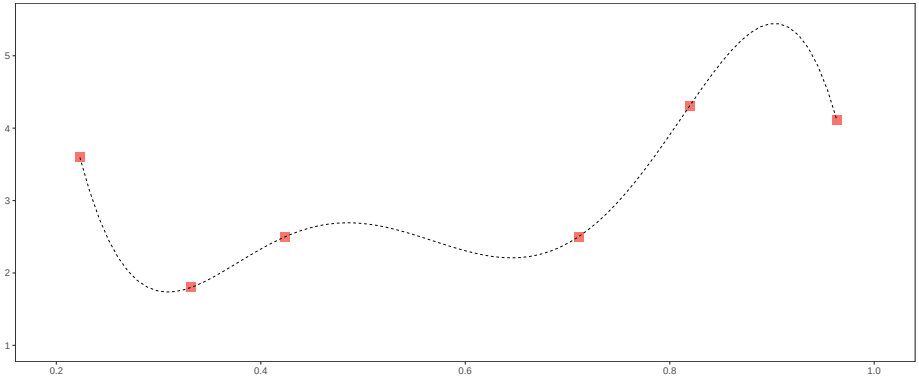
Overfit



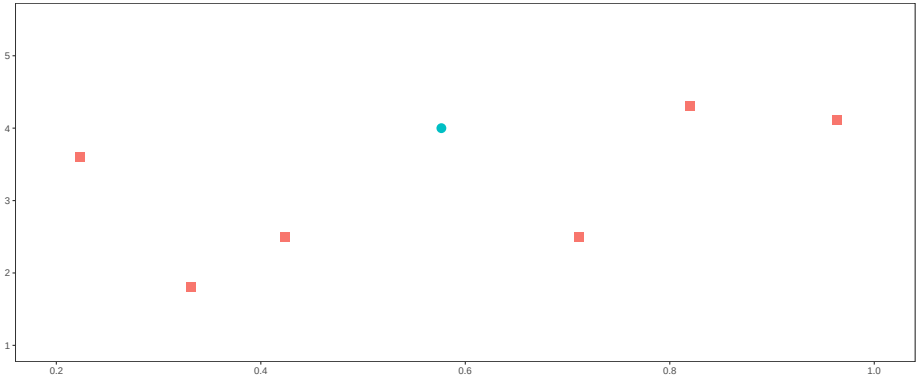
Overfit



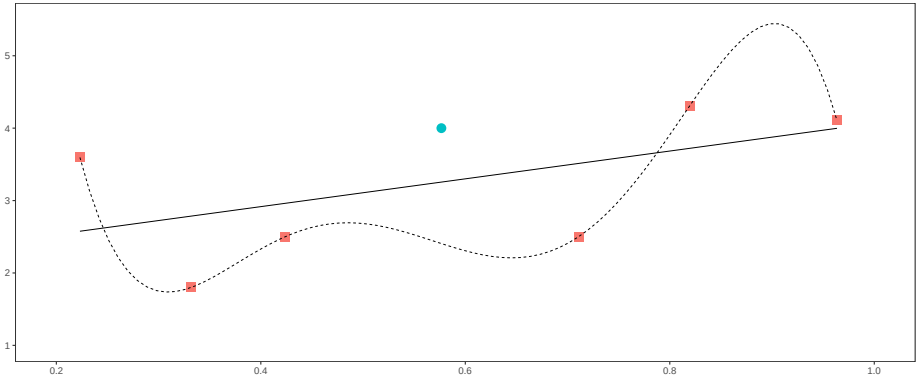
Overfit



Overfit



Overfit



Overfit

- En efecto si el modelo verdadero es $y = f(x) + u$
- donde f es un polinomio de grado p^*, with $E(u) = 0$ and $V(u) = \sigma^2$
- con p^* finito pero desconocido
- podemos ajustar polinomios de grados crecientes $p = 1, 2, ...$

$$Err(Y) = MSE(\hat{f}) + \sigma^2 \tag{4}$$

$$= Bias^2(\hat{f}) + V(\hat{f}) + Irreducible\ Error \tag{5}$$

Overfit

- Bias ?

$$\hat{f}(x) = X'\hat{\beta} = \sum_{s=0}^{p} x^s \hat{\beta}_s = x'\hat{\beta} \tag{6}$$

donde $X' = (1, x, , x^2, \ldots, x^p)$

Overfit

- Varianza:

$$V(\hat{f}(x)) = V(X'\hat{\beta}) = \sigma^2 \frac{p}{n} \tag{7}$$

Después de p^* aumentar la complejidad no reduce el sesgo, pero la varianza aumenta monotónicamente para σ^2 y n dados

Overfit 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: mide predicción dentro de muestra, es no decreciente en complejidad

Overfit y Predicción fuera de Muestra

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

$$Err_{\mathcal{T}est} = MSE[(Y, \hat{Y})|\mathcal{T}est] \tag{8}$$

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

$$Err_{\mathcal{T}rain} = MSE[(Y, \hat{Y})|\mathcal{T}rain] \tag{9}$$

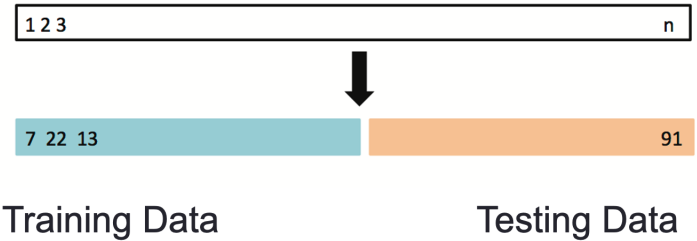
- Como elegimos $\mathcal{T}est$?

Qué son los Métodos de Remuestreo?

- Herramientas que implican extraer repetidamente muestras de un conjunto de entrenamiento y reajustar el modelo de interés en cada muestra para obtener más información sobre el modelo.
- Evaluación del modelo: estimar el error de predicción en la muestra de prueba
- Selección de modelo: seleccione el nivel apropiado de flexibilidad del modelo
- ¡Son computacionalmente costosos! Pero en estos días tenemos computadoras poderosas

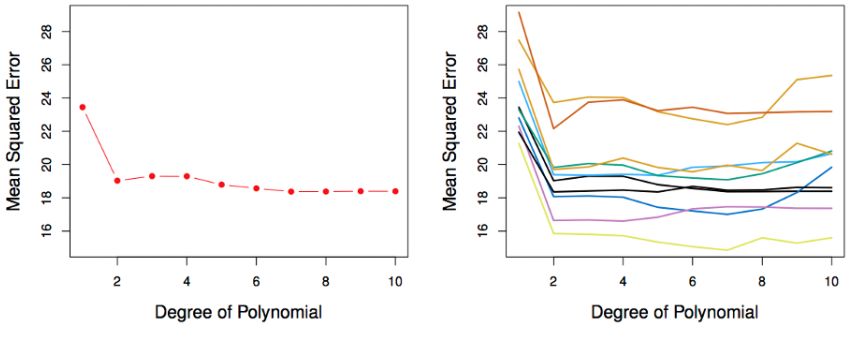
Enfoque de conjunto de validación

- Suponga que nos gustaría encontrar un conjunto de variables que den el menor error de predicción en la muestra de prueba (no de entrenamiento)
- Si tenemos muchos datos, podemos lograr este objetivo dividiendo aleatoriamente los datos en partes de entrenamiento y validación (prueba)
- Luego usaríamos la parte de entrenamiento para construir cada modelo posible (es decir, las diferentes combinaciones de variables) y elegimos el modelo que dio lel menor error de predicción en la muestra de prueba



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 monton de variabilidad. (Necesitamos algo mas estable)

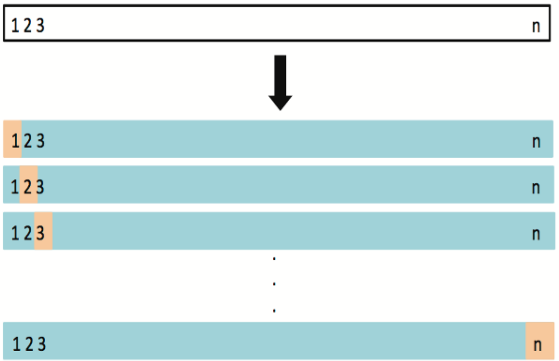


Enfoque de conjunto de validación

- 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 modelo (datos de entrenamiento). Los métodos estadísticos tienden a funcionar peor cuando se entrenan con pocas observaciones

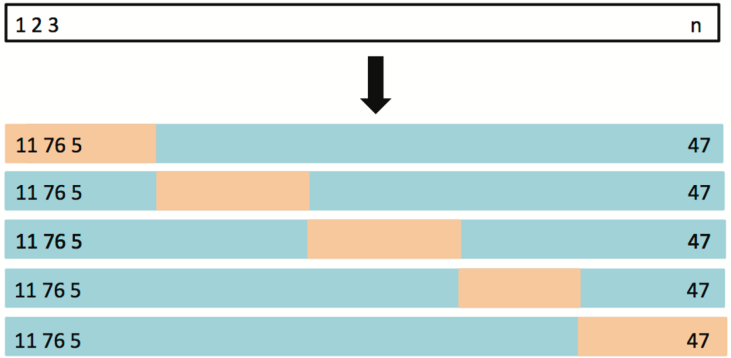
Leave-One-Out Cross Validation (LOOCV)

- 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

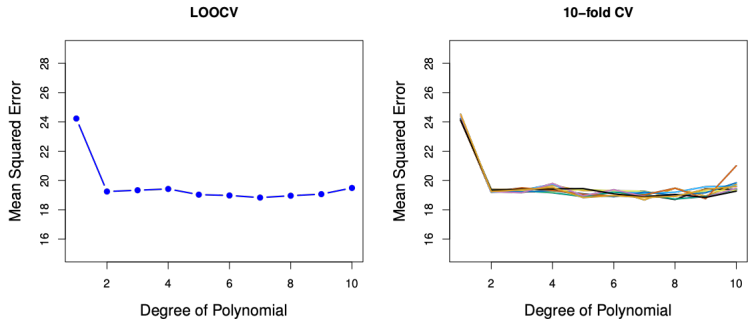
$$\bar{err}_j = MSE_j = \frac{1}{n_j} \sum L(y_j^k, \hat{y}_{-j}) \tag{10}$$

- Promediar

$$CV_{(k)} = \frac{1}{k} \sum_{j=1}^{k} \bar{err}_j = \frac{1}{k} \sum_{j=1}^{k} MSE_j \tag{11}$$

Validación cruzada en K-partes

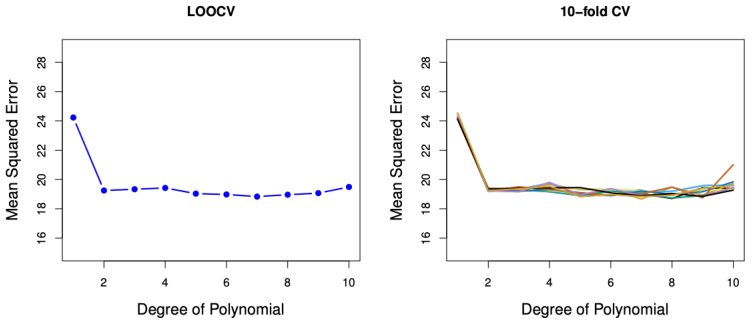
- Izquierda: LOOCV error
- Derecha: 10-fold CV
- LOOCV es caso especial de k-fold, donde k = n
- Ambos son estables, pero LOOCV (generalmente) es mas intensivo computacionalmente!



Validación cruzada en K-partes para selección de modelos

- Supongamos que α parametriza la complejidad del modelo (en nuestro ejemplo el grado del polinomio)
- Primero calculamos el CV error para un grupo de modelos (α), y elegimos el minimo

$$\min_{\alpha} CV_{(k)}(\alpha) \tag{12}$$



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 $\rightarrow$ 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 $\rightarrow$ 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 prediccion que no sufren ni de un sesgo excesivamente alto ni de una varianza muy alta Kohavi (1995)

Review & Next Steps

- Today:
  - Overfit and out of Sample Prediction
  - Metodos de Resampleo
    - Enfoque de Validación
    - LOOCV
    - K-fold Cross-Validation (Validación Cruzada)

Further Readings

- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).
- Lovelace, R., Nowosad, J., & Muenchow, J. (2019). Geocomputation with R. CRC Press. (Chapters 2 & 6)

Volvemos en 5 min con Python