Lecture 2: Sobreajuste & Validación Cruzada

Aprendizaje y Minería de Datos para los Negocios

Ignacio Sarmiento-Barbieri

Universidad de los Andes

October 22, 2021

Agenda

1. Recap
   - La máquina de aprender y Modelos Lineales
2. Overfit
   - Overfit y Predicción fuera de Muestra
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
3. Métodos de Remuestreo
   - Enfoque de conjunto de validación
   - LOOCV
   - Validación cruzada en K-partes
4. Break
5. `R para ML`

¿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..."

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:

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

- donde $f(X)$ es cualquier función,
- 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 $\hat{y} = \hat{f}(X)$
- La medida de cuan bien funciona nuestro modelo es $MSE(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2$
- Notemos que podemos descomponer el MSE en dos partes

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

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

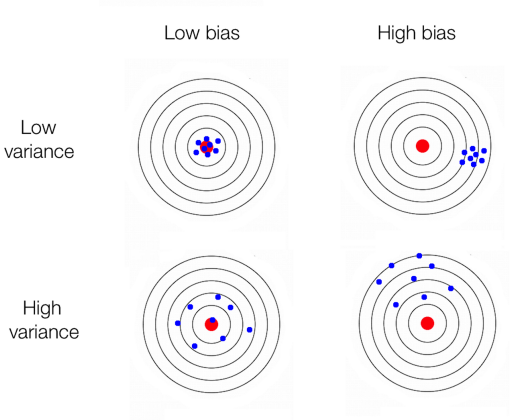
Predicción y Error Predictivo

- Descomponiendo un poco más:

$$Err(Y) = MSE(\hat{f}) + \sigma^2 \tag{3}$$
$$= Bias^2(\hat{f}) + V(\hat{f}) + Irreducible\,Error \tag{4}$$

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

Prediction Error



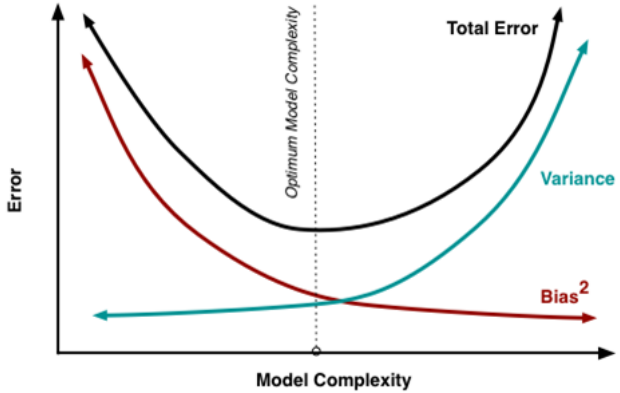
Source: https://tinyurl.com/y4lvjxpc

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:

$$y = f(X) + u \tag{5}$$

- proponemos que:

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{6}$$

- El problema se reduce a encontrar los βs
  - Un camino es OLS

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:

$$
\begin{aligned}
E(X\hat{\beta}) &= E(\hat{\beta}_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p) && (7)\\
&= E(\hat{\beta}_1) + E(\hat{\beta}_2) X_2 + \cdots + E(\hat{\beta}_p) X_p && (8)\\
&= X\beta && (9)
\end{aligned}
$$

- $MSE(\hat{y})$ 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:

$$y = \beta_1 X_1 + u_1$$

- $\hat{\beta}_1^{(1)}$ el estimador de OLS y on X_1
- La predicción es:

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

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

- $\hat{\beta}_1^{(2)}$ 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:

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

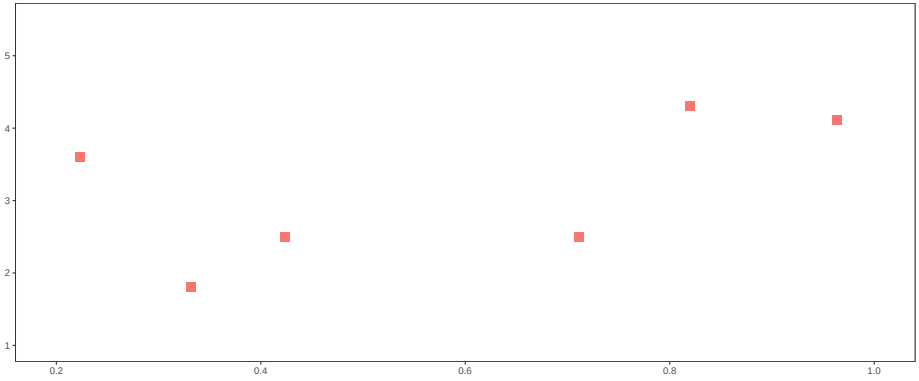
Complejidad y compensación de varianza/sesgo

- 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.
  - Si el modelo (2) se 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!!!

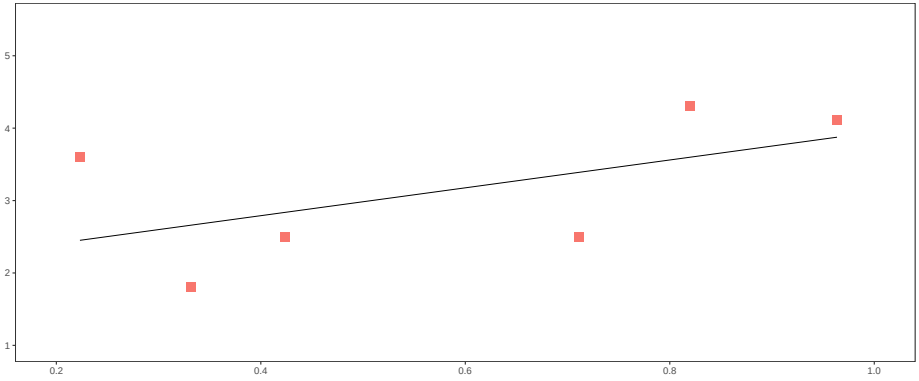
Complejidad y compensación de varianza/sesgo

- 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.

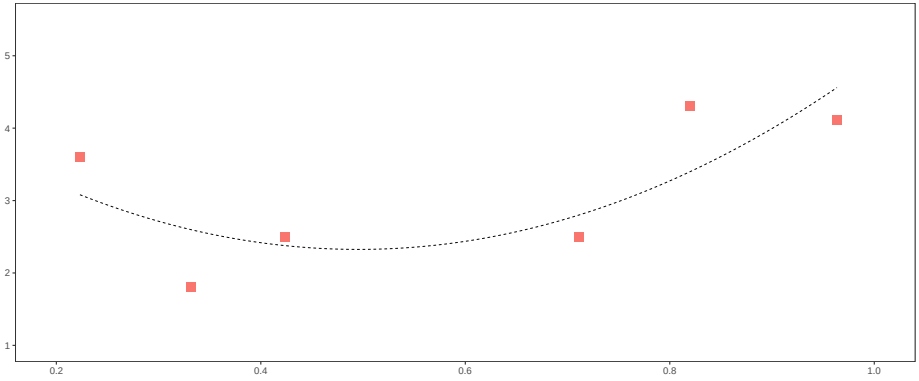
Overfit



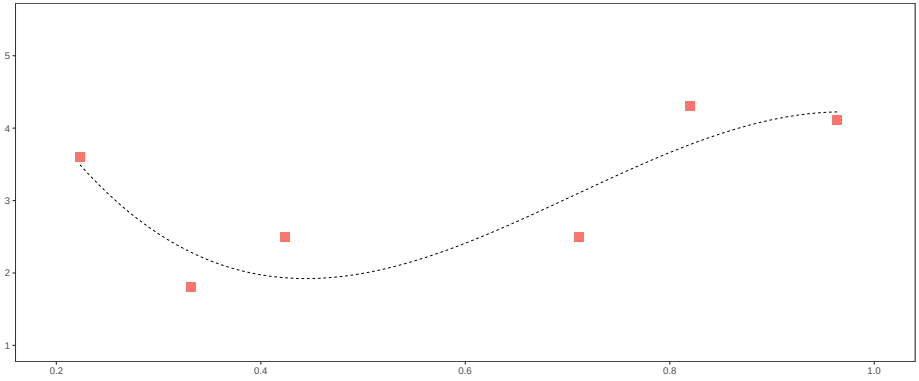
Overfit



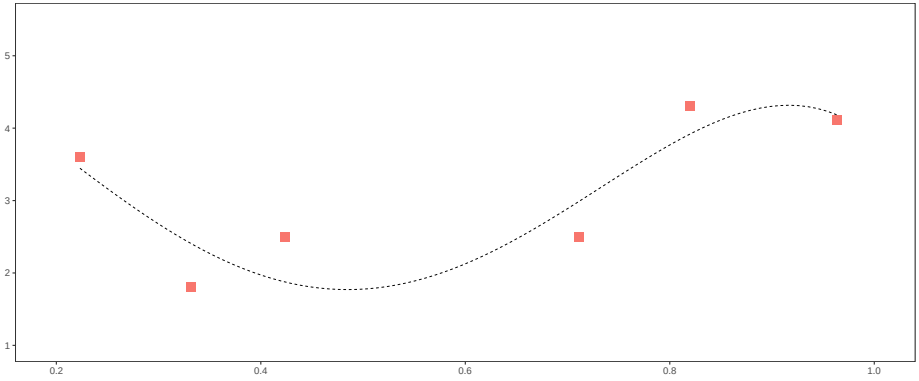
Overfit



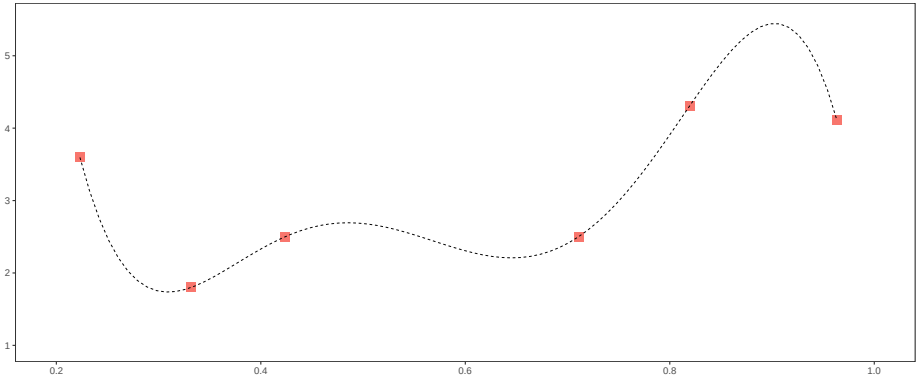
Overfit



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 $p = 1, 2, ...$

$$Err(Y) = MSE(\hat{f}) + \sigma^2 \tag{10}$$

$$= Bias^2(\hat{f}) + V(\hat{f}) + Irreducible\ Error \tag{11}$$

Overfit

- Sesgado ?

$$\hat{f}(x) = X'\hat{\beta} = \sum_{s=0}^{p} x^s \hat{\beta}_s = x'\hat{\beta} \tag{12}$$

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

Overfit

- Varianza:

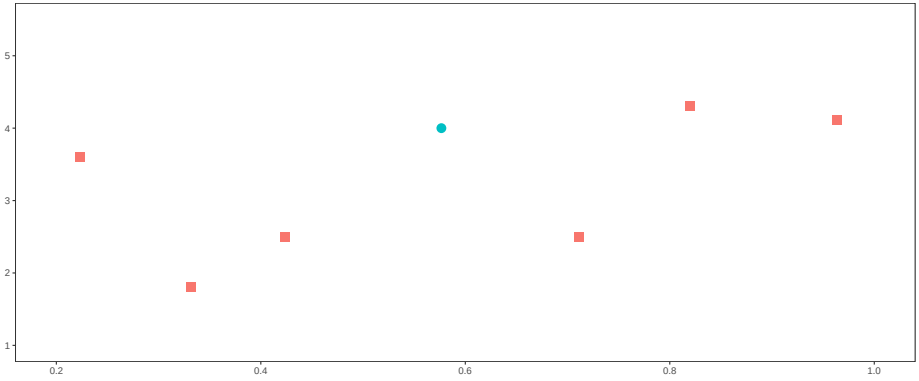
$$V(\hat{f}(x)) = V(X'\hat{\beta}) = \sigma^2 \frac{p}{n} \tag{13}$$

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

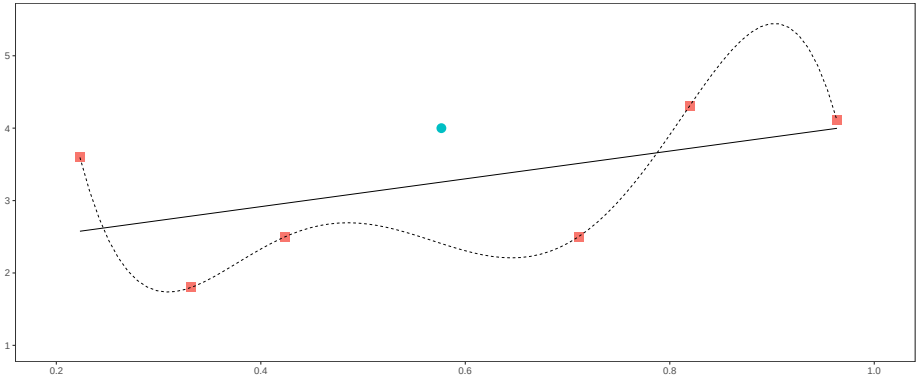
Overfit y Predicción fuera de Muestra

- 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: se concentra en la muestra y es no decreciente en complejidad

Overfit



Overfit



AIC

- 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

$$AIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - p_j \tag{14}$$

BIC

- 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 $n \to \infty$, 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:

$$SIC(j) = log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2\right) - \frac{1}{2}p_j log(n) \tag{15}$$

AIC vs BIC

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

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

- Note que

$$\frac{1}{2}log(n) > 1 \ \ for \ \ n > 8 \tag{18}$$

- 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.

Métodos de resampleo

- Los métodos de resampleo son una herramienta indispensable de la estadística moderna.
- Estos envuelven sacar muestras aleatorias de nuestra muestra y reajustar el modelo de interés en cada muestra para obtener información adicional del modelo.
- Quizás el método más conocido por ustedes es el de bootstrap.
- Nosotros vamos a discutir la validación cruzada (cross-validation)

Error de Prueba y de Entrenamiento

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

$$Err_{\mathcal{T}est} = MSE[(y, \hat{y})|\mathcal{T}est] \tag{19}$$

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

$$Err_{\mathcal{T}rain} = MSE[(y, \hat{y})|\mathcal{T}rain] \tag{20}$$

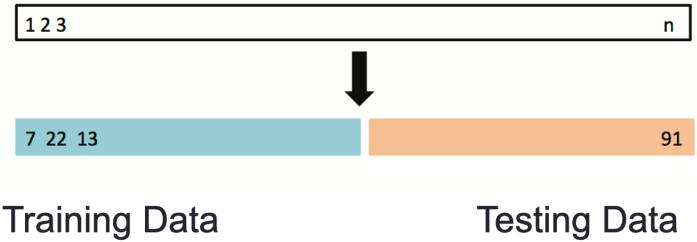
- Cómo elegimos $\mathcal{T}est$?

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

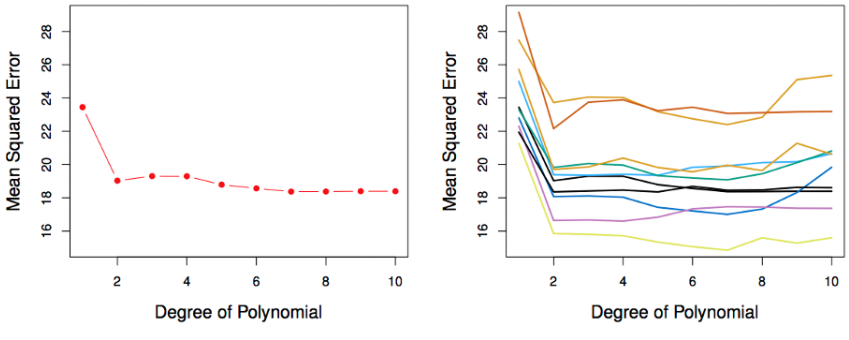
Enfoque de conjunto de validación

- 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 montón de variabilidad. (Necesitamos algo mas estable)

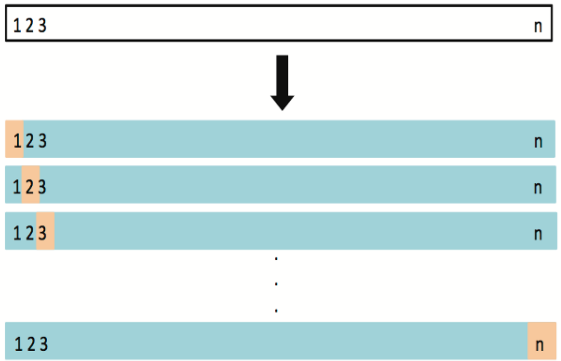


Enfoque de conjunto de validación

- 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

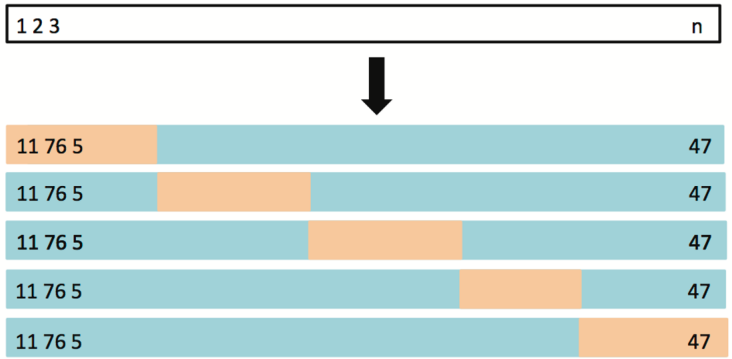
Leave-One-Out Cross Validation (LOOCV)

- 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

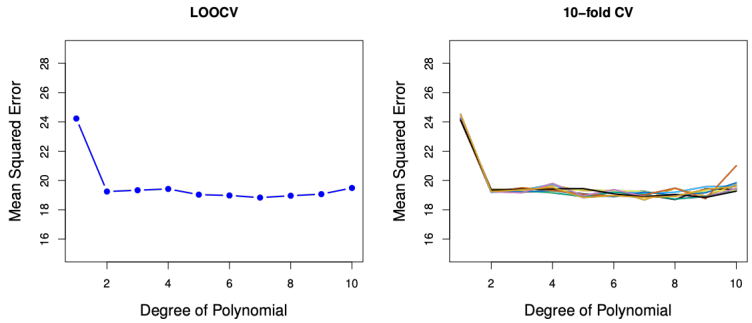
$$MSE_j = \frac{1}{n_j} \sum (y_j^k - \hat{y}_{-j})^2 \tag{21}$$

- Promediar

$$CV_{(k)} = \frac{1}{k} \sum_{j=1}^{k} MSE_j \tag{22}$$

Validación cruzada en K-partes

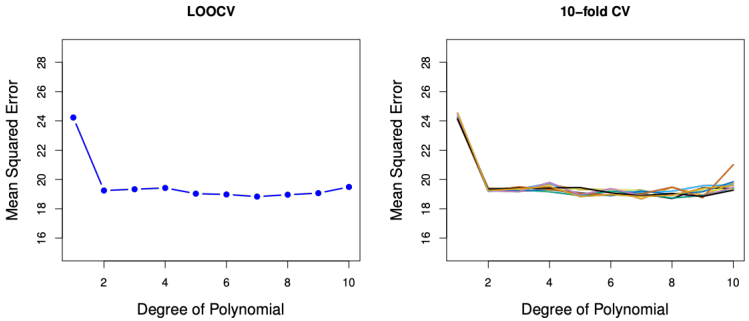
- Izquierda: LOOCV error
- Derecha: 10-fold CV
- LOOCV es caso especial de k-fold, donde k = n
- Ambos son estables, pero LOOCV (generalmente) es mas intensivo computacionalmente!



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 mínimo

$$\min_{\alpha} CV_{(k)}(\alpha) \tag{23}$$



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)

Revisión & Próximos Pasos

Hoy

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

Volvemos en 15 mins con R

R para ML

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