Fundamentos de la Programación Estadística en R

Algunas nociones fundamentales de Machine Learning

Germán Rosati
german.rosati@gmail.com

UNTREF - UNSAM - Digital House

12 de Marzo de 2018

¿Qué es un modelo?

- Básicamente: una manera de proponer hipótesis sobre la forma en que se combinan variables
- En general, vamos a estar tratando de generar modelos de esta forma

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

- Todo el problema es estimar $f(X)$, es decir, de qué forma(s) se combinan las X para generar un output
- Una posibilidad es suponer que Y es una combinación lineal de las X

¿Qué es un modelo?

Las dos culturas (Breiman, 2001)

"Todos los modelos son equivocados. Algunos son útiles."
George Box

- Podemos pensar al mundo como un productor de outputs en base a features

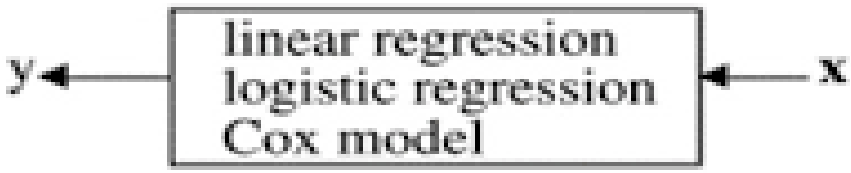
- Caja negra en las que se combinan las X de alguna forma que desconocemos
- Los problemas surgen cuando queremos estimar cuál es la manera en que el mundo produce resultados

¿Qué es un modelo?

Las dos culturas (Breiman, 2001)

Modelado estadístico

- Se comienza asumiendo que dentro de la caja negra hay un modelo estadístico
- Una forma común es asumir que los datos son generados por extracciones independientes de
  $output = f(predictores, ruido, parametros)$
- Los parámetros son estimados con los datos y luego se realizan las predicciones. La caja se llena con



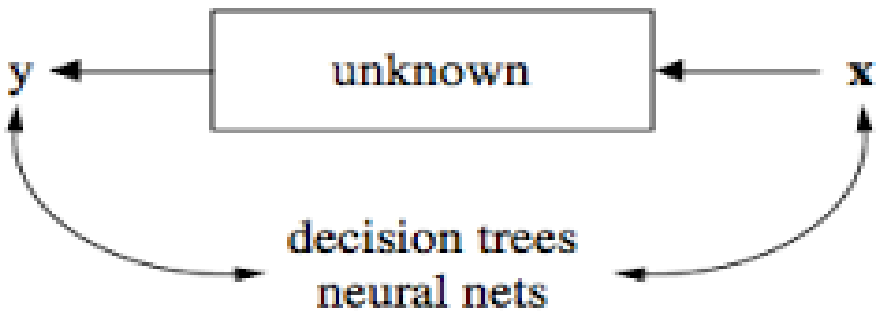
- Validación: Sí-No, usando medidas de bondad de ajuste y análisis de residuos

¿Qué es un modelo?

Las dos culturas (Breiman, 2001)

Modelado algorítmico (o Machine Learning, Data Mining, etc.)

- Se comienza asumiendo lo que pasa dentro de la caja es demasiado complejo y desconocido
- El enfoque es encontrar una función $f(x)$ -un algoritmo- que opera sobre las x para predecir las y. La caja negra tiene esta forma



- Validación: medido a través de la cuantificación de la precisión predictiva

¿Qué es un modelo?

¿Cómo evaluar un modelo?

- Ahora bien, ¿qué es un buen modelo?
- Desde la cultura del **modelado estadístico** un buen modelo es un modelo que ajusta bien a los datos y cuyos parámetros cumplen algunas propiedades "deseables"
  1. Ser insesgado
  2. Ser robusto
  3. Tener varianza mínima...
  4. Etc...

- El **modelado algorítimo** piensa sobre todo en la capacidad predictiva
- Pero... ¿sobre cuáles datos?
- Queremos modelos que funcionen bien -tengan bajo error- en datos que NO vimos, es decir, en datos "futuros", datos de test, *out of sample*
- Pero muchas veces esos datos no existen o tardan en aparecer
- $\Longrightarrow$ Separación en *Training Data* y *Test Data*
- Entreno-estimo-construyo el modelo sobre *Training Data* y evalúo sobre *Test Data*

- Que un modelo funcione bien en datos de entrenamiento no quiere decir que funcione bien en datos nuevos...
- En general, el error en datos de entrenamiento es más bajo que el error en datos de test

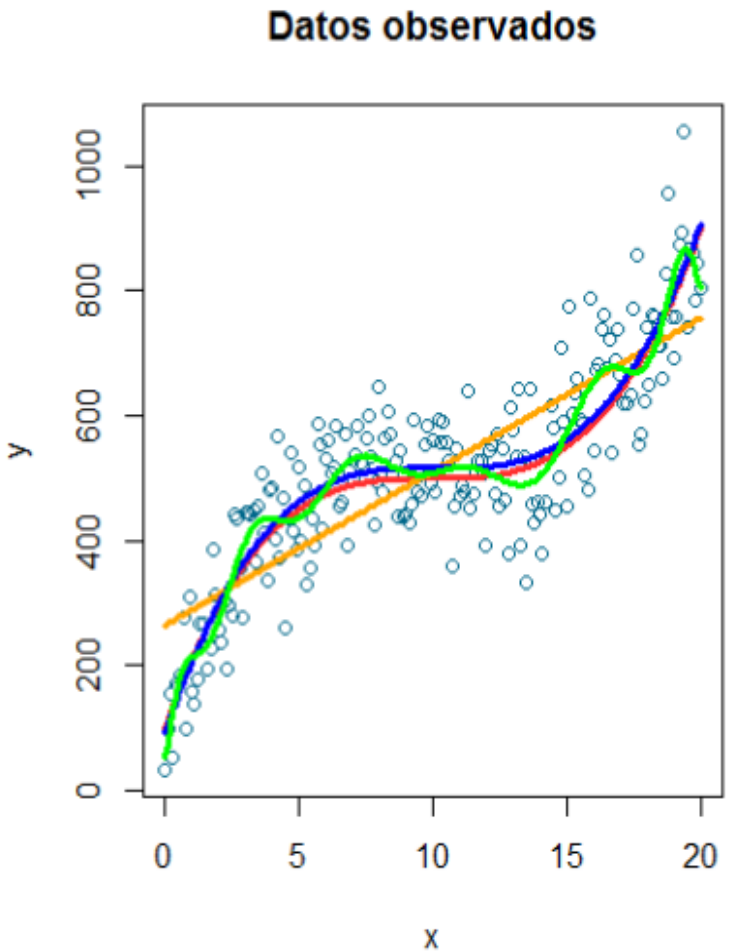
¿Qué es un modelo?

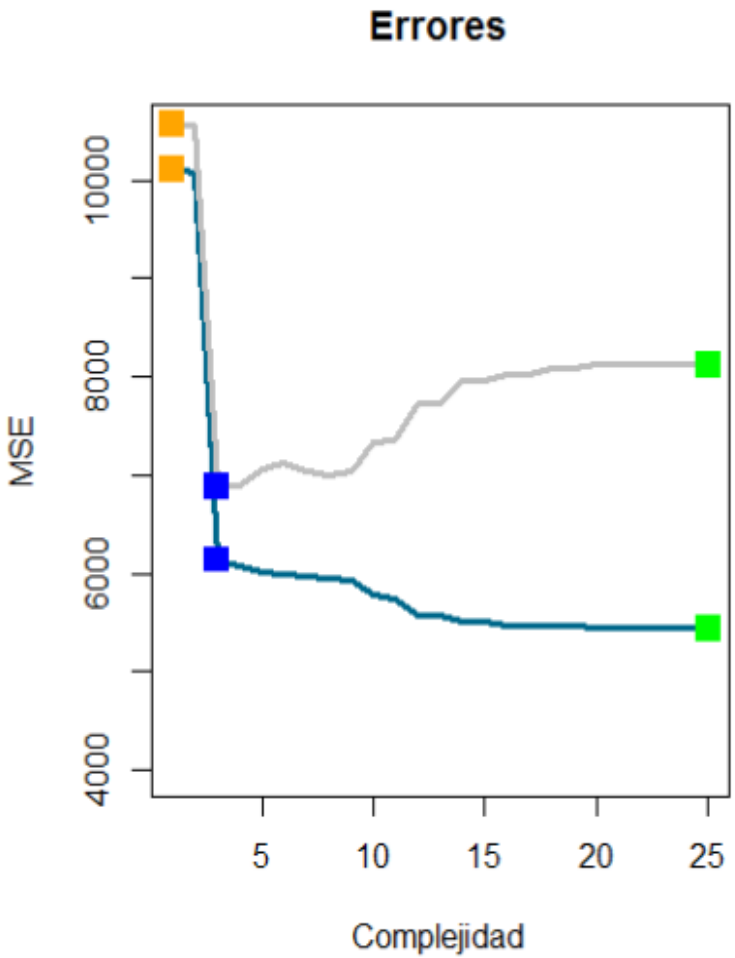
¿Cómo evaluar un modelo?

- Función original: $f(x_i) = 500 + 0{,}4X_i^3 + \epsilon_i$
- Modelo Lineal: $\hat{f}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- Modelo Cuadrático: $\hat{f}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2$
- Modelo Polinómico de orden 25:
  $\hat{f}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2 + ... + +\hat{\beta_{25}} X_i^{25}$

¿Qué es un modelo?

¿Cómo evaluar un modelo?





- TrS-error decrece constantemente: siempre es posible generar un modelo muy "complejo" como para que ajuste bien a los datos (¿cuáles?)
- TeS-error decrece hasta un punto y luego comienza a crecer nuevamente. Se produce "overfitting" (sobreajuste).
- El modelo "trabaja" demasiado para encontrar patrones en el TrS y tiende a confundir el verdadero patrón ($f(x)$ - el "proceso generador de los datos") con ruido (ϵ) que no existe en el TeS.

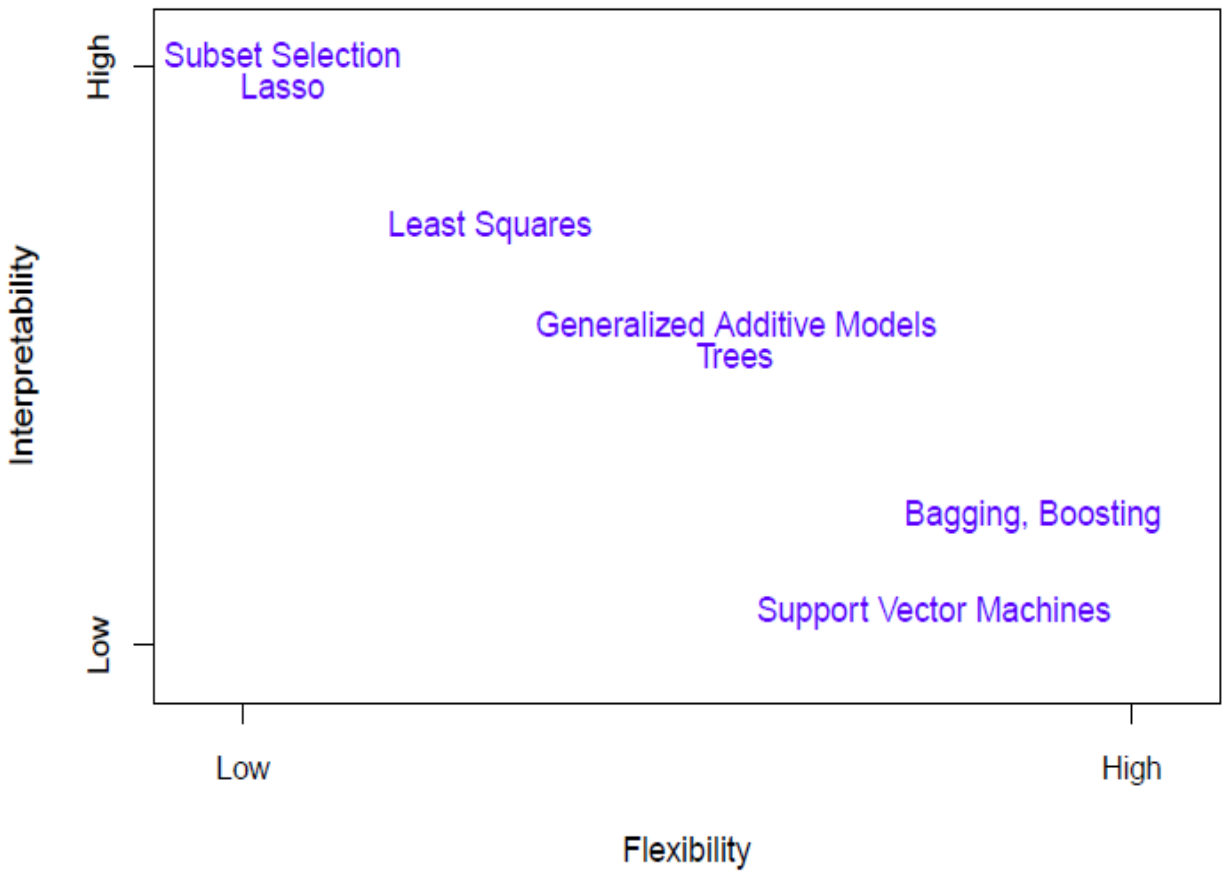
- El ECM puede descomponerse en tres partes

$$E[(y - \hat{f}(x))^2] = V(\hat{f}(x)) + bias^2 + \sigma^2 \quad (2)$$

- Error debido al sesgo: diferencia entre el valor esperado de nuestra predicción y el verdadero valor poblacional
- Error debido a la varianza: producido por la variabilidad de las predicciones del modelo en un punto determinado.
- El σ^2 es la parte ïrreductible"del error en el modelo

¿Qué es un modelo?

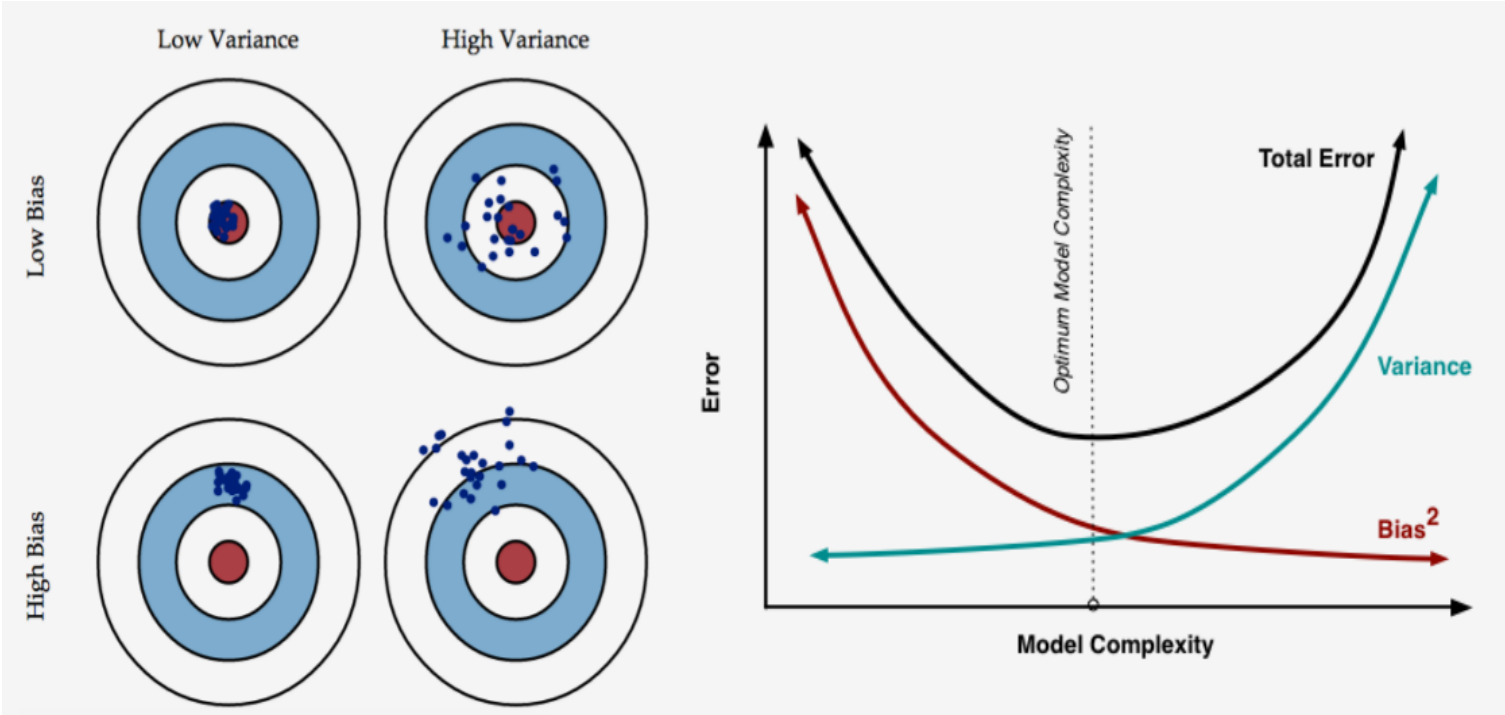
¿Cómo evaluar un modelo? - Algunos trade-offs



Fuente: James, Witten, Hastie y Tibshiraini (2013): 25

¿Qué es un modelo?

¿Cómo evaluar un modelo? - Balance Sesgo-Varianza



- Herramientas para estimar el error de generalización de un modelo -qué tan bien va a funcionar sobre datos "no vistos"
  1. *Validation Set Approach*
  2. *Cross Validation*
  3. *Bootsrap*
  4. *Etc.*

¿Qué es un modelo?

Validation Set Approach

- Dividimos el **aleatoriamente** *dataset* en *Training Set - TrS* y *Test Set - TeS*
- El modelo se ajusta en el TrS y el modelo ajustado se usa para predecir las observaciones correspondientes al TeS

¿Qué es un modelo?

Cross Validation

- Dividimos el **aleatoriamente** *dataset* en K porciones de igual tamaño
- Fiteamos el modelo dejando como TeS una de las K partes
- Computamos el error en la parte dejada afuera previamente
- Repetimos para $k = 1, 2, 3, ..., K$
- La estimación del error será el promedio de las K estimaciones de error

¿Qué es un modelo?)

Cross Validation

	Dataset Original				
Iteración 1	C1 (VaSet)	C2 (TrSet)	C3 (TrSet)	C4 (TrSet)	C5 (TrSet)
Iteración 2	C1 (TrSet)	C2 (VaSet)	C3 (TrSet)	C4 (TrSet)	C5 (TrSet)
Iteración 3	C1 (TrSet)	C2 (TrSet)	C3 (VaSet)	C4 (TrSet)	C5 (TrSet)
Iteración 4	C1 (TrSet)	C2 (TrSet)	C3 (TrSet)	C4 (VaSet)	C5 (TrSet)
Iteración 5	C1 (TrSet)	C2 (TrSet)	C3 (TrSet)	C4 (TrSet)	C5 (VaSet)

Comentarios finales

- La máxima de Box...
- Dado que todos los modelos son simplificaciones de la realidad, no podemos llegar a la "verdad" por complejidad creciente.
- Principio de Occam, caso contrario, *overfitting*
- ¿Modelado estadístico o algorítmico? Dependerá del problema en cuestión

Bibliografía recomendada