

¿Qué es Machine Learning?

Sesgo, varianza, overfitting y underfitting

German Rosati
german.rosati@gmail.com

UNSAM - CONICET

3 de octubre de 2022

¿Cómo evaluar un modelo?

- Métricas de error - Loss Functions
- Train y Test Data
- Overfitting - Underfitting
- Balance Sesgo-Varianza
- Estimando el error de generalización

Ahora sí... ¿Como evaluar un modelo?

Criterios I

- Ahora bien, ¿qué es un buen modelo?
- Desde la cultura del **modelado estadístico** (Breiman, 2001 [1]) 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...

¿Como evaluar un modelo?

Criterios II

- El **modelado algorítmico** (Breiman, 2001 [1]) 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
- \implies Separación en *Training Data* y *Test Data*
- Entreno-estimo-construyo el modelo sobre *Training Data* y evalúo sobre *Test Data*

¿Como evaluar un modelo?

Métricas de error - Loss Functions - Funciones de pérdida

Requisito: alguna medida que permita evaluar cómo funciona mi modelo

- Grande cuándo el modelo funciona "mal" pequeña cuando funciona "bien"
- Sirve para "tunear", calibrar los parámetros del modelo
- Muchas métricas: por ahora nombramos dos
 - Mean Squared Error para variables cuantitativas

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

- Error de clasificación para variables categóricas

$$err = Class = \frac{1}{N} \sum_{i=1}^N I(y_i \neq \hat{y}_i) \quad (2)$$

¿Como evaluar un modelo?

Train y 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

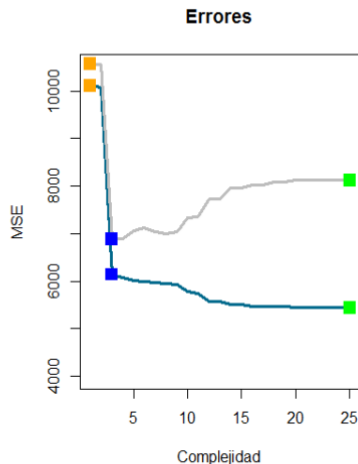
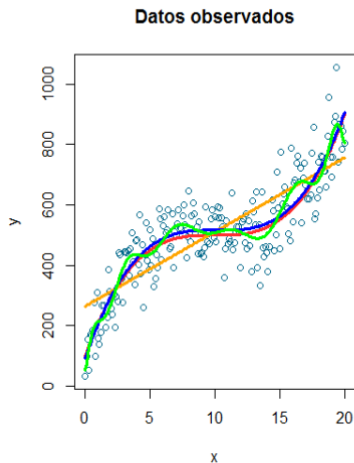
¿Como evaluar un modelo?

Ejemplo teórico

- 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:
 $\hat{f}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2 + \dots + \hat{\beta}_{25} X_i^{25}$

¿Como evaluar un modelo?

Ejemplo teórico



¿Como evaluar un modelo?

Overfitting - Underfitting

- 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.

¿Como evaluar un modelo?

Balance Sesgo-Varianza

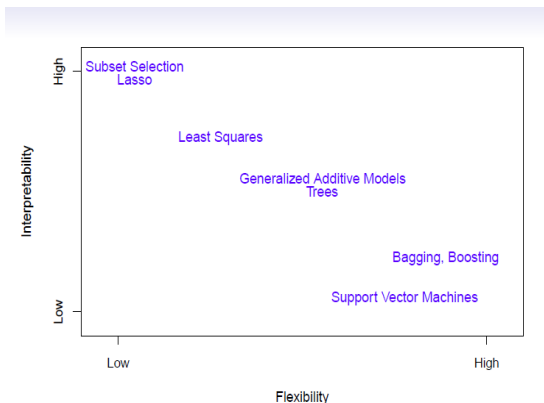
- El ECM puede descomponerse en tres partes

$$E[(y - \hat{f}(x))^2] = V(\hat{f}(x)) + bias^2 + \sigma^2 \quad (3)$$

- **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 “irreducible” del error en el modelo

¿Como evaluar un modelo?

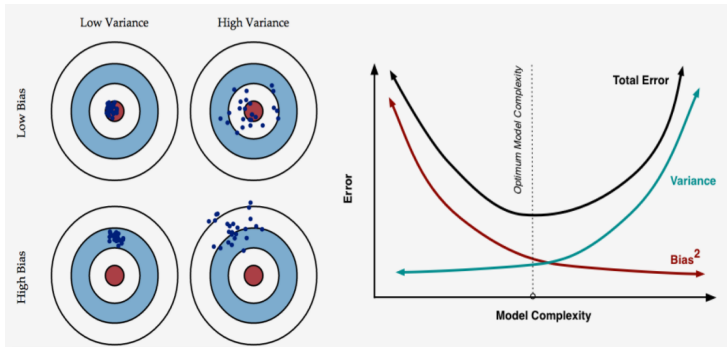
Algunos trade-offs



Fuente: (James, Witten, Hastie y Tibshirani, 2013 [2]) : 25

¿Como evaluar un modelo?

¿Cómo evaluar un modelo? - Balance Sesgo-Varianza



Fuente: [Scott Fortman's Blog](#)

¿Como evaluar un modelo?

Estimando el error de generalización

- 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.*

¿Como evaluar 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

¿Como evaluar un modelo?

Cross Validation

- 1 Dividimos el **aleatoriamente** *dataset* en K porciones de igual tamaño
- 2 Fiteamos el modelo dejando como TeS una de las K partes
- 3 Computamos el error en la parte dejada afuera previamente
- 4 Repetimos para $k = 1, 2, 3, \dots, K$

La estimación del error será el promedio de las K estimaciones de error

$$CV(\hat{f}) = \sum_{k=1}^K \frac{n_k}{N} err_k \quad (4)$$

¿Como evaluar 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)

¿Como evaluar un modelo?

Cross Validation

¿Cómo elegimos K ?

- K pequeño maximiza datos para estimar, sensible a valores extremos
- K grande maximza datos para evaluar, modelo estimado con menor precisión
- Regla del dígito pulgar oscilante: 5 o 10 (James, Witten, Hastie y Tibshirani, 2013 [2])

- 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

Referencias bibliográficas I



BREIMAN, L.

Statistical modeling: The two cultures (with comments and a rejoinder by the author).
Statistical Science 16, 3 (08 2001), 199–231.



JAMES, G., WITTEN, D., HASTIE, T., AND TIBSHIRANI, R.

An Introduction to Statistical Learning – with Applications in R, vol. 103 of *Springer Texts in Statistics*.
Springer, New York, 2013.