

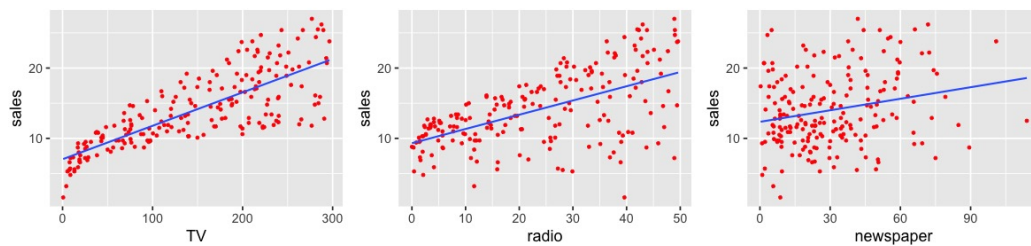
EST-25134: Aprendizaje Estadístico

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022.

Objetivo. Establecer las ideas básicas de aprendizaje estadístico. Ciertos criterios de optimalidad y descomposición del error. Discutiremos complejidad y compromiso entre sesgo y varianza.

Lectura sugerida: Capítulo 2, [2].

1. ¿QUÉ ES EL APRENDIZAJE ESTADÍSTICO?



Supongamos que tenemos datos de ventas de ciertas campañas de *marketing* en ciertos canales de distribución. Queremos estimar la relación

$$\text{Ventas} \approx f(\text{tv}, \text{radio}, \text{periodico}). \quad (1)$$

Lo podemos expresar como

$$Y = f(X) + \varepsilon. \quad (2)$$

1.1. ¿Por qué estimar f ?

- Podemos hacer predicciones.
- Podemos entender qué componentes de $X = (X_1, \dots, X_p)$ son importantes.
- Podemos tratar de entender la complejidad de f .

1.2. ¿Hay una f que sea óptima?

Podríamos utilizar

$$f(x) = \mathbb{E}[Y|x = 4], \quad (3)$$

que recibe el nombre **función de regresión**.

2. PROPIEDADES

La función de regresión (f) es **óptima** en términos del error cuadrático medio:

$$\mathbb{E}[(Y - g(x))^2 | X = x], \quad (4)$$

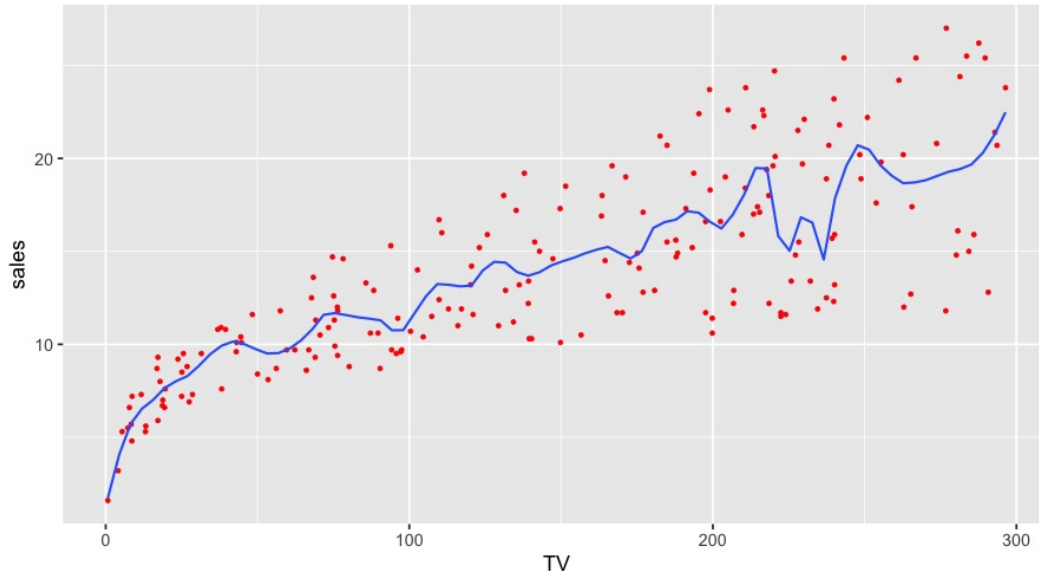


FIGURA 1. Ajuste por medio de promedios locales.

para cualquier función g evaluada en cualquier punto x .

El término $\varepsilon = Y - f(x)$ es el error **irreducible**,

Definir la función de pérdida. Usar probabilidad condicional. Y evaluar sólo $\mathbb{E}_{Y|X}$ en lugar de $\mathbb{E}_X \mathbb{E}_{Y|X}$.

2.1. Descomposición del error

Para cualquier estimador $\hat{f}(x)$ de $f(x)$ tenemos

$$\mathbb{E}[(Y - \hat{f}(x))^2 | X = x] = \underbrace{[f(x) - \hat{f}(x)]^2}_{\text{reducible}} + \underbrace{\mathbb{V}(\varepsilon)}_{\text{irreducible}}. \quad (5)$$

3. IMPORTANTE

Hasta ahora sólo hemos hablado de un procedimiento **predictivo**.

No hemos hablado de un procedimiento de **inferencia estadística**. Por ejemplo,

- Qué predictores están asociados con la respuesta?
- Qué tipo de relación tiene cada predictor con la respuesta?
- Se puede resumir la relación de manera lineal?

4. ¿CÓMO ESTIMAMOS F ?

- Tenemos datos observados para la x que nos interesa?
- Podemos calcular $\mathbb{E}[Y | X = x]$?
- Qué tal que relajamos:

$$\hat{f}(x) = \text{Promedio}(Y | X \in \mathcal{N}(x)). \quad (6)$$

4.1. Vecinos cercanos

Ejemplo de un modelo **no paramétrico**. Este modelo es bueno cuando p es pequeño y n es grande. Puede sobre-ajustar rápidamente.

4.2. Maldición de la dimensionalidad

Los vecinos... no son tan cercanos en dimensiones moderadas/altas.

5. MODELOS PARAMÉTRICOS

El modelo **lineal** es un modelo paramétrico de la forma

$$f_L(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \quad (7)$$

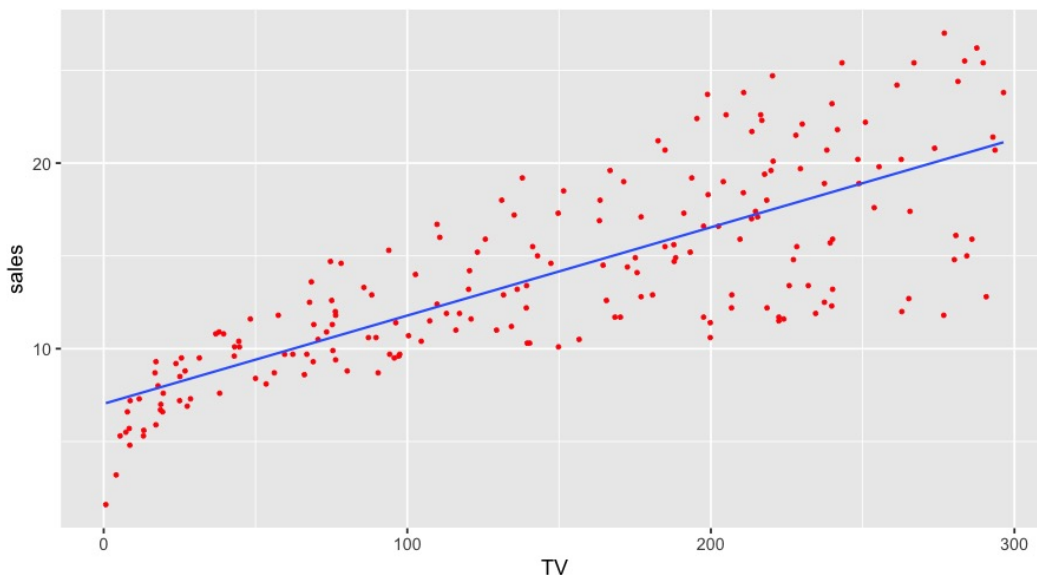


FIGURA 2. Ajuste lineal.

5.1. Compromisos

- El modelo lineal es "fácil" de interpretar. Sin embargo, puede no tener un buen desempeño.
- Hay un balance entre un *buen* ajuste y sobre(sub)-ajuste.
- Complejidad vs Simplicidad

Dificultad de interpretación cuando hay datos observacionales.

6. EVALUANDO LA PRECISIÓN DEL MODELO

Supongamos que entrenamos un modelo $\hat{f}(x)$ sobre \mathcal{D}_n . ¿Cómo evaluamos su desempeño bajo el conjunto que se utilizó para entrenar?

Función de pérdida / Error de entrenamiento / Error de prueba.

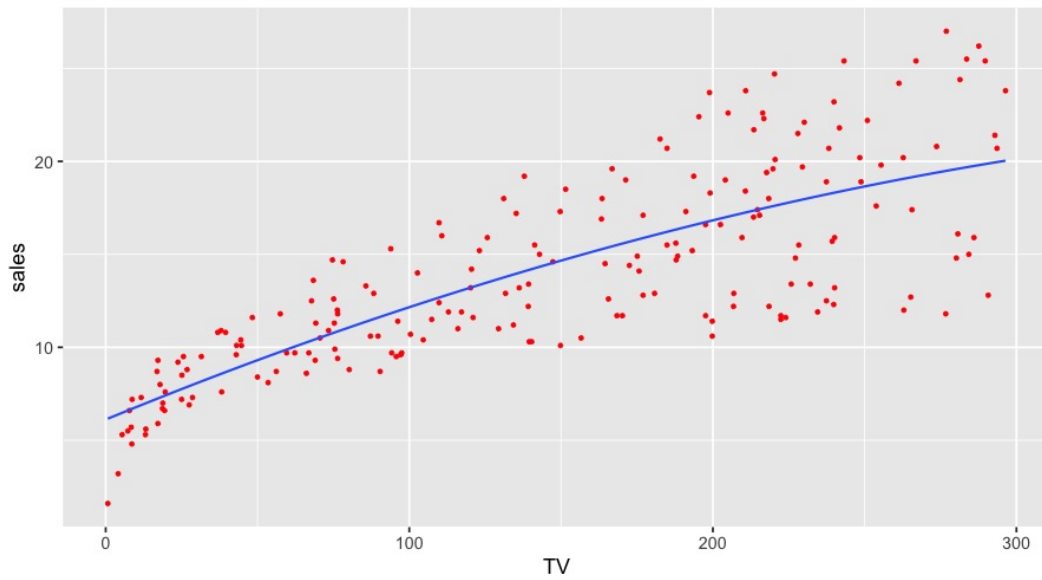


FIGURA 3. *Ajuste cuadrático.*

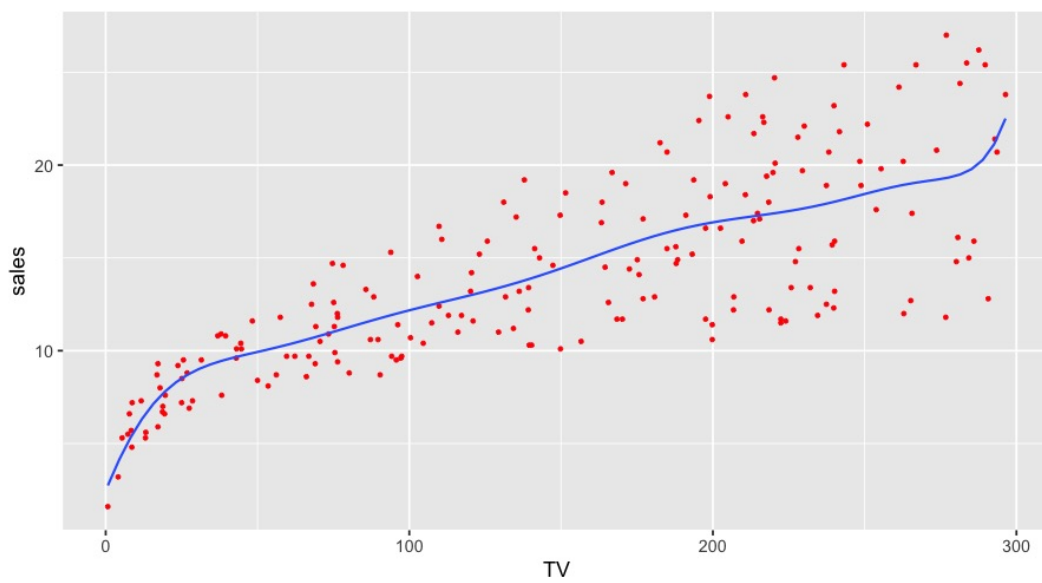


FIGURA 4. *Ajuste polinomial.*

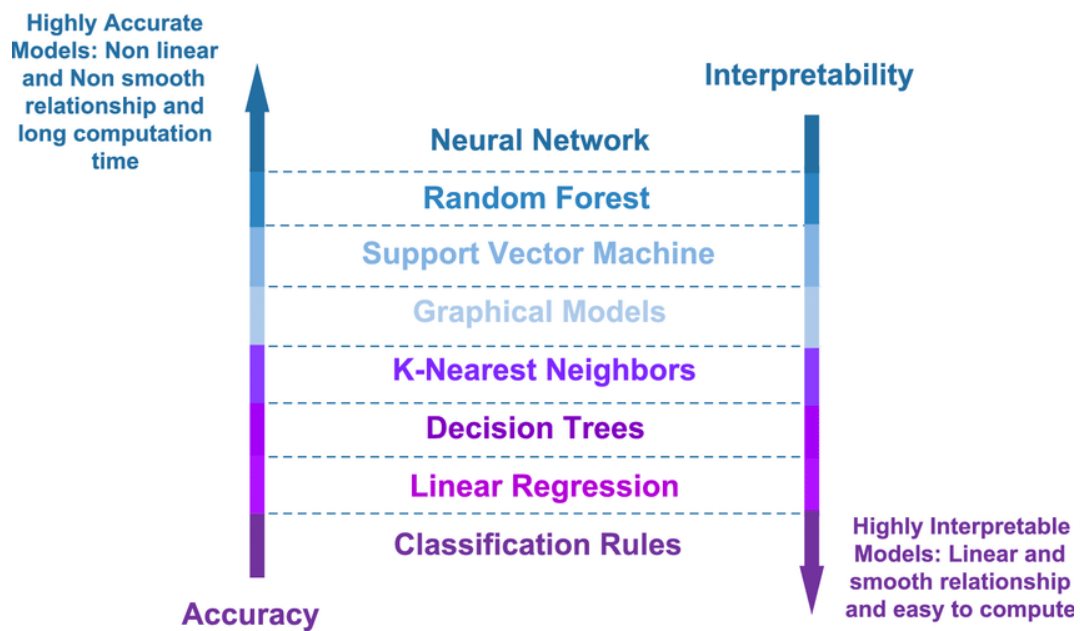


FIGURA 5. Tomado de [1]

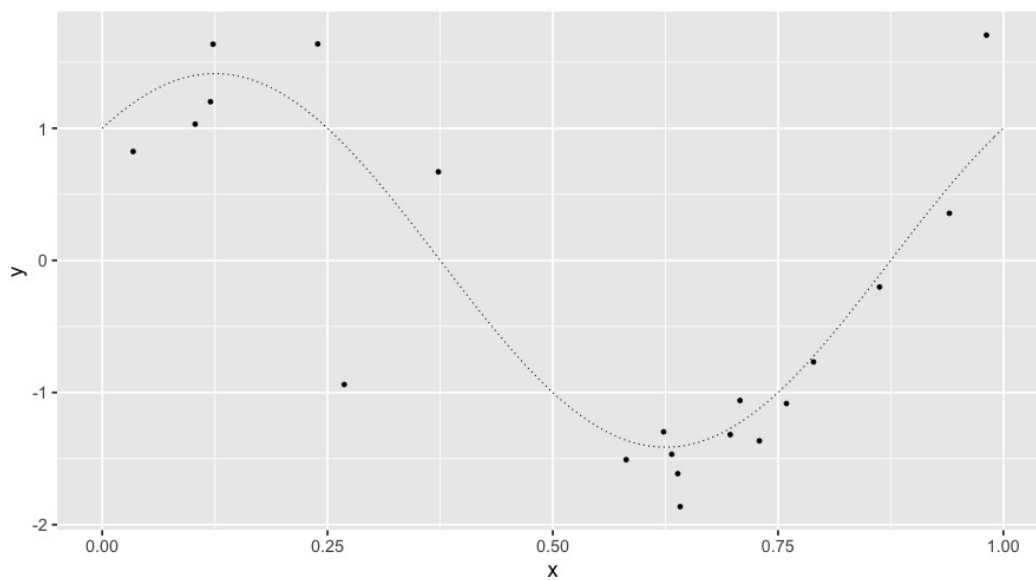


FIGURA 6. Función latente y observaciones.

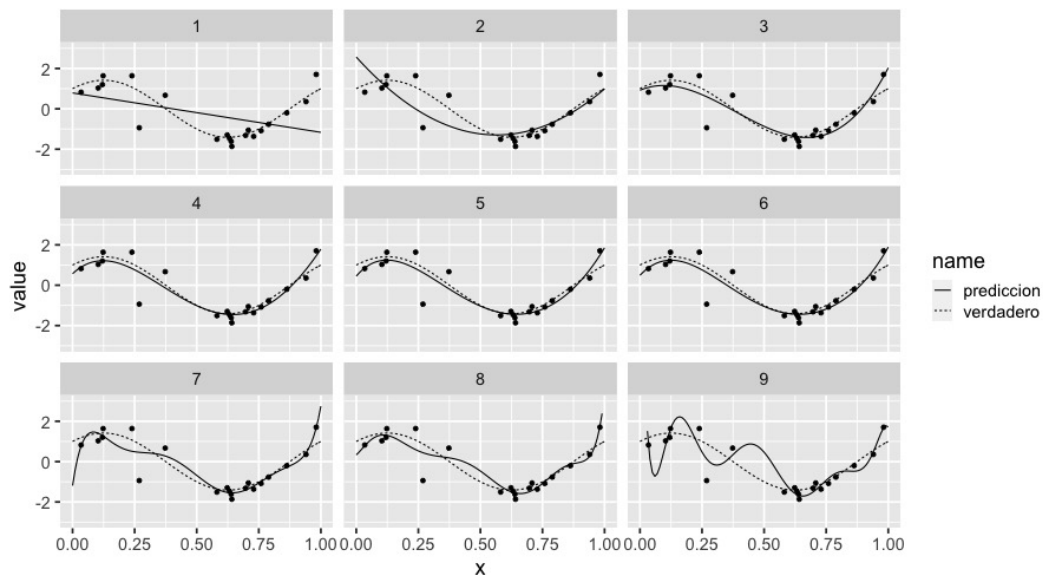


FIGURA 7. Ajuste bajo distintos grados del polinomio

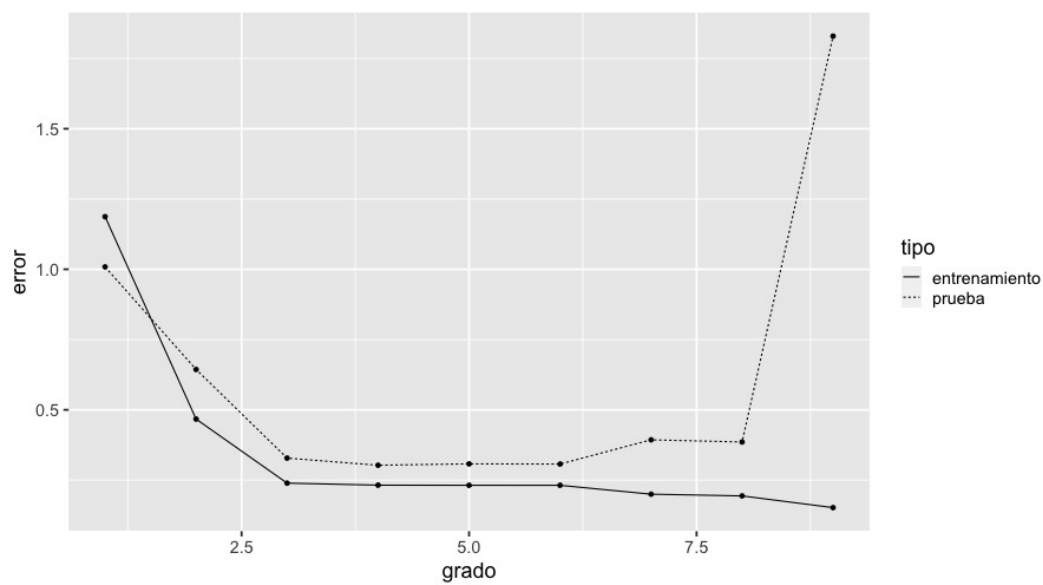


FIGURA 8. Errores de entrenamiento / prueba

6.1. Ejemplo (Regresión)

7. COMPROMISO ENTRE SESGO Y VARIANZA

Supongamos que ajustamos un modelo $\hat{f}(x)$ a un conjunto de datos \mathcal{D}_n . Sea (x_0, y_0) un punto no utilizado en el conjunto de entrenamiento. Si el modelo es $Y = f(X) + \varepsilon$. Entonces

$$\mathbb{E}[(y_0 - \hat{f}(x_0))^2] = \mathbb{V}(\hat{f}(x_0)) + [\text{Sesgo}(\hat{f}(x_0))]^2 + \mathbb{V}(\varepsilon). \quad (8)$$

Valor esperado. Definición de Sesgo. Figura descomposición.

8. PROBLEMAS DE CLASIFICACIÓN

La predicción es sobre una y_n que es cualitativa. Nos interesa el **error de clasificación**.

Definir función de pérdida. Es decir nos interesa

$$\frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i). \quad (9)$$

8.1. Objetivos

- Construir un clasificador $C(X)$.
- Medir la incertidumbre en la clase.
- Entender los roles de los predictores.

8.2. El clasificador óptimo

Supongamos que hay K clases en \mathcal{C} las cuales están numeradas. Sea

$$p_k(x) = \mathbb{P}(Y = k | X = x) \quad k = 1, \dots, K. \quad (10)$$

El clasificador óptimo Bayesiano es

$$C(x) = j \text{ si } p_j(x) = \max\{p_1(x), \dots, p_K(x)\}. \quad (11)$$

Prueba de optimalidad. Es el clasificador con menor error en la población. Se puede utilizar un modelo de **vecinos mas cercanos**.

REFERENCIAS

- [1] H. Fourati, R. Maaloul, and L. Chaari. A survey of 5G network systems: Challenges and machine learning approaches. *International Journal of Machine Learning and Cybernetics*, 12(2):385–431, feb 2021. ISSN 1868-808X. . 5
- [2] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. Springer US, New York, NY, 2021. ISBN 978-1-07-161417-4 978-1-07-161418-1. . 1