Ejercicios: sesión 2

Minería de datos II-Universidad Francisco de Vitoria

Pablo Hidalgo (pablo.hidalgo@ufv.es)

Ejercicio 1

La tabla siguiente recoge un conjunto de datos con **seis observaciones**, **tres predictores** (X_1, X_2, X_3) y **una variable respuesta** cuantitativa (Y).

Obs.	X_1	X_2	X_3	Y
1	0	3	0	1
2	2	0	0	1
3	0	4	3	1
4	0	5	12	0
5	-1	0	0	0
6	1	1	0	1

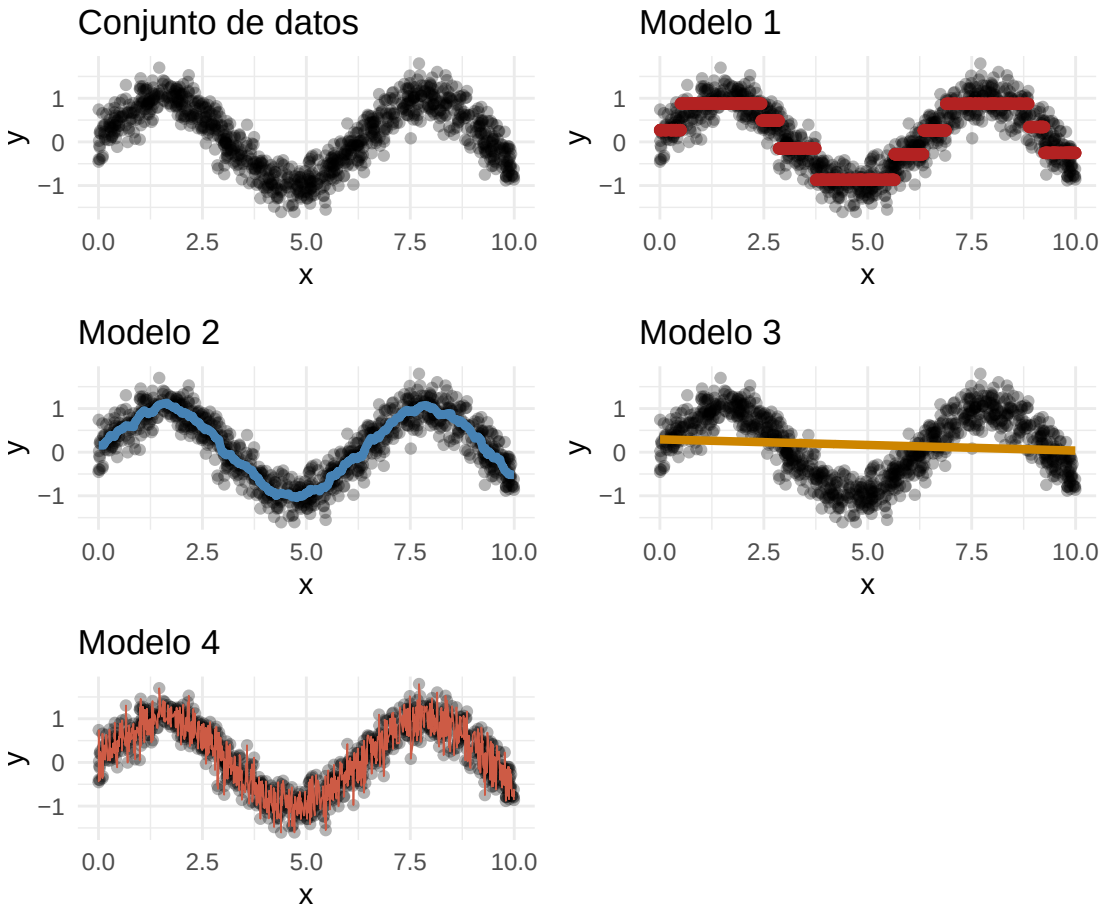
Queremos usar este conjunto de datos para hacer una predicción para Y cuando $X_1 = X_2 = X_3 = 0$ usando el modelo de k vecinos más cercanos (KNN).

(a) Calcula la distancia euclídea entre cada observación y el punto $X_1 = X_2 = X_3 = 0$.
(b) ¿Cuál es la predicción para $k = 1$?¿Por qué?
(c) ¿Cuál es la predicción para $k = 2$?¿Y para $k = 3$?¿Por qué?

Ejercicio 2

Supongamos que tenemos un conjunto de datos con dos variables: X la variable predictora e Y la variable que queremos predecir. Sobre este conjunto de datos se han ajustado **cuatro modelos**: regresión lineal, árbol de regresión y dos modelos KNN con un número k de vecinos distintos.

Al tratarse de un problema de dos dimensiones, se puede representar de forma gráfica fácilmente.



El gráfico de la esquina superior izquierda muestra los datos originales; en el resto de gráficos se ha representado en color la predicción dada por un modelo entrenado con esos datos.

a) Justifica con qué gráfico se corresponde cada modelo.
b) Hemos obtenido los RMSE en entrenamiento: 0.7185, 0.3083, 0.2807 y 0. ¿Con qué modelo se corresponde cada uno y por qué?
c) ¿Identificas *overfitting* o *underfitting* en algún modelo?