

REGRESIÓN KNN, LINEAL Y POLINOMIAL REGULARIZACIÓN LASSO Y RIDGE

CLASE 20

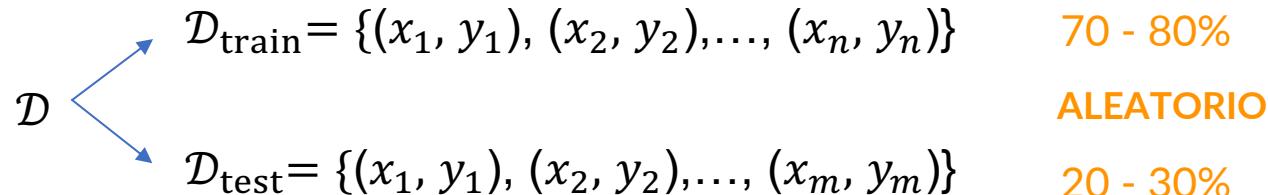
ALGORITMOS DE APRENDIZAJE DE MÁQUINA

- **Algoritmos** → procedimiento, método o conjunto de pasos o reglas para lograr una tarea.
- **Aprendizaje** → proceso de convertir experiencia (i.e. data) en habilidad o conocimiento (i.e. un modelo o programa capaz de realizar una tarea)
- **Algoritmos de aprendizaje de máquina (ML)** → métodos computacionales que utilizan data anterior (i.e. experiencia) para generar modelos o programas capaces de realizar tareas como predecir, clasificar, agrupar, ordenar o reducir dimensionalidad.
- Para una tarea dada, pueden proponerse múltiples algoritmos posibles
 - El éxito de un algoritmo de ML se evalúa en base a métricas de precisión, eficiencia y tiempo computacional.
 - La elección del algoritmo a usar dependerá de: contexto y complejidad del problema, suposiciones de base, tamaño y variedad de la data disponible.
 - Implementarlo

APRENDIZAJE SUPERVISADO

APRENDIZAJE SUPERVISADO

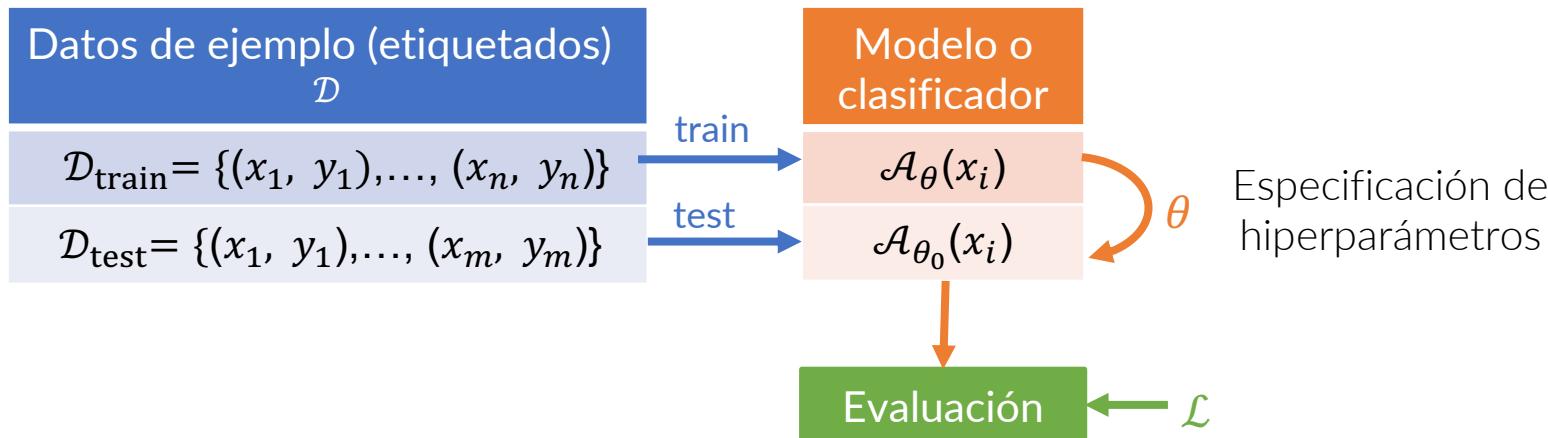
- El objetivo es realizar predicciones precisas para *nuevos datos* con características similares a los datos usados para construir el modelo → **generalización**
- **Entrenamiento y testeo:** el conjunto de datos de ejemplo \mathcal{D} se divide aleatoriamente en dos
 - **Datos de entrenamiento ($\mathcal{D}_{\text{train}}$)** → para entrenar el modelo
 - **Datos de prueba ($\mathcal{D}_{\text{test}}$)** → para evaluar qué tan bien funciona el modelo frente a datos que no conoce. No se usa para el entrenamiento.
 - **Etiquetas (y)** → valores o categorías asignadas a los datos de ejemplo



- **Función de pérdida (\mathcal{L})** → función que mide la diferencia o pérdida entre la predicción y el valor real de una etiqueta o valor.

APRENDIZAJE SUPERVISADO

- El objetivo es realizar predicciones precisas para *nuevos datos* con características similares a los datos usados para construir el modelo → **generalización**
- **Entrenamiento y testeo:**
 - **Hiperparámetros (θ)** → parámetros libres del modelo que no son determinados por el algoritmo, sino entregados como input



APRENDIZAJE SUPERVISADO

MODELOS DE REGRESIÓN

REGRESIÓN EN sklearn

<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html>

https://scikit-learn.org/stable/modules/linear_model.html

Procedimiento general:

- Dividir datos en conjuntos train/test
- Definir modelo o estimador (regresión, u otro)
- Fit → predict
- Evaluar métricas (score, R², etc)

Methods

<code>fit(X, y[, sample_weight])</code>	Fit linear model.
<code>get_params([deep])</code>	Get parameters for this estimator.
<code>predict(X)</code>	Predict using the linear model.
<code>score(X, y[, sample_weight])</code>	Return the coefficient of determination of the prediction.
<code>set_params(**params)</code>	Set the parameters of this estimator.

`fit(X, y, sample_weight=None)`

Fit linear model.

Parameters:	X : {array-like, sparse matrix} of shape (n_samples, n_features) Training data.
y : array-like of shape (n_samples,) or (n_samples, n_targets)	Target values. Will be cast to X's dtype if necessary.
sample_weight : array-like of shape (n_samples,), default=None	Individual weights for each sample.
Returns:	self : object Fitted Estimator.

PREDICTORES Y OUTCOMES

- En un problema de regresión, buscamos predecir el valor de una variable a partir del valor de otras variables.
- Ejemplo:
 - Predecir el consumo de combustible de un auto a partir de sus características de diseño.

p predictores : j=1,2,...,p

n observaciones i=1,2,...n	car_name	mpg	X_i									
			cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
0	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
1	Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
2	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
3	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
4	Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2

$Y = y_1, \dots, y_n$

outcome / variable dependiente
respuesta

$X = X_1, \dots, X_p$

$X_j = x_{1j}, \dots, x_{ij}, \dots, x_{nj}$

predictores /variable independiente
covariates / features

REGRESIÓN

- Para predecir Y , asumimos que se relaciona con X mediante una función desconocida f .
- Problema de inferencia \Rightarrow encontrar \hat{f} , la estimación de f
 - Ej: análisis exploratorio
 - ¿cuáles son las variables que determinan Y ?
 - ¿cuál es la contribución de cada una de ellas?
 - ¿qué función analítica puede representar a f ?
¿lineal?
- Problema de **predicción** \Rightarrow usar la data de entrenamiento para predecir o generalizar a situaciones no observadas.
 - No nos interesa la forma de f , sino sólo las **predicciones** \hat{y}_i



$$Y = f(X)$$

$f \rightarrow$ función desconocida
 $\hat{f} \rightarrow$ estimación de f

Predicciones $\rightarrow \hat{y}_i = \hat{f}(x_{i,1}, \dots, x_{i,p})$

REGRESIÓN - kNN

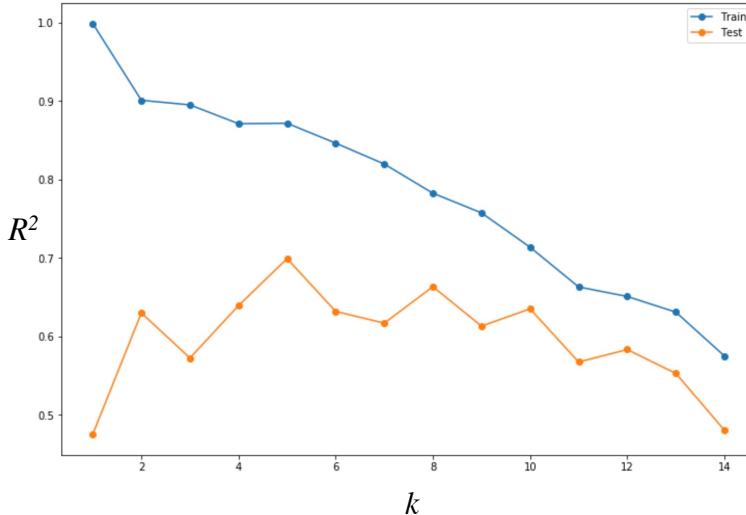
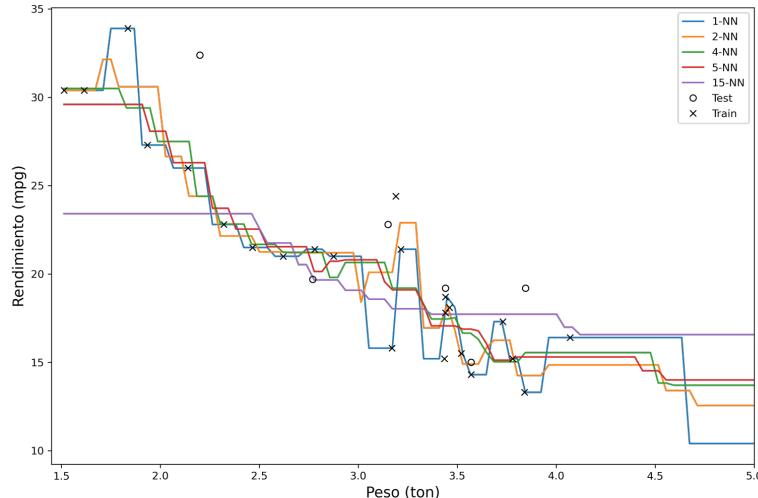
¿Cuál es la forma más simple de predecir \hat{y}_i ?

Usamos el promedio de las respuestas a otras observaciones más cercanas a ella

→ los “k vecinos más cercanos” o **k-nearest neighbors**:

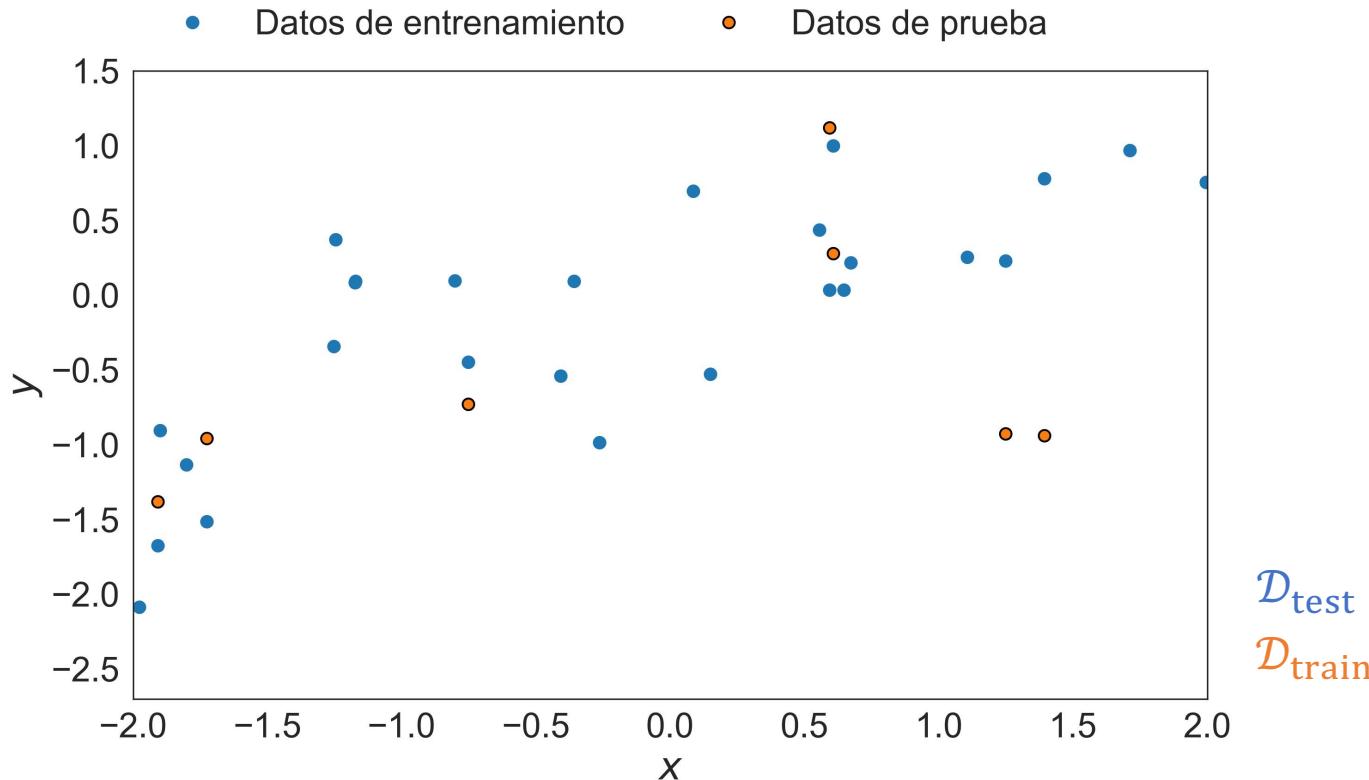
$$\hat{y}_i = \frac{1}{k} \sum_{j=1}^k y_{ij}$$

Donde $\{x_{i1}, \dots, x_{ik}\}$ son las k observaciones más similares (cercanas) a x_i



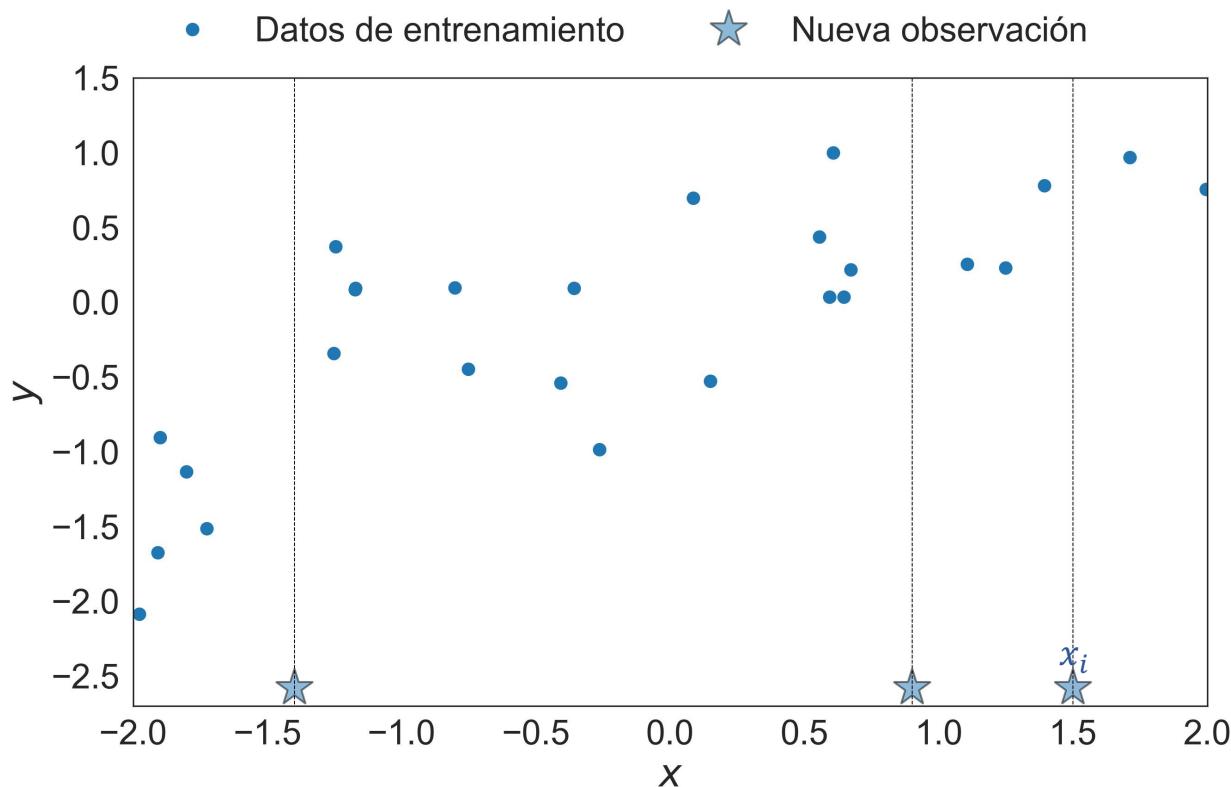
REGRESIÓN kNN

Datos etiquetados (\mathcal{D})



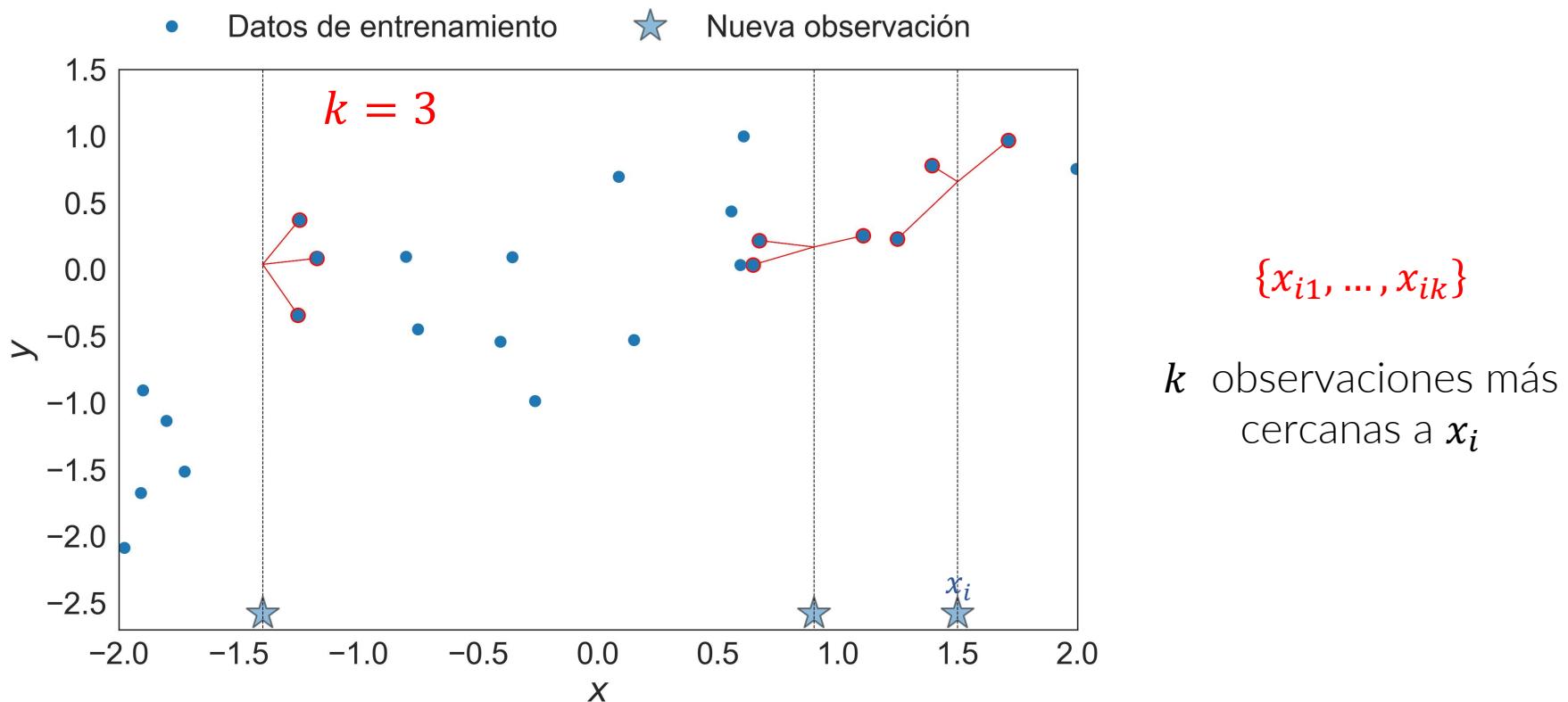
REGRESIÓN kNN

¿Cómo estimar y_i para una nueva observación x_i ?



REGRESIÓN kNN

Identificar los k vecinos más cercanos a cada valor de x



REGRESIÓN kNN

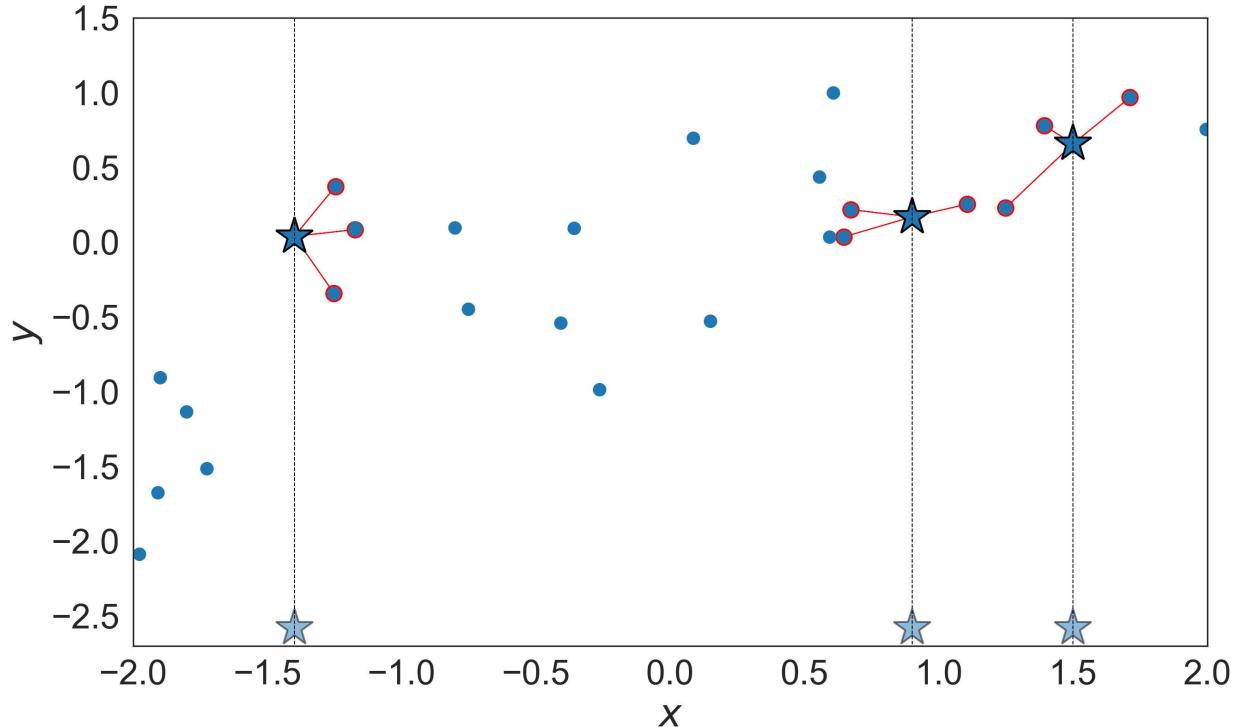
- Datos de entrenamiento



- Nueva observación



- Predicción

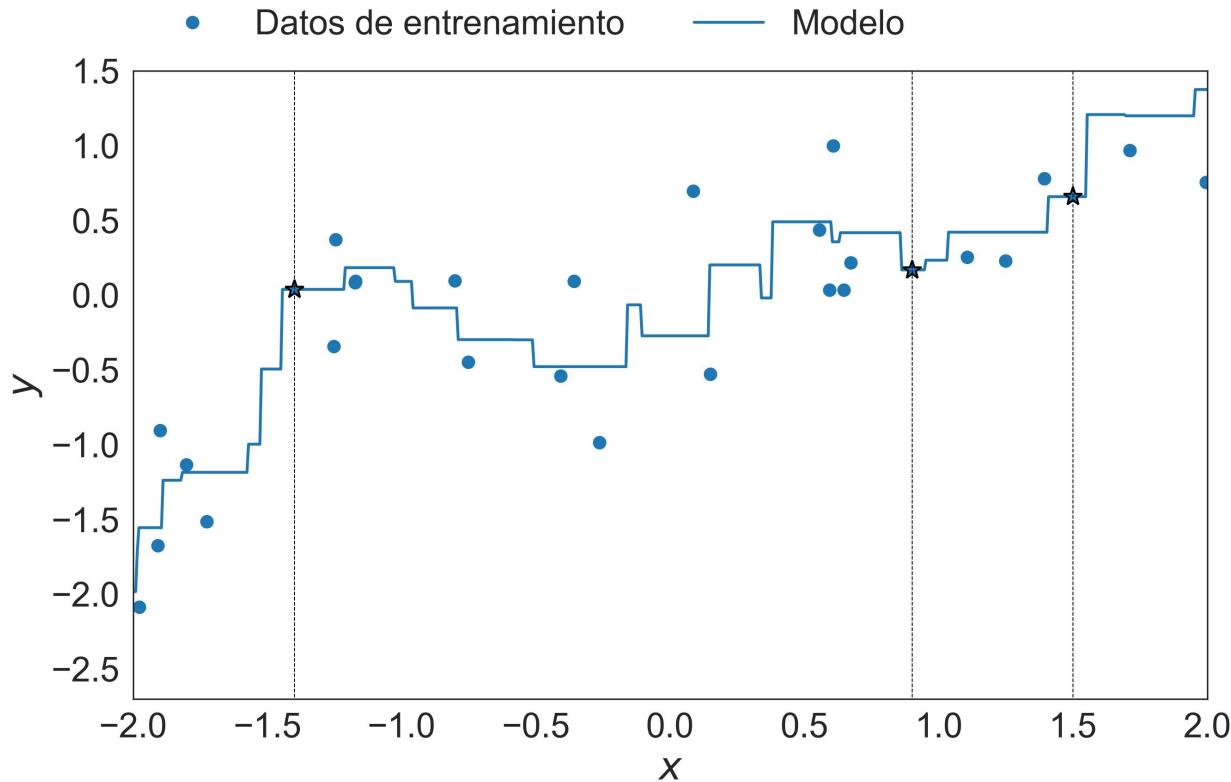


Predicción:

$$\hat{y}_i = \frac{1}{k} \sum_{j=1}^k y_{i_j}$$

REGRESIÓN kNN

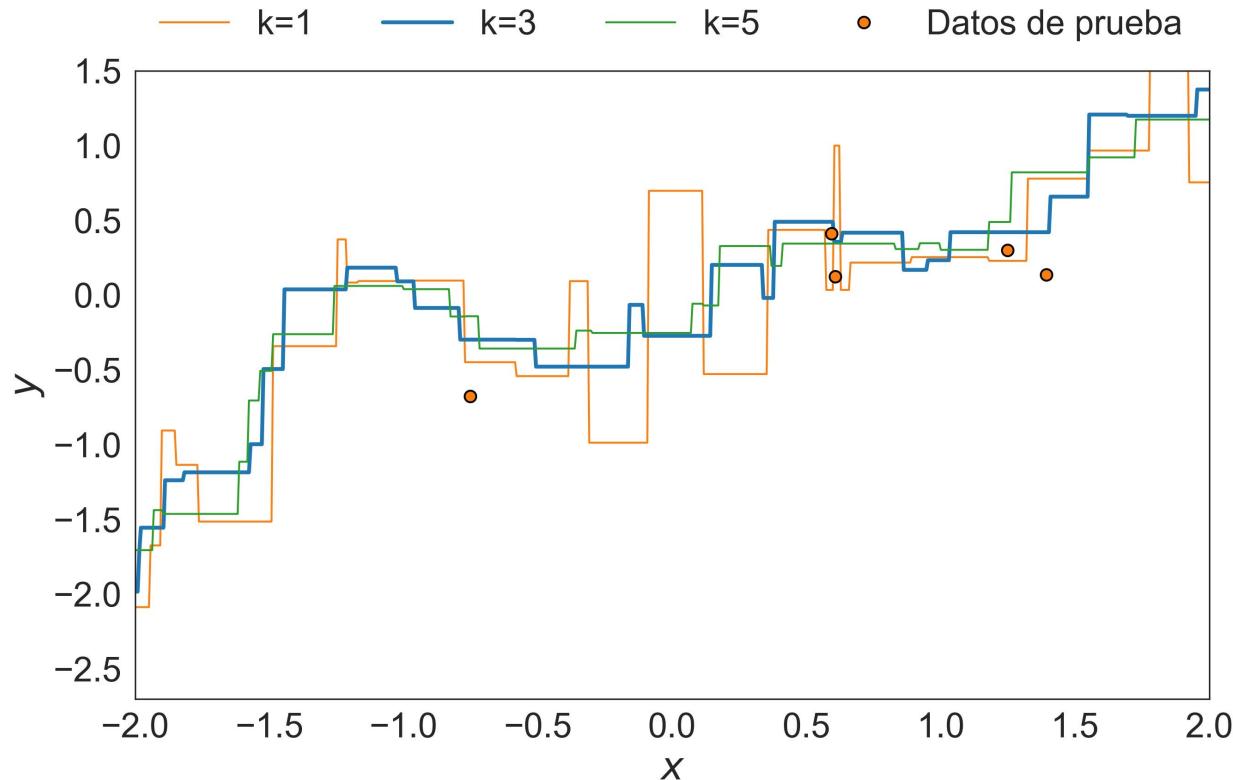
Modelo entrenado para $k = 3$



¿Qué pasa si cambiamos k ?

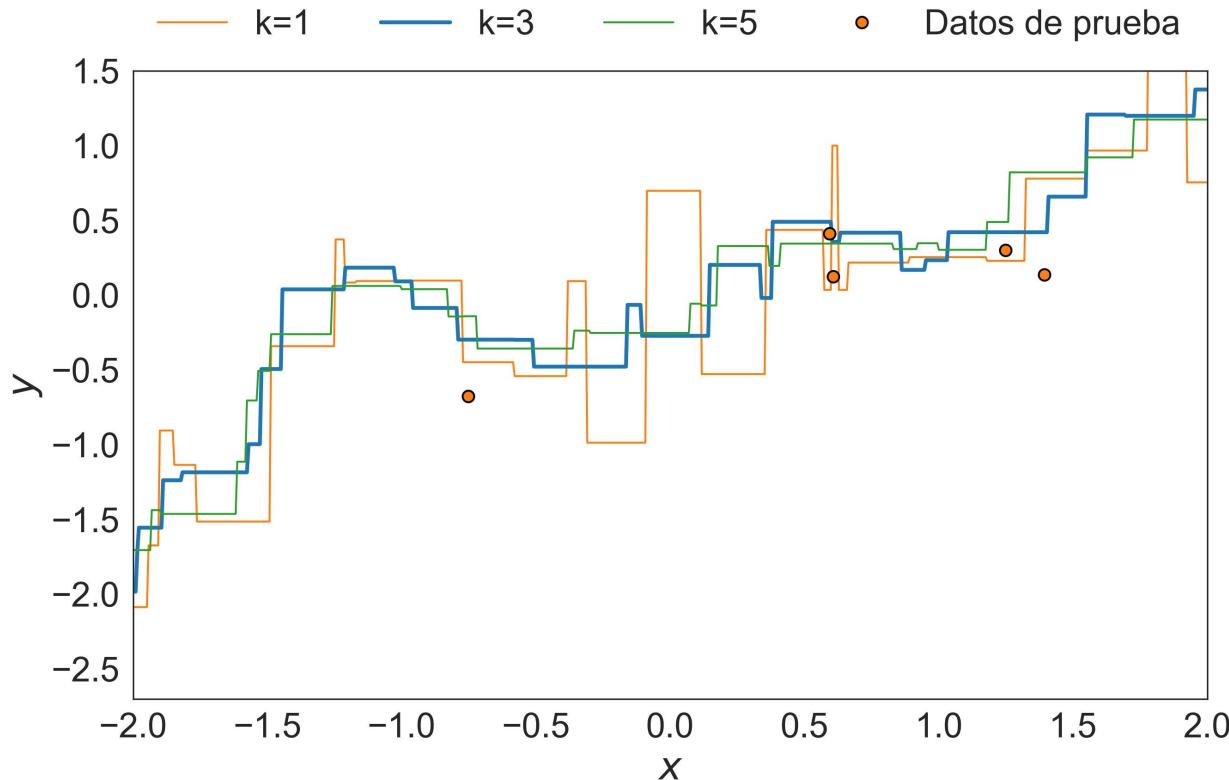
REGRESIÓN kNN

Ajuste de hiperparámetro: ¿cuál es la mejor elección de k ?



REGRESIÓN kNN

Ajuste de hiperparámetro: ¿cuál es la mejor elección de k ?



Funciones de pérdida:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

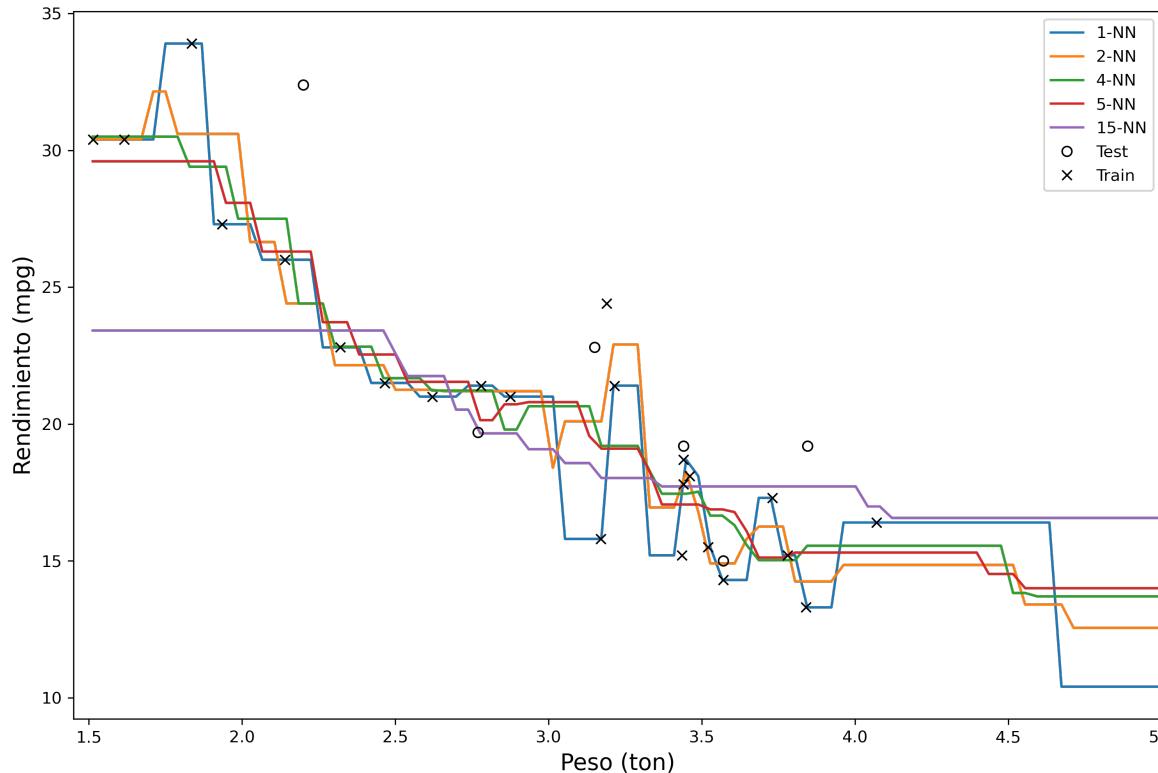
$$\text{RMSE} = \sqrt{\text{MSE}}$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$\Rightarrow k = 3$

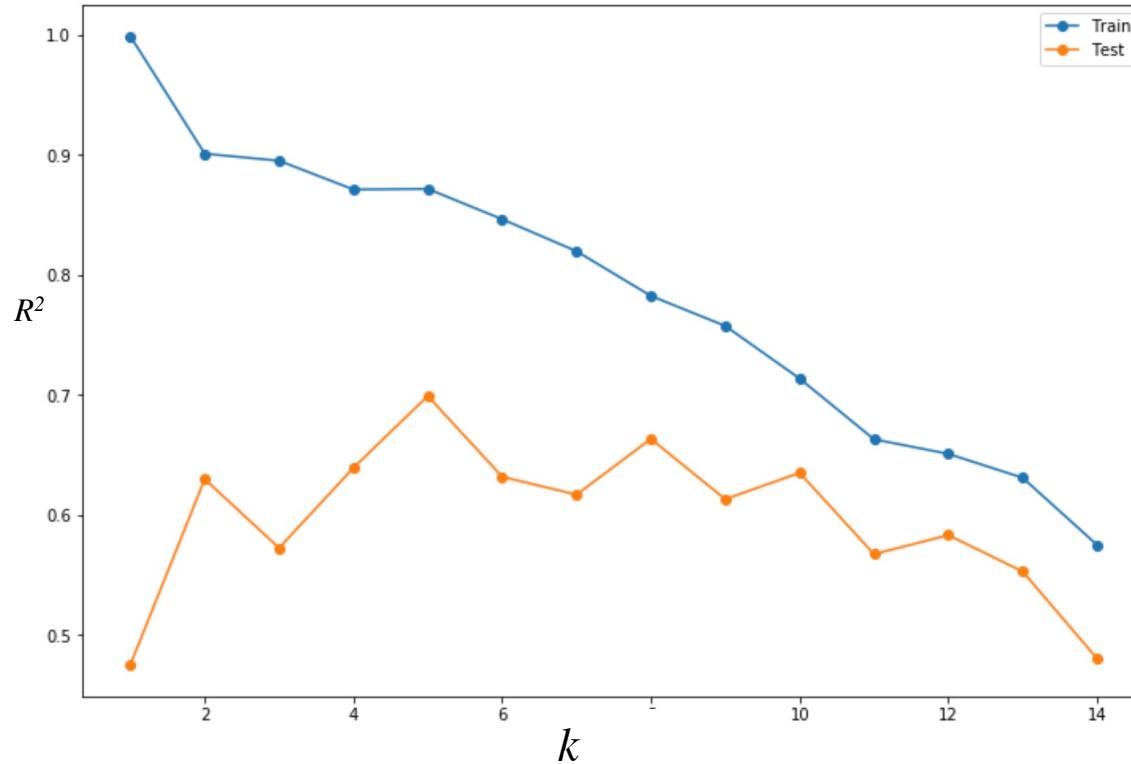
REGRESIÓN kNN

Ejemplo: predicción de rendimiento de vehículos



REGRESIÓN kNN

Ejemplo: predicción de rendimiento de vehículos



RREGRESIÓN kNN: NORMALIZACIÓN

Si hay múltiples predictores: se define una medida de distancia multidimensional para identificar las observaciones más similares o “vecinos”.

- Distancia euclídea: $D(\mathbf{x}_i, \mathbf{x}_0) = \sqrt{\sum_{j=1}^P (x_{i,j} - x_{0,j})^2}$
- Si los predictores tienen diferentes escalas y variabilidad → se introducen **efectos de escala** en la medición de distancia.
- Por lo tanto, **para $p > 1$, es necesario estandarizar los predictores.**
- **Normalización z:** se resta la media, y se divide por la desviación estándar.

$$\rightarrow x_{scaled} = \frac{x - \mu}{\sigma}$$

REGRESIÓN LINEAL

- En la regresión kNN, obtenemos predicciones pero no la forma de la función f .
 - ¿qué pasa con y si duplico el valor de x ?
- Otra opción, es construir un modelo asumiendo una forma simple para f :

$$Y = f(X) + \epsilon = \beta_0 + \beta_1 X + \epsilon$$

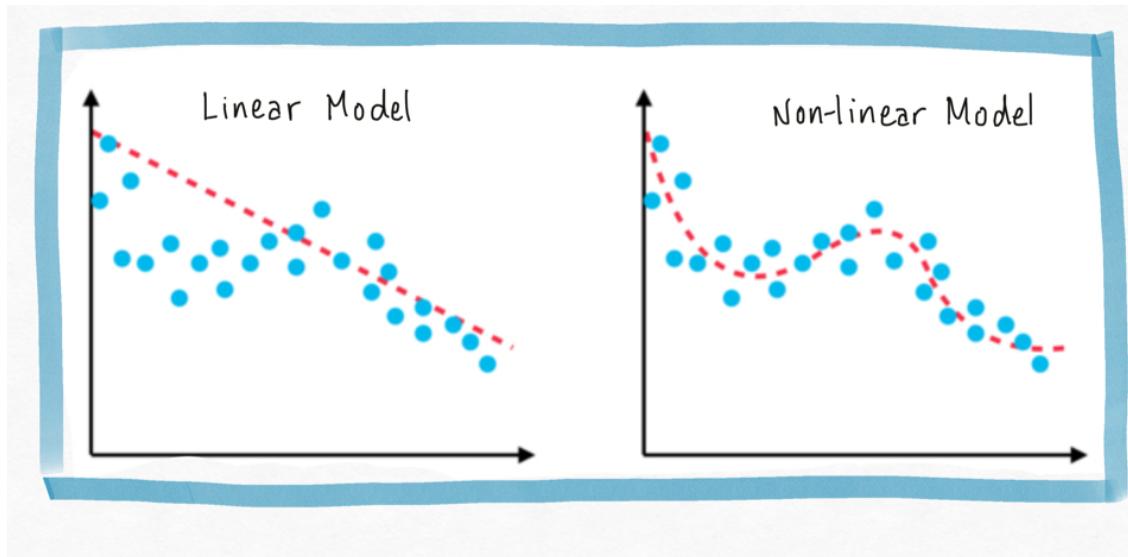
- ¿Cómo estimamos los coeficientes de regresión?

$$\mathcal{L}(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 X))^2$$

$$\hat{\beta}_0, \hat{\beta}_1 = \operatorname{argmin} \mathcal{L}(\beta_0, \beta_1) \Rightarrow \begin{aligned} \frac{\partial \mathcal{L}}{\partial \beta_0} &= 0 \\ \frac{\partial \mathcal{L}}{\partial \beta_1} &= 0 \end{aligned}$$

REGRESIÓN POLINOMIAL

- Los modelos lineales pueden entrenarse para conjuntos de datos con muchos predictores. Pero, la relación entre predictores y outcomes no siempre es lineal.
- En este caso, buscamos un modelo de la forma $Y = f_{\beta}(X)$, donde f es una función no-lineal y β es el vector de parámetros de f .



REGRESIÓN POLINOMIAL

- La forma no-lineal más simple que podemos considerar es un polinomio de grado M :

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_M x^M$$

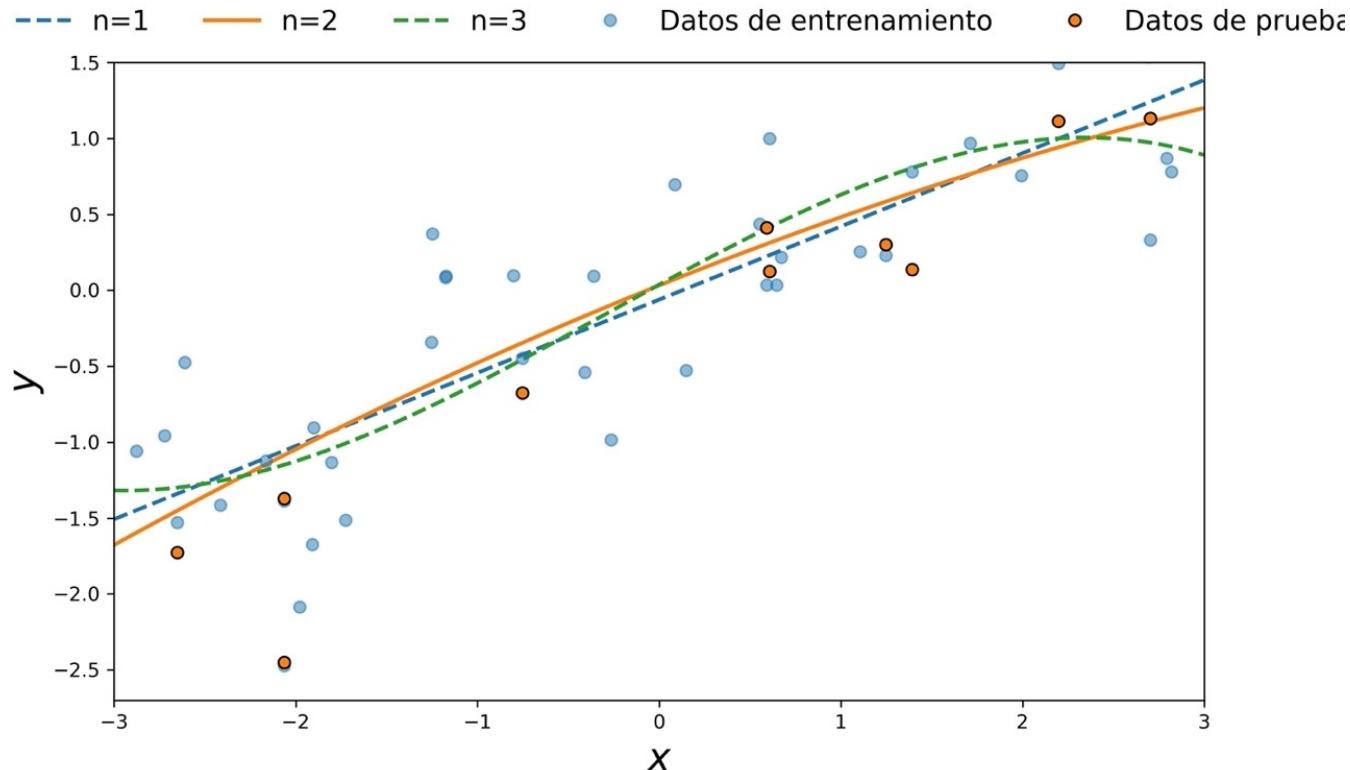
- En este caso, tratamos cada variable x^m como un nuevo predictor, y escribimos el problema como un modelo lineal de la forma:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1^1 & \dots & x_1^M \\ 1 & x_2^1 & \dots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^1 & \dots & x_n^M \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_M \end{pmatrix}.$$

- Una vez construidos los vectores (Y, X) , podemos resolver el problema de la misma forma que una regresión lineal.

$$\Rightarrow \mathbf{Y} = \boldsymbol{\beta} \mathbf{X}$$

REGRESIÓN LINEAL Y POLINOMIAL - EJEMPLO

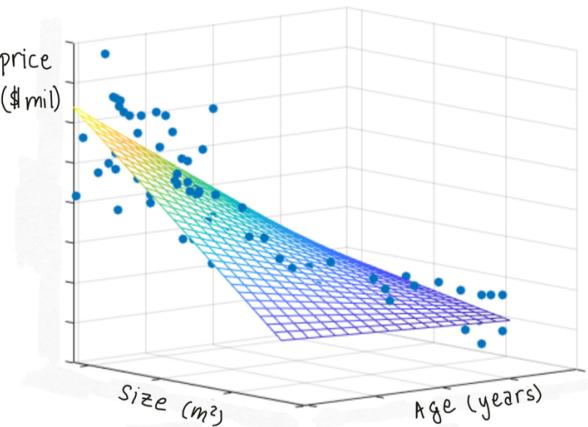


REGRESIÓN MULTILINEAL

- Usualmente, la variable de respuesta Y no depende sólo de una variable predictora, sino de un conjunto de ellas: $f(X_1, \dots, X_J)$.

$$Y = y_1, \dots, y_n, \quad X = X_1, \dots, X_J$$

- Por lo tanto, el modelo a entrenar tiene la forma: $f(X_1, \dots, X_J) = \beta_0 + \beta_1 X_1 + \dots + \beta_J X_J$

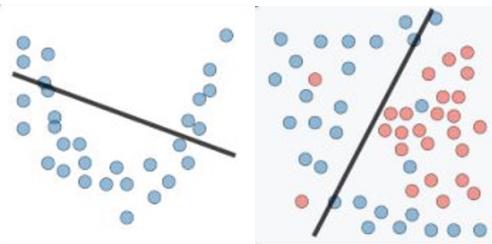


$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_y \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,J} \\ 1 & x_{2,1} & \dots & x_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,J} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_J \end{pmatrix},$$

$$\Rightarrow \mathbf{Y} = \boldsymbol{\beta}\mathbf{X}$$

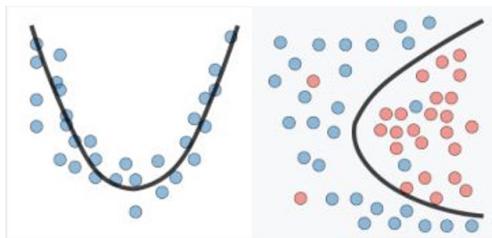
APRENDIZAJE SUPERVISADO: OVERFITTING

- Si el modelo se ajusta muy cercanamente a la data de entrenamiento, pero falla al generalizar o predecir la data de prueba → **overfitting**



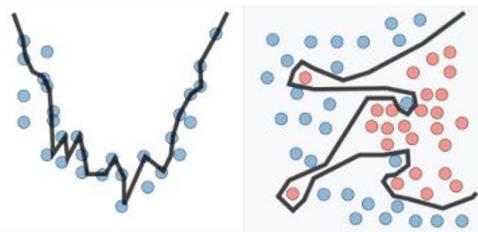
Underfitting

Alto error de entrenamiento
Error de prueba similar a
error de entrenamiento



Óptimo

Error de entrenamiento
levemente más bajos que
error de prueba



Overfitting

Muy bajo error de entrenamiento
Error de prueba mucho mayor a
error de entrenamiento

Regresión

Clasificación

APRENDIZAJE SUPERVISADO: OVERFITTING

- Si el modelo se ajusta muy cercanamente a la data de entrenamiento, pero falla al generalizar o predecir la data de prueba → **overfitting**
- Factores que influyen en overfitting:
 - **Complejidad del modelos (d)**
 - Demasiado simple → underfitting
 - Demasiado complejo → overfitting
 - **Nº de datos de entrenamiento (N)**
 - A mayor cantidad y variedad de datos, más complejo puede ser el modelo sin caer en overfitting
 - **Magnitud del ruido**

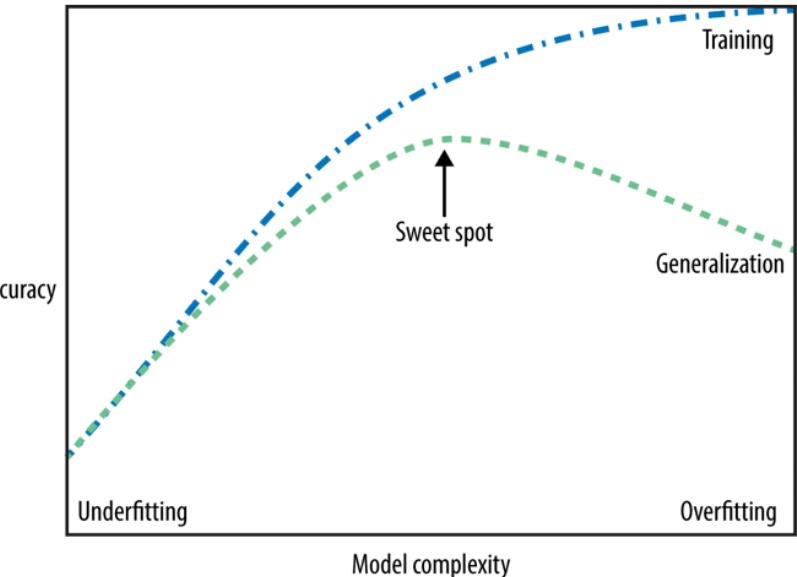
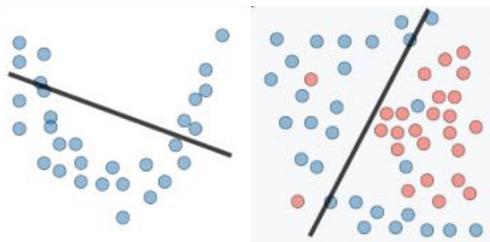


Figure 2-1. Trade-off of model complexity against training and test accuracy

APRENDIZAJE SUPERVISADO: OVERFITTING

- Si el modelo se ajusta muy cercanamente a la data de entrenamiento, pero falla al generalizar o predecir la data de prueba → **overfitting**

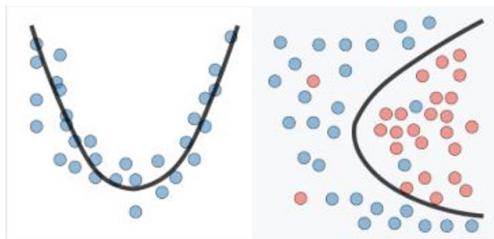


Underfitting

Alto error de entrenamiento
Error de prueba similar a
error de entrenamiento

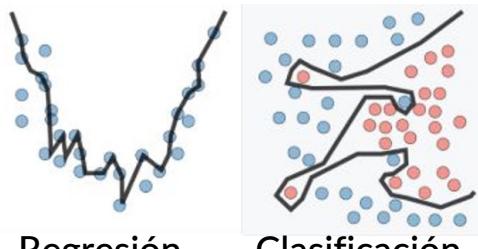


Complejizar modelo
Agregar features



Óptimo

Error de entrenamiento
levemente más bajos que
error de prueba



Overfitting

Muy bajo error de entrenamiento
Error de prueba mucho mayor a
error de entrenamiento



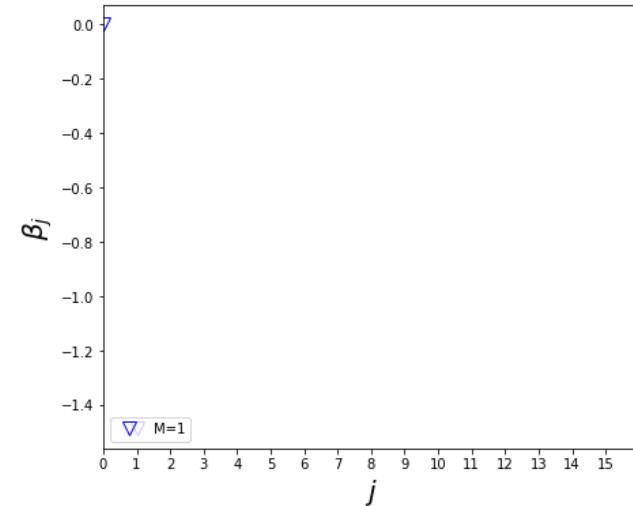
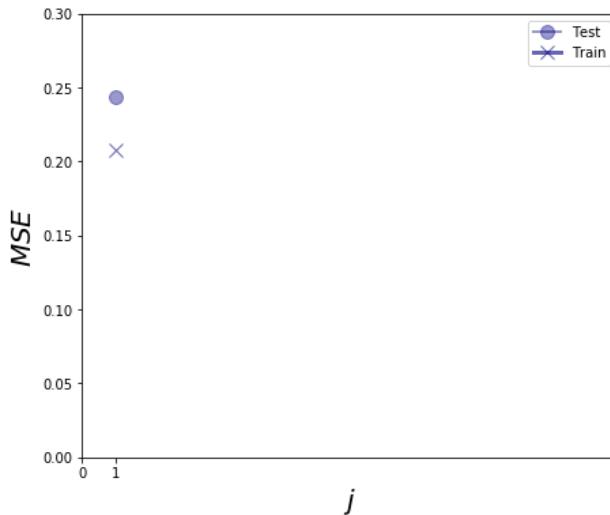
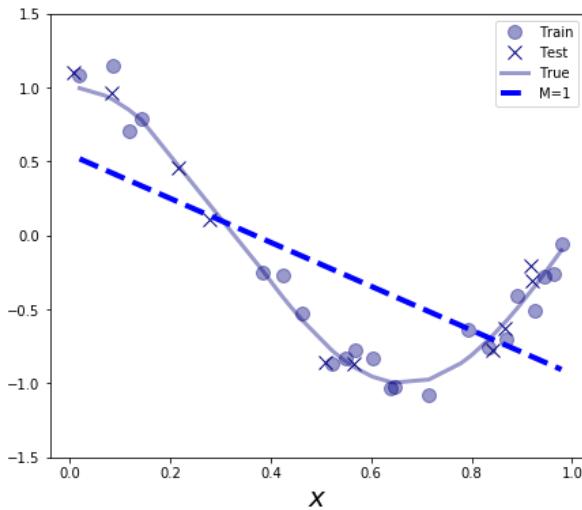
Simplificar modelo
Buscar más datos
Regularización

Regresión

Clasificación

REGRESIÓN Y OVERFITTING

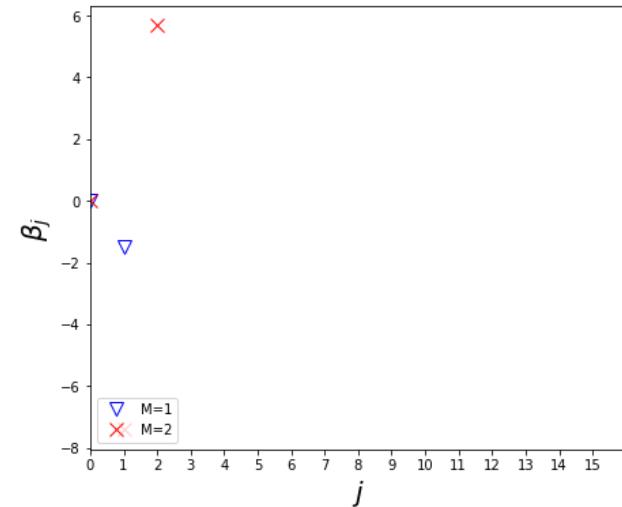
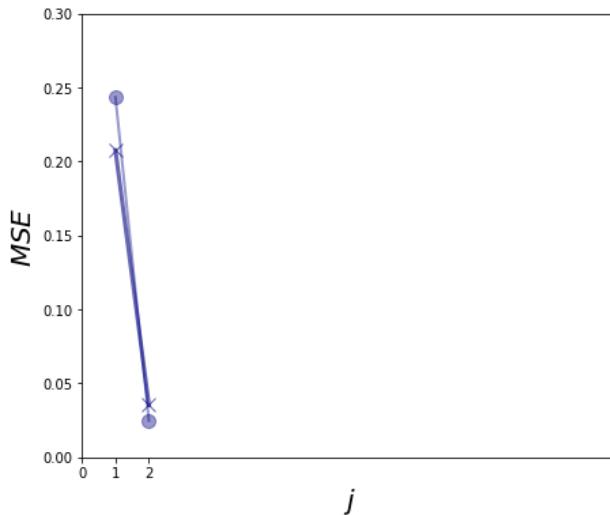
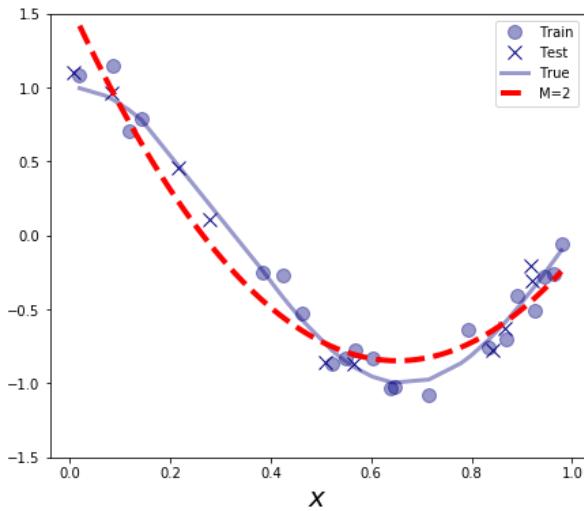
- Supongamos un modelo de regresión polinomial para un conjunto de $n=30$ datos



Ajuste lineal → underfitting

REGRESIÓN Y OVERFITTING

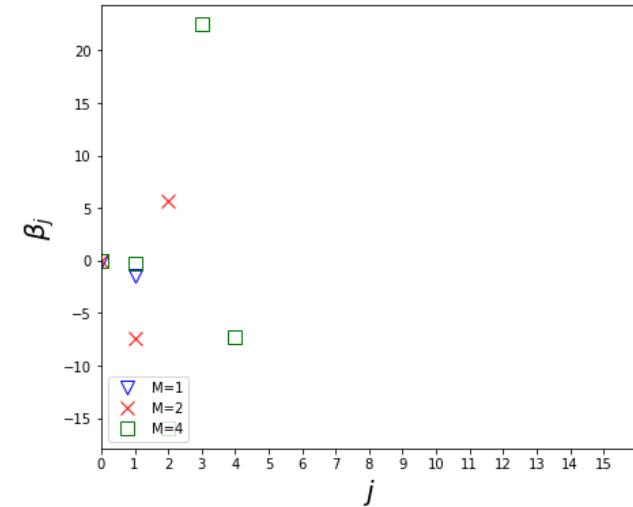
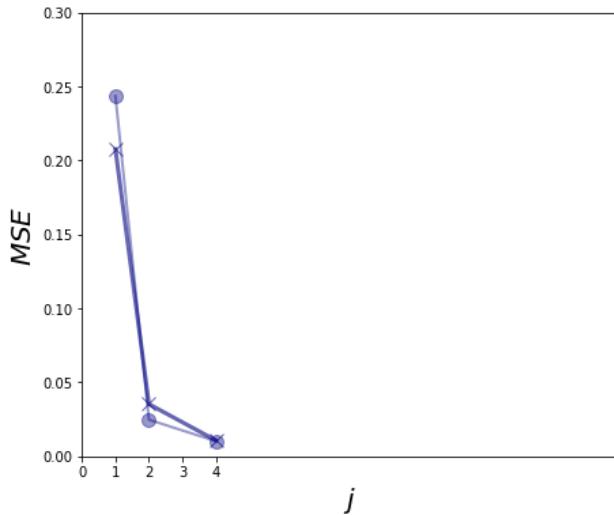
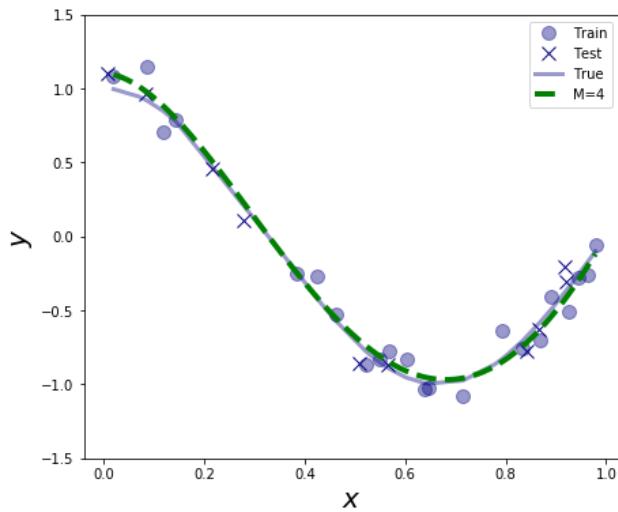
- Supongamos un modelo de regresión polinomial para un conjunto de $n=30$ datos



Disminuye el error 😊

REGRESIÓN Y OVERFITTING

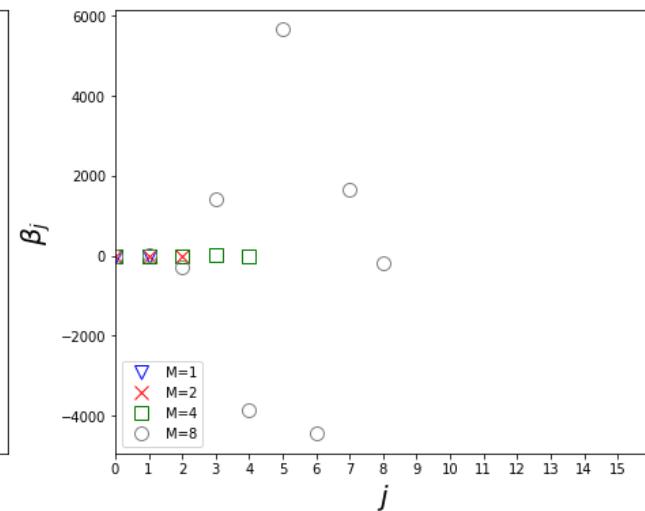
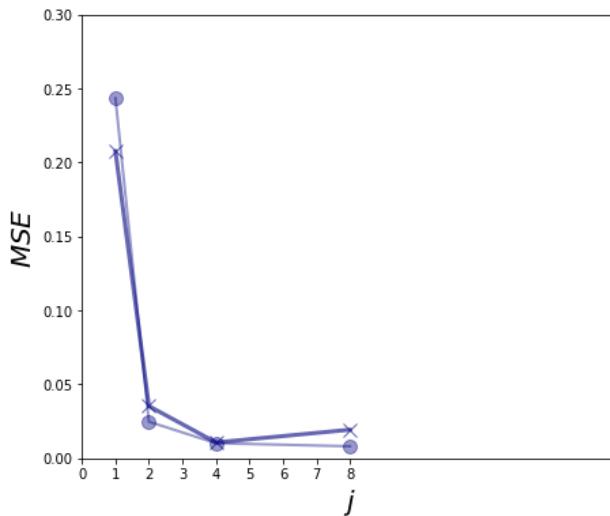
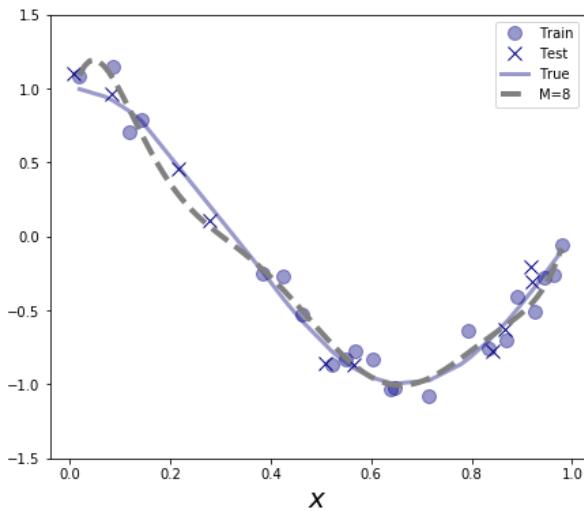
- Supongamos un modelo de regresión polinomial para un conjunto de $n=30$ datos



Disminuye el error 😊

REGRESIÓN Y OVERFITTING

- Supongamos un modelo de regresión polinomial para un conjunto de $n=30$ datos

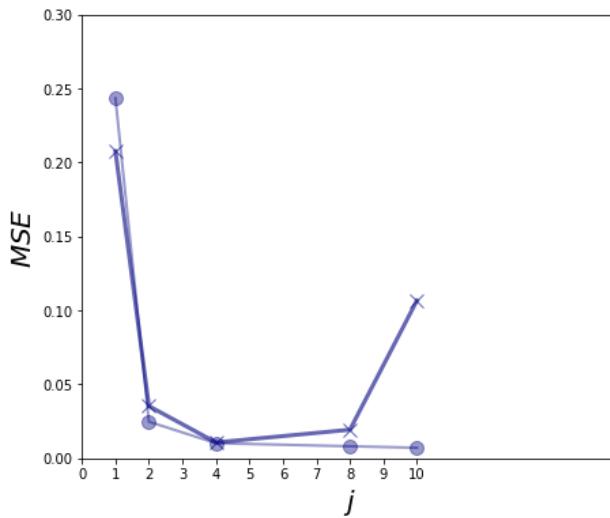
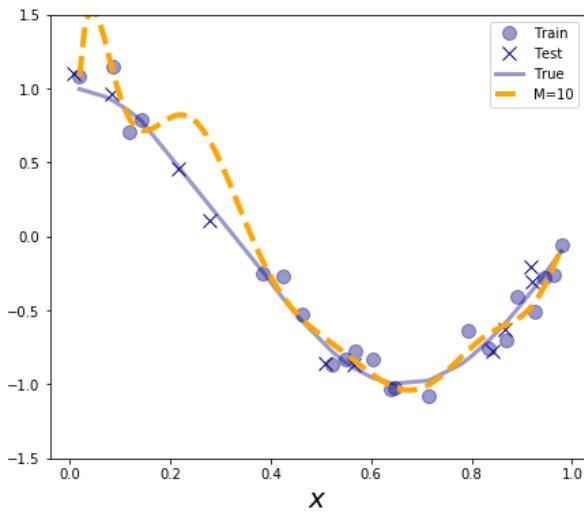


Aumenta el error de
validación → comienza el
overfitting... ($M=8$) 😞

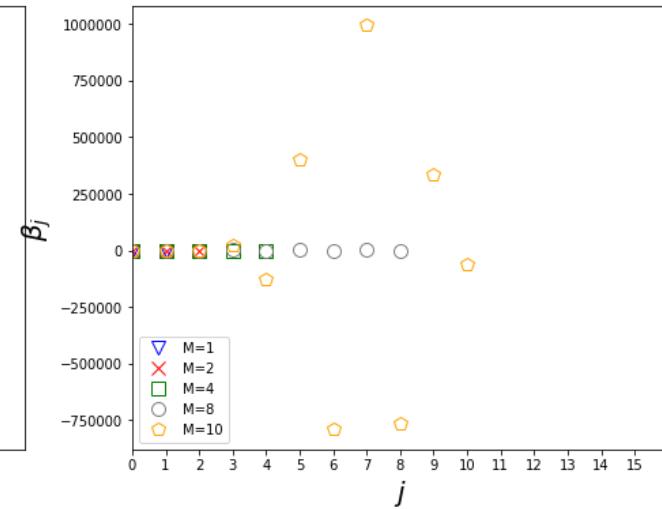
Valores de los
coeficientes se hacen
extremos 😞

REGRESIÓN Y OVERFITTING

- Supongamos un modelo de regresión polinomial para un conjunto de $n=30$ datos



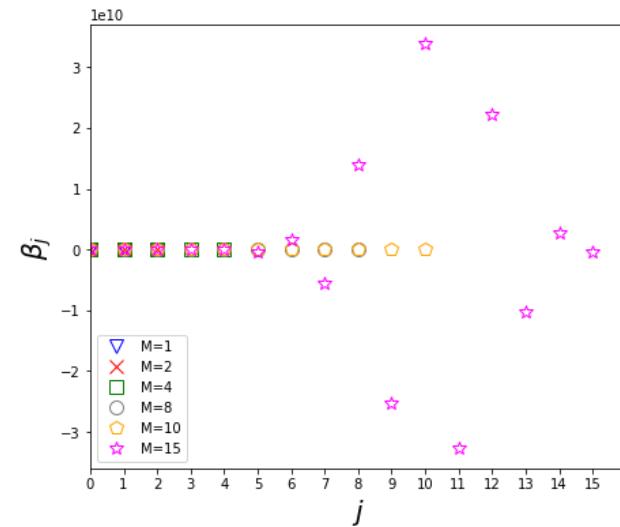
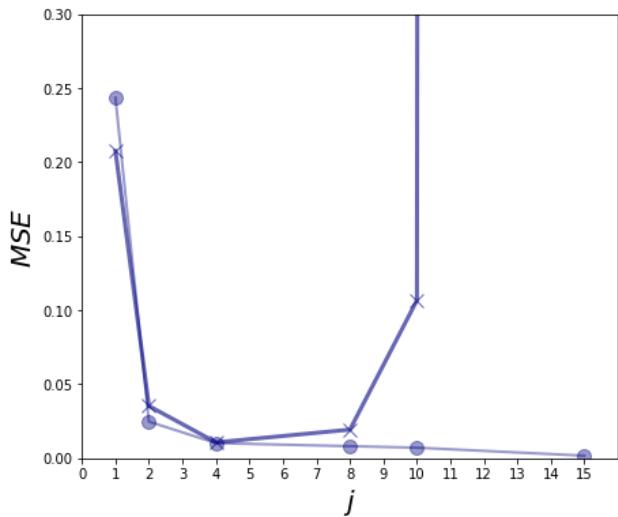
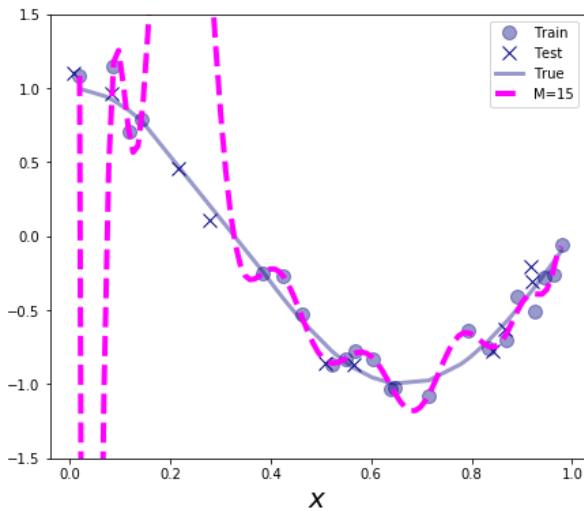
Overfitting!! 😥 😥



Valores de los
coeficientes se hacen
más extremos 😥

REGRESIÓN Y OVERFITTING

- Supongamos un modelo de regresión polinomial para un conjunto de $n=30$ datos



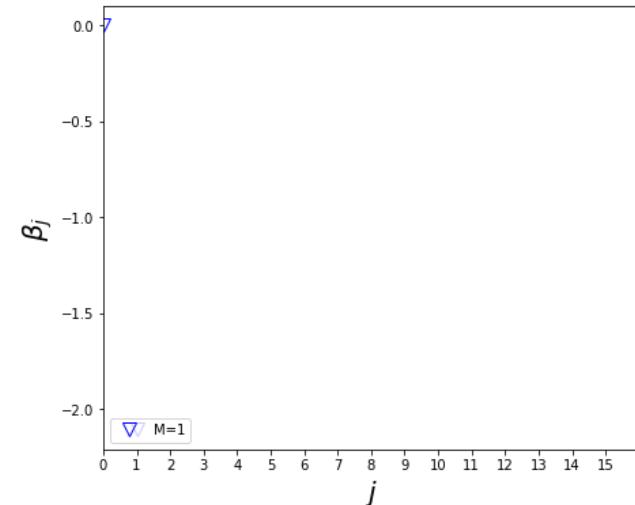
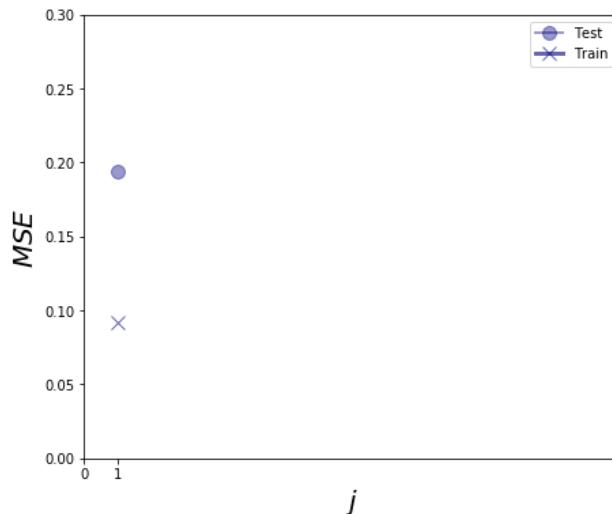
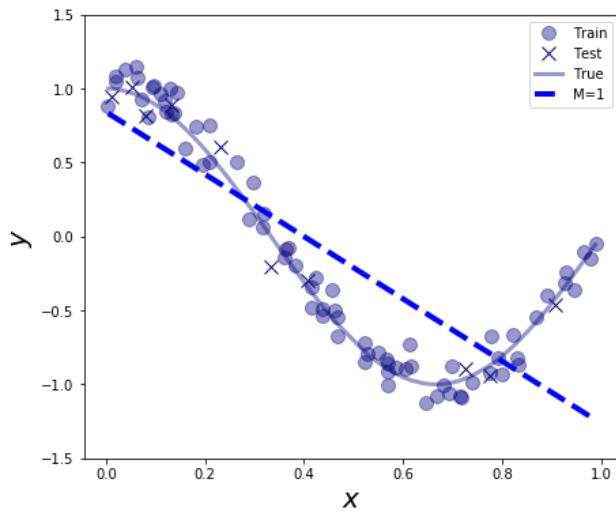
Mega overfitting!! 😞 😞 😞

Valores de los
coeficientes se hacen
muy extremos 😞

¿Qué pasa si agregamos más datos?

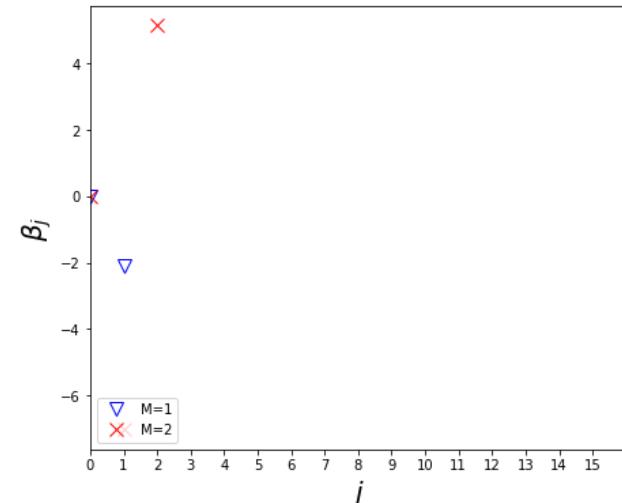
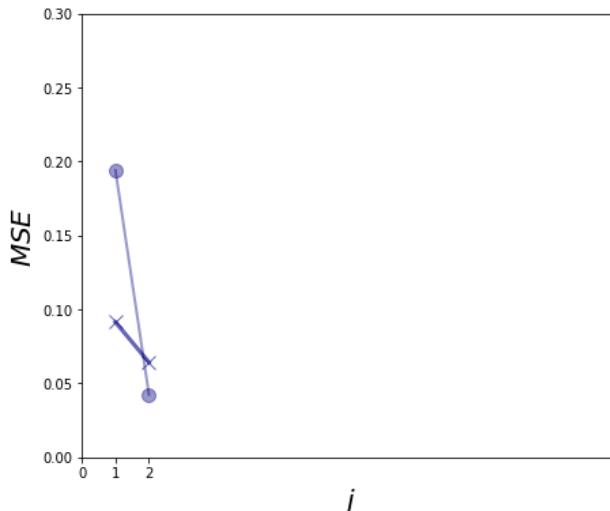
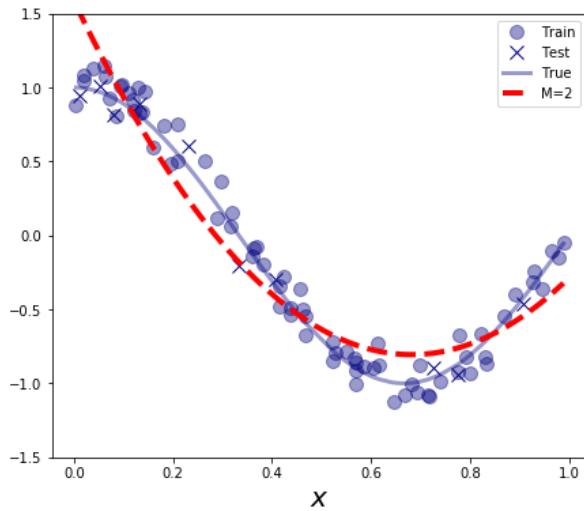
REGRESIÓN Y OVERFITTING

- Supongamos un modelo de regresión polinomial para un conjunto de **n=100 datos**



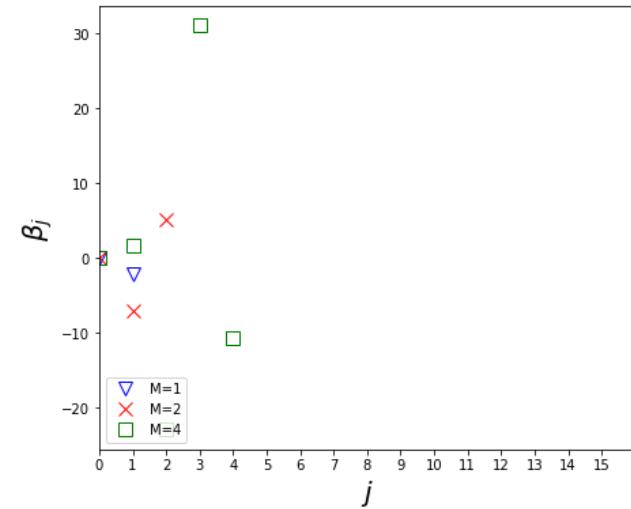
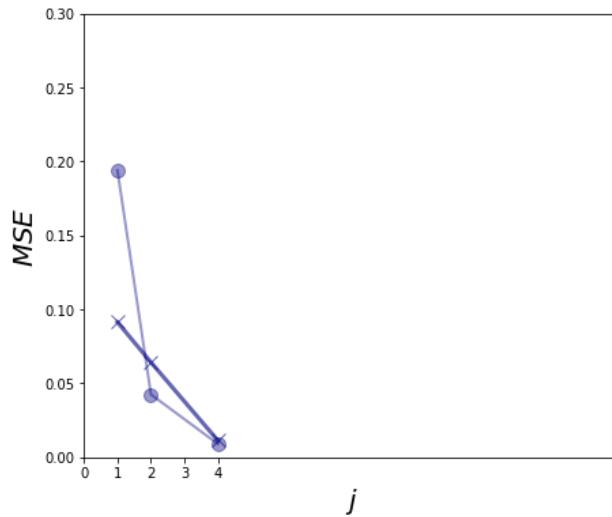
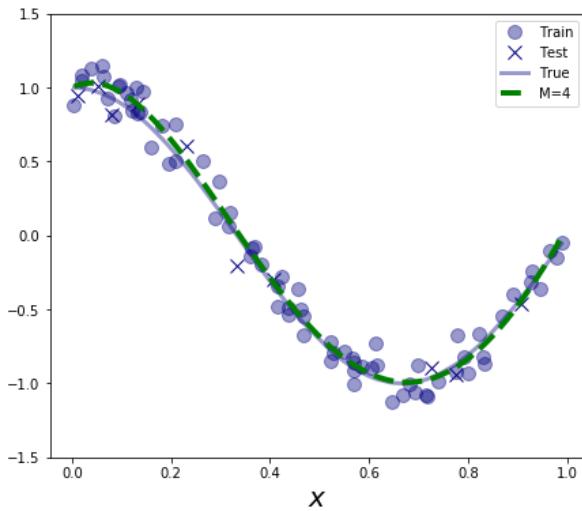
REGRESIÓN Y OVERFITTING

- Supongamos un modelo de regresión polinomial para un conjunto de **n=100 datos**



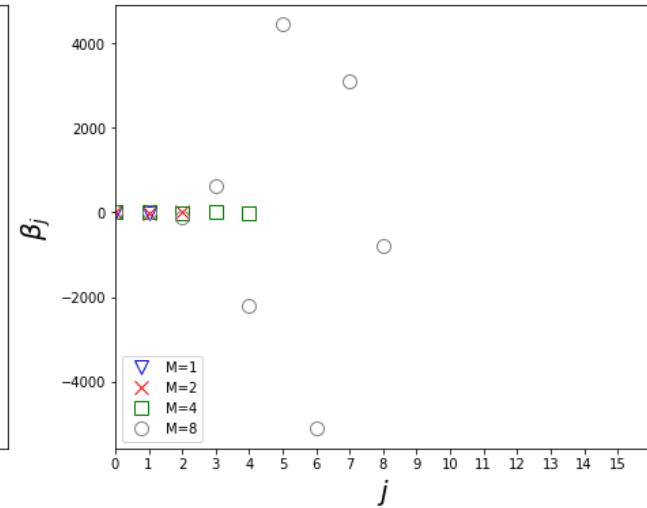
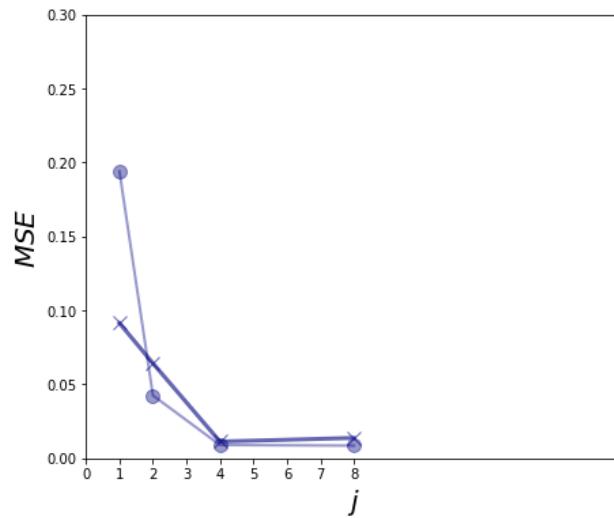
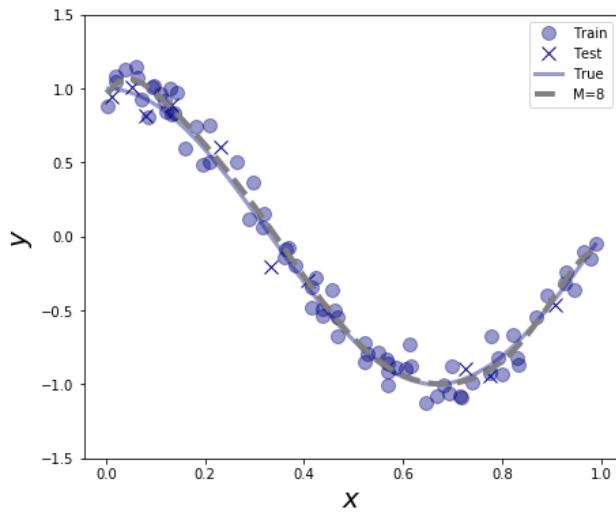
REGRESIÓN Y OVERFITTING

- Supongamos un modelo de regresión polinomial para un conjunto de **n=100 datos**



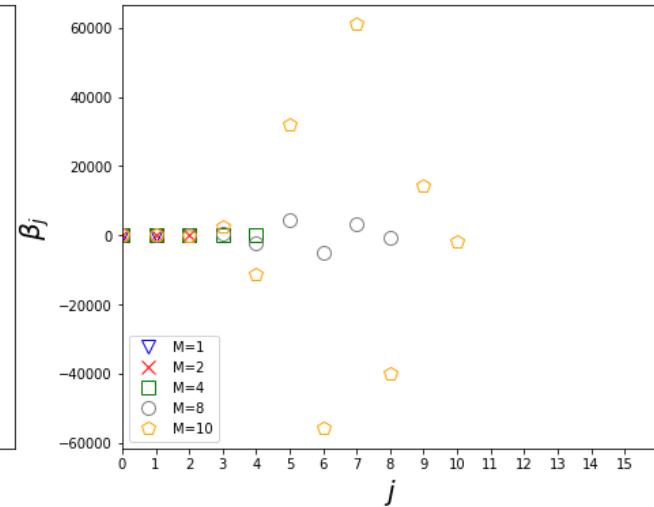
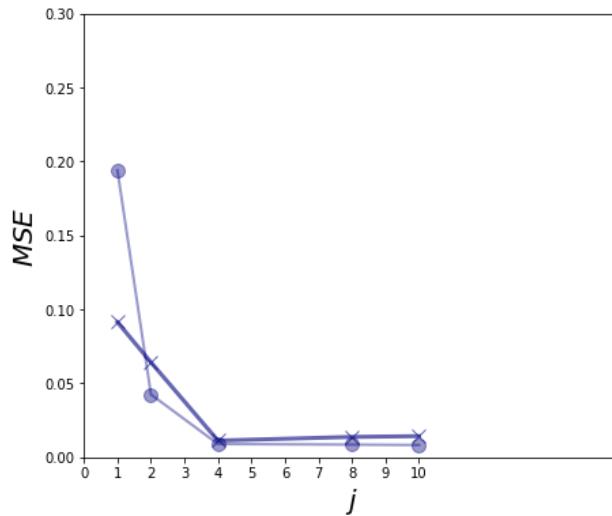
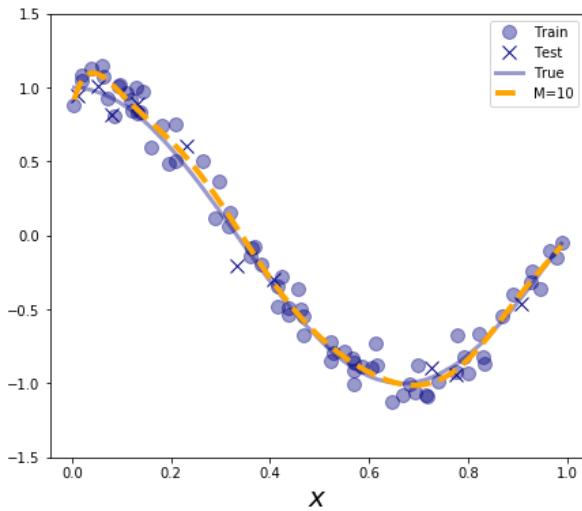
REGRESIÓN Y OVERFITTING

- Supongamos un modelo de regresión polinomial para un conjunto de **n=100 datos**



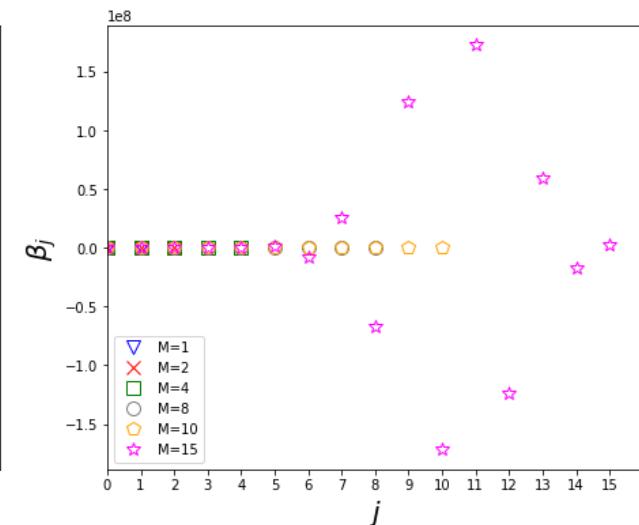
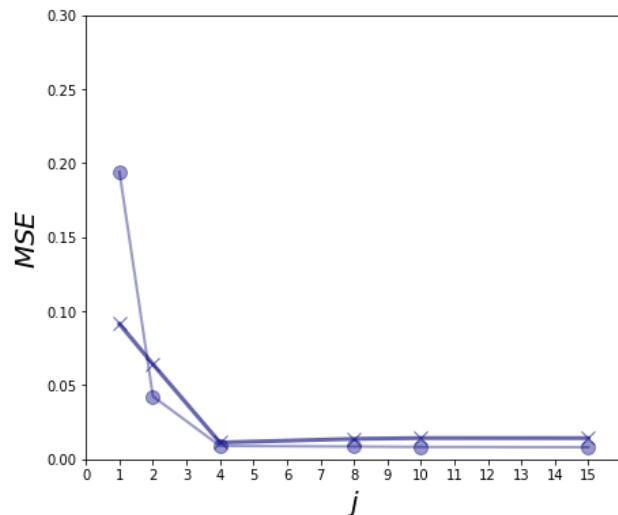
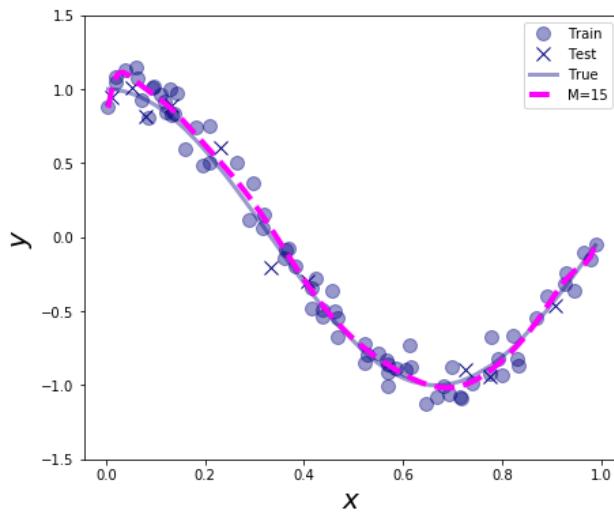
REGRESIÓN Y OVERFITTING

- Supongamos un modelo de regresión polinomial para un conjunto de **n=100 datos**



REGRESIÓN Y OVERFITTING

- Supongamos un modelo de regresión polinomial para un conjunto de **n=100 datos**



→ Con más datos, puedo entrenar un modelo más complejo, con menor riesgo de caer en overfitting.

REGULARIZACIÓN

- En una regresión lineal, minimizamos una función de pérdida $\mathcal{L}(\beta)$ para determinar los coeficientes β que acompañan a cada variable predictora X_j .
 - Ej: regresión lineal

$$\mathcal{L}(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 X_i))^2 = \frac{1}{n} \sum_{i=1}^n |y_i - \beta^T X_i|^2$$

$$\rightarrow \hat{\beta} = \operatorname{argmin} \mathcal{L}(\beta)$$

- Cuando tenemos modelos complejos o muchas variables, podemos caer en overfitting.
 - Algunos coeficientes de la regresión alcanzan valores extremos.

REGULARIZACIÓN

- La regularización consiste en **modificar la función de pérdida $\mathcal{L}(\beta)$** , agregando un término que **penalize** alguna propiedad dada de los parámetros del modelo:

$$L_{reg}(\beta) = L(\beta) + \boxed{\alpha R(\beta)} \rightarrow \text{regularización}$$

- El término α es un escalar que determina el peso o importancia que se le da al término de regularización.
- Al ajustar el modelo usando la función de pérdida modificada $L_{reg}(\beta)$, obtenemos parámetros del modelo con ciertas propiedades deseables.
 - Por ejemplo, queremos coeficientes **valores que no sean extremos**(porque como vimos en el ejemplo anterior, esto conduce a overfitting).

REGRESIÓN LASSO

- Como queremos desmotivar valores extremos en los coeficientes de la regresión, podemos elegir un término de regularización que **penalize la magnitud** de los coeficientes β .

$$L_{LASSO}(\beta) = L(\beta) + \alpha \sum_{m=1}^M |\beta_m|$$

$$L_{LASSO}(\beta) = \frac{1}{n} \sum_{i=1}^n |y_i - \beta^T X_i|^2 + \alpha \sum_{m=1}^M |\beta_m|$$

- El término α es un hiperparámetro que debemos testear y elegir antes de entrenar el modelo final.
- $\sum_{m=1}^M |\beta_m|$ corresponde a la norma l_1 del vector de coeficientes $\beta \rightarrow$ regularización l_1

REGRESIÓN RIDGE

- Otra opción es elegir un término de regularización que **penalize el cuadrado de la magnitud** de los coeficientes β .

$$L_{Ridge}(\beta) = L(\beta) + \alpha \sum_{m=1}^M \beta_m^2$$

$$L_{Ridge}(\beta) = \frac{1}{n} \sum_{i=1}^n |y_i - \beta^T X_i|^2 + \alpha \sum_{m=1}^M \beta_m^2$$

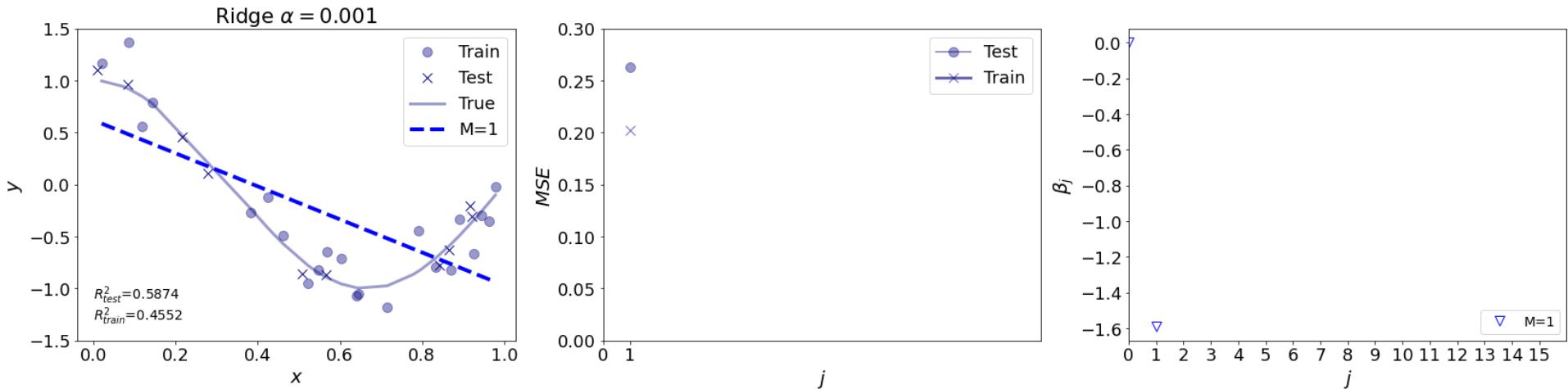
- El término α es un hiperparámetro que debemos testear y elegir antes de entrenar el modelo final.
- $\alpha \sum_{m=1}^M \beta_m^2$ corresponde a la norma l_2 del vector de coeficientes β → regularización l_2

REGRESIÓN LASSO Y RIDGE

- Tanto en Lasso como en Ridge, mientras mayor es el parámetro α , más se penalizan valores extremos de β .
 - $\alpha \sim 0 \rightarrow$ volvemos a la función de pérdida original (MSE), y es lo mismo que una regresión ordinaria.
 - $\alpha \gg 0 \rightarrow$ el término MSE en la función de pérdida regularizada será insignificante comparada con el término de regularización, y se fuerza que los coeficientes β sean cercanos a cero.
- Para elegir un valor razonable de α : debemos probar distintos valores y usar alguna forma de validación.
- Por lo tanto, los pasos son:
 - Elegir un valor de α
 - Ajustar el modelo de regresión a los datos y registrar el MSE o score para el dataset de prueba.
 - Repetir, y encontrar el α que entrega el menor MSE para el dataset de prueba.

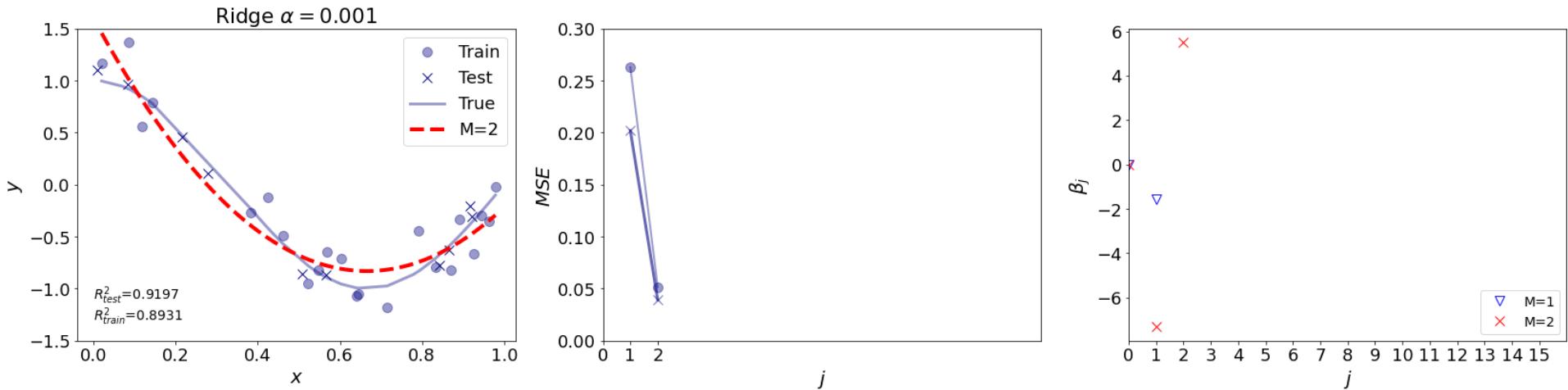
REGRESIÓN RIDGE

- Veamos el mismo caso anterior, pero ahora aplicando una regresión tipo Ridge, con distintos valores del hiperparámetro $\alpha = [0.001, 1.0, 10]$



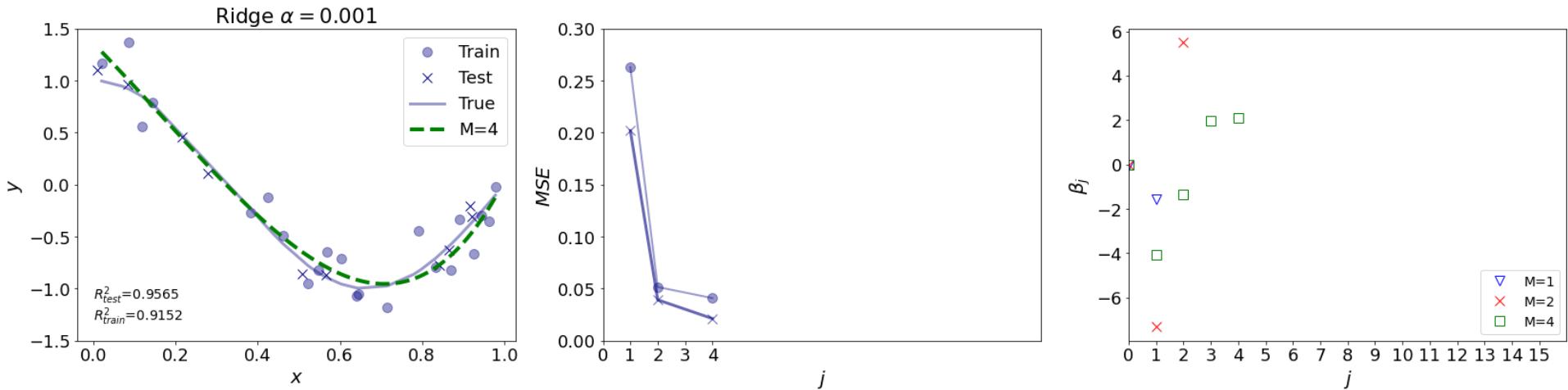
REGRESIÓN RIDGE

- Veamos el mismo caso anterior, pero ahora aplicando una regresión tipo Ridge, con distintos valores del hiperparámetro $\alpha = [0.001, 1.0, 10]$



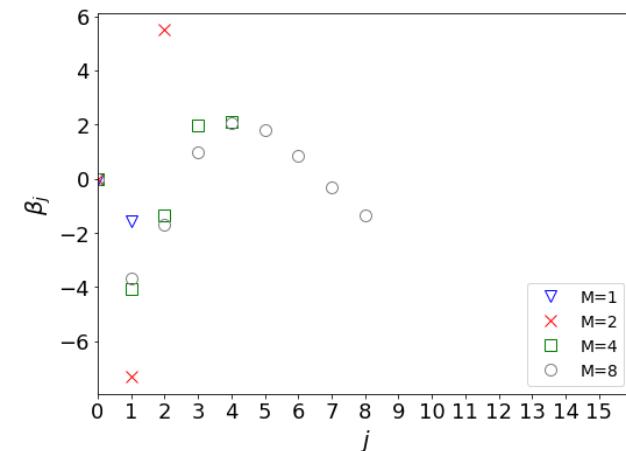
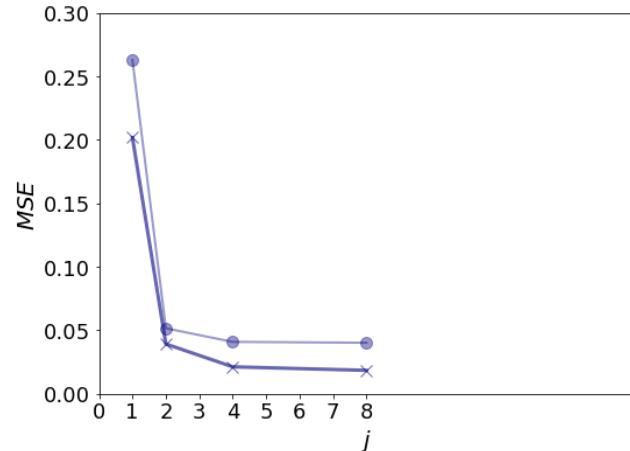
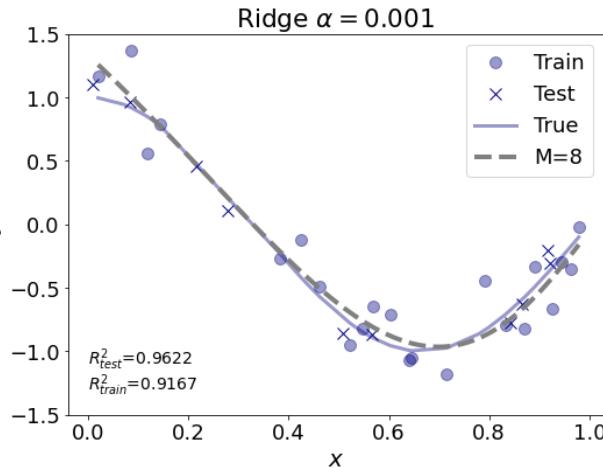
REGRESIÓN RIDGE

- Veamos el mismo caso anterior, pero ahora aplicando una regresión tipo Ridge, con distintos valores del hiperparámetro $\alpha = [0.001, 1.0, 10]$



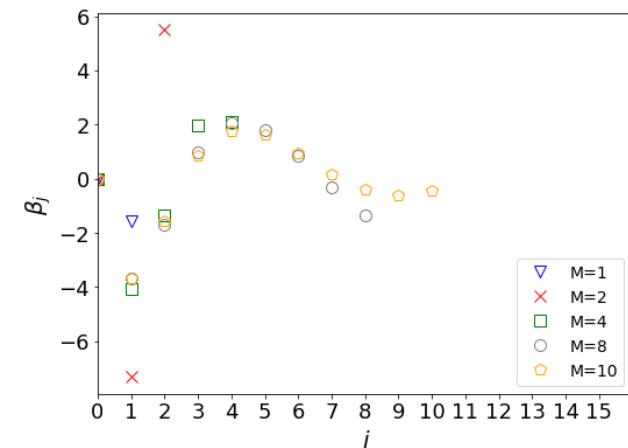
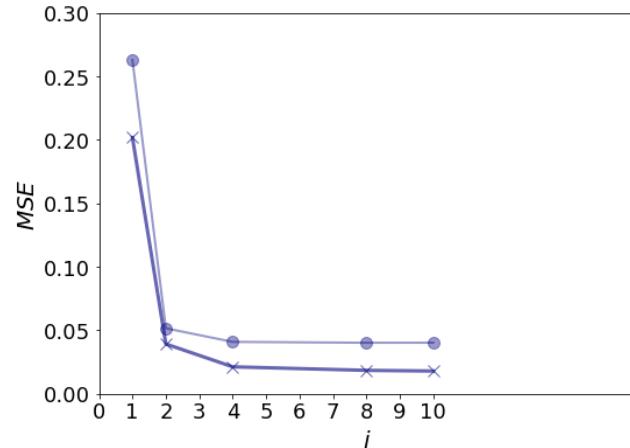
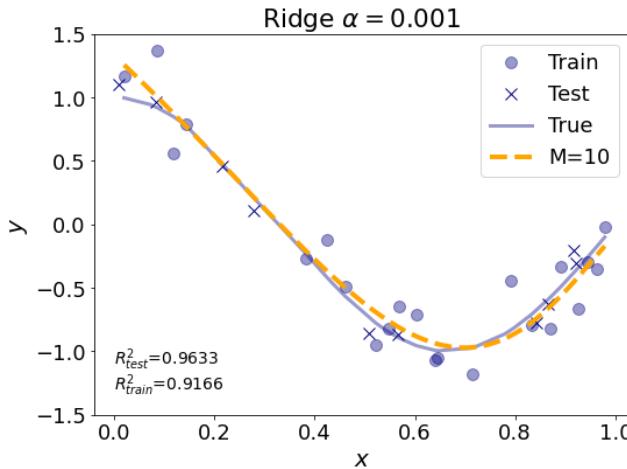
REGRESIÓN RIDGE

- Veamos el mismo caso anterior, pero ahora aplicando una regresión tipo Ridge, con distintos valores del hiperparámetro $\alpha = [0.001, 1.0, 10]$



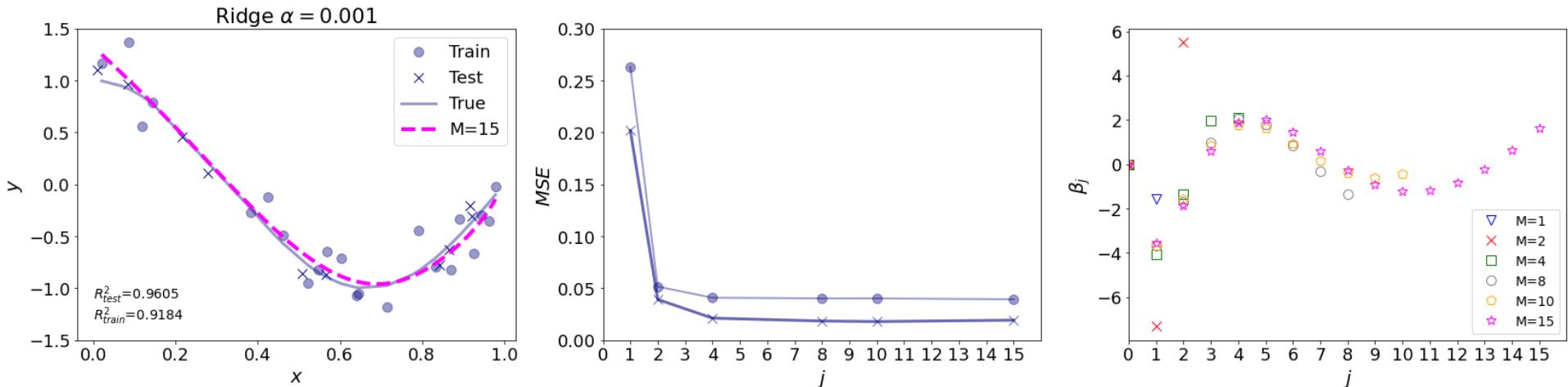
REGRESIÓN RIDGE

- Veamos el mismo caso anterior, pero ahora aplicando una regresión tipo Ridge, con distintos valores del hiperparámetro $\alpha = [0.001, 1.0, 10]$



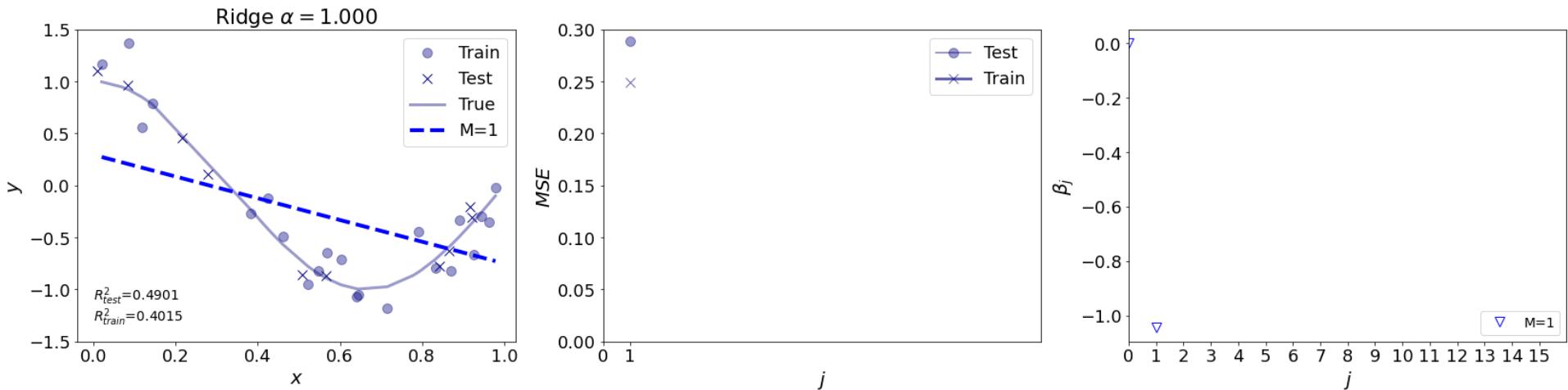
REGRESIÓN RIDGE

- Veamos el mismo caso anterior, pero ahora aplicando una regresión tipo Ridge, con distintos valores del hiperparámetro $\alpha = [0.001, 1.0, 10]$



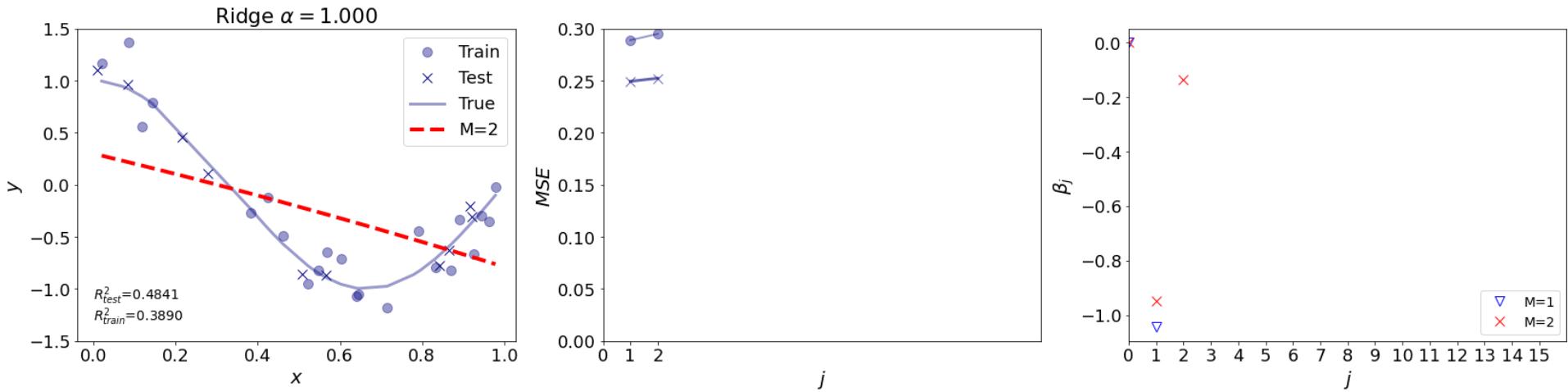
REGRESIÓN RIDGE

- Veamos el mismo caso anterior, pero ahora aplicando una regresión tipo Ridge, con distintos valores del hiperparámetro $\alpha = [0.001, 1.0, 10]$



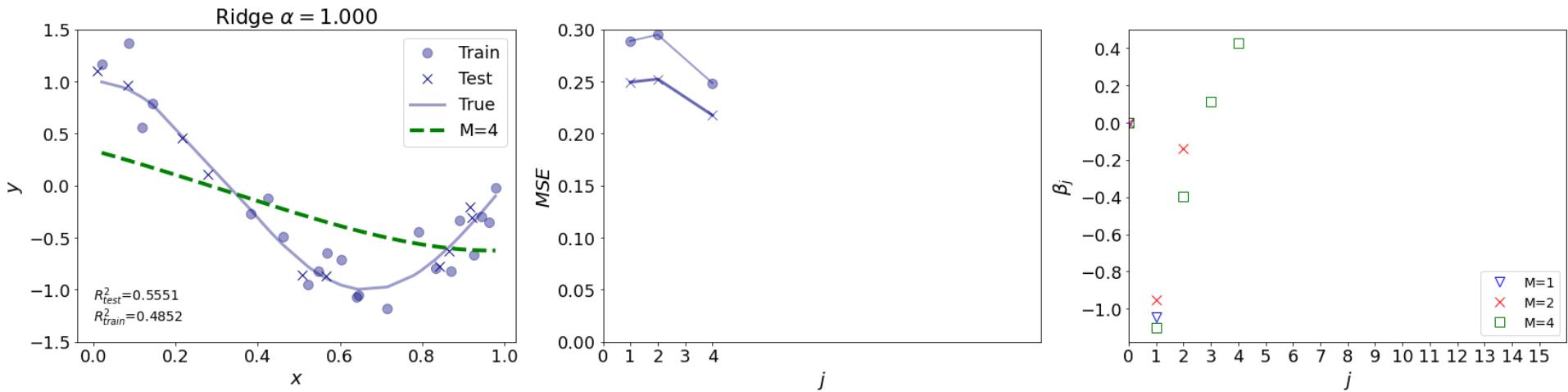
REGRESIÓN RIDGE

- Veamos el mismo caso anterior, pero ahora aplicando una regresión tipo Ridge, con distintos valores del hiperparámetro $\alpha = [0.001, 1.0, 10]$



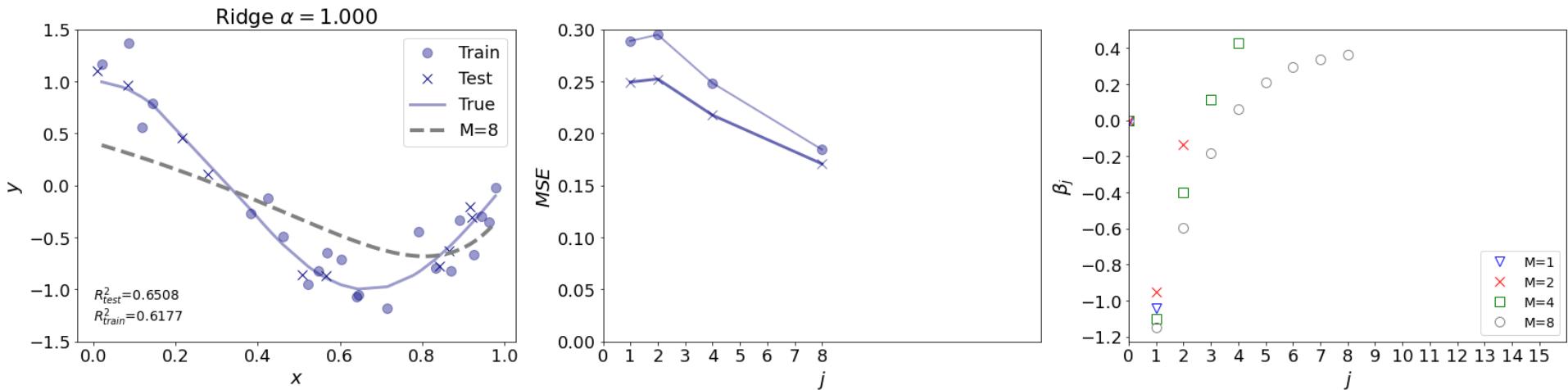
REGRESIÓN RIDGE

- Veamos el mismo caso anterior, pero ahora aplicando una regresión tipo Ridge, con distintos valores del hiperparámetro $\alpha = [0.001, 1.0, 10]$



REGRESIÓN RIDGE

- Veamos el mismo caso anterior, pero ahora aplicando una regresión tipo Ridge, con distintos valores del hiperparámetro $\alpha = [0.001, 1.0, 10]$



REGRESIÓN RIDGE

- Veamos el mismo caso anterior, pero ahora aplicando una regresión tipo Ridge, con distintos valores del hiperparámetro $\alpha = [0.001, 1.0, 10]$

