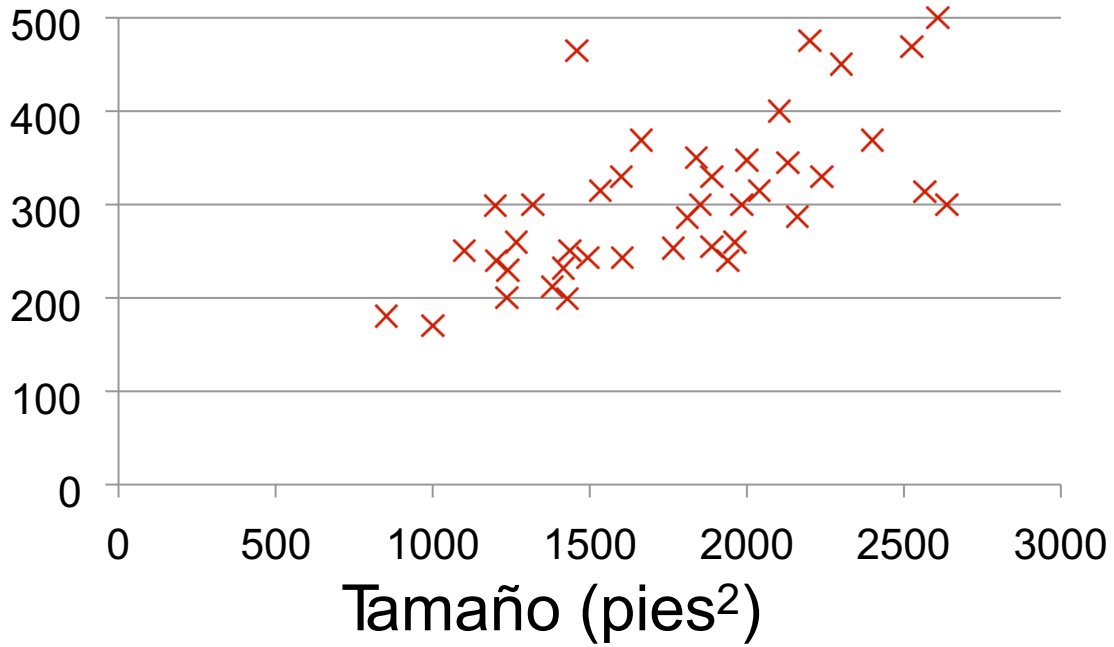
TEMA 9. TÉCNICAS DE REGRESIÓN

Contenidos

1. Introducción
2. Regresión lineal
3. Árboles de regresión
4. Redes neuronales
5. Ejercicio de regresión

1. Introducción

Precios casas (1000s USD)



Aprendizaje Supervisado

Dada la **"respuesta correcta"** para los **ejemplos del entrenamiento**

Regresión

Predecir la salida que es un **valor real**

1. Introducción

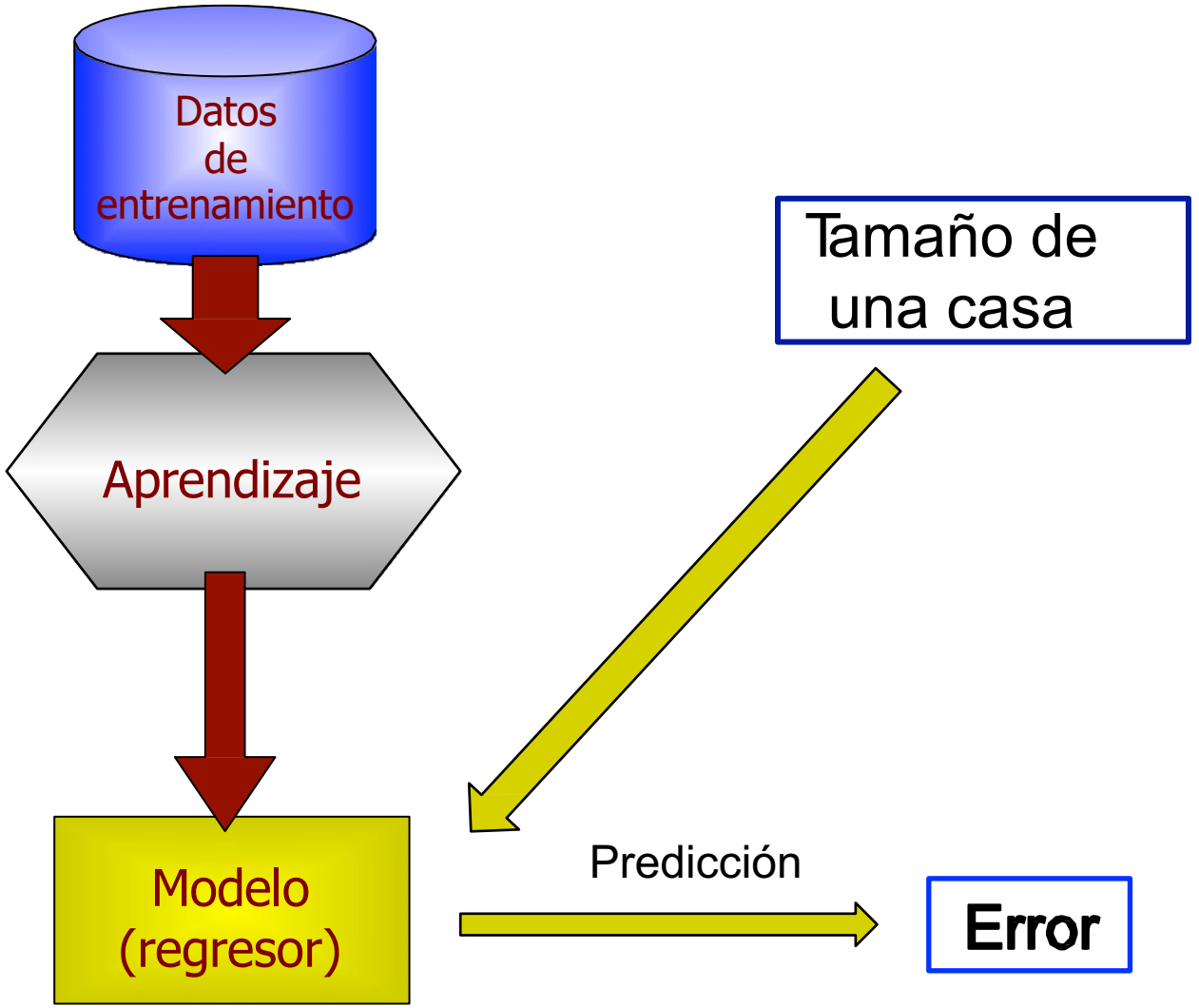
Conjunto de entrenamiento →

Tamaño (x)	**Precio (y)**
2104	460
1416	232
1534	315
852	178
…	…

Notación:

m = Número de ejemplos de entrenamiento
x's = variables de entrada / atributos
y's = variable de salida / variable "target"

1. Introducción



1. Introducción

x_1	x_3	x_3	x_4	x_5
Tamaño	**#Habitaciones**	**#Plantas**	**Edad**	**Precio**
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
…	…	…	…	…

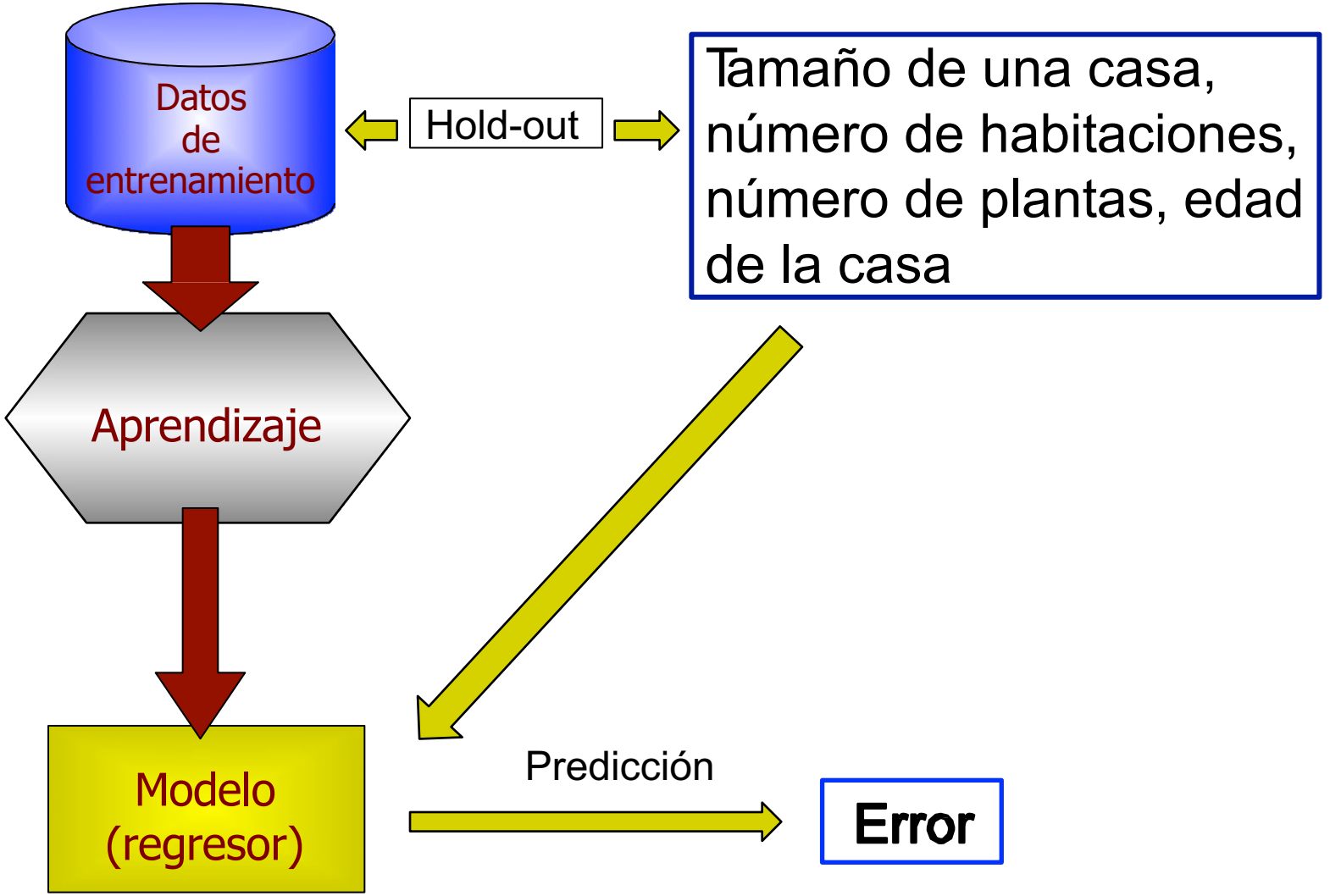
Notación:

n = número de atributos

x_i= atributos i-ésimo de entrenamiento

$x_i^{(j)}$= valor j del atributo de entrenamiento i-ésimo

1. Introducción



1. Introducción

Evaluación de la regresión (I)

- **Error absoluto medio (MAE):**

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

- **Error relativo medio(MAPE):**

$$\mathrm{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

1. Introducción

Evaluación de la regresión (y II)

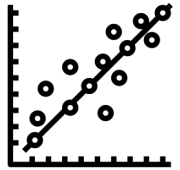
- **Raíz del error cuadrático medio (RMSE):**

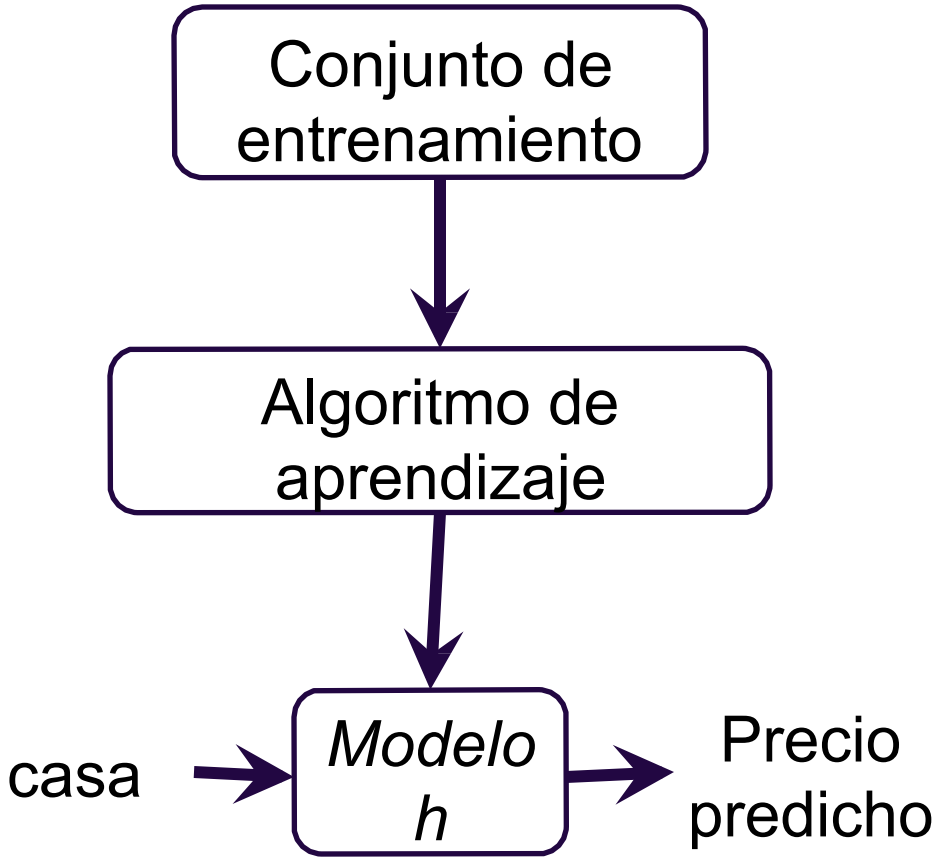
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

- **Coeficiente de correlación lineal (CC o R^2):**

$$CC_{XY} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

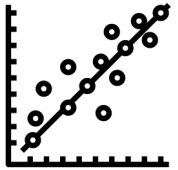
2. Regresión lineal





¿Cómo representamos *h* ?

2. Regresión lineal



¿Cómo representamos *h* ?

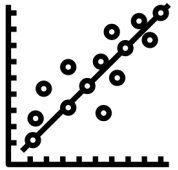
Recta—*Regresión lineal univariable*

$$h_\theta(x) = \theta_0 + \theta_1 x$$

Recta—*Regresión lineal multivariable*

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

2. Regresión lineal



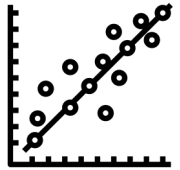
Calcular el modelo

Calcular los parámetros θ_i usando el conjunto de entrenamiento

$$h_\theta(x) = \theta_0 + \theta_1 x$$

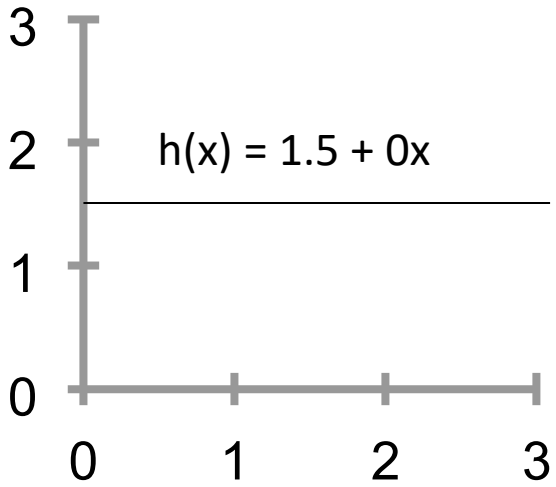
$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

2. Regresión lineal



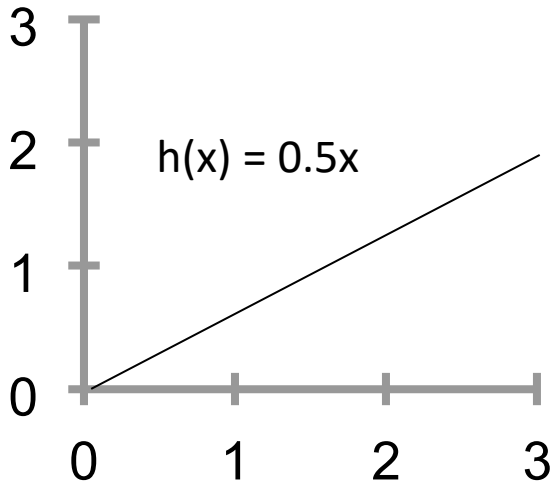
$h_\theta(x) = \theta_0 + \theta_1 x$

θ_i's: Parámetros



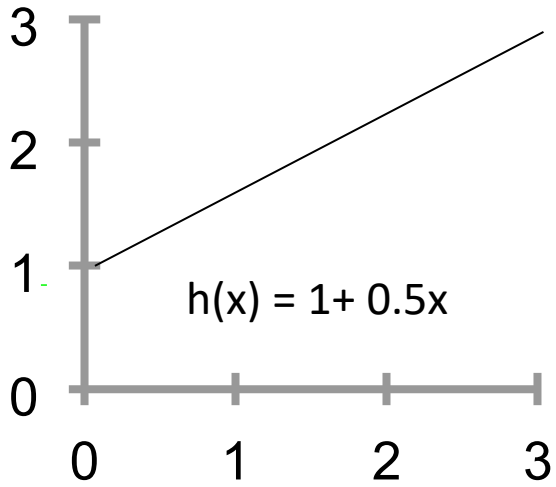
$\theta_0 = 1.5$

$\theta_1 = 0$



$\theta_0 = 0$

$\theta_1 = 0.5$

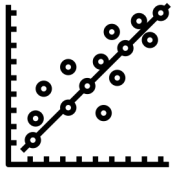


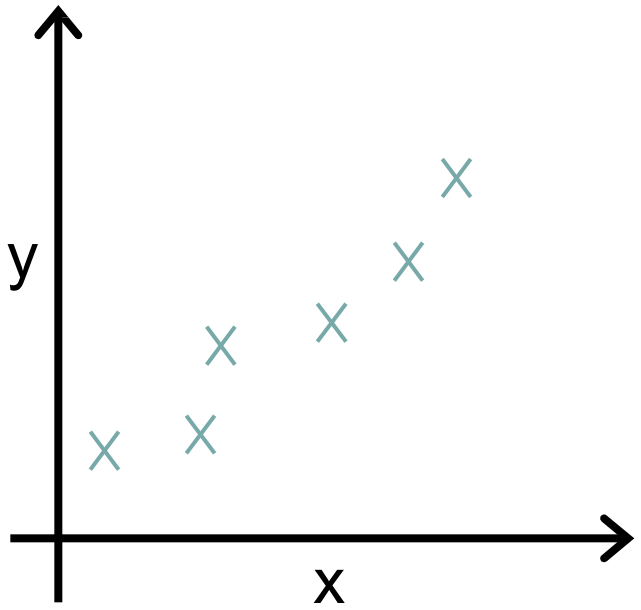
$\theta_0 = 1$

$\theta_1 = 0.5$

Cómo elegir θ_i's ?

2. Regresión lineal

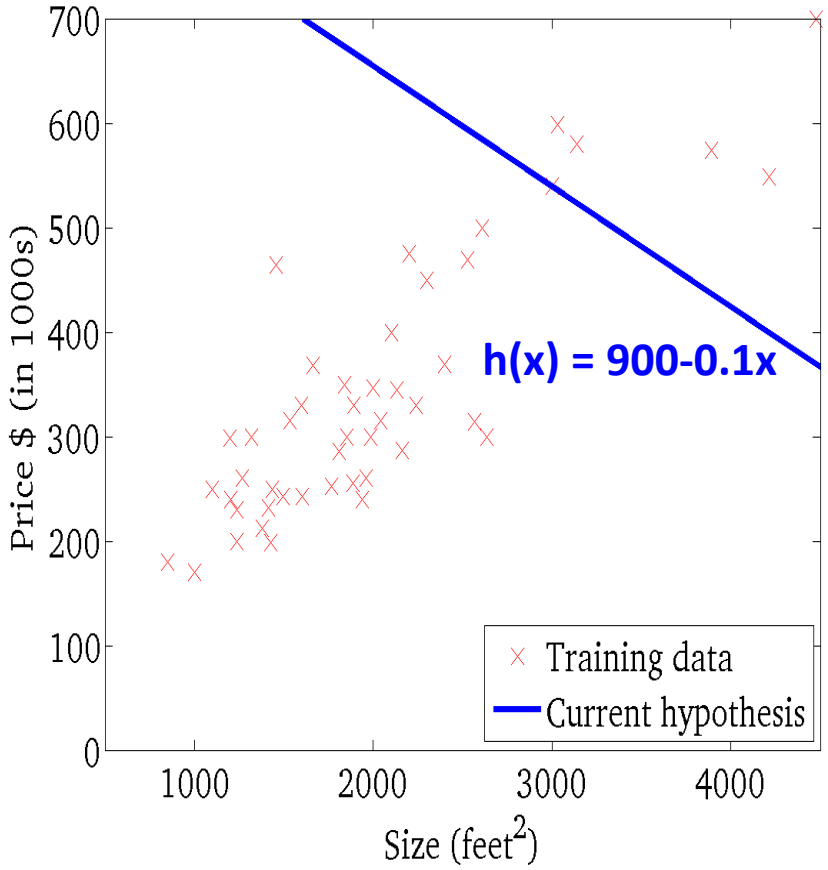




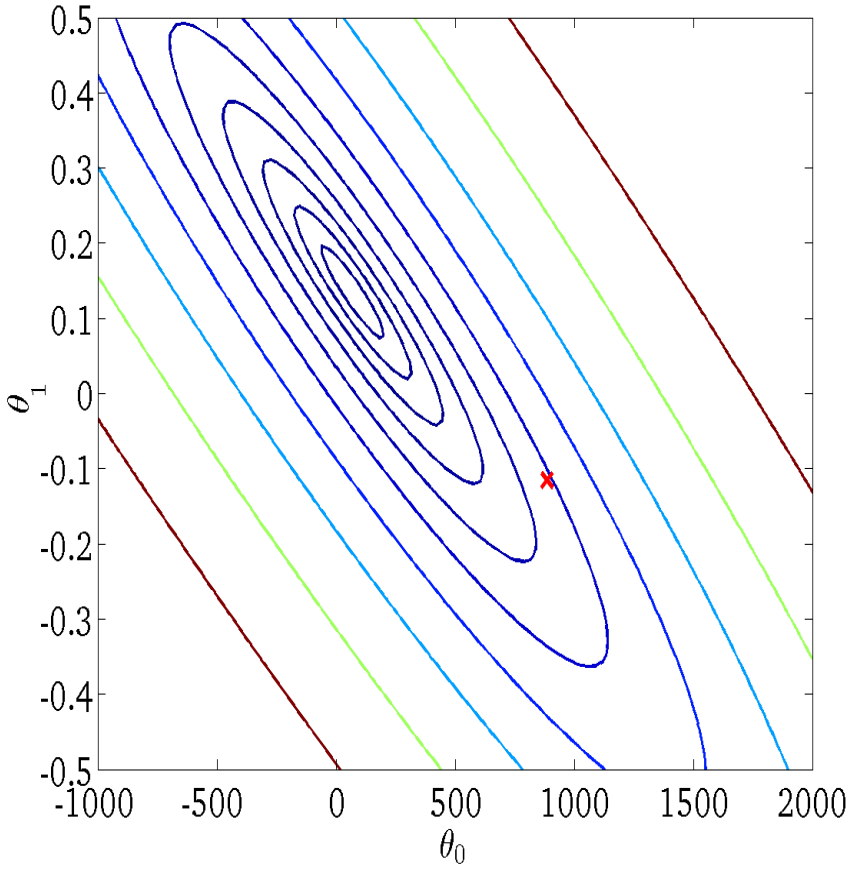
Idea:
elegir θ_0, θ_1 tal que $h_\theta(x)$ esté cerca de y para los ejemplos de entrenamiento (x, y)

2. Regresión lineal

$h_\theta(x)$



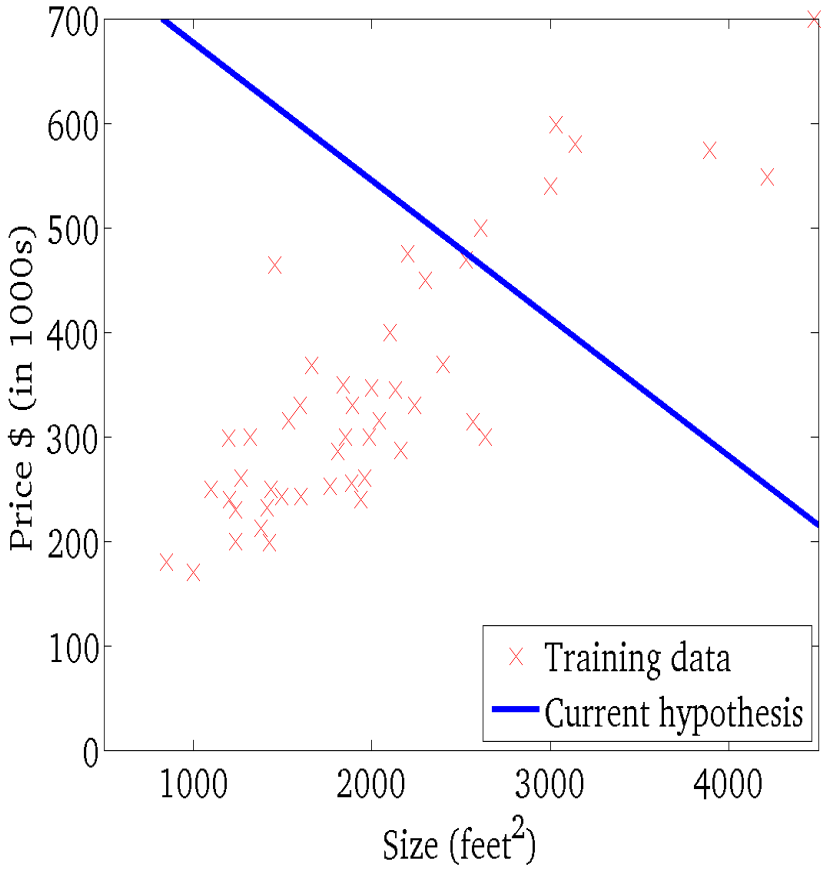
parámetros θ_0, θ_1

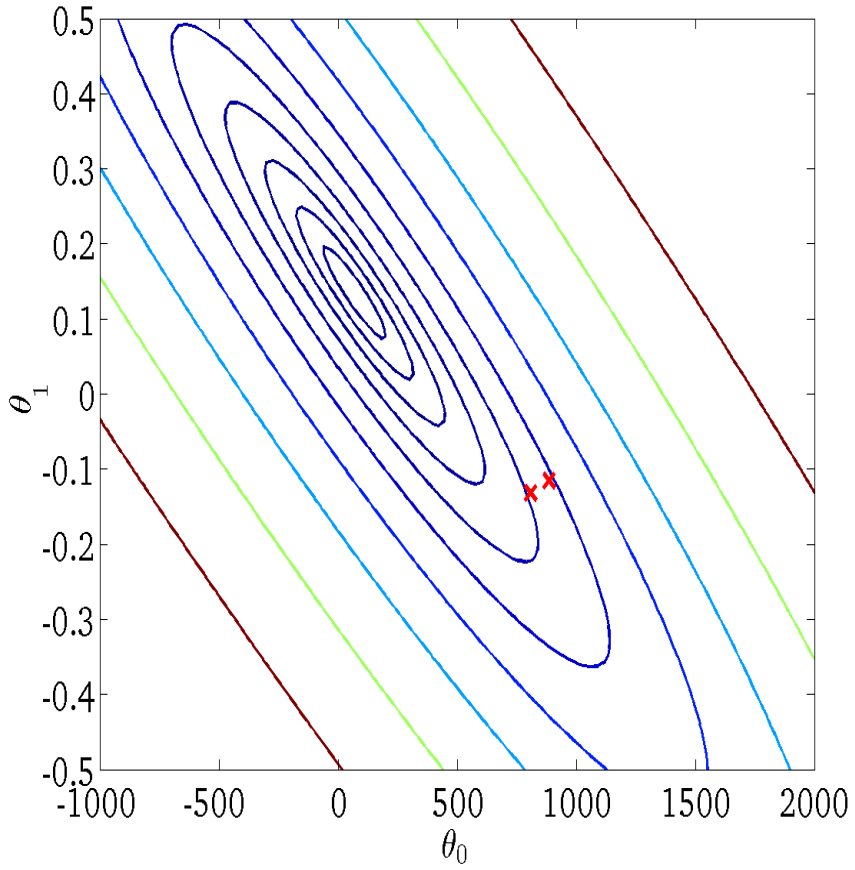


2. Regresión lineal

$h_\theta(x)$

parámetros θ_0, θ_1

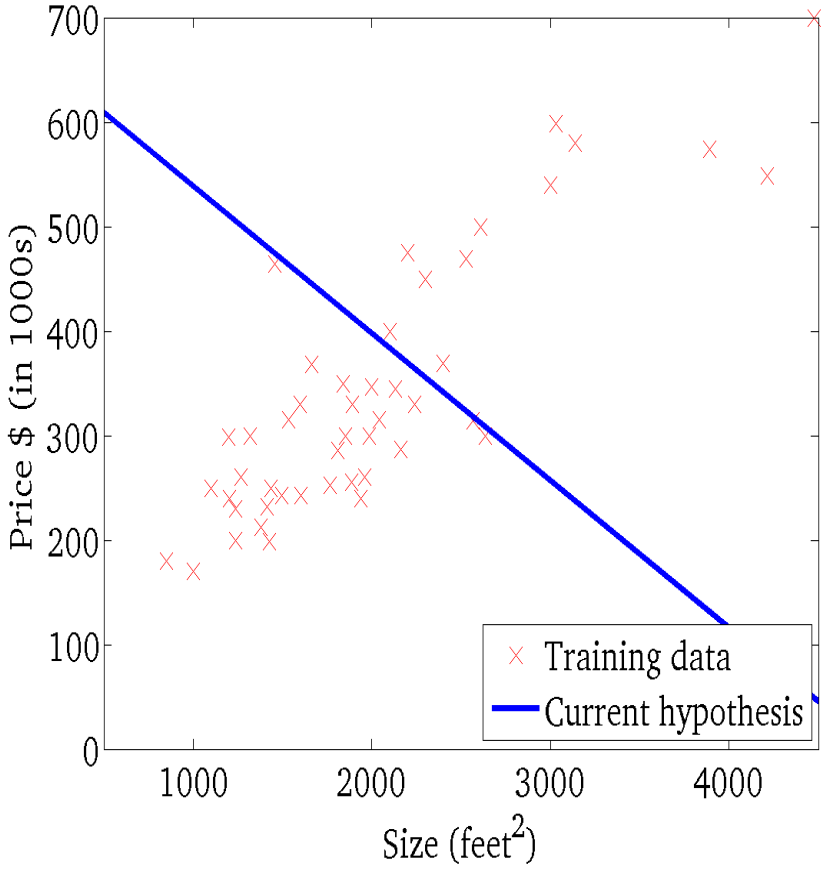


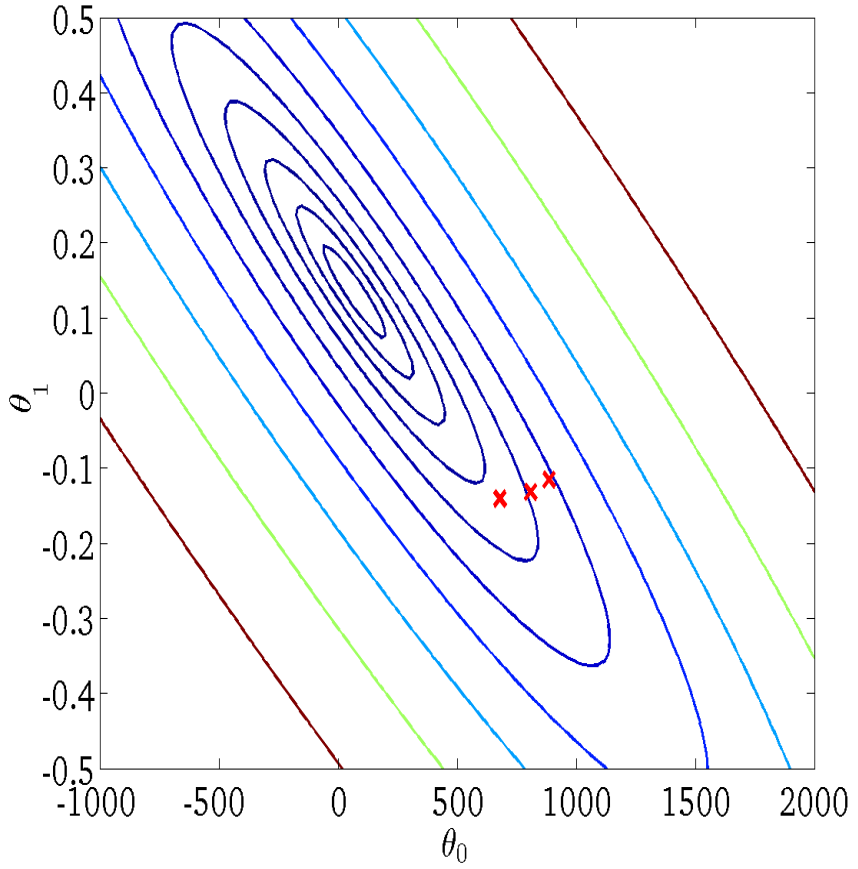


2. Regresión lineal

$h_\theta(x)$

parámetros θ_0, θ_1

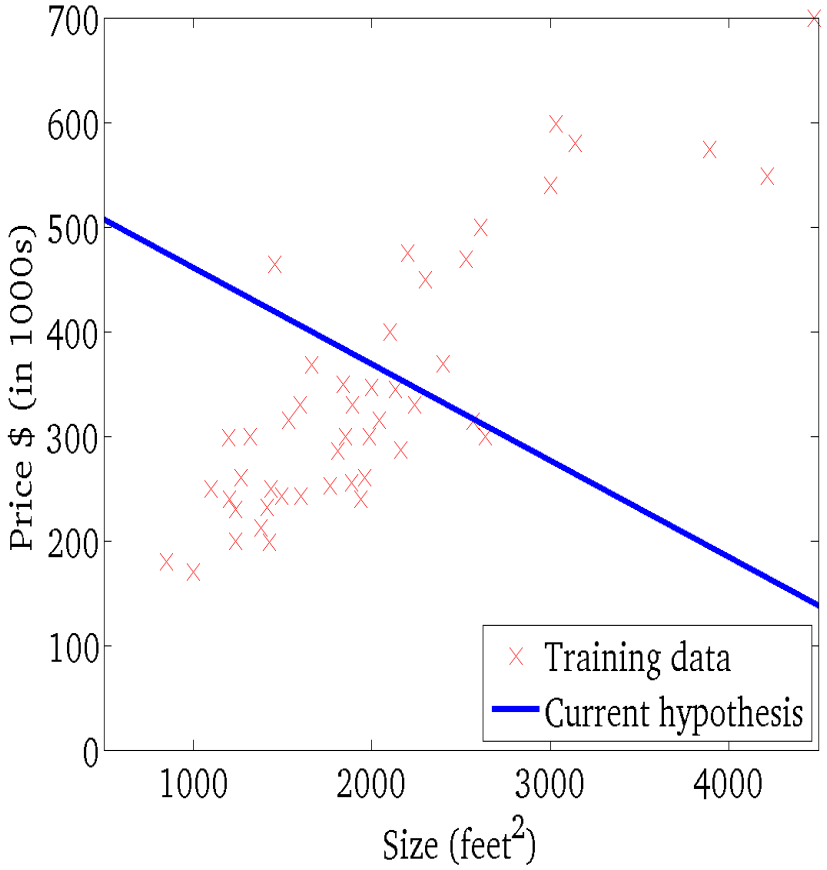


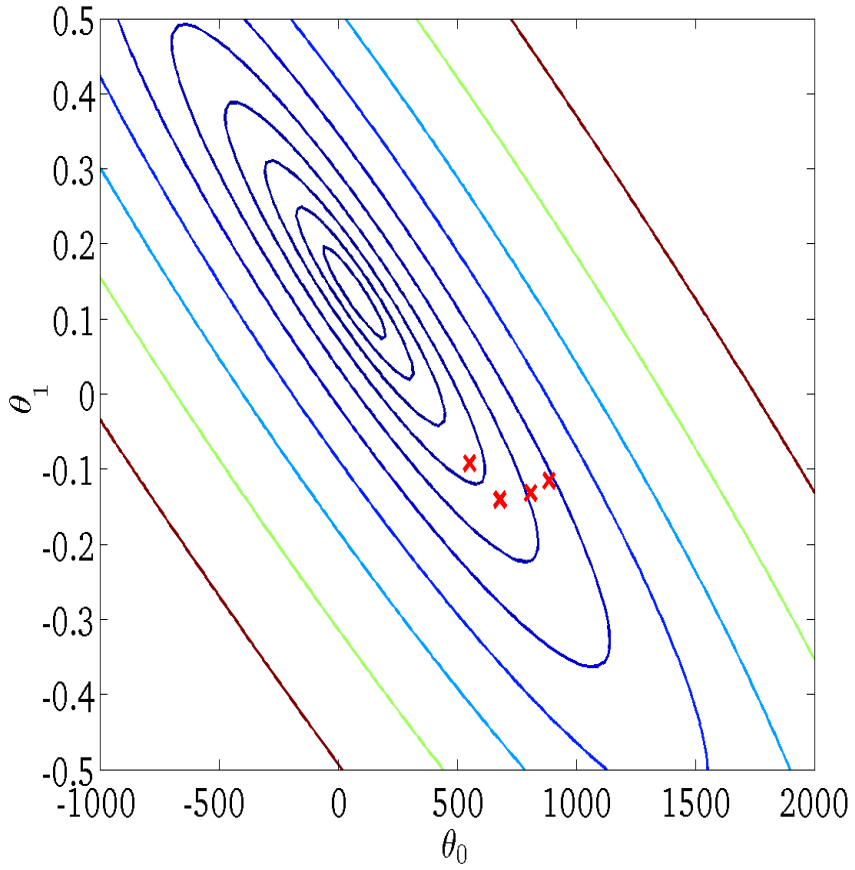


2. Regresión lineal

$h_\theta(x)$

parámetros θ_0, θ_1

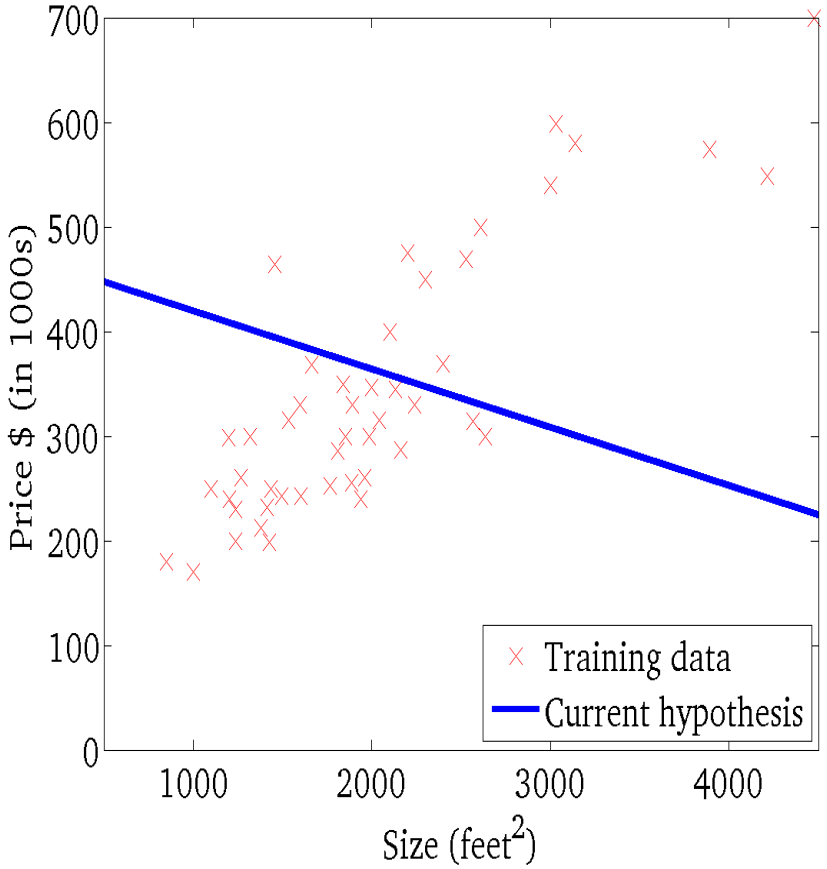


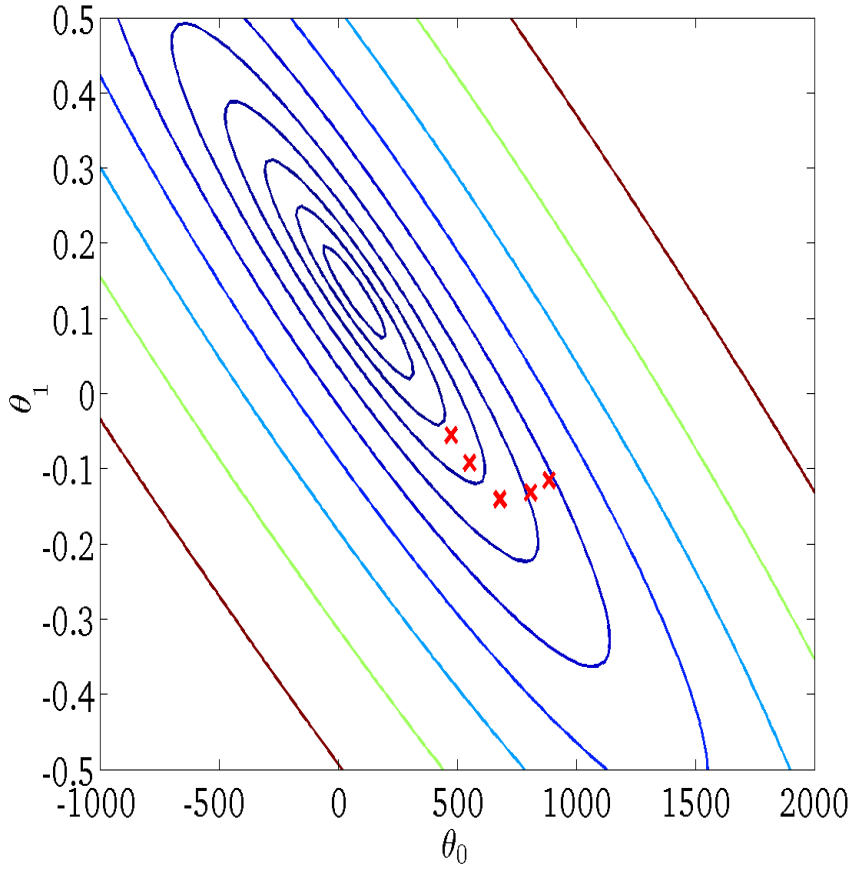


2. Regresión lineal

$h_\theta(x)$

parámetros θ_0, θ_1

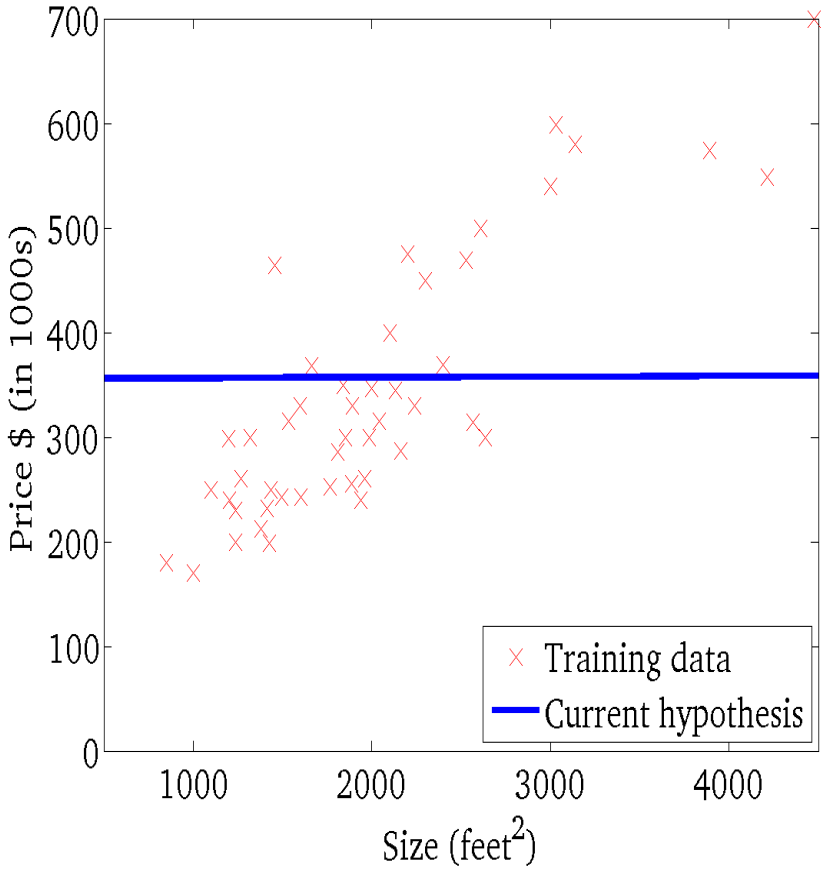


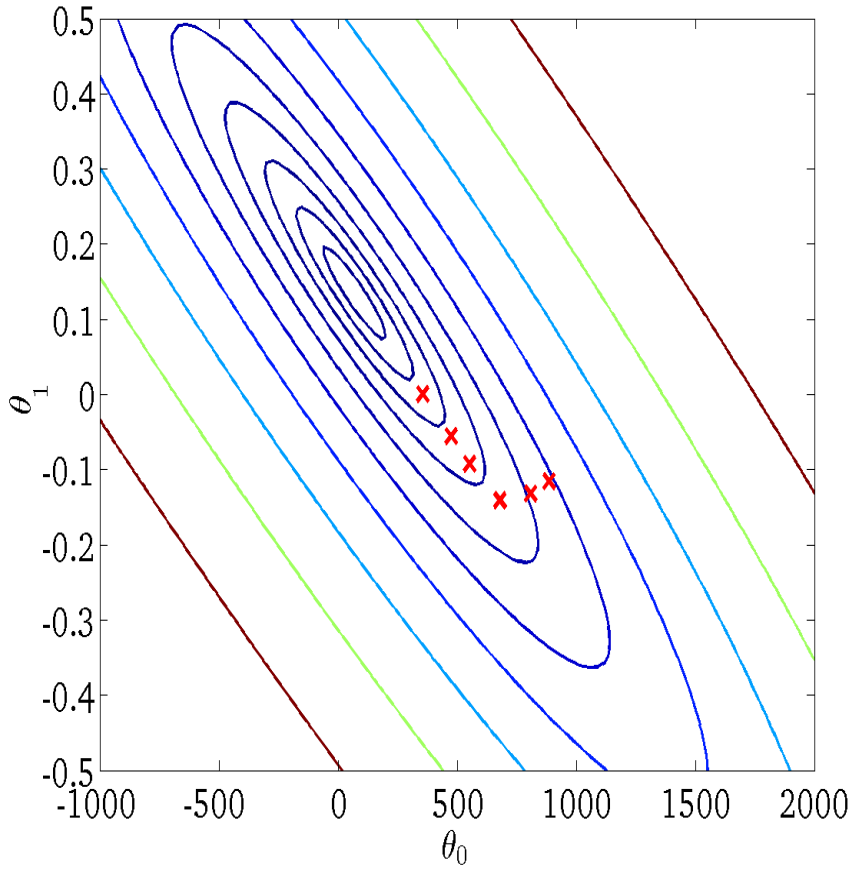


2. Regresión lineal

$h_\theta(x)$

parámetros θ_0, θ_1

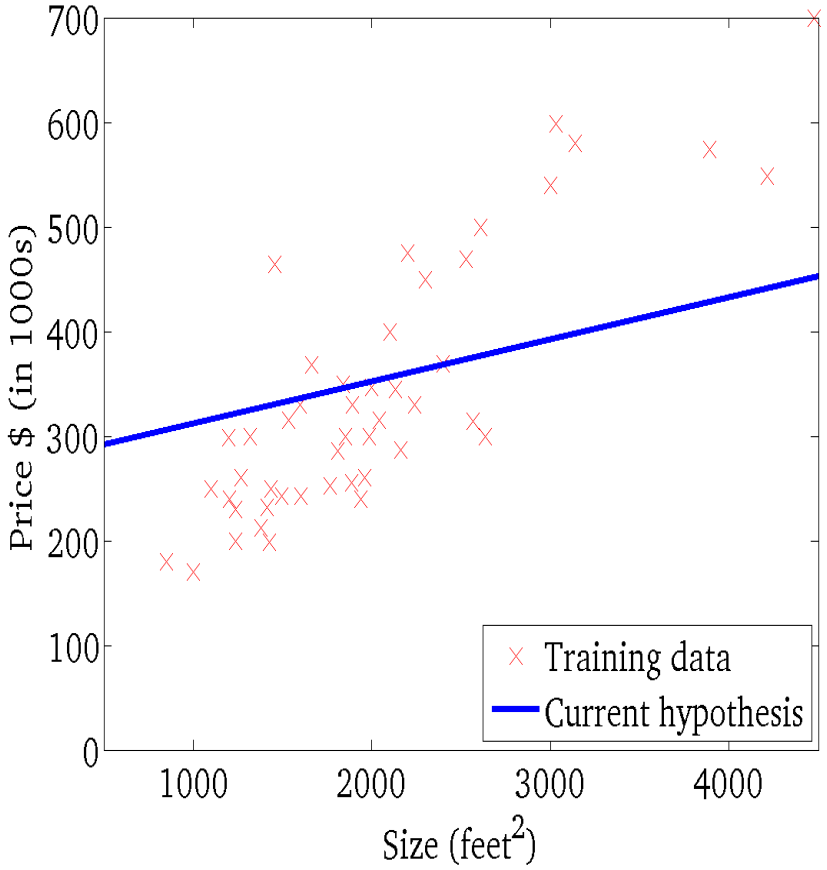


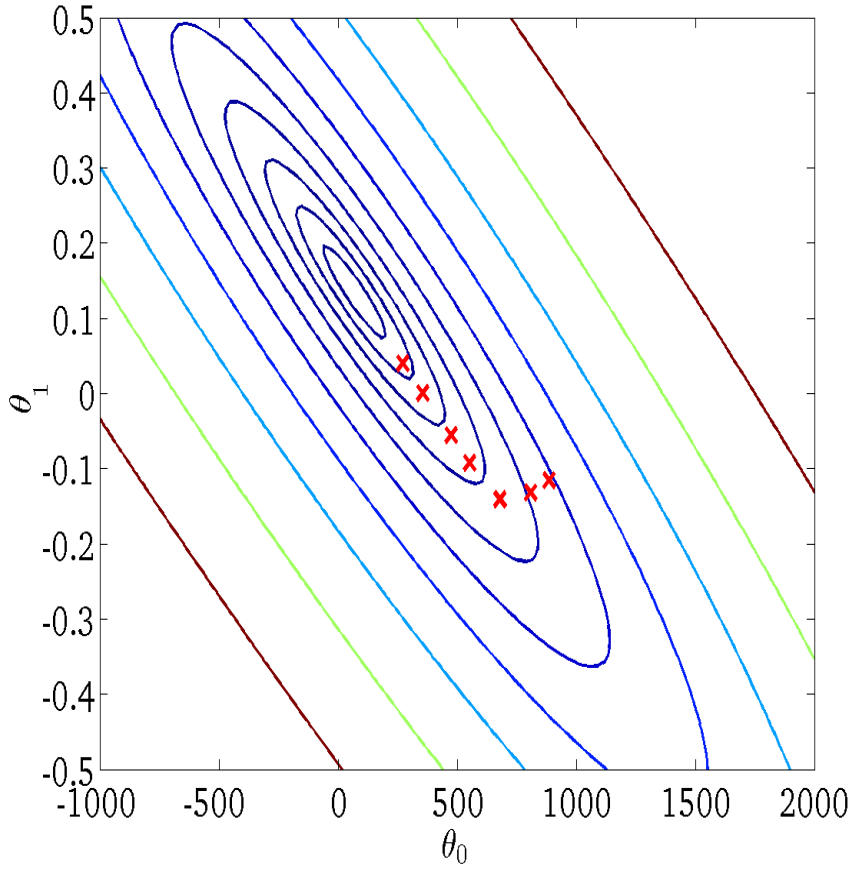


2. Regresión lineal

$h_\theta(x)$

parámetros θ_0, θ_1

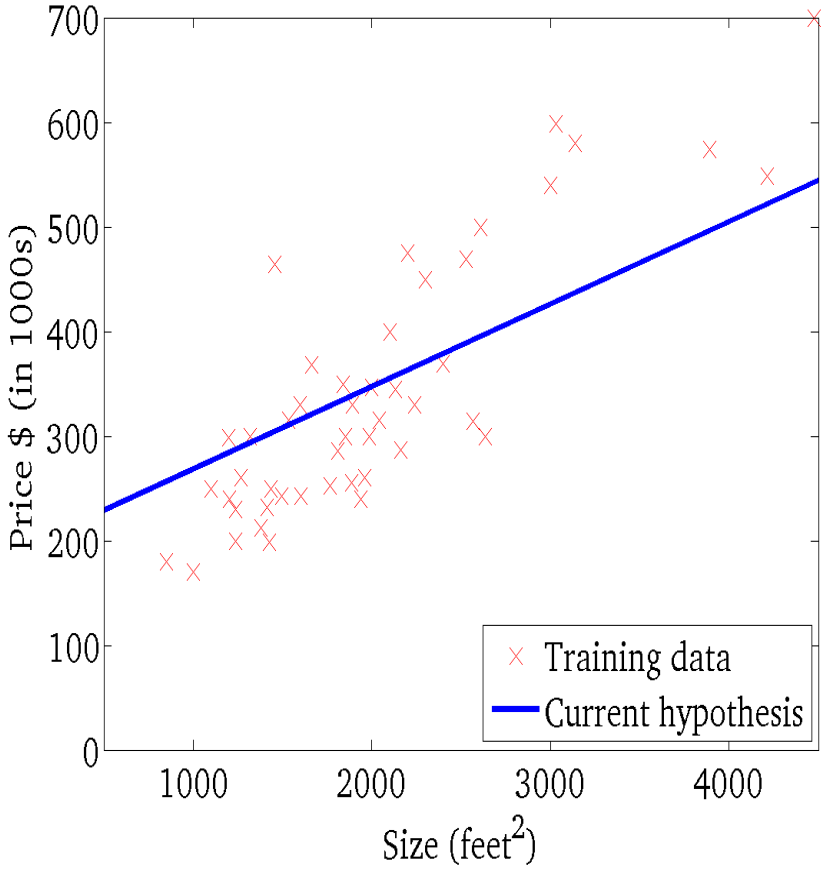


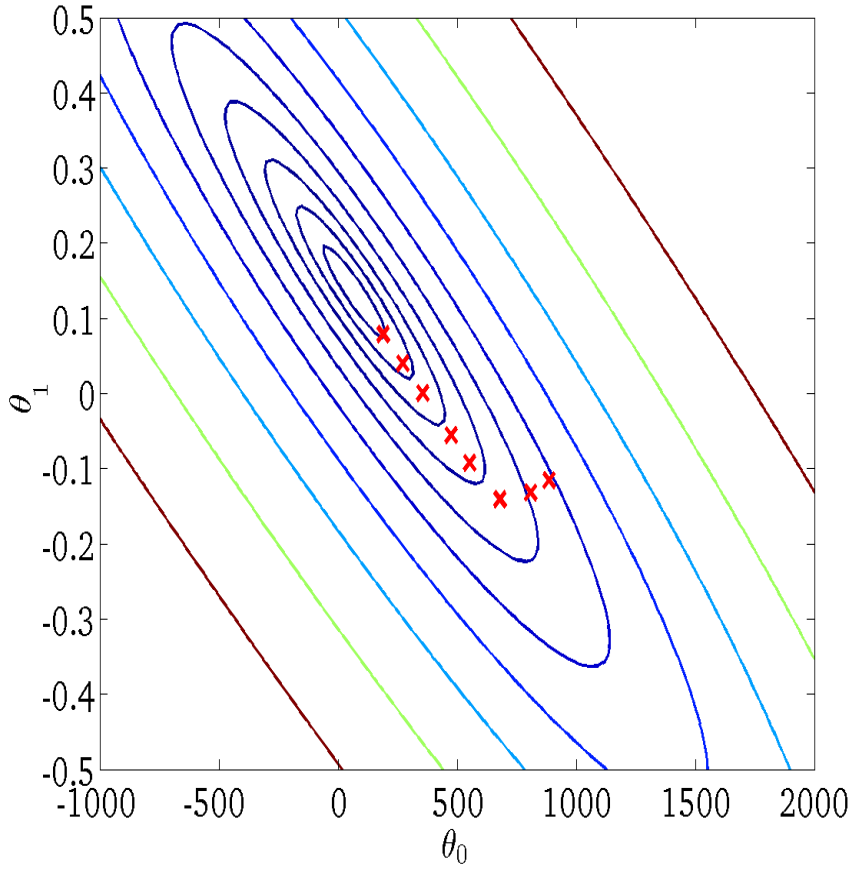


2. Regresión lineal

$h_\theta(x)$

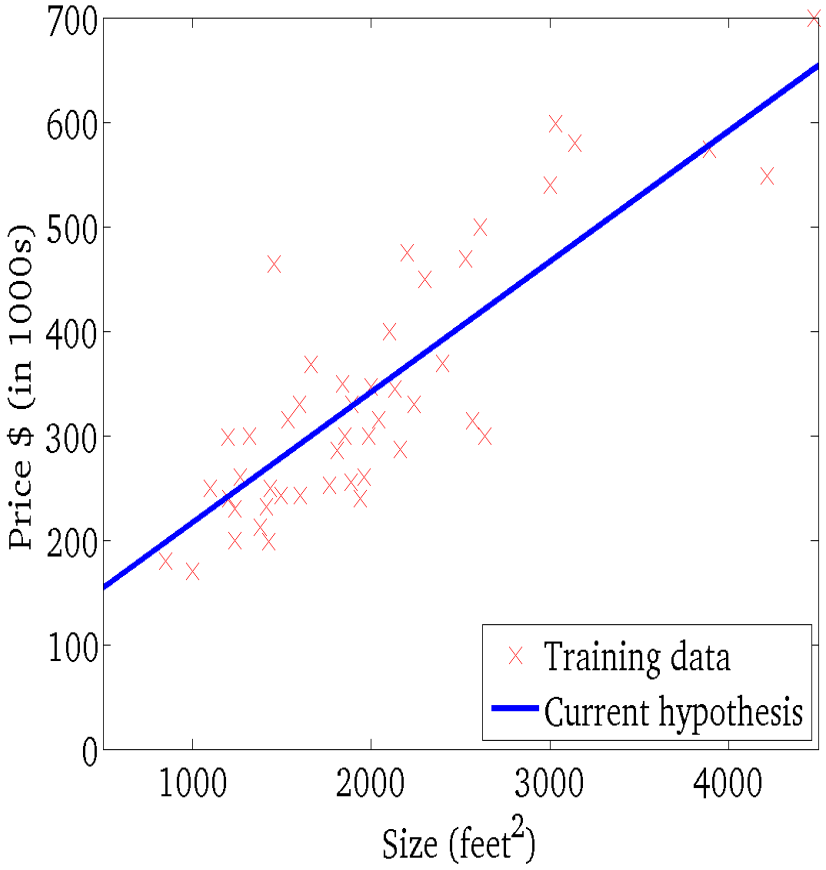
parámetros θ_0, θ_1



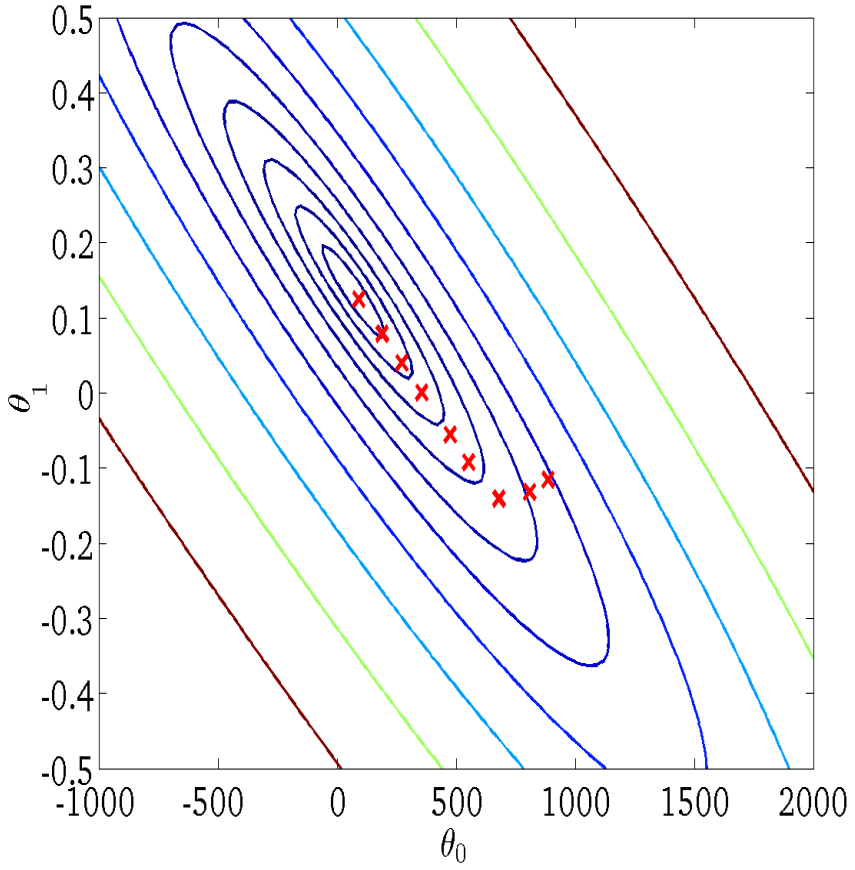


2. Regresión lineal

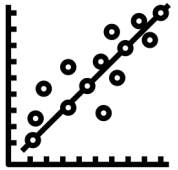
$h_\theta(x)$

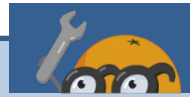


parámetros θ_0, θ_1



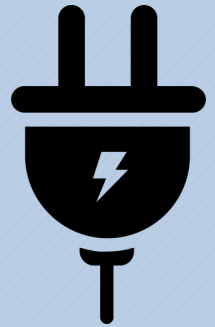
2. Regresión lineal





Práctica con Orange

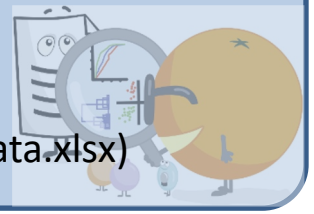
- **Objetivo: estimación del consumo energético**
  - Atributos y características del dataset:



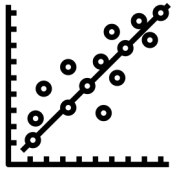
Attribute	Attribute name
X1	Relative compactness
X2	Surface area
X3	Wall area
X4	Roof area
X5	Overall height
X6	Orientation
X7	Glazing area
X8	Glazing area distribution
Y1	Heating load
Y2	Cooling load

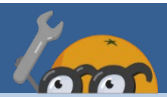
Data Set Characteristics:	Multivariate	**Number of Instances:**	768	**Area:**	Computer
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	8	**Date Donated**	2012-11-30
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	161295

https://archive.ics.uci.edu/ml/datasets/Energy+efficiency
(http://archive.ics.uci.edu/ml/machine-learning-databases/00242/ENB2012_data.xlsx)



2. Regresión lineal



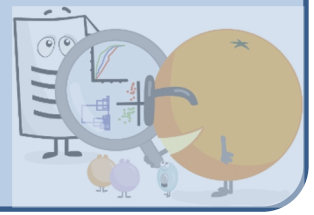


Práctica con Orange

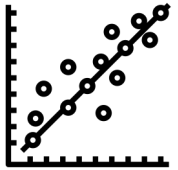
- **Objetivo: estimación del consumo energético**
  - Previsualización de los datos:

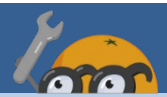
	A	B	C	D	E	F	G	H	I	J
1	X1	X2	X3	X4	X5	X6	X7	X8	Y1	Y2
2	0,98	514,50	294,00	110,25	7,00	2	0,00	0	15,55	21,33
3	0,98	514,50	294,00	110,25	7,00	3	0,00	0	15,55	21,33
4	0,98	514,50	294,00	110,25	7,00	4	0,00	0	15,55	21,33
5	0,98	514,50	294,00	110,25	7,00	5	0,00	0	15,55	21,33
6	0,90	563,50	318,50	122,50	7,00	2	0,00	0	20,84	28,28
7	0,90	563,50	318,50	122,50	7,00	3	0,00	0	21,46	25,38
8	0,90	563,50	318,50	122,50	7,00	4	0,00	0	20,71	25,16
9	0,90	563,50	318,50	122,50	7,00	5	0,00	0	19,68	29,60
10	0,86	588,00	294,00	147,00	7,00	2	0,00	0	19,50	27,30
11	0,86	588,00	294,00	147,00	7,00	3	0,00	0	19,95	21,97
12	0,86	588,00	294,00	147,00	7,00	4	0,00	0	19,34	23,49
13	0,86	588,00	294,00	147,00	7,00	5	0,00	0	18,31	27,87
14	0,82	612,50	318,50	147,00	7,00	2	0,00	0	17,05	23,77
15	0,82	612,50	318,50	147,00	7,00	3	0,00	0	17,41	21,46
16	0,82	612,50	318,50	147,00	7,00	4	0,00	0	16,95	21,16
17	0,82	612,50	318,50	147,00	7,00	5	0,00	0	15,98	24,93
18	0,79	637,00	343,00	147,00	7,00	2	0,00	0	28,52	37,73
19	0,79	637,00	343,00	147,00	7,00	3	0,00	0	29,90	31,27
20	0,79	637,00	343,00	147,00	7,00	4	0,00	0	29,63	30,93
21	0,79	637,00	343,00	147,00	7,00	5	0,00	0	28,75	39,44

https://archive.ics.uci.edu/ml/datasets/Energy+efficiency



2. Regresión lineal

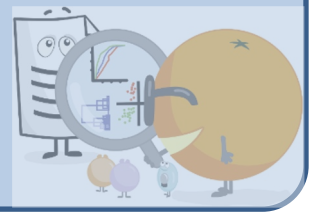




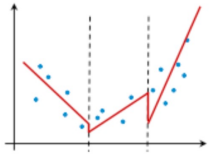
Práctica con Orange

- **Datasets de estudio**
  - Consumo eléctrico (.XLSX)
- **Algoritmo**
  - Regresión lineal
- **Visualización**
  - Gradient Descent (Add-on Educational)

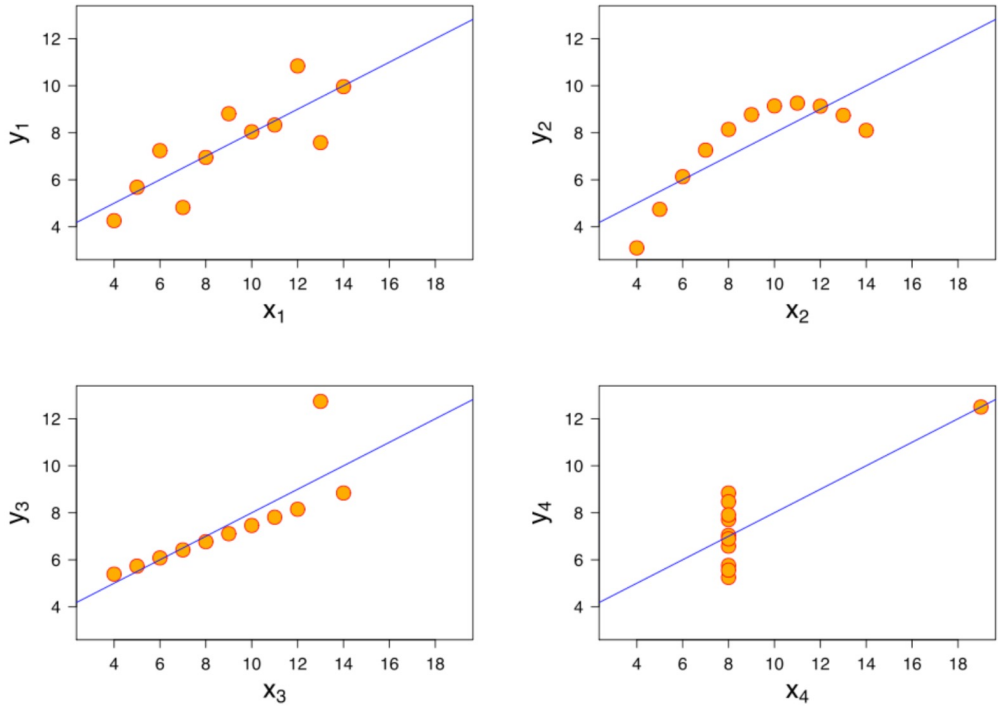
- **Métricas de evaluación**
  - MAE
  - RMSE
  - R2
- **Validación**
  - Validación cruzada con 10 bolsas



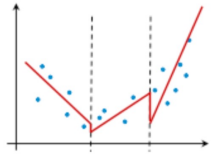
3. Árboles de Regresión



- Falta de expresividad de un modelo basado en regresión lineal múltiple:



3. Árboles de Regresión

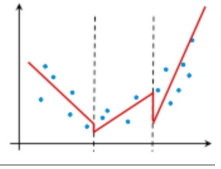


- Modelo de árboles multi-respuesta
- Genera un árbol asociando a cada hoja una regresión lineal o un valor numérico concreto.
- La división de nodos del árbol suele hacerse por reducción de varianza de la clase.

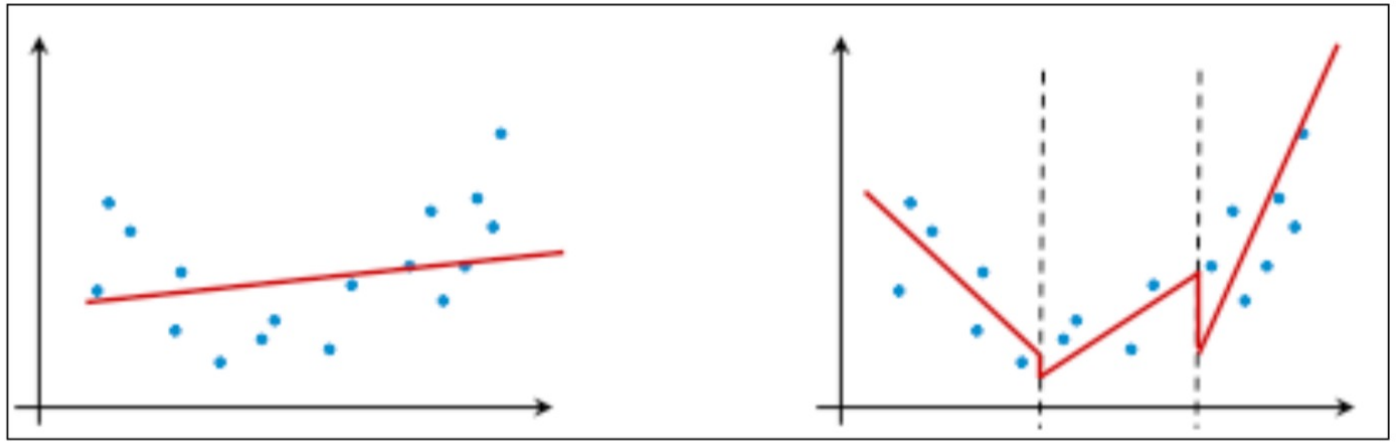
https://towardsdatascience.com/tree-based-methods-regression-trees-4ee5d8db9fe9

http://www2.stat.duke.edu/~rcs46/lectures_2017/08-trees/08-tree-regression.pdf

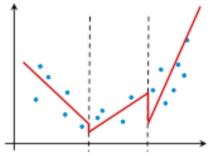
3. Árboles de Regresión



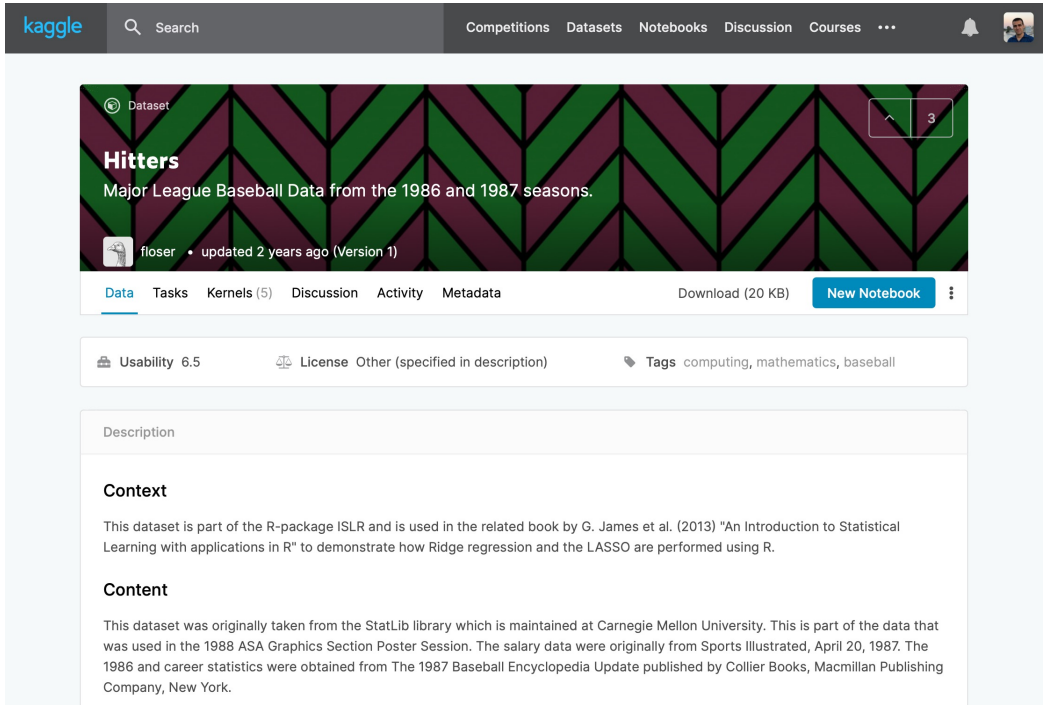
- Se aprenden regresiones lineales por tramos



3. Árboles de Regresión



- Ejemplo: predecir el salario de un jugador de béisbol

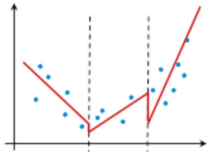


Hitters Dataset

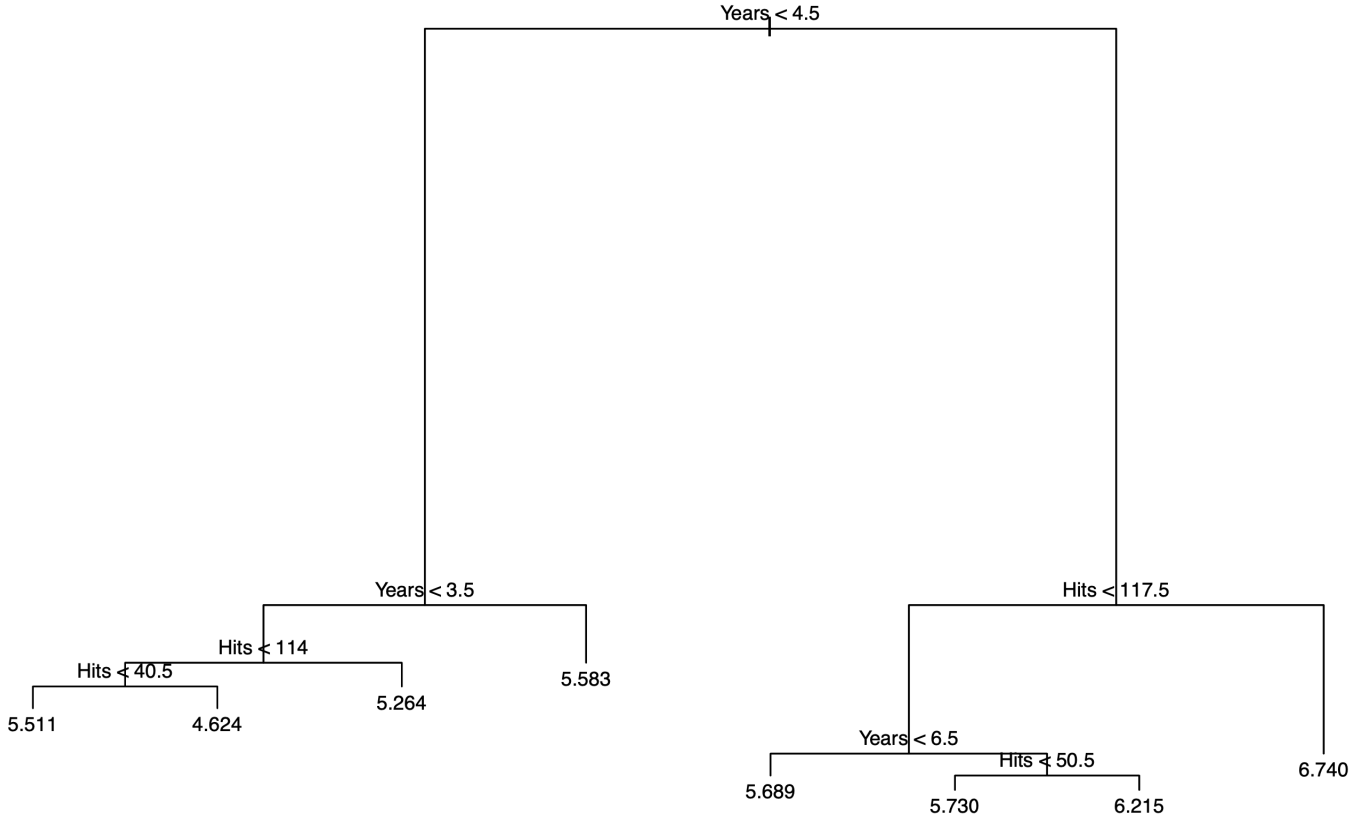
https://www.kaggle.com/floser/hitters

https://gist.github.com/keeganhines/59974f1ebef97bbaa44fb19143f90bad

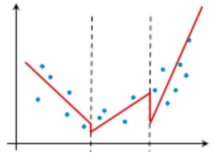
3. Árboles de Regresión



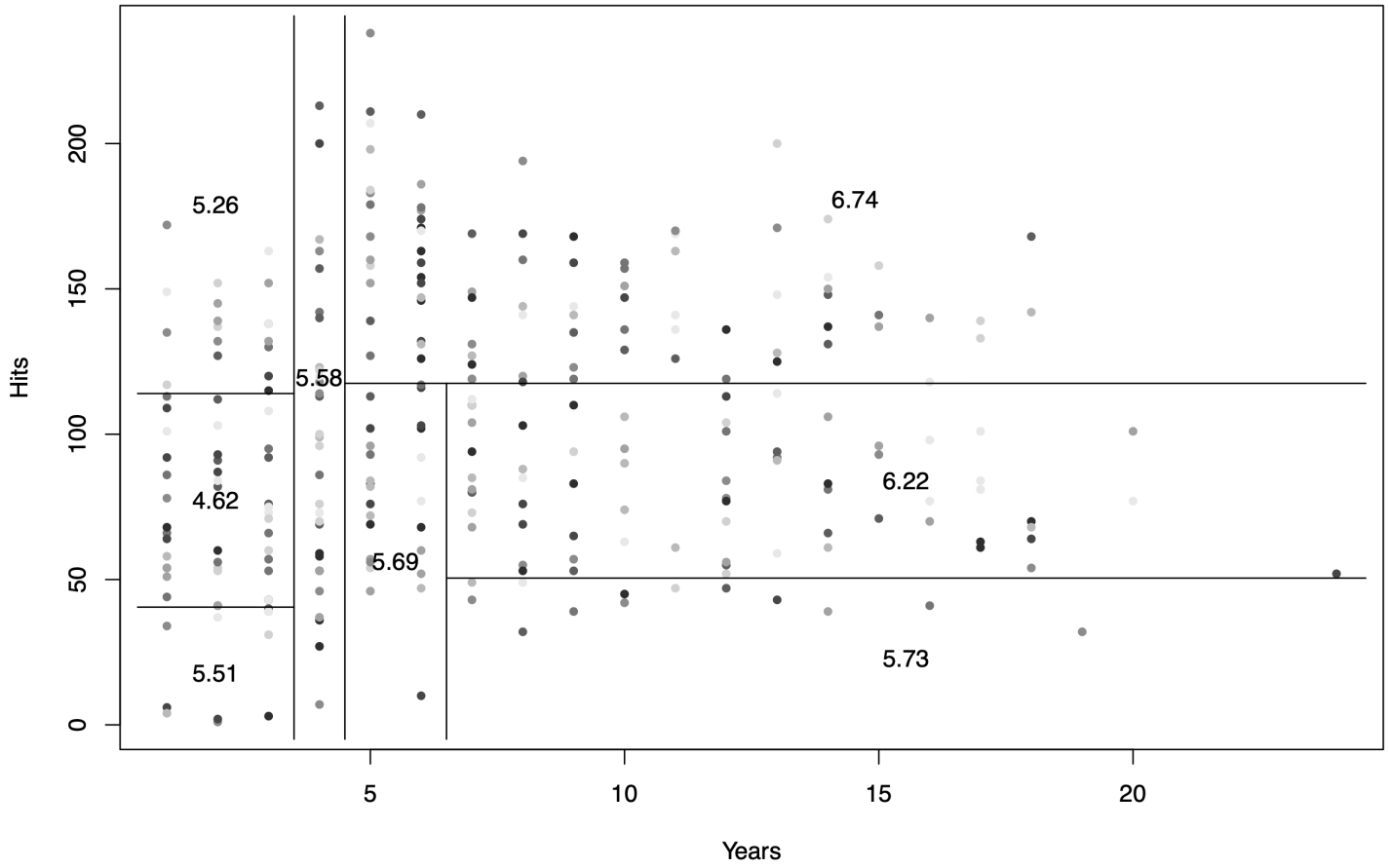
- Ejemplo: predecir el salario de un jugador de béisbol



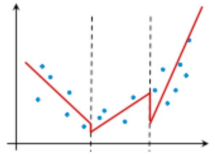
3. Árboles de Regresión

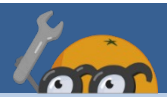


- Ejemplo: predecir el salario de un jugador de béisbol



3. Ejercicio de árboles de regresión

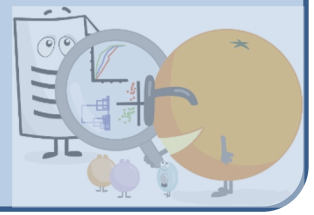




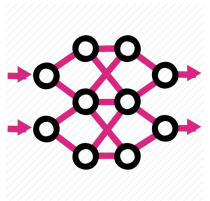
Práctica con Orange

- **Datasets de estudio**
  - Consumo eléctrico
  - Hitters

    "regression-tree-hitters-estudio.ows"
- **Algoritmo**
  - Árbol de regresión (Tree)
- **Visualización**
  - Tree Viewer
  - Pythagorean Tree
- **Métricas de evaluación**
  - MAE
  - RMSE
  - R2
- **Validación**
  - Validación cruzada con 10 bolsas

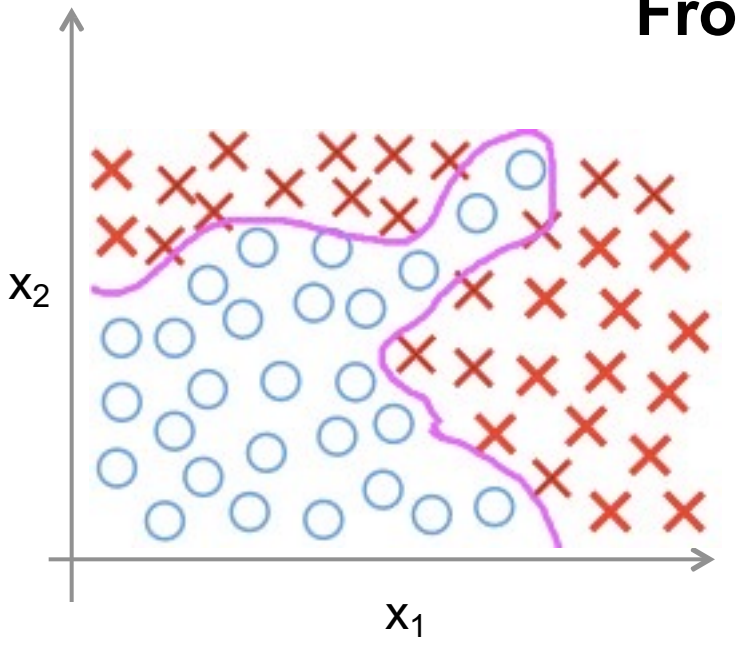


4. Redes neuronales

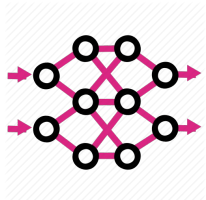


Válidas para clasificación y regresión

Fronteras de decisión no lineales



4. Redes neuronales



Ej. Visión por computador: Detección de coches

Coches






No coches

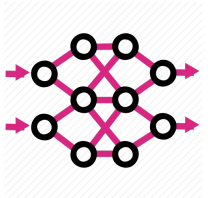





Testing:

Qué es esto?

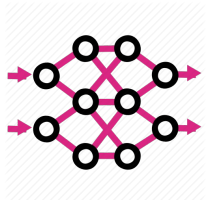
4. Redes neuronales



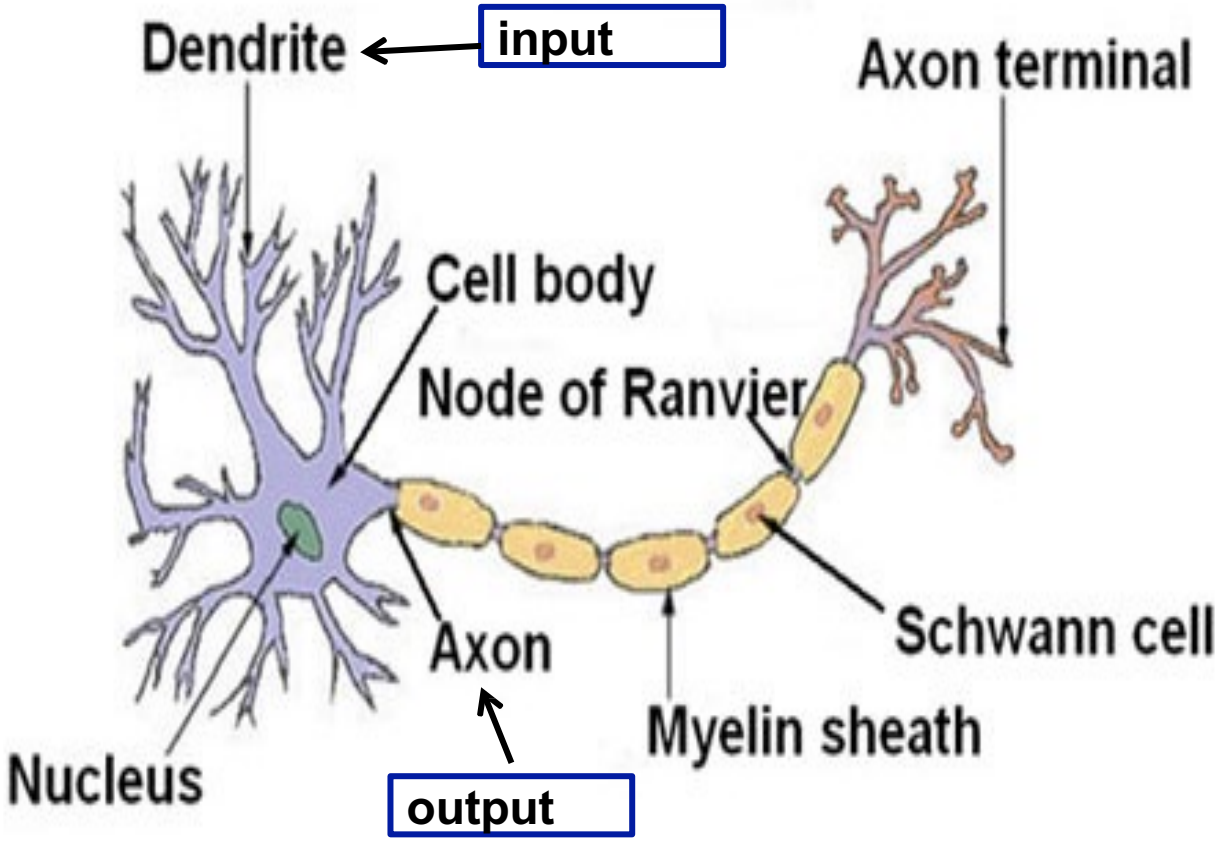
Neuronas y el cerebro

- Orígenes: Algoritmos que intentan imitar el cerebro.
- Se usó mucho en los 80 y principios de los 90 pero su popularidad disminuyó a final de los 90.
- Reciente resurgir: Deep Learning.

4. Redes neuronales

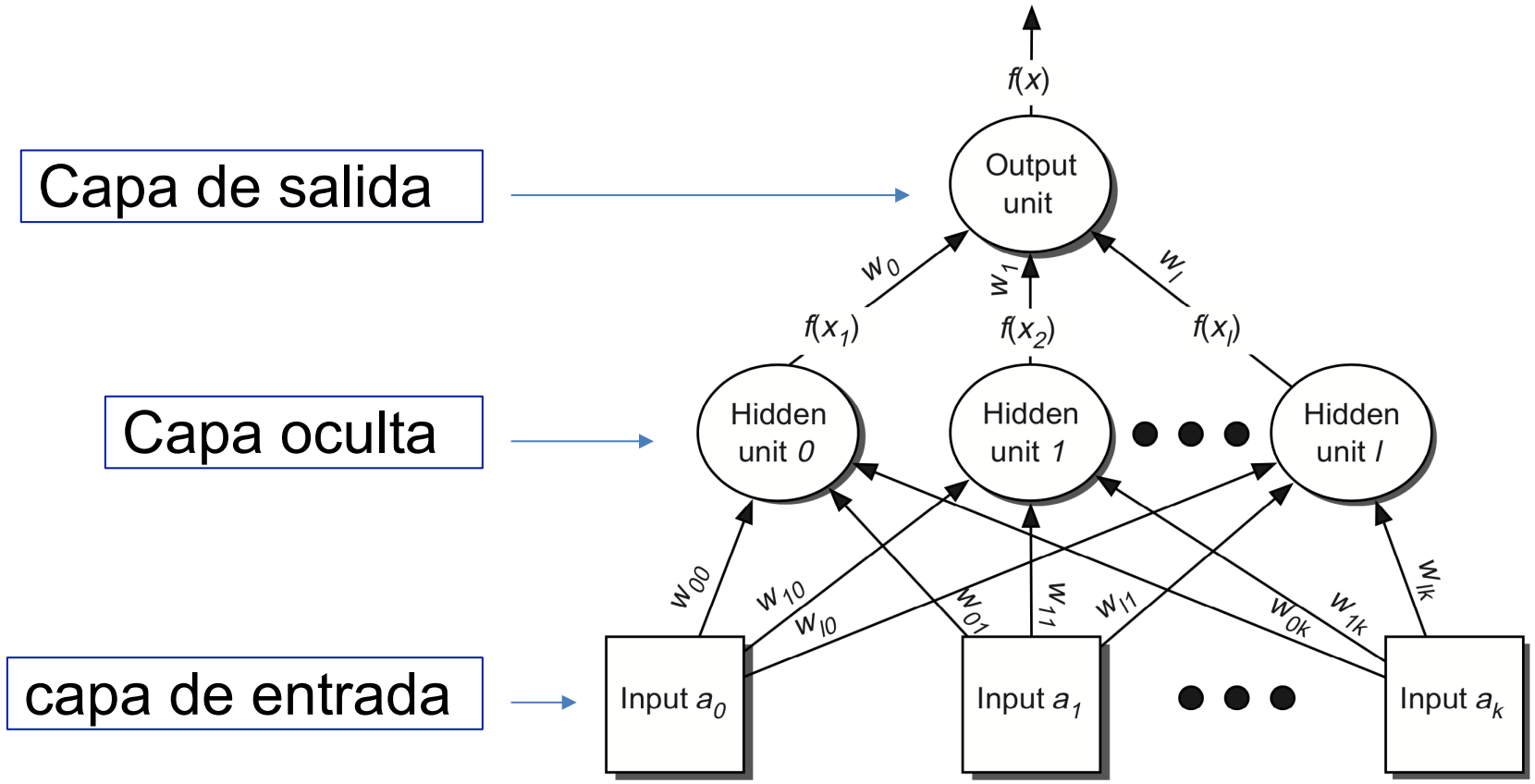


Neurona en el cerebro

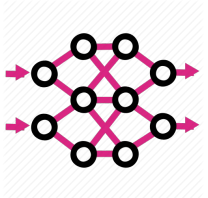


4. Redes neuronales

- Arquitectura de una red neuronal

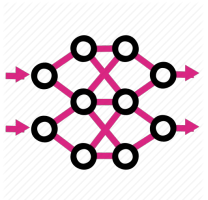


4. Redes neuronales



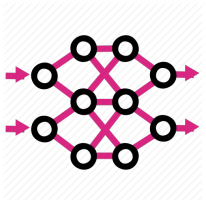
- Cada conexión entre neuronas es un peso que hay que determinar en el entrenamiento.
- El número de neuronas en la capa de entrada es el número de atributos del conjunto de datos.
- El número de neuronas de la capa oculta es un parámetro a determinar.

4. Redes neuronales



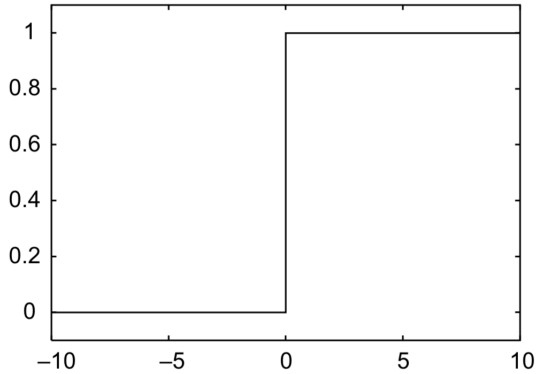
- Cada conexión entre neuronas es un peso que hay que determinar en el entrenamiento.
- Pesos:
  - Aquellos entre la capa de entrada y la capa oculta.
  - Aquellos entre la capa oculta y la de salida.

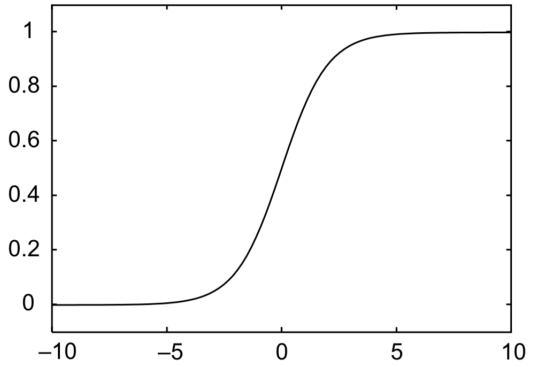
4. Redes neuronales



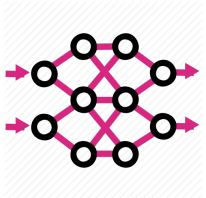
- Función de activación: Sigmoide

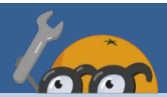
$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$





4. Ejercicio de redes neuronales

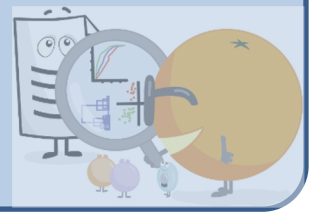




Práctica con Orange

- **Datasets de estudio**
  - Consumo eléctrico
  - Forest Fires (https://archive.ics.uci.edu/ml/datasets/Forest+Fires)

  "regression-neuralnetwork-forestfires-estudio.ows"
- **Algoritmo**
  - Neural Network
- **Métricas de evaluación**
  - MAE
  - RMSE
  - R2
- **Validación**
  - Validación cruzada con 10 bolsas



5. Ejercicio de regresión

- **Problema**
  - Determinar la **calidad de vinos** a partir de sus características físico-químicas.
  - Se tienen dos conjuntos de datos, uno de vino tinto y otro de vino blanco, ambos obtenidos a partir de la variedad llamada “Vinho Verde” con origen en el norte de Portugal.
- **Referencias**
  - https://archive.ics.uci.edu/ml/datasets/Wine+Quality
  - P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. **Modeling wine preferences by data mining from physicochemical properties**. *Decision Support Systems*, 47(4):547-553, 2009.
- **Características del conjunto de datos**

Data Set Characteristics:	Multivariate	**Number of Instances:**	4898	**Area:**	Business
Attribute Characteristics:	Real	**Number of Attributes:**	12	**Date Donated**	2009-10-07
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	478686