

Ingeniería de Conocimiento: Técnicas de Analítica de Datos

Jose Aguilar

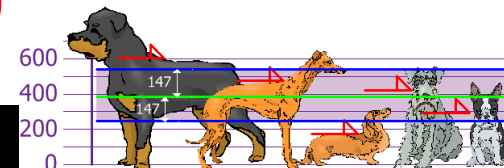
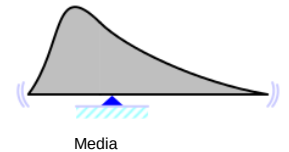
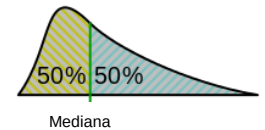
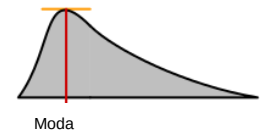
- La minería de datos
- El análisis estadístico
- El análisis predictivo
- La Correlación
- La Regresión
- Pronosticar
- Modelado de procesos
- Optimización
- Simulación

Dos categorías principales:

- * Estadísticas descriptivas
- * Estadística inferencial

- Usar **medidas de resumen** para describir la tendencia central de una distribución (media, moda, mediana)
- Utilizar la **dispersión o variabilidad** (desviación estándar, varianza, y el rango) para saber cómo se extienden los datos alrededor de la media.

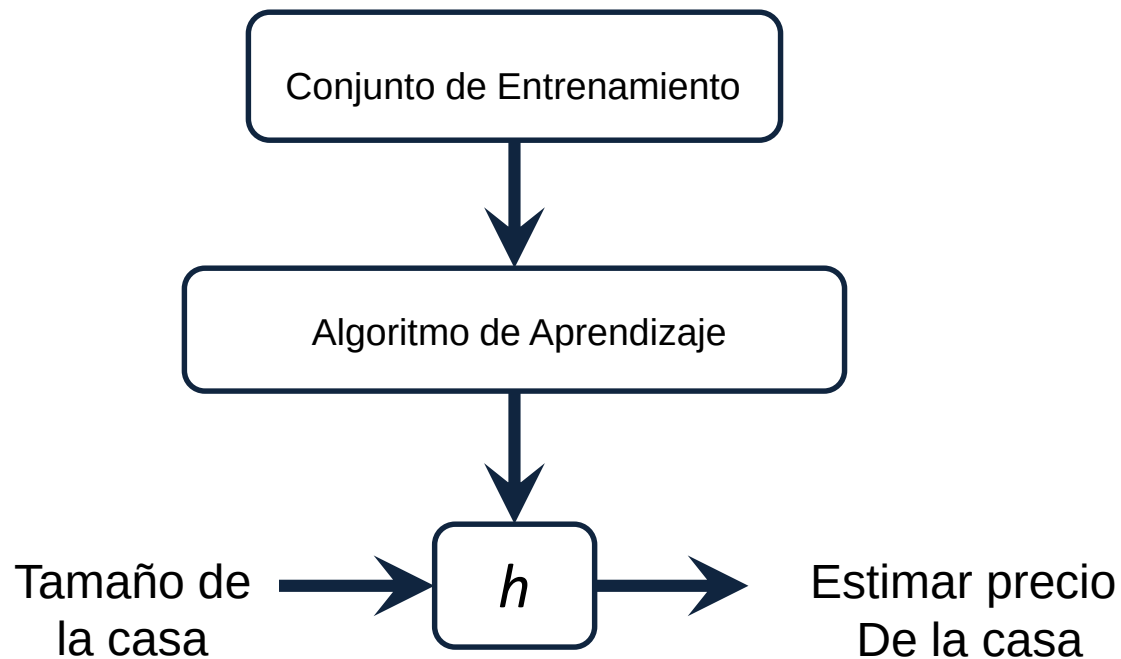
- Frecuencias (contar)
- Porcentaje
- Media (suma de todos los valores $\div$ no. de valores)
- Moda (valor más frecuente)
- Mediana (valor medio o posición central)
- Rango (intervalo entre el valor máximo y mínimo)
- Desviación estándar (variación esperada con respecto a la media)
- Varianza (la esperanza del cuadrado de la desviación)
- Ranqueo (clasificar, ordenar)



Compradores	Número
Hombre	
Viejo	6
Joven	4
Mujer	
Vieja	10
Joven	15

- **Más compradores femeninos que compradores masculinos**
- **Más jóvenes compradores femeninos que los compradores varones jóvenes**
- **Compradores masculinos jóvenes no están interesados en comprar en el centro comercial**

Construcción de modelos



ALGORITMOS DE APRENDIZAJE

1. SUPERVISADOS: predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos. Produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema.

- Clasificación, Predicción

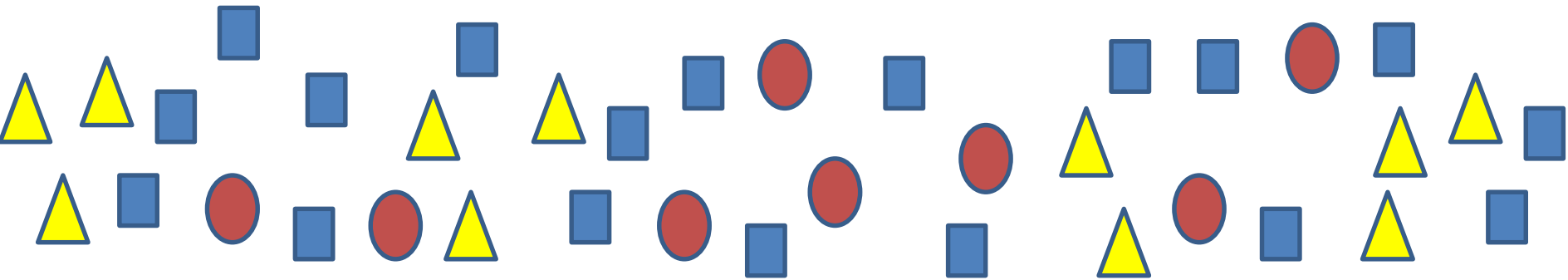
2. NO SUPERVISADOS: descubren patrones y tendencias en los datos, sin tener ningún tipo de conocimiento previo acerca de cuales son los patrones y categorías buscadas.

- Clustering, Análisis de enlace, Análisis de frecuencia

3. OTROS: Aprendizaje semisupervisado, Aprendizaje por refuerzo, Transducción, Aprendizaje multi-tarea, etc.

Aprendizaje supervisado

El **proceso de modelado** se realiza sobre un conjunto de **ejemplos** formado por **entradas al sistema** y la **respuesta que debería dar** para cada entrada.



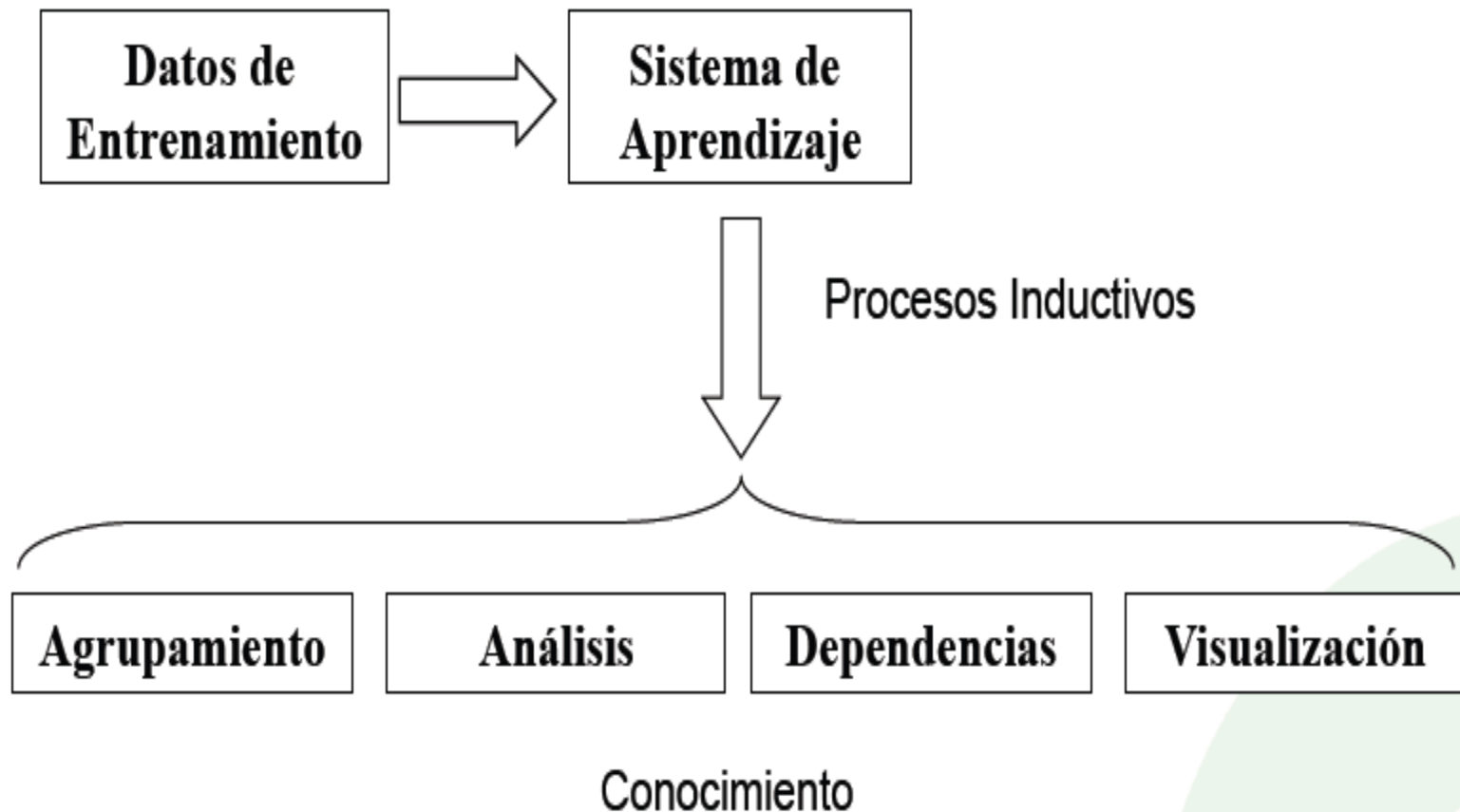
Aprendizaje no supervisado

Todo el **proceso de modelado** se lleva a cabo sobre un **conjunto de ejemplos formado tan sólo por entradas** al sistema.

No se tiene información sobre las **categorías de esos ejemplos**.

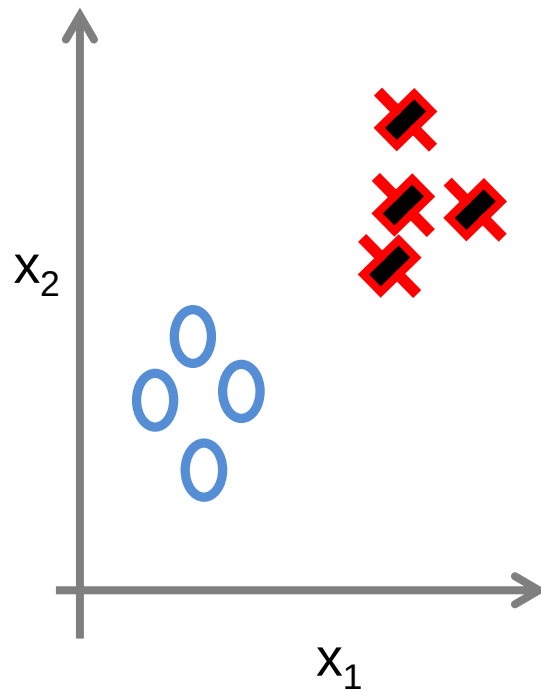
Por lo tanto, en este caso, el sistema tiene que ser capaz de **reconocer patrones** para poder etiquetar las nuevas entradas.

Aprendizaje no supervisado

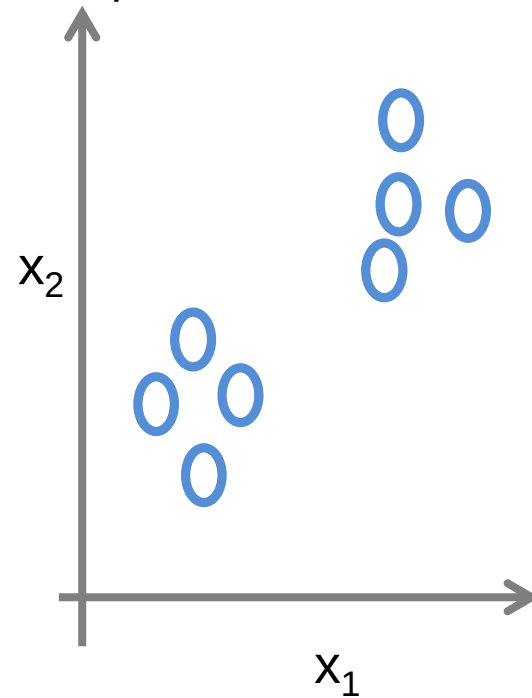


Aprendizaje no supervisado

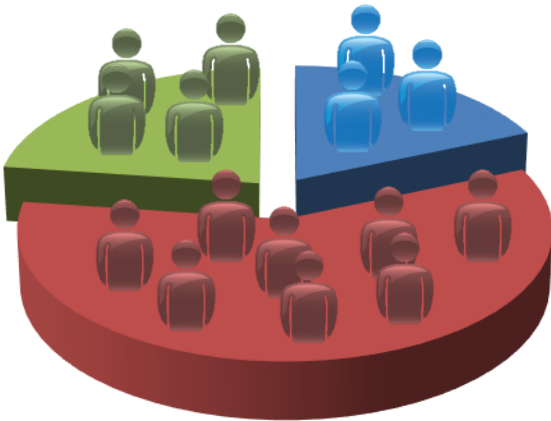
Aprendizaje supervisado



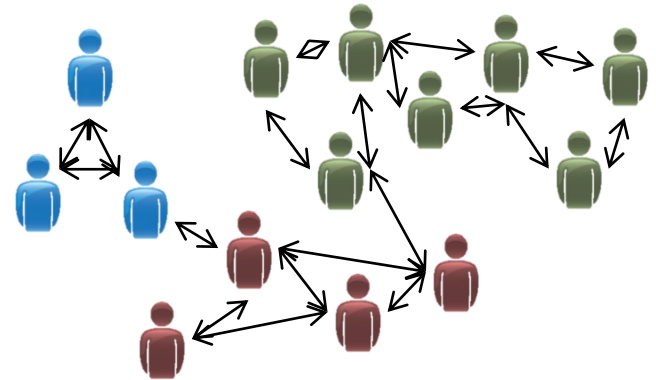
Aprendizaje no supervisado



Aprendizaje no supervisado



Segmentar mercado



Análisis de Redes
Sociales

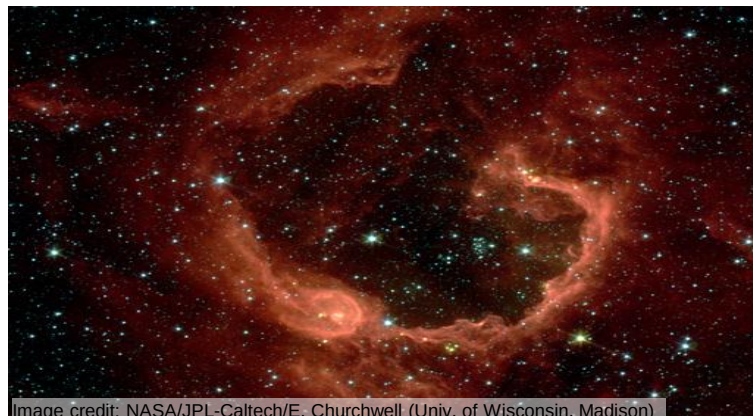


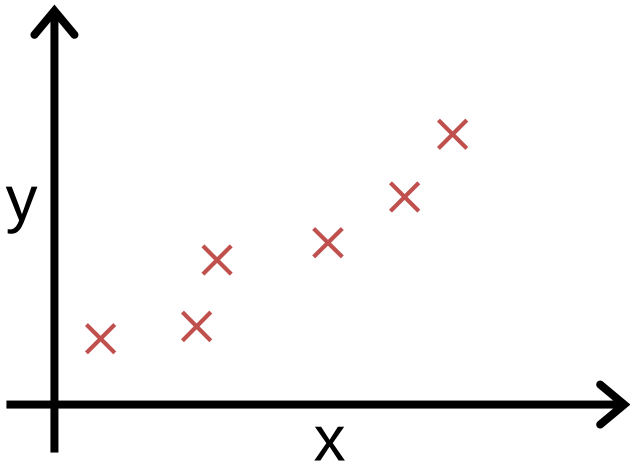
Image credit: NASA/JPL-Caltech/E. Churchwell (Univ. of Wisconsin, Madison)

Análisis datos Astronómicos

La Hipótesis

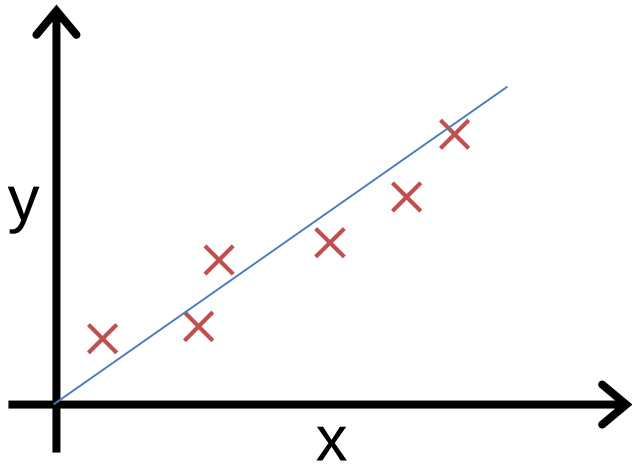
$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

La Hipótesis



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

La Hipótesis



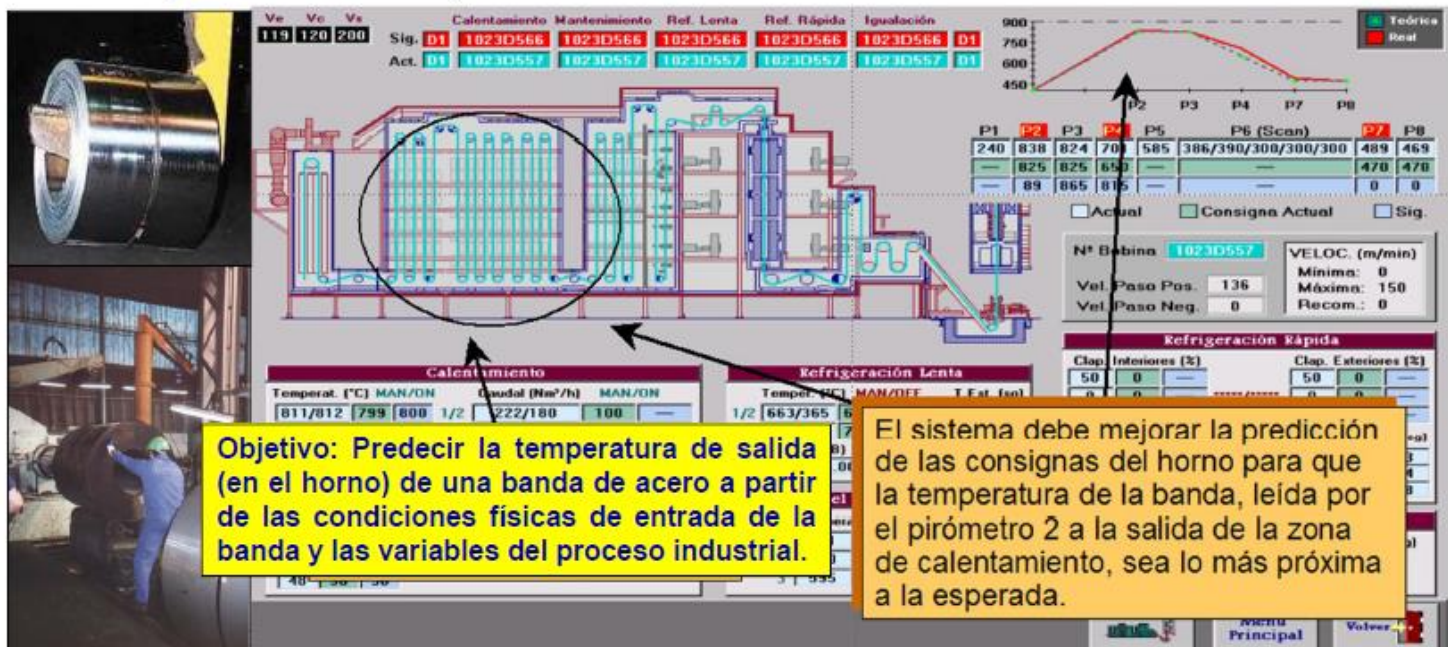
$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

Idea: Escoger θ_0, θ_1 para que $h_{\theta}(x)$
se acerque a y con el set de entrenamiento (x, y)

Regresión

Ante nuevos valores de variables independientes, el modelo obtenido debe permitir prever el valor numérico de la variable dependiente,

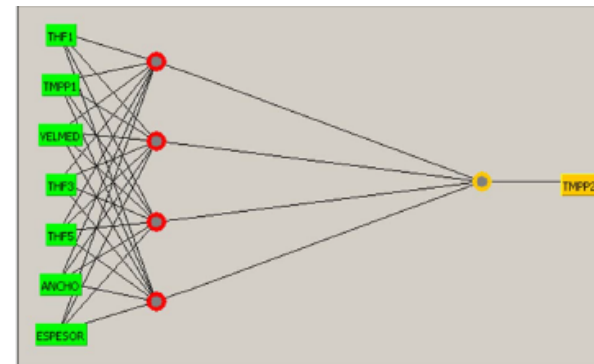
Por ejemplo:



Regresión

Ejemplo: A partir de la temperatura de entrada (TMPP1) de la banda de acero en el horno, el espesor(ESPEJOR), ancho(ANCHO) y temperaturas en diferentes zonas del horno (THF1, THF3,THC5) y velocidad de la banda dentro del horno(VELMED), se pretende predecir la temperatura de salida de la banda (TMPP2)

No.	THF1 Numeric	TMPP1 Numeric	TMPP2 Numeric	VELMED Numeric	THF3 Numeric	THF5 Numeric	ANCHO Numeric	ESPEJOR Numeric
1	770.0	212.0	754.0	145.0	800.0	821.0	1250.0	0.60000...
2	772.0	210.0	749.0	145.0	801.0	822.0	1250.0	0.60000...
3	778.0	211.0	761.0	138.0	808.0	833.0	1250.0	0.60000...
4	758.0	221.0	755.0	145.0	788.0	811.0	1250.0	0.60000...
5	752.0	223.0	752.0	105.0	783.0	806.0	1250.0	0.60000...
6	766.0	213.0	753.0	132.0	796.0	814.0	1250.0	0.60000...
7	799.0	216.0	786.0	131.0	829.0	848.0	1250.0	0.60000...
8	812.0	244.0	830.0	100.0	842.0	864.0	1350.0	0.67000...
9	808.0	248.0	830.0	80.0	838.0	856.0	1350.0	0.67000...
10	804.0	250.0	826.0	80.0	834.0	850.0	1350.0	0.67000...
11	803.0	249.0	824.0	80.0	833.0	848.0	1350.0	0.67000...
12	800.0	248.0	821.0	89.0	830.0	847.0	1350.0	0.67000...
13	802.0	248.0	823.0	90.0	832.0	852.0	1350.0	0.67000...
14	798.0	250.0	822.0	81.0	830.0	847.0	1350.0	0.67000...
15	782.0	222.0	789.0	98.0	812.0	822.0	1250.0	0.60000...
16	779.0	225.0	780.0	115.0	809.0	818.0	1250.0	0.60000...
17	781.0	216.0	774.0	120.0	811.0	823.0	1250.0	0.60000...
18	795.0	224.0	791.0	120.0	826.0	841.0	1250.0	0.60000...
19	811.0	244.0	814.0	120.0	840.0	864.0	1350.0	0.67000...
20	815.0	247.0	823.0	120.0	845.0	870.0	1350.0	0.67000...
21	807.0	244.0	810.0	120.0	837.0	855.0	1250.0	0.60000...
22	805.0	216.0	803.0	120.0	835.0	850.0	1250.0	0.60000...



A partir de los datos de 1979 bobinas, se desarrolla una red neuronal con cuatro neuronas en la capa oculta y que predice la temperatura de salida de la banda de acero con un error medio absoluto de 7,36 grados centígrados.

Regresión

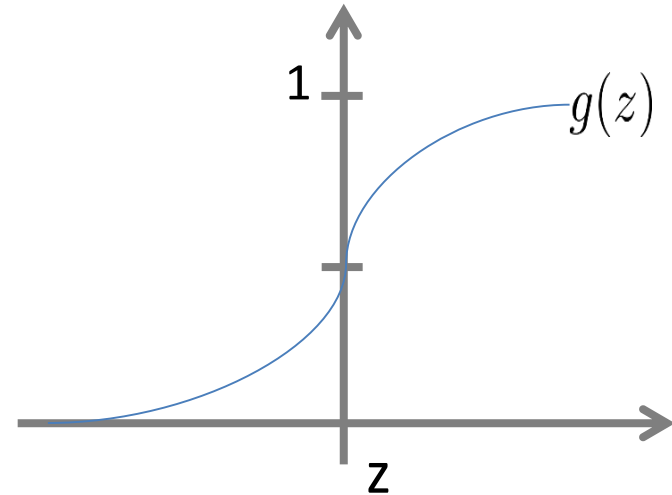
Regresión Lógica : $0 \leq h_{\theta}(x) \leq 1$

Cercano a clasificación

Regresión Lógica

$$h_{\theta}(x) = g(\theta^T x)$$

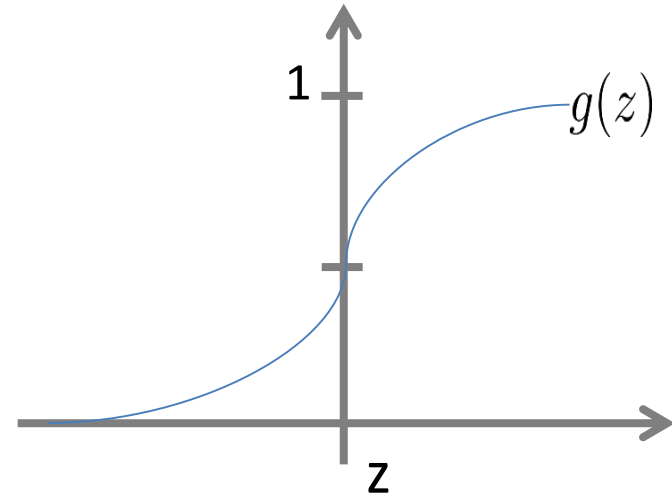
$$g(z) = \frac{1}{1+e^{-z}}$$



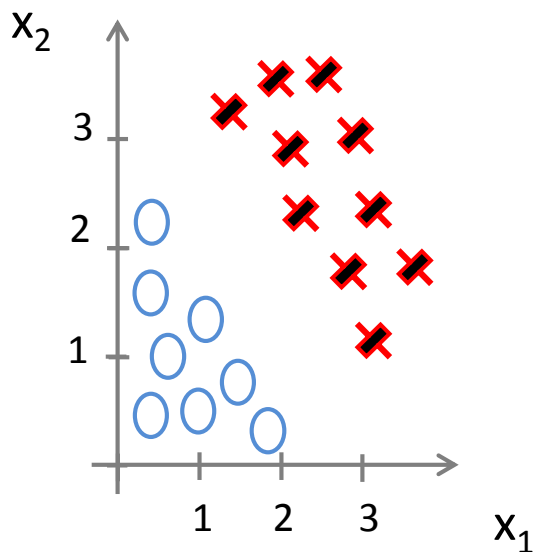
Regresión Lógica: Barrera de decisión

predice “y=1” si $h_{\theta}(x) \geq 0.5$

predice “y=0” si $h_{\theta}(x) < 0.5$

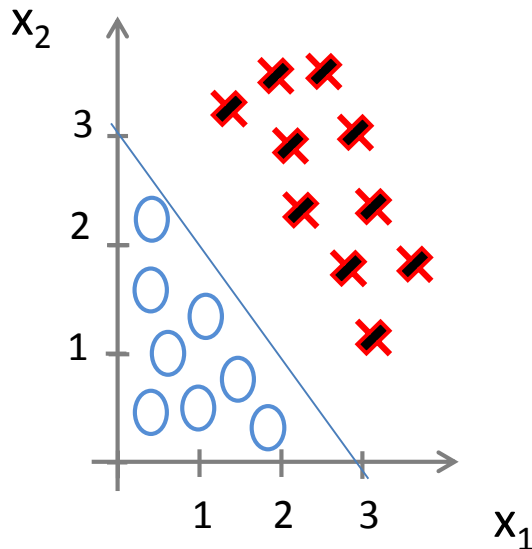


Regresión Lógica: Barrera de decisión



$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

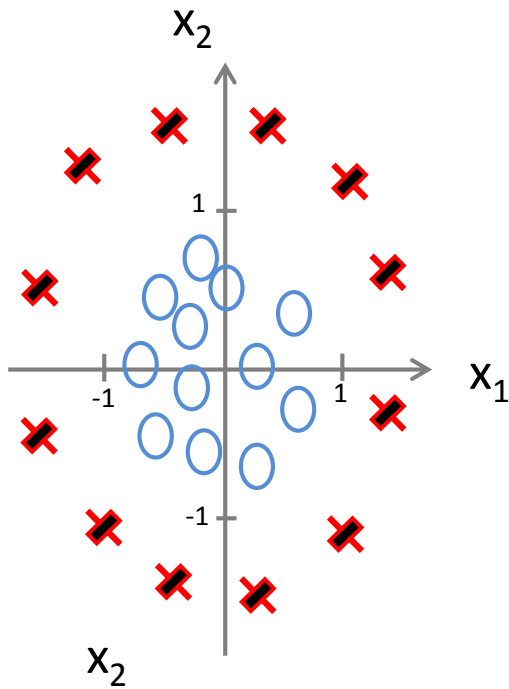
Regresión Lógica: Barrera de decisión



$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

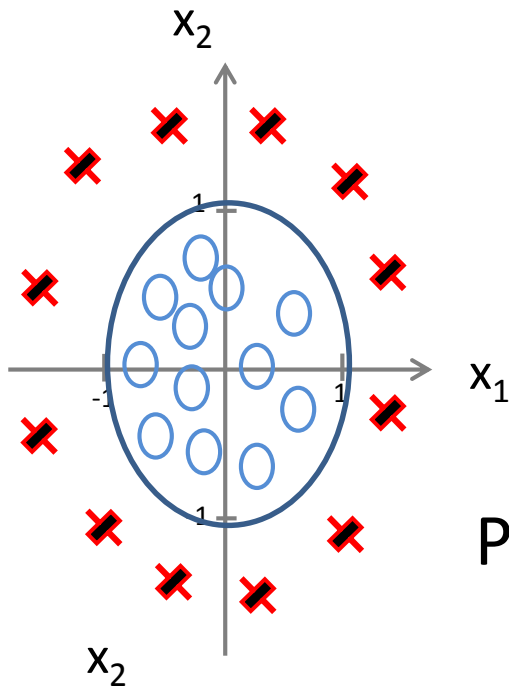
Predice “ $y=1$ ” si $-3 + x_1 + x_2 \geq 0$

Regresión Lógica: Barrera de decisión



$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

Regresión Lógica: Barrera de decisión



$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

Predice “y=1” si $-1 + x_1^2 + x_2^2 \geq 0$

Regresión Lógica:

Entrenamiento y Función de costos

Conjunto de
entrenamiento

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$$

:

n ejemplos

$$x \in \begin{bmatrix} x_0 \\ x_1 \\ \dots \\ x_n \end{bmatrix}$$

$$y \in \{0, 1\}$$

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

Como escoger el parámetro θ ?

Regresión Lógica: Función de costos

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(h_{\theta}(x^{(i)}), y^{(i)})$$

$$\text{Cost}(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases}$$

$$\min_{\theta} J(\theta)$$

Gradiente descendiente

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^m y^{(i)} \log h_{\theta}(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_{\theta}(x^{(i)})) \right]$$

Queremos $\min_{\theta} J(\theta)$:

Repeat {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \quad (\text{simultaneamente actualizar } \theta_j)$$

}

Problema de Optimización!!!

Dado θ , queremos calcular

- $J(\theta)$
- $\frac{\partial}{\partial \theta_j} J(\theta)$

(for $j=0, \dots, n$) {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

}

Función de costos

Regresión lineal

Hipótesis:

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

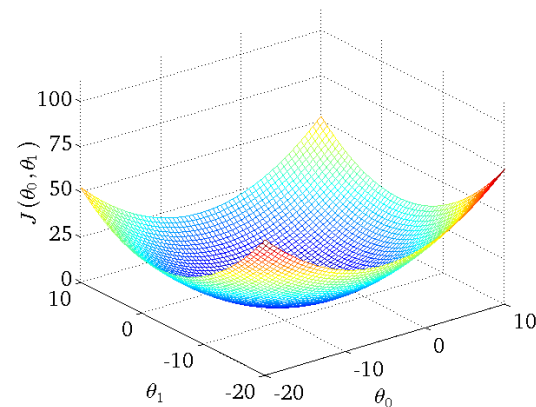
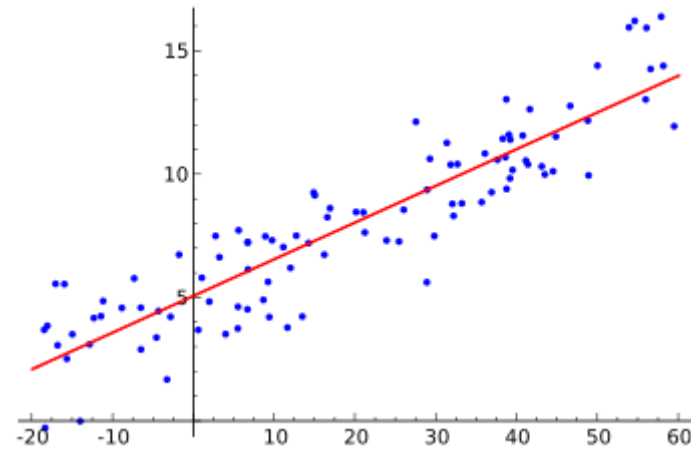
Parámetros:

$$\theta_0, \theta_1$$

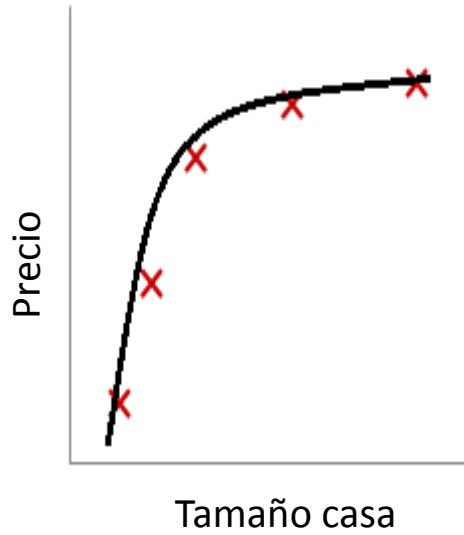
Función de Costo:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

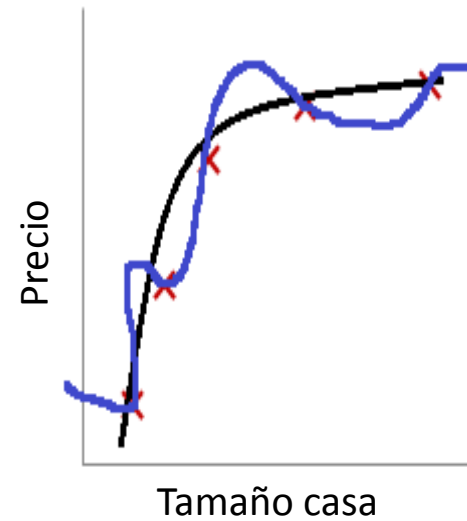
Objetivo: minimize $J(\theta_0, \theta_1)$
 θ_0, θ_1



Función de costo



$$\theta_0 + \theta_1 x + \theta_2 x^2$$



$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

Penalizar $\theta_3 \theta_4$

$$\min_{\theta} \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

Gradiente descendiente con regresión lineal

Gradiente descendiente

Se tiene la función de costos $J(\theta_0, \theta_1)$

Se quiere $\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$

En general:

- Comenzar con algún valor para $J(\theta_0, \theta_1)$
- Cambiar θ_0, θ_1 para reducir $J(\theta_0, \theta_1)$ hasta encontrar el mínimo.

Gradiente descendiente con regresión lineal

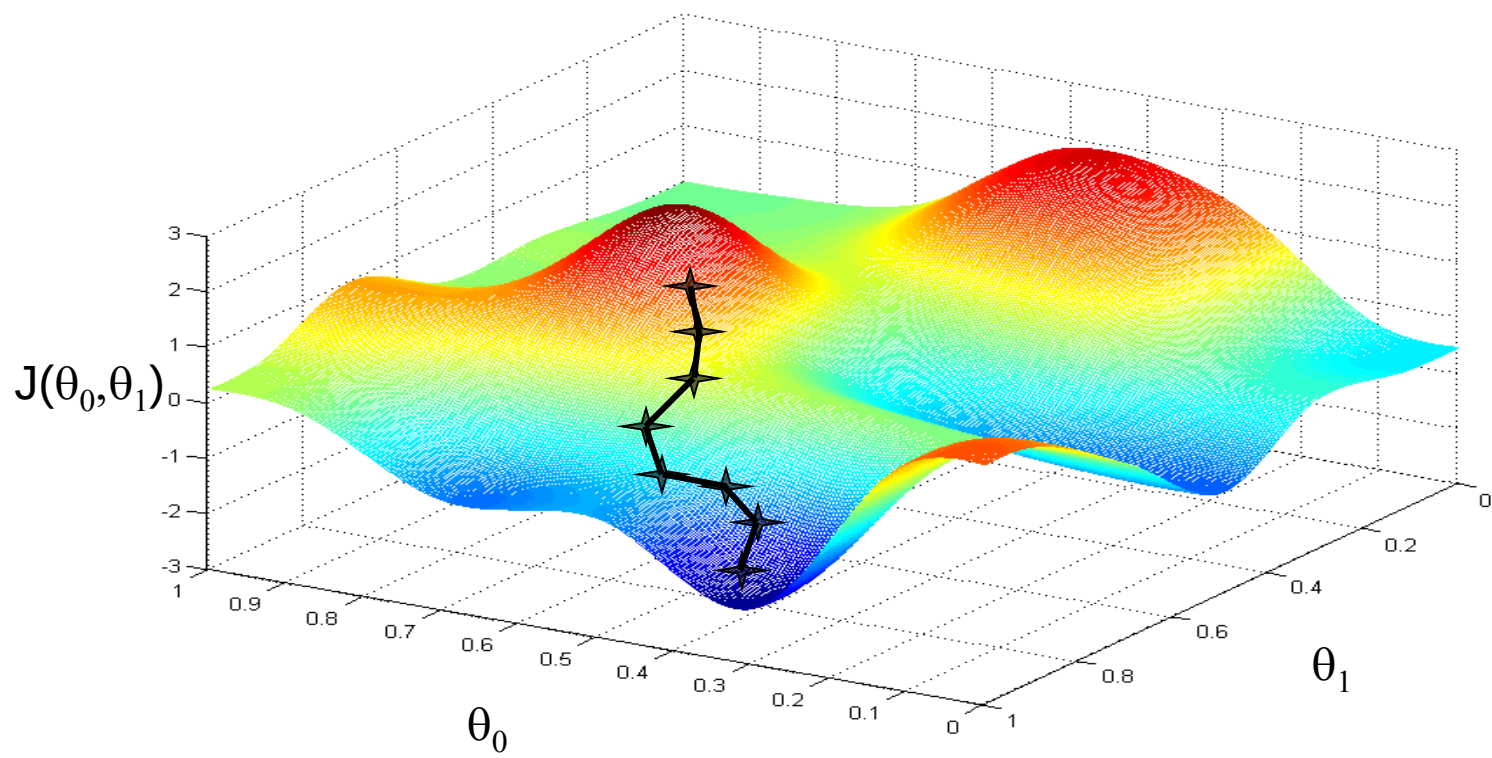
Algoritmo del Gradiente Descendiente

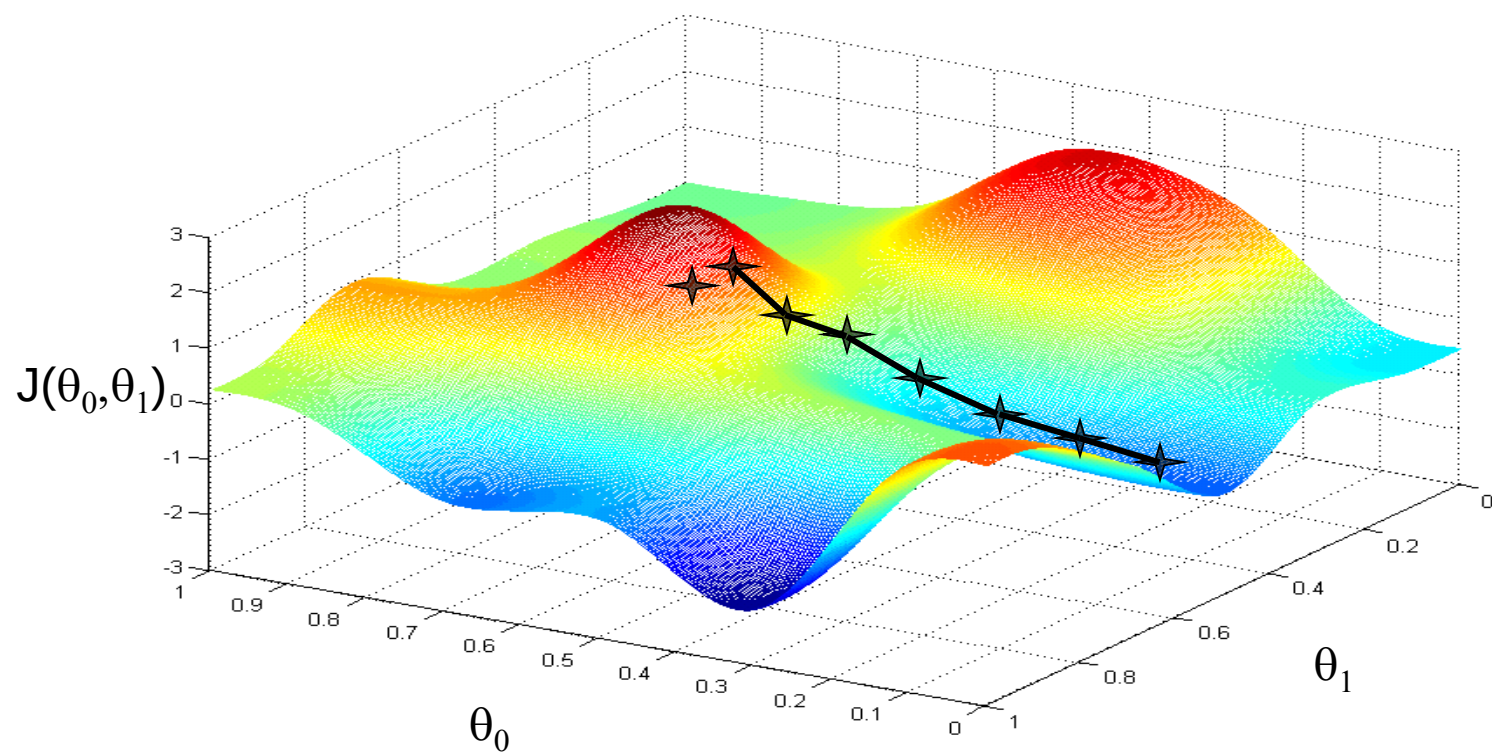
repeat until convergence {
 $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$
 (for $j = 1$ and $j = 0$)
}

Modelo de Regresión Lineal

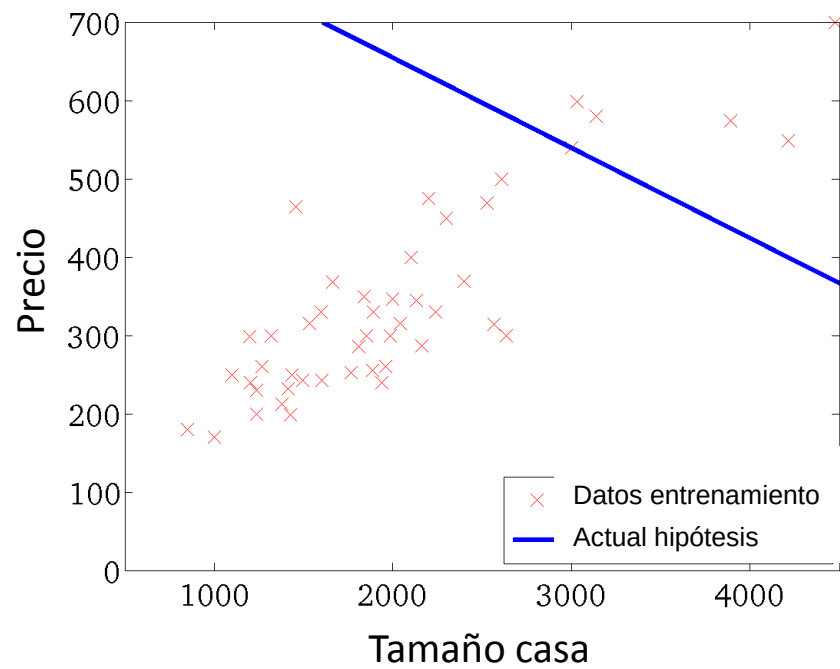
$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

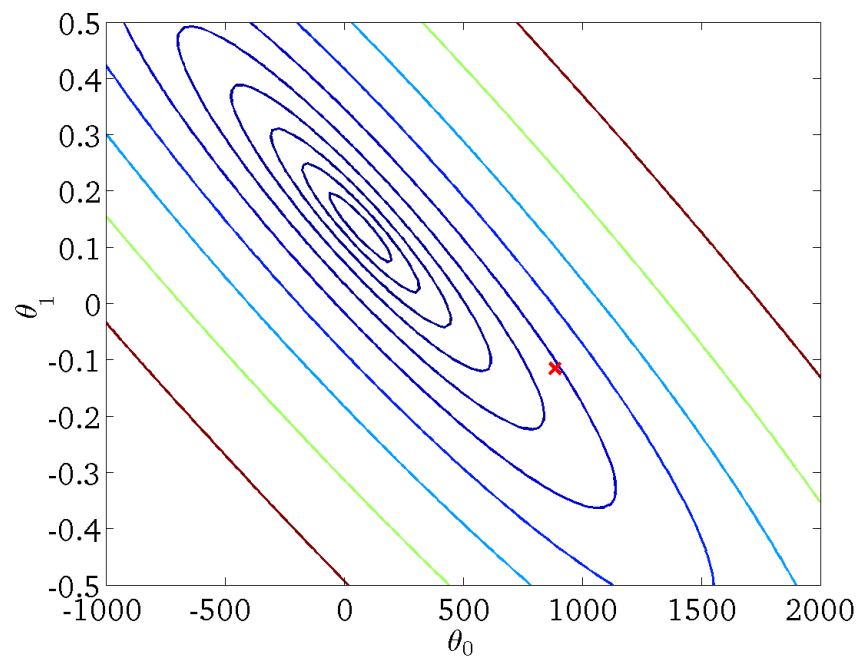




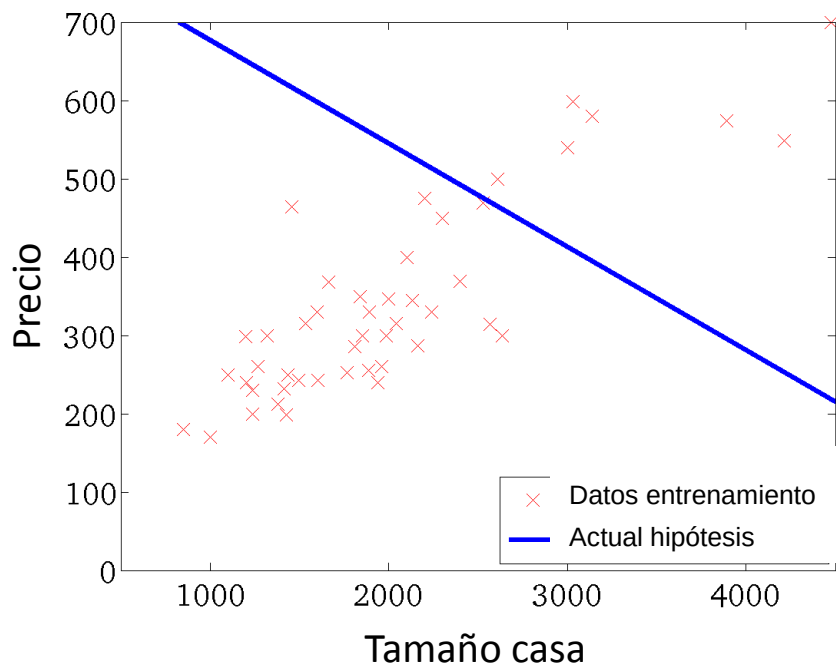
$$h_{\theta}(x)$$



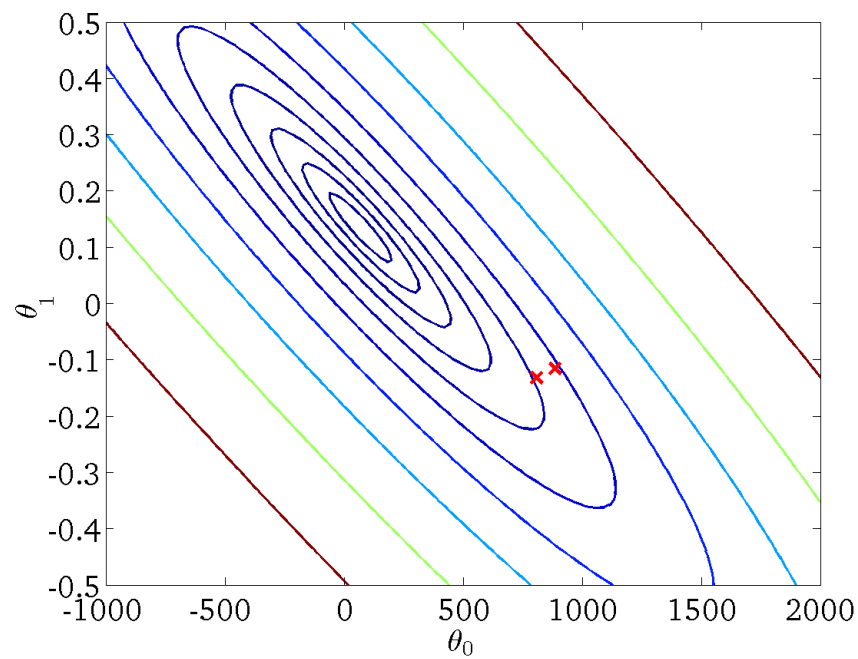
$$J(\theta_0, \theta_1)$$



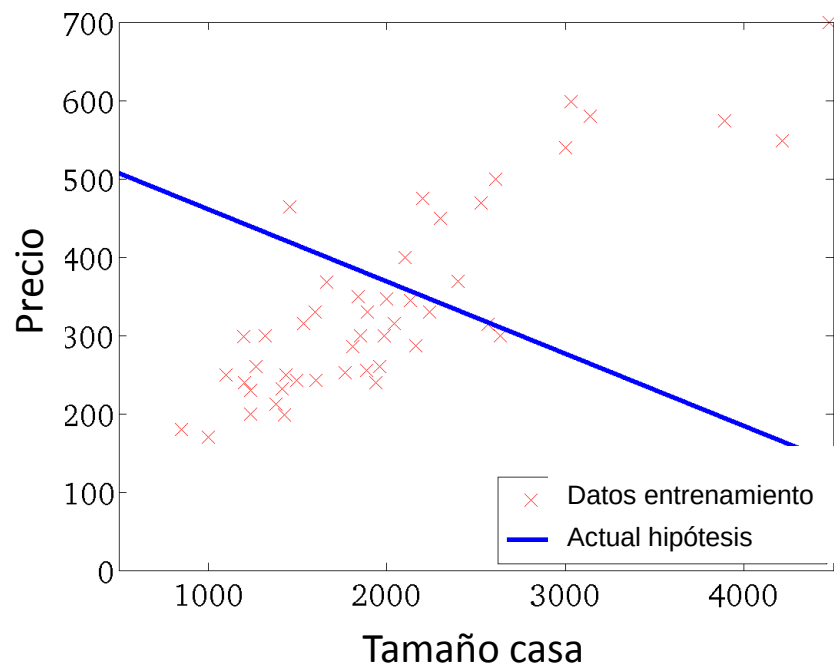
$$h_{\theta}(x)$$



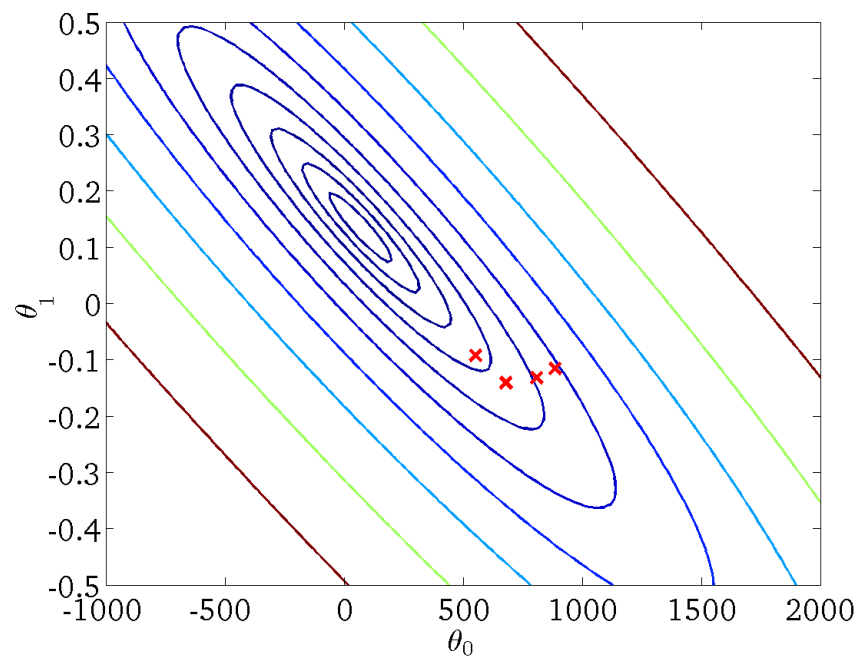
$$J(\theta_0, \theta_1)$$



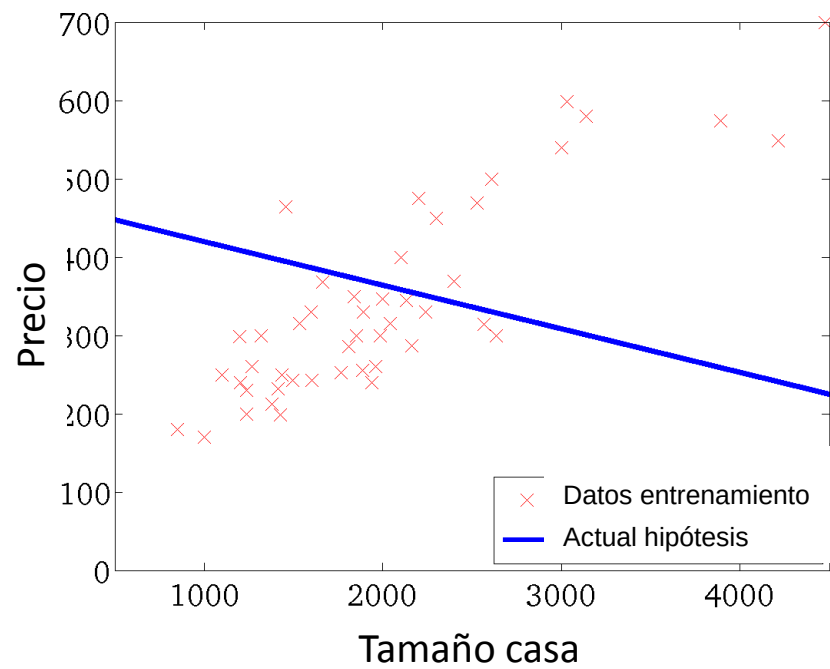
$$h_{\theta}(x)$$



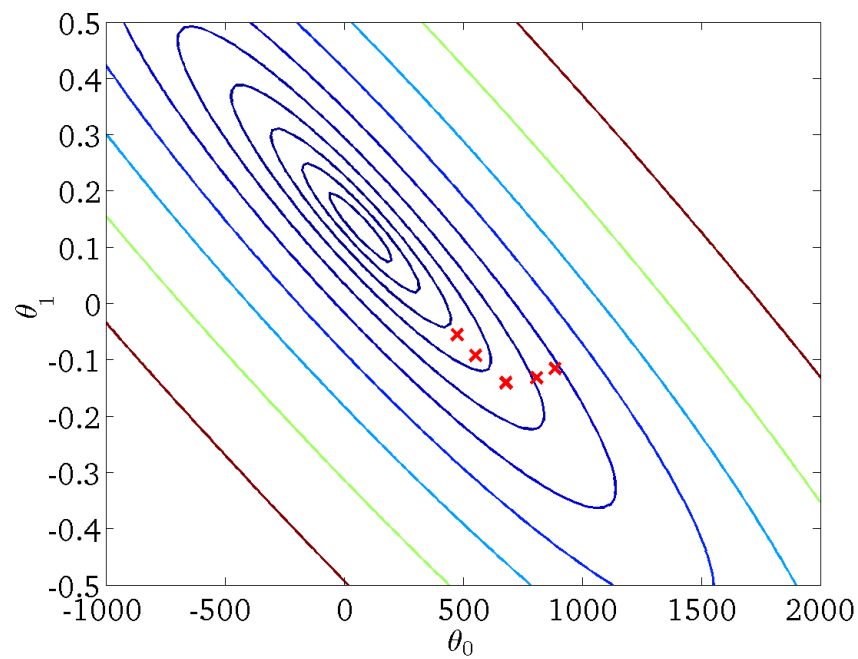
$$J(\theta_0, \theta_1)$$



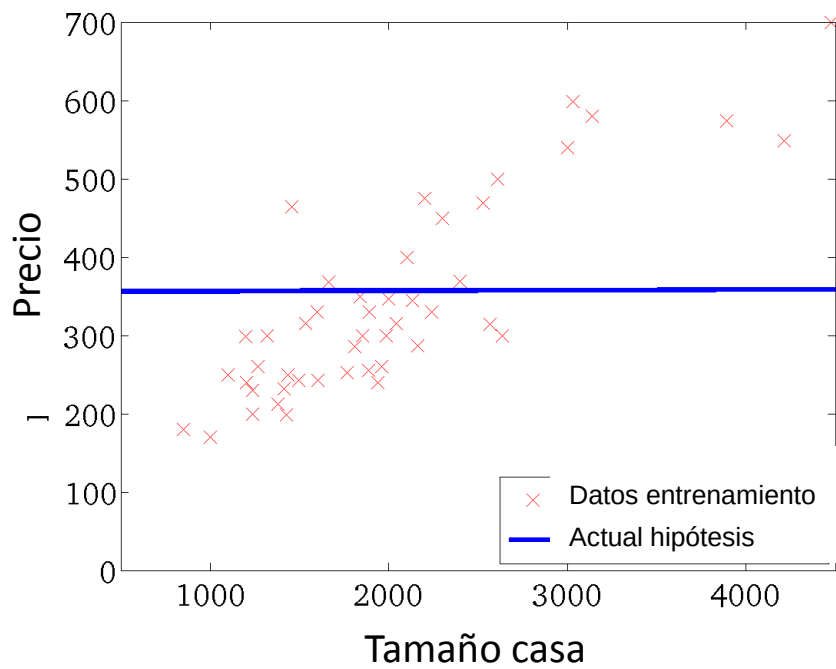
$$h_{\theta}(x)$$



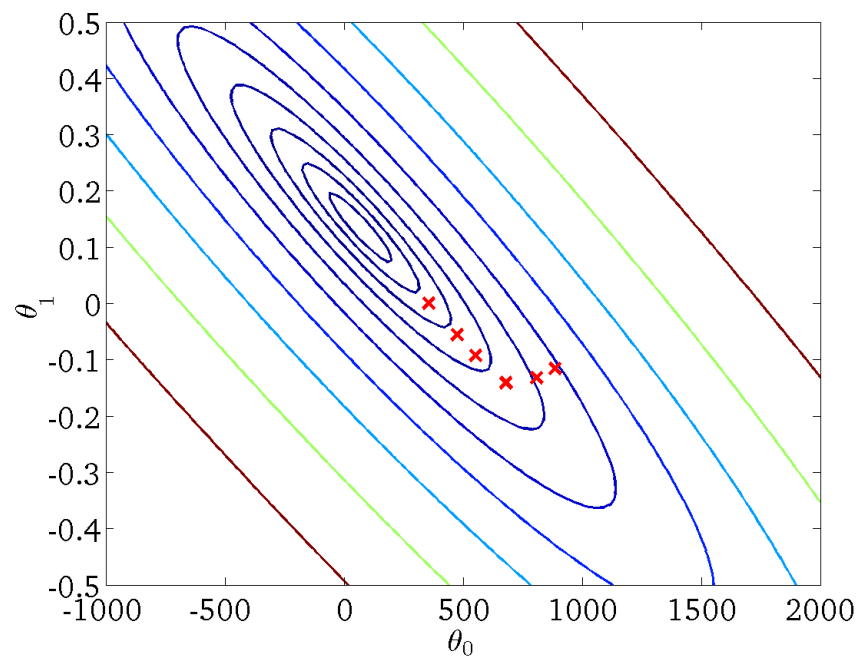
$$J(\theta_0, \theta_1)$$



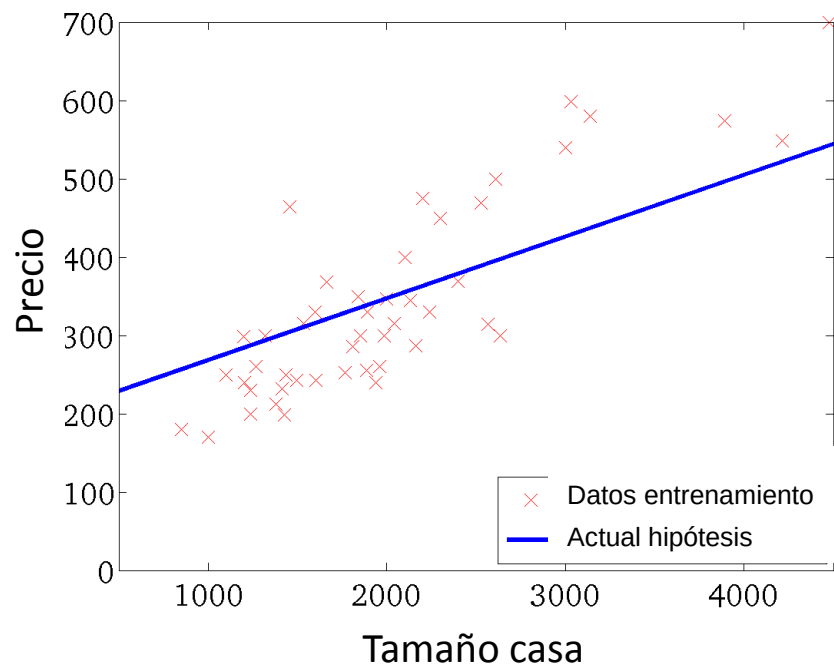
$$h_{\theta}(x)$$



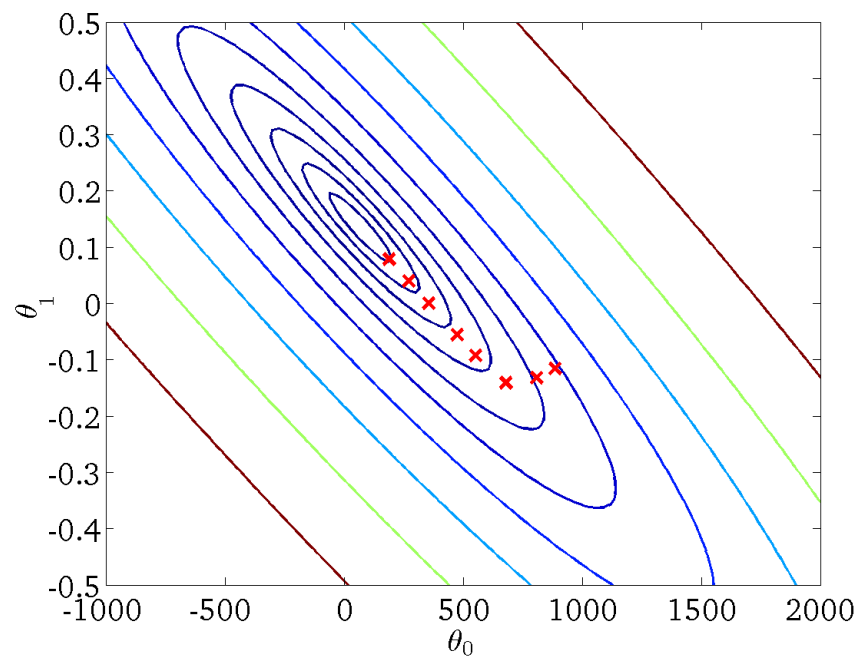
$$J(\theta_0, \theta_1)$$



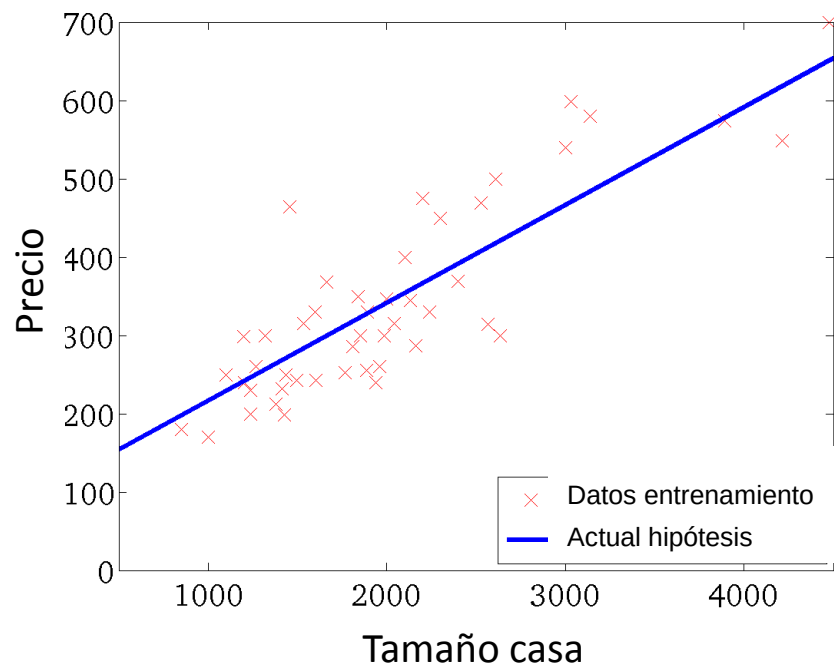
$$h_{\theta}(x)$$



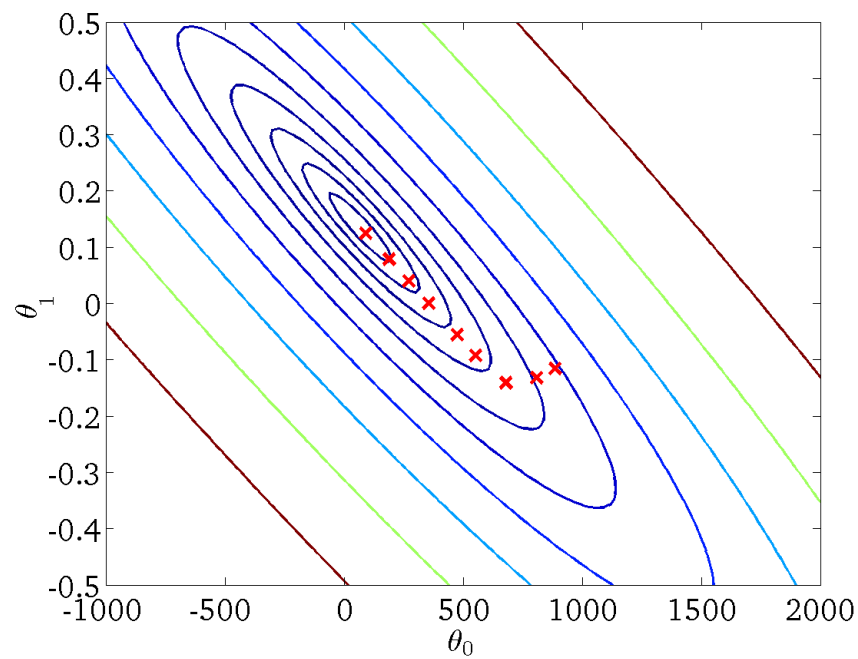
$$J(\theta_0, \theta_1)$$



$$h_{\theta}(x)$$

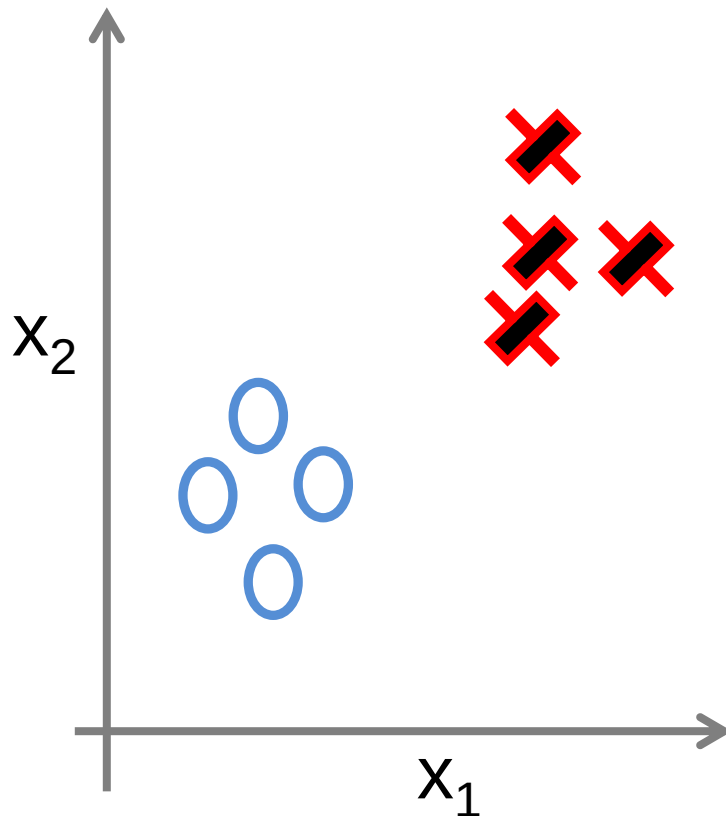


$$J(\theta_0, \theta_1)$$

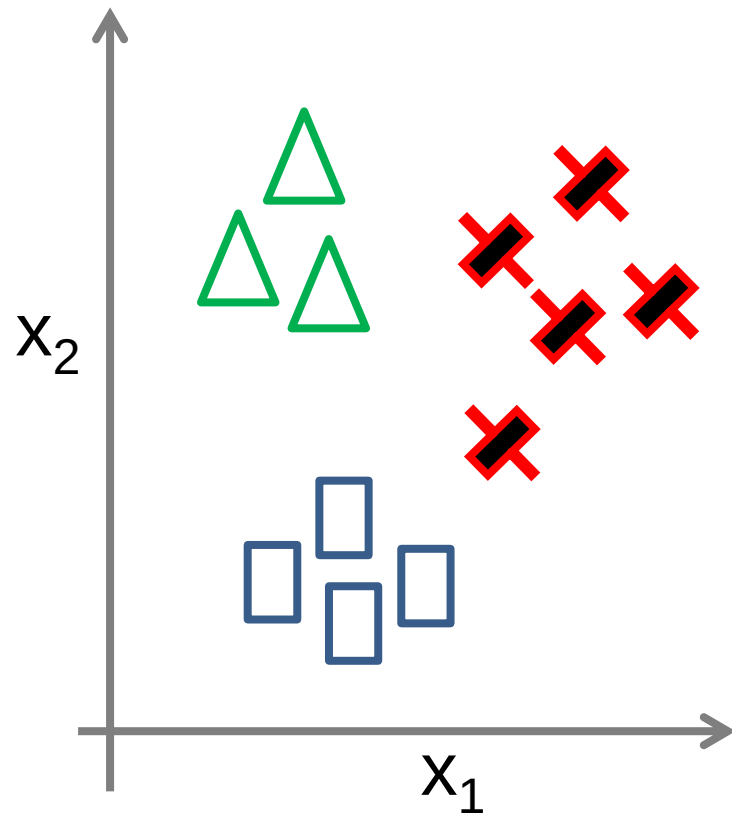


Clasificación multi-clases

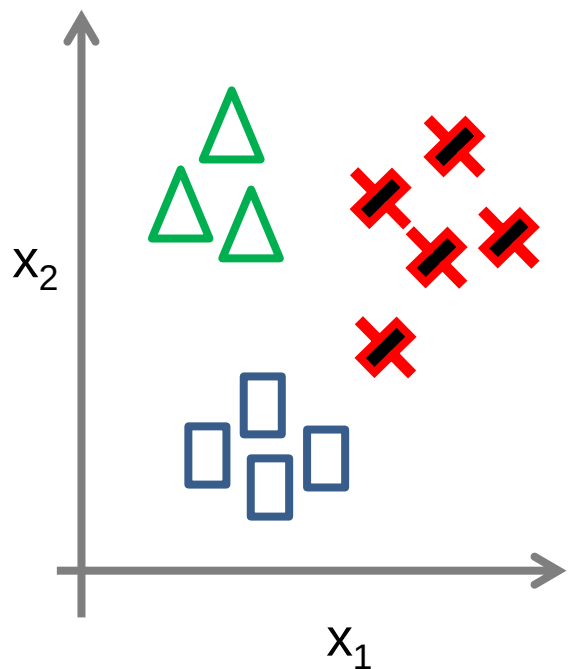
Clasificación Binaria



Clasificación Multi-clases



Una-vs-todos (uno-vs-resto):

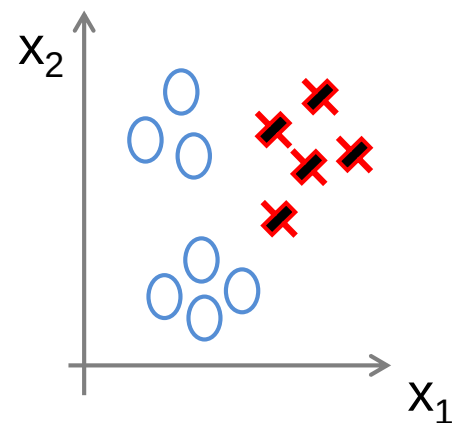
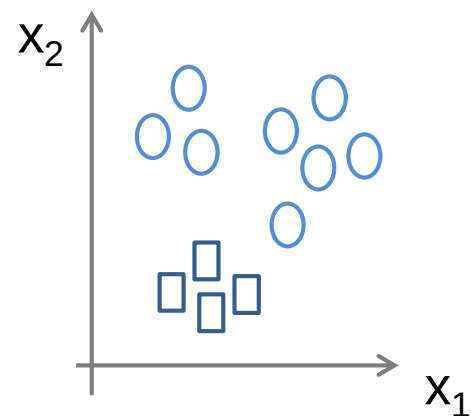
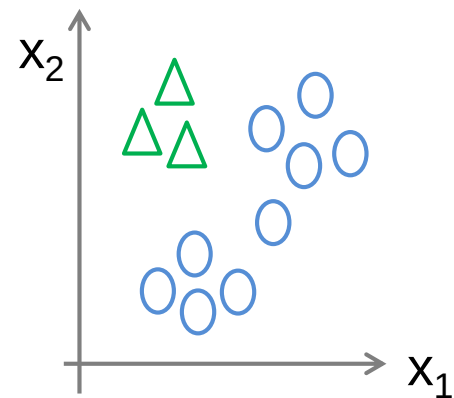


Clase 1: 

Clase 2: 

Clase 3: 

$$h_{\theta}^{(i)}(x) = P(y = i|x; \theta) \quad (i = 1, 2, 3)$$



Uno vs Todo

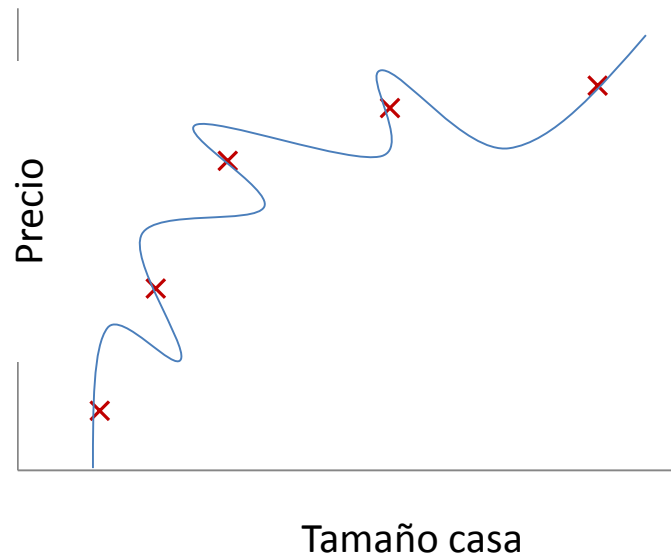
Entrenar a un clasificador de regresión lógica $h_{\theta}^{(i)}(x)$
por cada clase i para predecir la probabilidad de
 $y = i$

Con **una nueva entrada** x , para hacer **predicción** se
toma la clase i que maximice

$$\max_i h_{\theta}^{(i)}(x)$$

Evaluando una hipótesis

$$h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

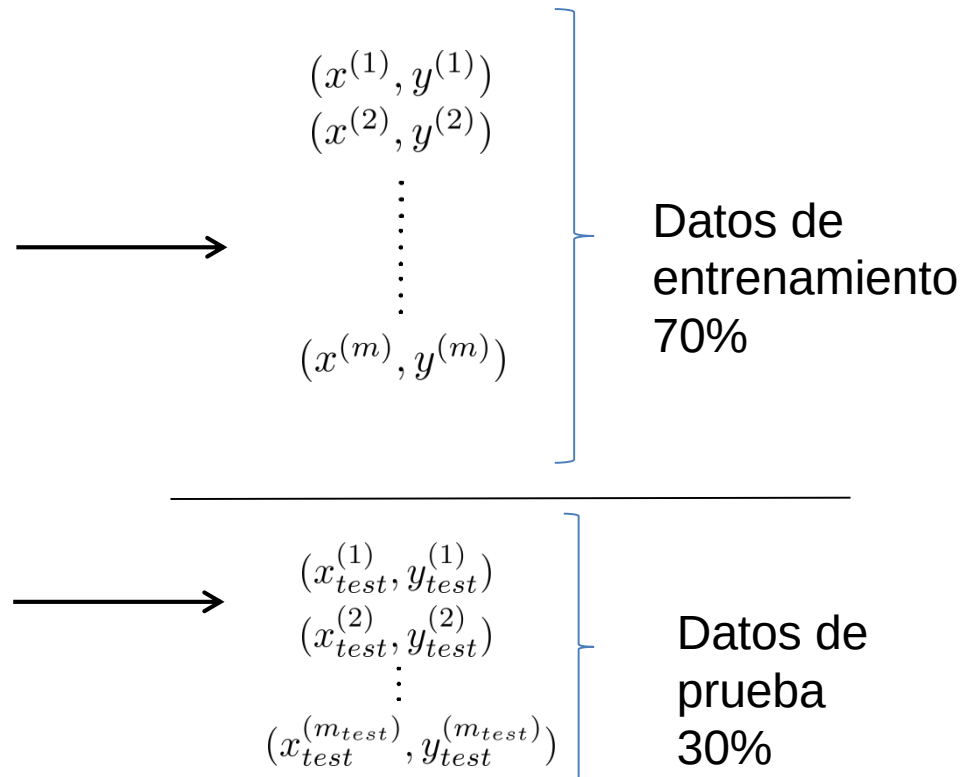


Selección del modelo/entrenamiento/validación/ set de pruebas

- Una vez que los parámetros $\theta_0, \theta_1, \dots, \theta_4$ se ajustaron a un conjunto de datos (conjunto de entrenamiento)
- El error se mide con otros datos que se apartaron del conjunto de datos de entrada.

Evaluando la hipótesis

Tamaño	Precio
2104	400
1600	330
2400	369
1416	232
3000	540
1985	300
1534	315
1427	199
1380	212
1494	243



Procedimiento para el entrenamiento/prueba de regresión lineal

- Obtener el parámetro θ con los datos de entrenamiento (minimizando el error de entrenamiento $J(\theta)$)
- Calcular el error de prueba $J_{test}(\theta)$

$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_1^{m_{test}} \left(h_{\theta} \left(x^{(i)}_{test} \right) - y^{(i)}_{test} \right)^2$$

Procedimiento para el entrenamiento/prueba de regresión logística

- Obtener el parámetro θ
- Calcular el error de prueba $J_{test}(\theta)$

$$J_{test}(\theta) = -\frac{1}{m_{test}} \sum_{i=1}^{m_{test}} y_{test}^{(i)} \log h_{\theta}(x_{test}^{(i)}) + (1 - y_{test}^{(i)}) \log h_{\theta}(x_{test}^{(i)})$$

- Error de clasificación errónea

$$err(h_{\theta}(x), y) = \begin{cases} 1 & \text{if } h_{\theta}(x) \geq 0.5, \ y = 0 \text{ or} \\ & \text{if } h_{\theta}(x) \leq 0.5, \ y = 1 \\ 0 & \text{de otra manera} \end{cases}$$

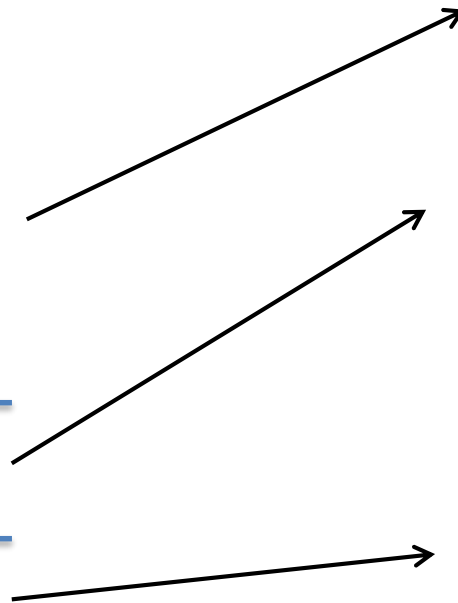
$$err(h_{\theta}(x), y) = \frac{1}{m_{test}} \sum_1^{m_{test}} err(h_{\theta}(x), y)$$

Selección del modelo/entrenamiento/validación/ set de pruebas

Tamaño	Precio
2104	400
1600	330
2400	369
1416	232
3000	540
1985	300
1534	315
1427	199
1380	212
1494	243

Selección del modelo/entrenamiento/validación/ set de pruebas

Tamaño	Precio
2104	400
1600	330
2400	369
1416	232
3000	540
1985	300
1534	315
1427	199
1380	212
1494	243



$$(x^{(1)}, y^{(1)})$$

$$(x^{(2)}, y^{(2)})$$

$$\vdots$$

$$(x^{(m)}, y^{(m)})$$

$$(x_{cv}^{(1)}, y_{cv}^{(1)})$$

$$(x_{cv}^{(2)}, y_{cv}^{(2)})$$

$$\vdots$$

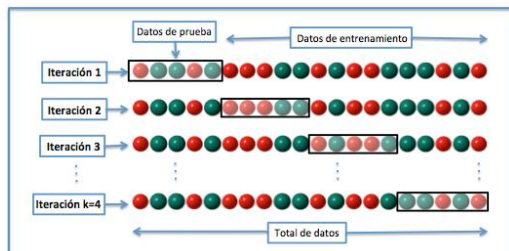
$$(x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})$$

$$(x_{test}^{(1)}, y_{test}^{(1)})$$

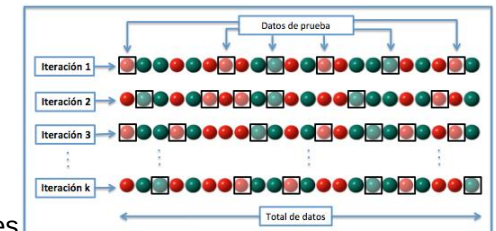
$$(x_{test}^{(2)}, y_{test}^{(2)})$$

$$\vdots$$

$$(x_{test}^{(m_{test})}, y_{test}^{(m_{test})})$$



Validación cruzada de K iteraciones con K=4.



Validación cruzada aleatoria con k iteraciones

Selección del modelo/entrenamiento/validación/ set de pruebas

Error entrenamiento: $J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$

Error Validación cruzada: $J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} (h_{\theta}(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$

Error Test: $J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} (h_{\theta}(x_{test}^{(i)}) - y_{test}^{(i)})^2$

Precision/ Recall

	1	0
1	Verdaderos positivos	Falsos Positivos
0	Falsos negativos	Verdaderos negativos

Precisión:

(de los datos que fueron predichos $y = 1$, que fracción en realidad tiene ese valor)

$$\text{precision} = \frac{\text{Verdaderos pos}}{\text{no. de predichos pos}} = \frac{\text{Verdaderos pos}}{\text{Verdaderos pos} + \text{Falsos pos}}$$

Recall:

(de los datos que realmente $y = 1$, que fracción detecto correctamente el modelo)

$$\text{recall} = \frac{\text{Verdaderos pos}}{\text{no. de actual pos}} = \frac{\text{Verdaderos pos}}{\text{Verdaderos pos} + \text{Falsos neg}}$$

Negociando entre Precisión & Recall

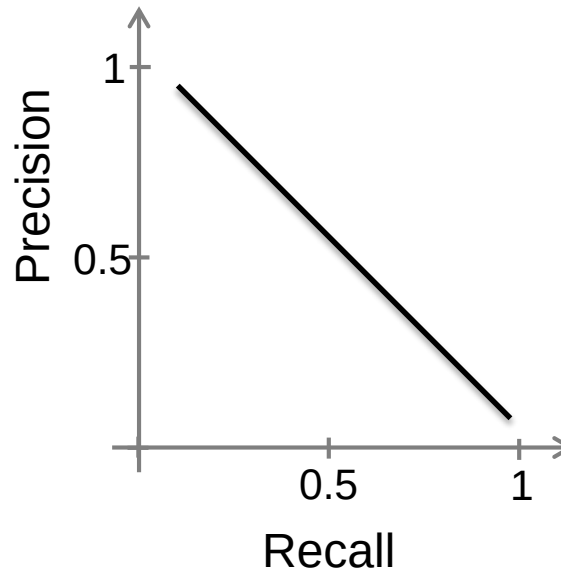
Regresión Lógica: $0 \leq h_{\theta}(x) \leq 1$

Predice 1 si $h_{\theta}(x) \geq 0.5$

Predice 0 si $h_{\theta}(x) < 0.5$

$$\text{precision} = \frac{\text{Verdaderos pos}}{\text{no. de predichos pos}}$$

$$\text{recall} = \frac{\text{Verdaderos pos}}{\text{no. de actual pos}}$$



Negociando entre Precisión & Recall

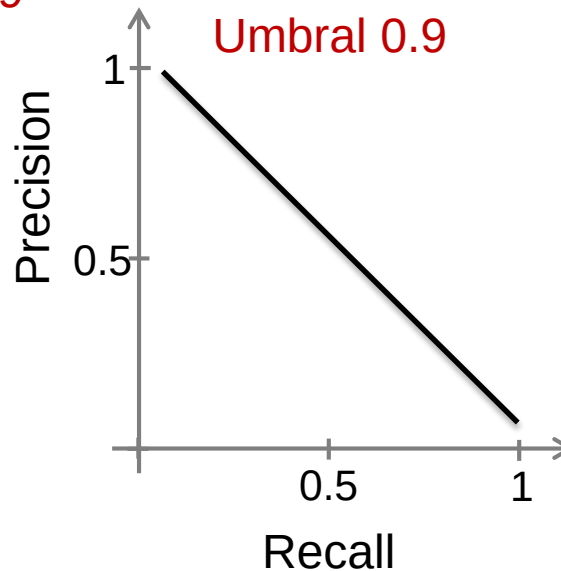
Regresión Lógica: $0 \leq h_{\theta}(x) \leq 1$

Predice 1 si $h_{\theta}(x) \geq \cancel{0.5}$ **0.9**

Predice 0 si $h_{\theta}(x) < \cancel{0.5}$ **0.9**

$$\text{precision} = \frac{\text{Verdaderos pos}}{\text{no. de predichos pos}}$$

$$\text{recall} = \frac{\text{Verdaderos pos}}{\text{no. de actual pos}}$$



Negociando entre Precisión & Recall

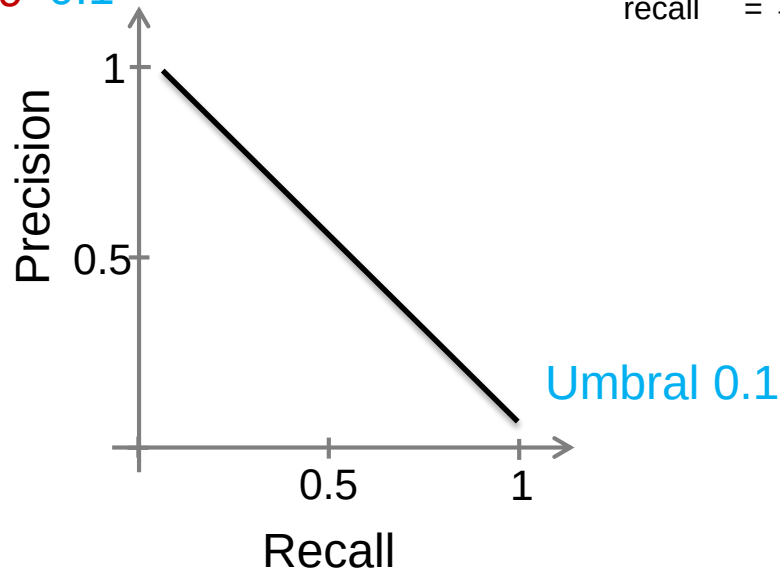
Regresión Lógica: $0 \leq h_{\theta}(x) \leq 1$

Predice 1 si $h_{\theta}(x) \geq \cancel{0.5}$ ~~0.9~~ 0.1

Predice 0 si $h_{\theta}(x) < \cancel{0.5}$ ~~0.9~~ 0.1

$$\text{precision} = \frac{\text{Verdaderos pos}}{\text{no. de predichos pos}}$$

$$\text{recall} = \frac{\text{Verdaderos pos}}{\text{no. de actual pos}}$$



Negociando entre Precisión & Recall

Average: $\frac{P+R}{2}$

F₁ Score: $2 \frac{PR}{P+R}$

Negociando entre Precisión & Recall

Average: $\frac{P+R}{2}$

F₁ Score: $2 \frac{PR}{P+R}$

	Precision(P)	Recall (R)	Average	F ₁ Score
Algoritmo 1	0.5	0.4	0.45	0.444
Algoritmo 2	0.7	0.1	0.4	0.175
Algoritmo 3	0.02	1.0	0.51	0.0392

Sobre-ajustamiento (Overfitting)

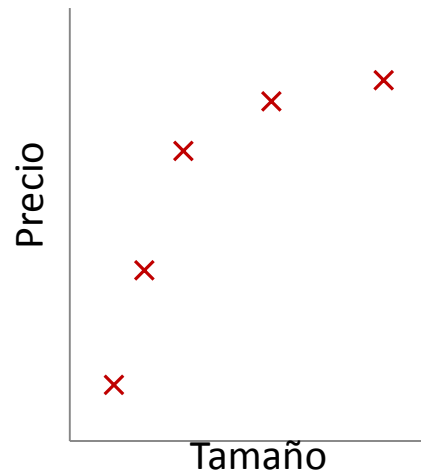
- Es el **efecto de sobre-entrenar** un algoritmo de aprendizaje con unos ciertos datos para los que se conoce el resultado deseado.
- El algoritmo de aprendizaje debe alcanzar un estado en el que **sea capaz de predecir el resultado en otros casos a partir de lo aprendido con los datos de entrenamiento**, **generalizando** para poder resolver situaciones distintas a las definidas en el entrenamiento.
- Sin embargo, cuando un sistema **se entrena demasiado** (se sobre-entrena) **o se entrena con datos extraños**, el algoritmo de aprendizaje puede quedar ajustado a unas características muy específicas de los datos de entrenamiento.

Durante la fase de sobre-ajuste el éxito al responder las muestras de entrenamiento sigue incrementándose mientras que su actuación con muestras nuevas va empeorando.

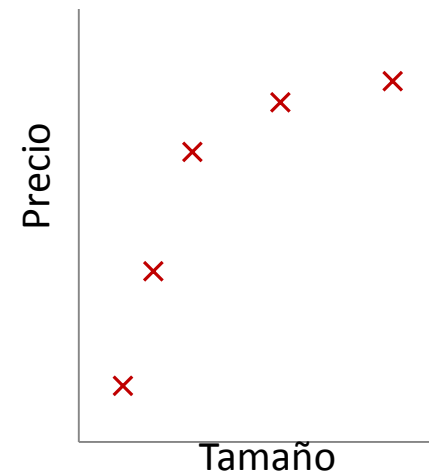
Sobre-ajustamiento (Overfitting)



$$\theta_0 + \theta_1 x$$

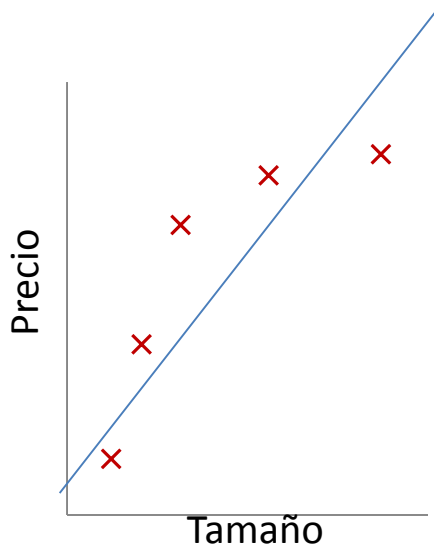


$$\theta_0 + \theta_1 x + \theta_2 x^2$$

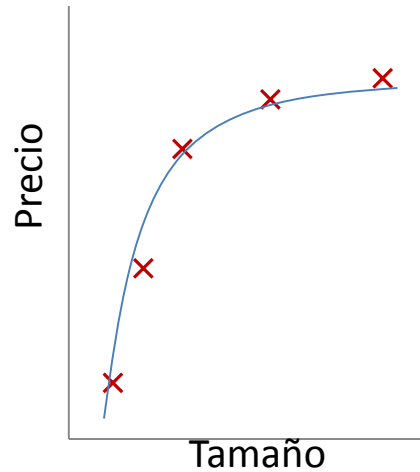


$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

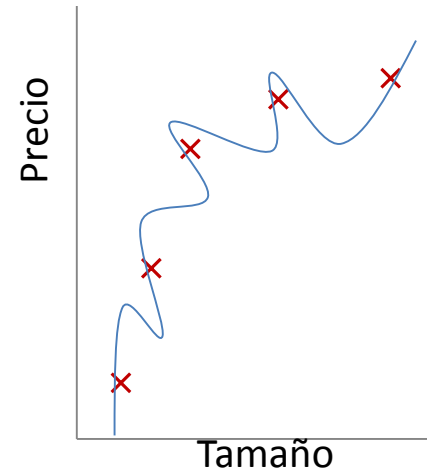
Sobre-ajustamiento (Overfitting)



$$\theta_0 + \theta_1 x$$

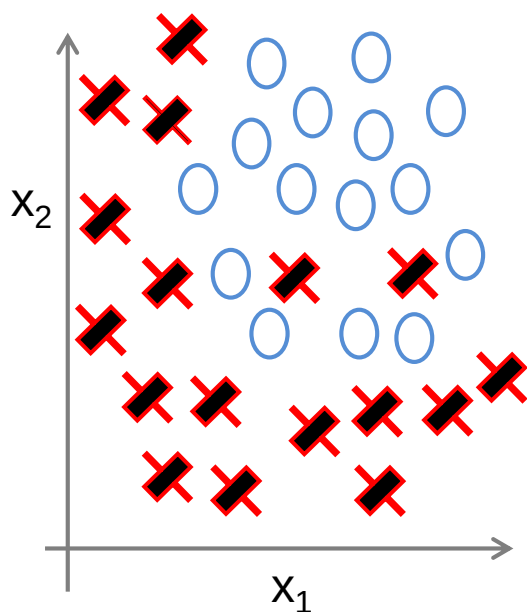


$$\theta_0 + \theta_1 x + \theta_2 x^2$$



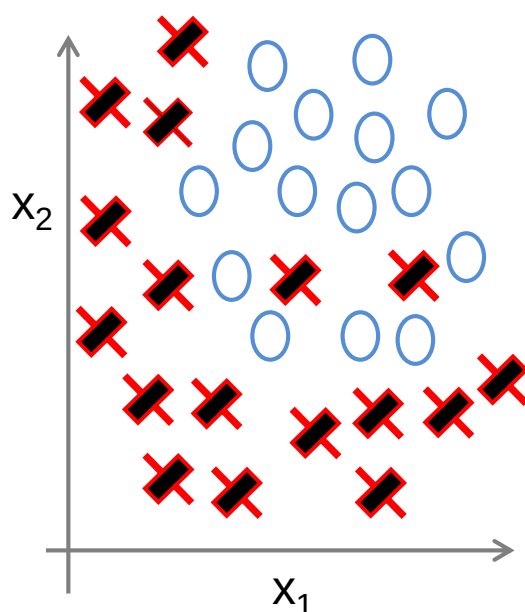
$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

Sobre-ajustamiento (Overfitting)

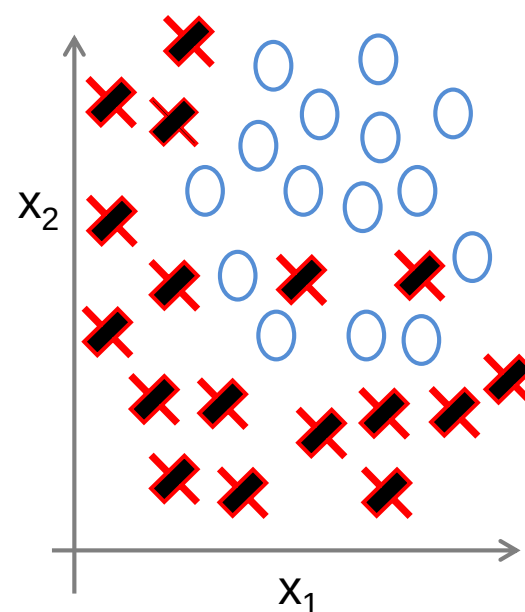


$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

g (= función
sigmoidal)

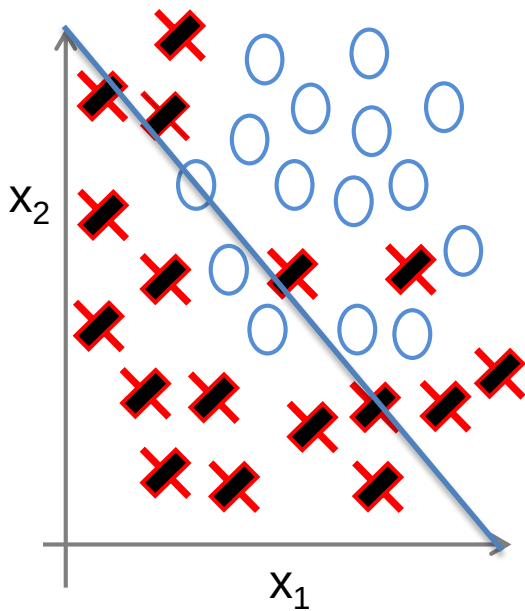


$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$$

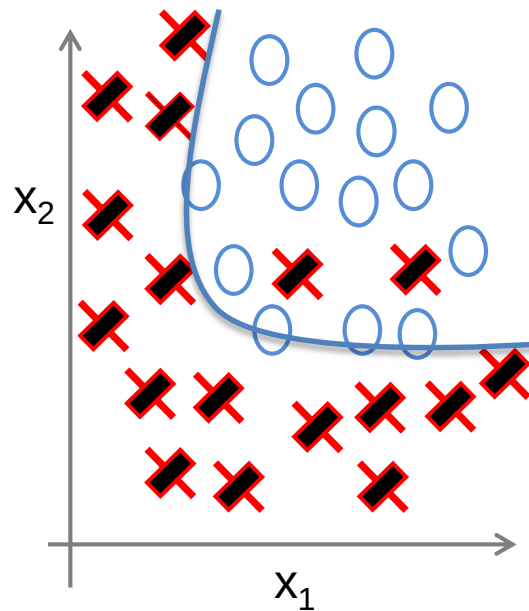


$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots)$$

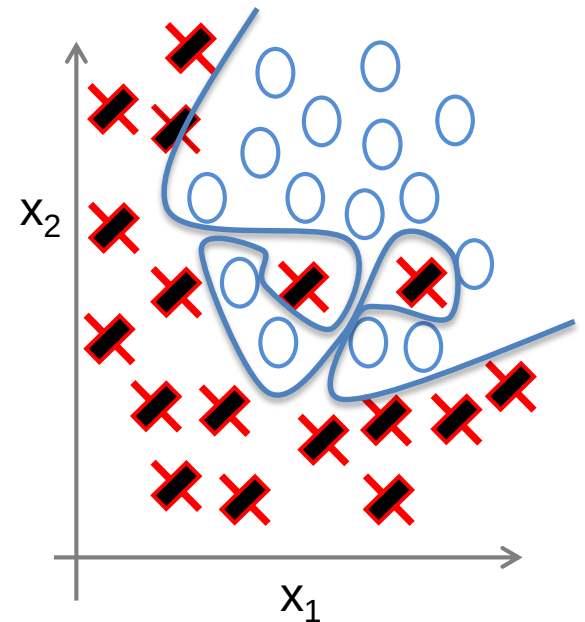
sigmoidal



$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

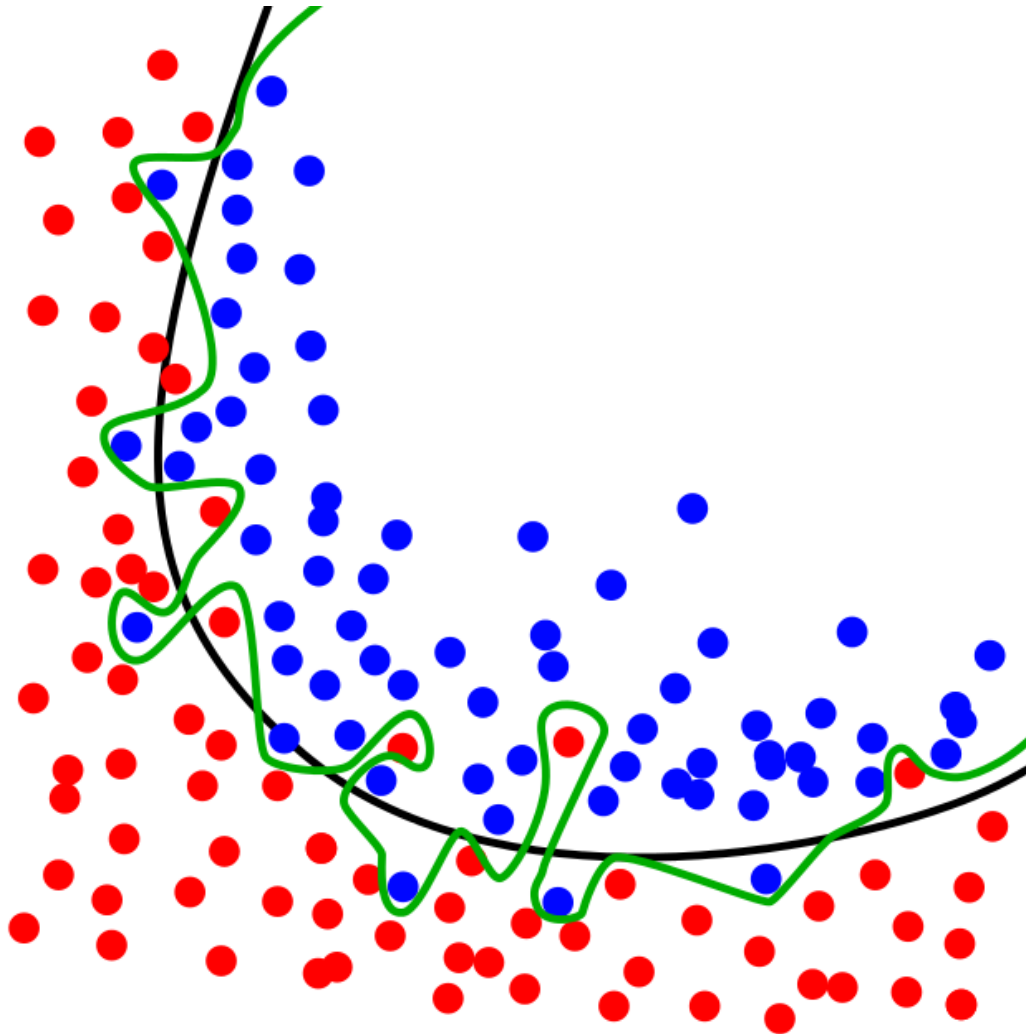


$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$$



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots)$$

Sobre-ajustamiento (Overfitting)



Emplear la **línea verde** como clasificador se adapta mejor a los datos con los que hemos entrenado al clasificador, pero está *demasiado* adaptada a ellos, de forma que ante nuevos datos probablemente arrojará más errores que la clasificación usando la **línea negra**.

Sobre-ajustamiento (Overfitting)

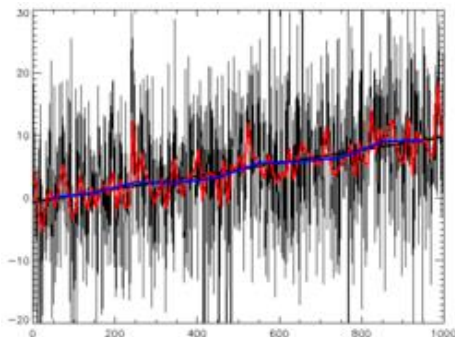
Opciones:

- **Reducir el número de características.**
 - Seleccionar manualmente las características que desea conservar.
- **Regularización.**
 - Mantener todas las características, pero reducir la magnitud/valores de los parámetros.
 - Funciona bien cuando tenemos una gran cantidad de características, y cada una contribuye un poco a la predicción.

Minería en Secuencia de Datos

Una **serie temporal** o **cronológica** es una secuencia de datos, observaciones o valores, medidos en determinados momentos y ordenados cronológicamente. Los datos pueden estar espaciados a intervalos iguales o desiguales

- Buscar **similitudes** en serie temporal de datos
 - **Análisis de Tendencias** en series temporales de datos
 - Minería de **Patrones Secuenciales** en cadenas simbólicas
- Clasificación** de Secuencia



**Ejemplo: Alineación de secuencias biológicas
(bioinformática)**

DEFINICION BÁSICA DE SERIE DE TIEMPO

Una *serie de tiempo* es una colección o conjunto de mediciones de cierto fenómeno o experimento registrados secuencialmente en el tiempo, en forma *equiespaciada* (a intervalos de tiempo iguales) .

Las observaciones de una serie de tiempo serán denotadas por

$$Y(t_1), Y(t_2), \dots, Y(t_n)$$

donde $Y(t_i)$ es el valor tomado por el proceso en el instante t_i .

SERIES TEMPORALES

- La serie puede ser simple o multiple según se disponga de un valor para cada instante del tiempo o varios valores.
- Puede ser continua o discreta, según la naturaleza de la variable dependiente.
- Tiene que estar perfectamente ordenado.
- Los intervalos de tiempo considerado, han de tener la misma amplitud.
- Los datos han de ser homogéneos, no se puede cambiar de criterios, metodología, etc..

Ejemplos de series de tiempo

1. Economía: Precios de un artículo, tasas de desempleo, tasa de inflación, índice de precios, precio del dólar, precio del cobre, precios de acciones, ingreso nacional bruto, etc.
2. Meteorología: Cantidad de agua caída, temperatura máxima diaria, Velocidad del viento (energía eólica), energía solar, etc.
3. Geofísica: Series sismológicas.
4. Química: Viscosidad de un proceso, temperatura de un proceso.
5. Demografía: Tasas de natalidad, tasas de mortalidad.

Ejemplos de series de tiempo

6. Medicina: Electrocardiograma, electroencefalograma.

7. Marketing: Series de demanda, gastos, utilidades, ventas, ofertas.

8. Telecomunicaciones: Análisis de señales.

9. Transporte: Series de tráfico.

...y muchos otros.

Patrones Secuenciales

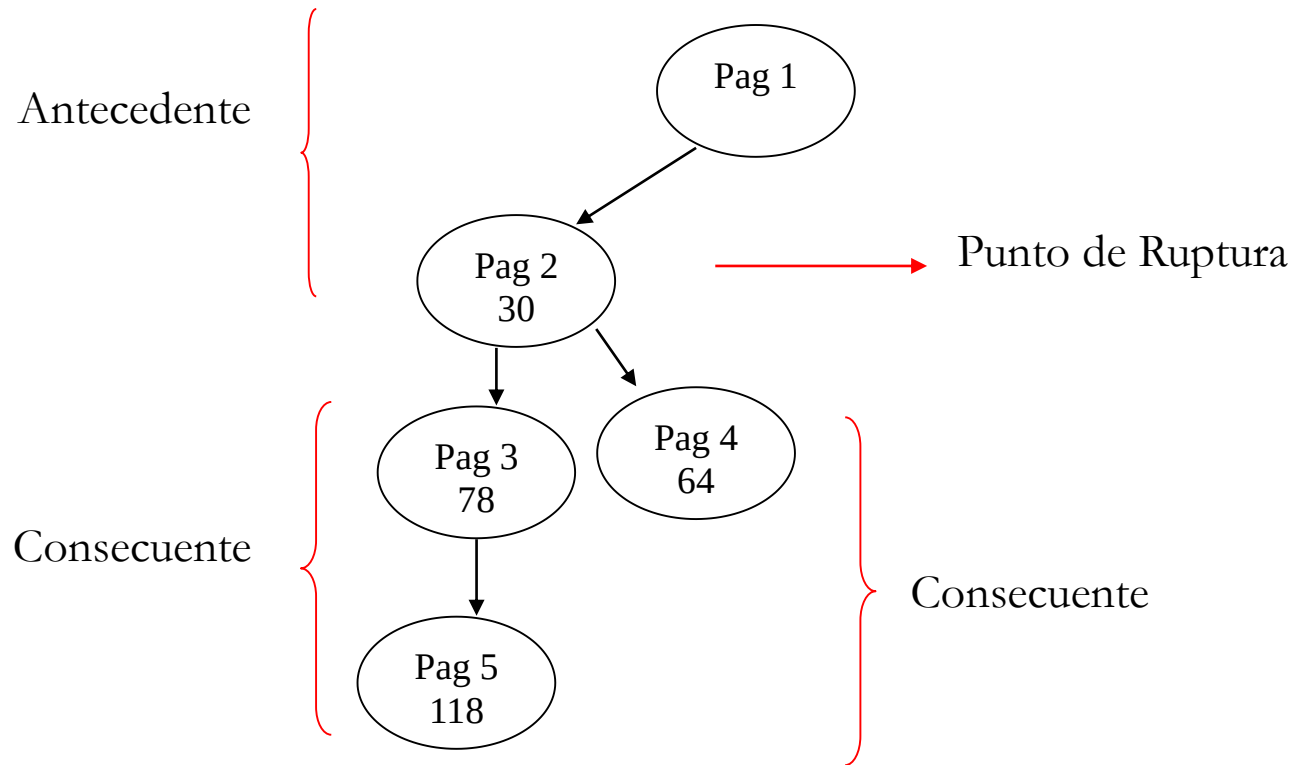
- Descubrir patrones en los cuales la presencia de un conjunto de ítems es seguido por otro ítem en orden temporal.
- Ejemplo: Encontrar y predecir el comportamiento de los visitantes de un sitio Web con respecto al tiempo.

$$[x1 \rightarrow x2 \rightarrow x3] \rightarrow [y1 \rightarrow y2] \text{ en } t \text{ días}$$

/public/team.jsp ->
/public/findUsers.jsp->
/private/mycourses/website/folders/assignment/assignment_view.jsp->
/public/portalDocument.js

Patrones Secuenciales

Generación FBP-Árbol (Matriz FTM, Lista de Caminos)



Patrones Secuenciales

Algoritmo Patrones (FBP-Arbol, soporte, confianza)

La confianza de una *regla de comportamiento-frecuente* se representa como $conf(PIND \rightarrow PDEP)$ y define la probabilidad de recorrer el camino PDEP una vez se ha recorrido el camino PIND.

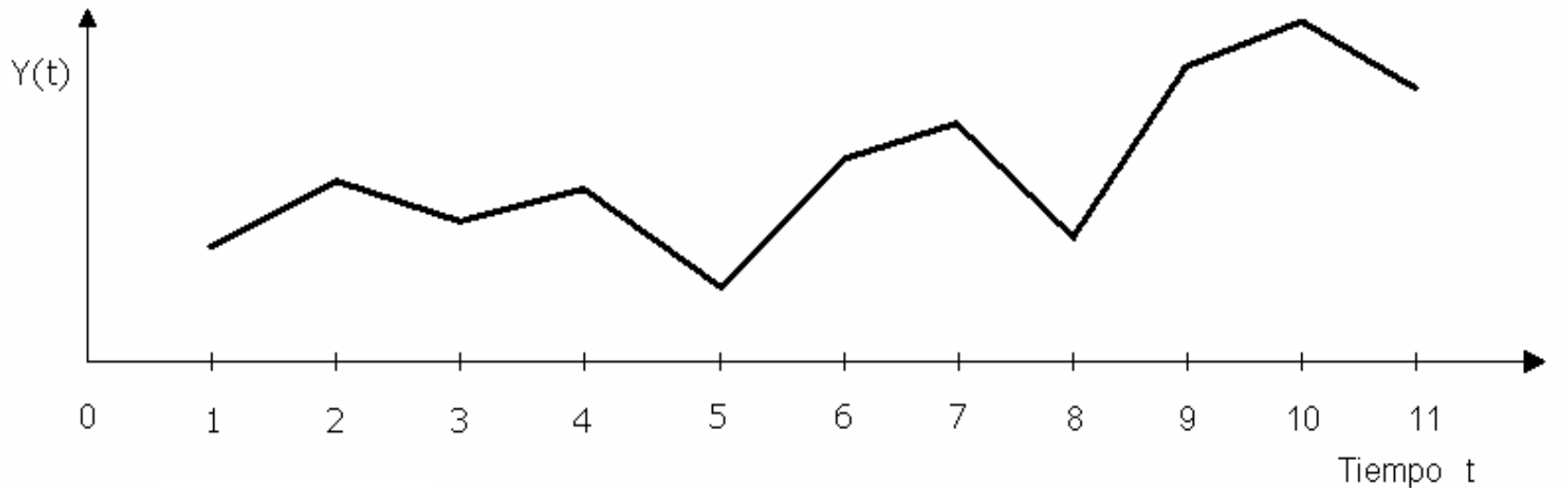
- Se recorre el árbol desde las hojas al nodo raíz.
- Teniendo en cuenta el soporte de cada camino, las reglas son calculados como sigue.
- Buscar en hojas el punto de ruptura.
 - Si la hoja no es Punto ruptura, ir a hoja anterior.
 - Si la hoja es Punto Ruptura, calcular $conf$.
 - Si $conf > confianza$, genera Patrón
 - Si $conf < confianza$, podar rama de árbol.

SERIES TEMPORALES

- **Método que se centran en el dominio del tiempo: evolución temporal.**
 - Análisis Clásico.
 - Metodología Box-Jenkins
- **Método que se centra en el estudio de las frecuencias.**
 - Análisis espectral.

ANALISIS GRAFICO DE UNA SERIE DE TIEMPO

Por muy simple que parezca, el paso más importante en el análisis de series de tiempo consiste en graficar la serie.



Esto debe hacerse siempre, independiente de cuán simples o complejos sean los procedimientos que se emplean posteriormente.

Predicción por descomposición de series

Descomposición de series:

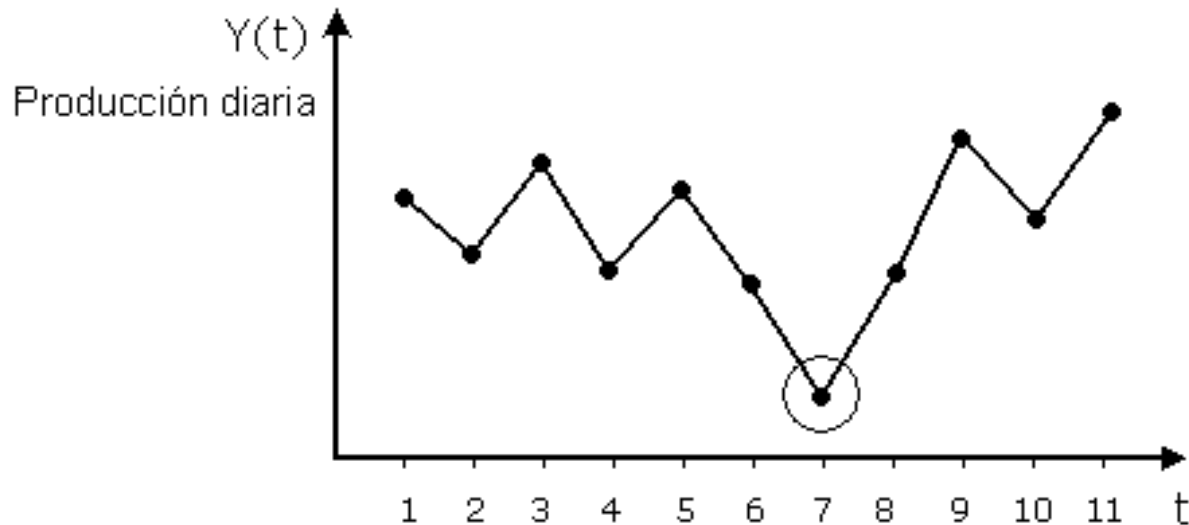
Sobre una serie temporal Y_t podemos identificar una serie de componentes básicos que se denominan respectivamente como:

- TENDENCIA: T_t Movimientos de larga duración que se mantienen durante todo el periodo de observación.
- CICLO: C_t Oscilaciones alrededor de la tendencia producidos por períodos alternativos de prosperidad y depresión.
- ESTACIONALIDAD: S_t Movimiento que se produce, dentro de un periodo anual, por motivos no estrictamente económicos (climáticos, sociales, etc.)
- IRREGULARIDAD: I_t Movimientos erráticos generados por causas ajenas al fenómeno económico y no repetidos en el tiempo

Predicción por descomposición de series

a) Valores Atípicos:

Se refiere a puntos de la serie que se escapan de lo normal.



Movimientos que no responden a ningún patrón que son resultado de factores fortuitos o aleatorios.

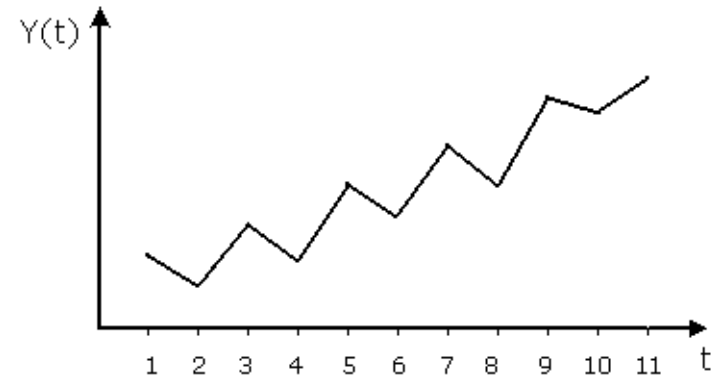
Puntos de ruptura, interrupciones en la evolución de la serie.

Predicción por descomposición de series

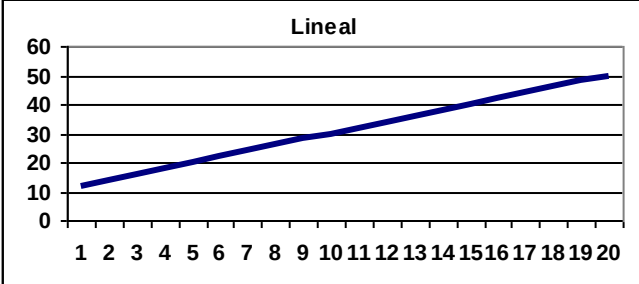
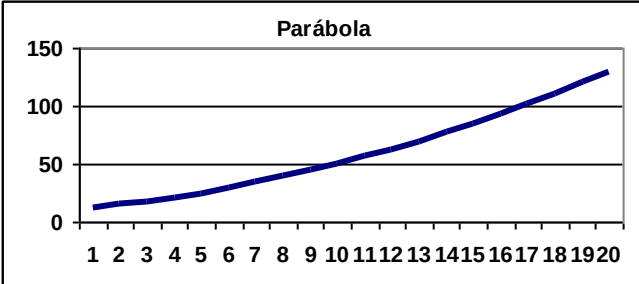
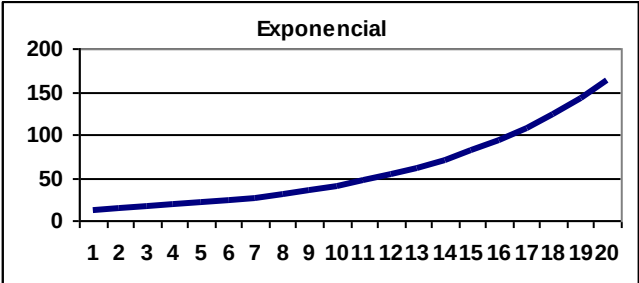
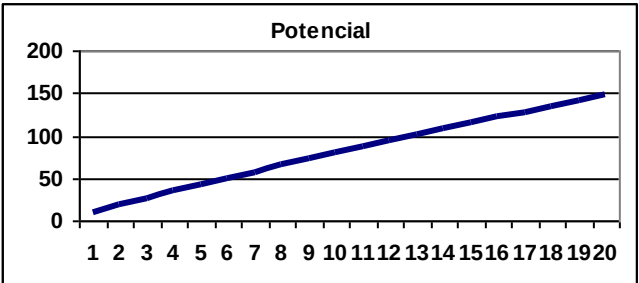
b) Tendencias

La tendencia representa el comportamiento predominante de la serie.

- Comportamiento de la serie (+10 años)
- Mientras más larga sea la serie mejor.
- Nos muestra si la serie es estacionaria o evolutiva.
- El comportamiento tendencial puede ser:
 - Lineal
 - Exponencial
 - Parabólica
 -



Predicción por descomposición de series: Tendencias

Lineal	$Y_t = a + b * t$	 A line graph titled 'Lineal' showing a linear trend. The x-axis represents time (t) from 1 to 20, and the y-axis represents the value (Y_t) from 0 to 60. The line starts at approximately (1, 10) and ends at (20, 50). <table><tr><th>t</th><th>Y_t</th></tr><tr><td>1</td><td>10</td></tr><tr><td>2</td><td>12</td></tr><tr><td>3</td><td>14</td></tr><tr><td>4</td><td>16</td></tr><tr><td>5</td><td>18</td></tr><tr><td>6</td><td>20</td></tr><tr><td>7</td><td>22</td></tr><tr><td>8</td><td>24</td></tr><tr><td>9</td><td>26</td></tr><tr><td>10</td><td>28</td></tr><tr><td>11</td><td>30</td></tr><tr><td>12</td><td>32</td></tr><tr><td>13</td><td>34</td></tr><tr><td>14</td><td>36</td></tr><tr><td>15</td><td>38</td></tr><tr><td>16</td><td>40</td></tr><tr><td>17</td><td>42</td></tr><tr><td>18</td><td>44</td></tr><tr><td>19</td><td>46</td></tr><tr><td>20</td><td>50</td></tr></table>	t	Y_t	1	10	2	12	3	14	4	16	5	18	6	20	7	22	8	24	9	26	10	28	11	30	12	32	13	34	14	36	15	38	16	40	17	42	18	44	19	46	20	50
t	Y_t																																											
1	10																																											
2	12																																											
3	14																																											
4	16																																											
5	18																																											
6	20																																											
7	22																																											
8	24																																											
9	26																																											
10	28																																											
11	30																																											
12	32																																											
13	34																																											
14	36																																											
15	38																																											
16	40																																											
17	42																																											
18	44																																											
19	46																																											
20	50																																											
Parábola	$Y_t = a + b * t + c * t^2$	 A line graph titled 'Parábola' showing a quadratic trend. The x-axis represents time (t) from 1 to 20, and the y-axis represents the value (Y_t) from 0 to 150. The curve starts at approximately (1, 10) and ends at (20, 130). <table><tr><th>t</th><th>Y_t</th></tr><tr><td>1</td><td>10</td></tr><tr><td>2</td><td>14</td></tr><tr><td>3</td><td>19</td></tr><tr><td>4</td><td>24</td></tr><tr><td>5</td><td>30</td></tr><tr><td>6</td><td>36</td></tr><tr><td>7</td><td>43</td></tr><tr><td>8</td><td>50</td></tr><tr><td>9</td><td>58</td></tr><tr><td>10</td><td>66</td></tr><tr><td>11</td><td>75</td></tr><tr><td>12</td><td>84</td></tr><tr><td>13</td><td>94</td></tr><tr><td>14</td><td>104</td></tr><tr><td>15</td><td>114</td></tr><tr><td>16</td><td>124</td></tr><tr><td>17</td><td>134</td></tr><tr><td>18</td><td>144</td></tr><tr><td>19</td><td>154</td></tr><tr><td>20</td><td>164</td></tr></table>	t	Y_t	1	10	2	14	3	19	4	24	5	30	6	36	7	43	8	50	9	58	10	66	11	75	12	84	13	94	14	104	15	114	16	124	17	134	18	144	19	154	20	164
t	Y_t																																											
1	10																																											
2	14																																											
3	19																																											
4	24																																											
5	30																																											
6	36																																											
7	43																																											
8	50																																											
9	58																																											
10	66																																											
11	75																																											
12	84																																											
13	94																																											
14	104																																											
15	114																																											
16	124																																											
17	134																																											
18	144																																											
19	154																																											
20	164																																											
Exponencial	$Y_t = a * b^t$ $Ln(Y_t) = c + d * t$	 A line graph titled 'Exponencial' showing an exponential trend. The x-axis represents time (t) from 1 to 20, and the y-axis represents the value (Y_t) from 0 to 200. The curve starts at approximately (1, 10) and ends at (20, 160). <table><tr><th>t</th><th>Y_t</th></tr><tr><td>1</td><td>10</td></tr><tr><td>2</td><td>12</td></tr><tr><td>3</td><td>14</td></tr><tr><td>4</td><td>16</td></tr><tr><td>5</td><td>18</td></tr><tr><td>6</td><td>20</td></tr><tr><td>7</td><td>22</td></tr><tr><td>8</td><td>24</td></tr><tr><td>9</td><td>26</td></tr><tr><td>10</td><td>28</td></tr><tr><td>11</td><td>30</td></tr><tr><td>12</td><td>32</td></tr><tr><td>13</td><td>34</td></tr><tr><td>14</td><td>36</td></tr><tr><td>15</td><td>38</td></tr><tr><td>16</td><td>40</td></tr><tr><td>17</td><td>42</td></tr><tr><td>18</td><td>44</td></tr><tr><td>19</td><td>46</td></tr><tr><td>20</td><td>50</td></tr></table>	t	Y_t	1	10	2	12	3	14	4	16	5	18	6	20	7	22	8	24	9	26	10	28	11	30	12	32	13	34	14	36	15	38	16	40	17	42	18	44	19	46	20	50
t	Y_t																																											
1	10																																											
2	12																																											
3	14																																											
4	16																																											
5	18																																											
6	20																																											
7	22																																											
8	24																																											
9	26																																											
10	28																																											
11	30																																											
12	32																																											
13	34																																											
14	36																																											
15	38																																											
16	40																																											
17	42																																											
18	44																																											
19	46																																											
20	50																																											
Potencial	$Y_t = a * t^b$ $Ln(Y_t) = c + d * Ln(t)$	 A line graph titled 'Potencial' showing a power-law trend. The x-axis represents time (t) from 1 to 20, and the y-axis represents the value (Y_t) from 0 to 200. The curve starts at approximately (1, 10) and ends at (20, 150). <table><tr><th>t</th><th>Y_t</th></tr><tr><td>1</td><td>10</td></tr><tr><td>2</td><td>12</td></tr><tr><td>3</td><td>14</td></tr><tr><td>4</td><td>16</td></tr><tr><td>5</td><td>18</td></tr><tr><td>6</td><td>20</td></tr><tr><td>7</td><td>22</td></tr><tr><td>8</td><td>24</td></tr><tr><td>9</td><td>26</td></tr><tr><td>10</td><td>28</td></tr><tr><td>11</td><td>30</td></tr><tr><td>12</td><td>32</td></tr><tr><td>13</td><td>34</td></tr><tr><td>14</td><td>36</td></tr><tr><td>15</td><td>38</td></tr><tr><td>16</td><td>40</td></tr><tr><td>17</td><td>42</td></tr><tr><td>18</td><td>44</td></tr><tr><td>19</td><td>46</td></tr><tr><td>20</td><td>50</td></tr></table>	t	Y_t	1	10	2	12	3	14	4	16	5	18	6	20	7	22	8	24	9	26	10	28	11	30	12	32	13	34	14	36	15	38	16	40	17	42	18	44	19	46	20	50
t	Y_t																																											
1	10																																											
2	12																																											
3	14																																											
4	16																																											
5	18																																											
6	20																																											
7	22																																											
8	24																																											
9	26																																											
10	28																																											
11	30																																											
12	32																																											
13	34																																											
14	36																																											
15	38																																											
16	40																																											
17	42																																											
18	44																																											
19	46																																											
20	50																																											

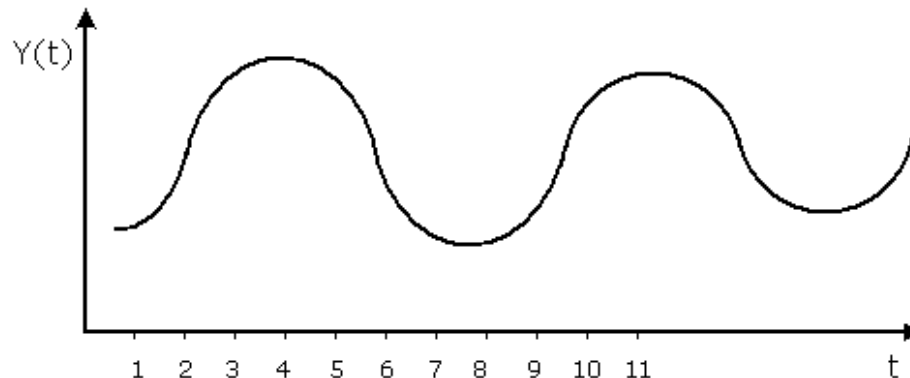
Predicción por descomposición de series

c) Variaciones cíclicas o estacionales

Estas variaciones representan un movimiento *periódico* de la serie de tiempo.

- La **variación estacional** ocurre con períodos identificables, como la estacionalidad del empleo, o de la venta de ciertos productos, cuyo período es un año.
- El término **variación cíclica** se suele referir a ciclos grandes, cuyo período no es atribuible a alguna causa.

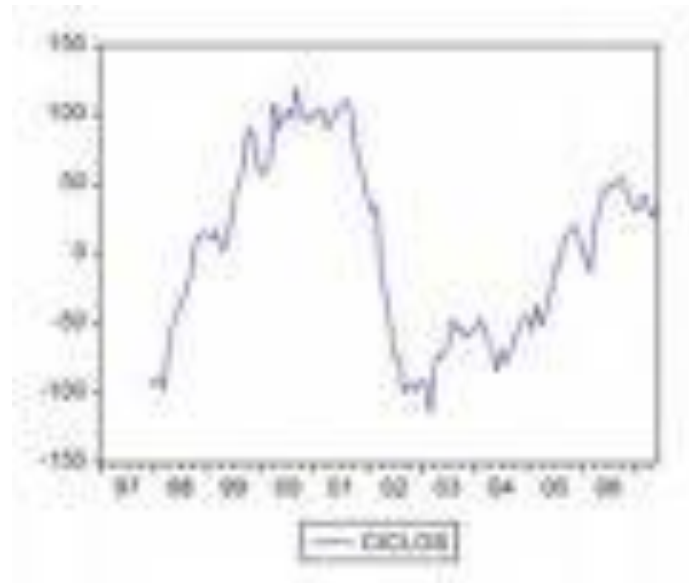
La duración del período puede ser un año, un trimestre, un mes, un día, etc.



Predicción por descomposición de series

VARIACIONES CÍCLICAS (C)

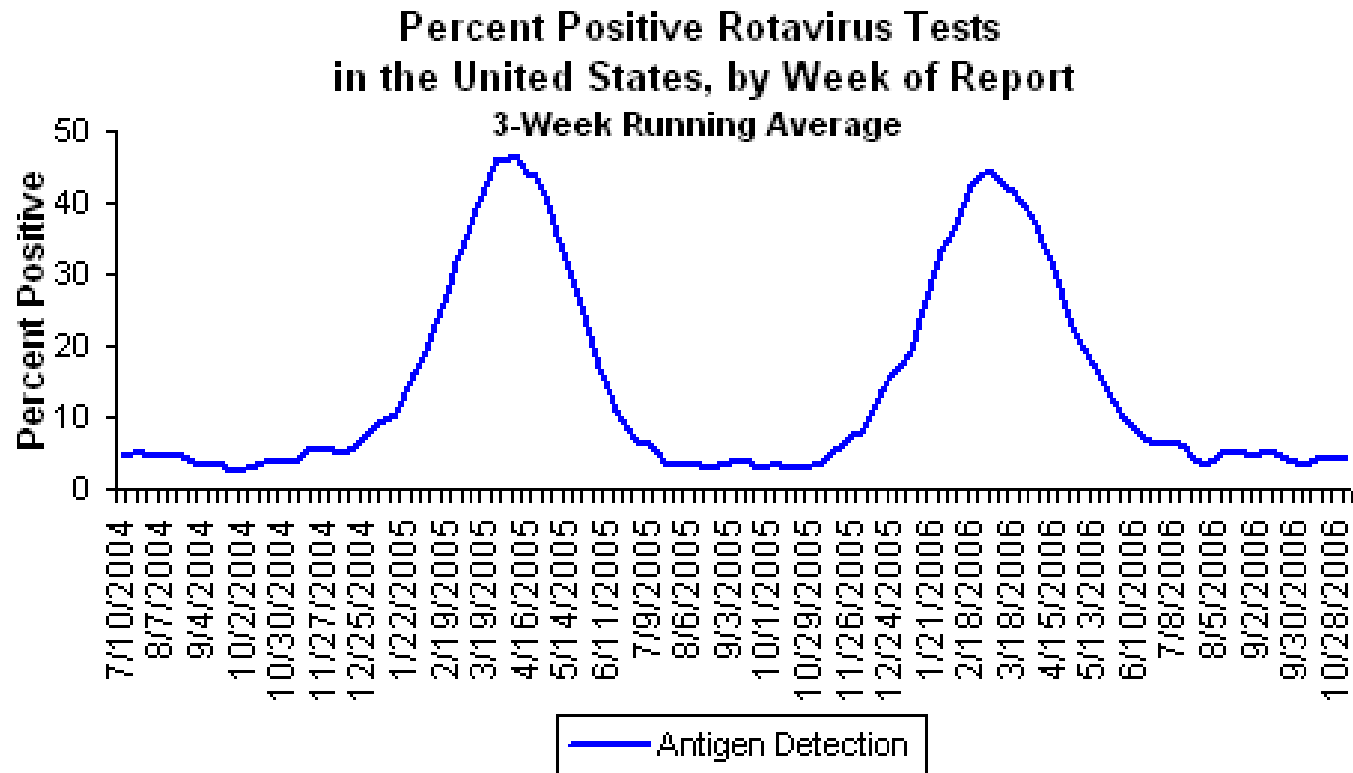
- Son movimientos a medio plazo que corresponden normalmente con las fases expansivas y recesivas de la economía. (Entre 2 y 10 años).
- Son difíciles de detectar, salvo que se tenga un periodo largo de años.



Predicción por descomposición de series

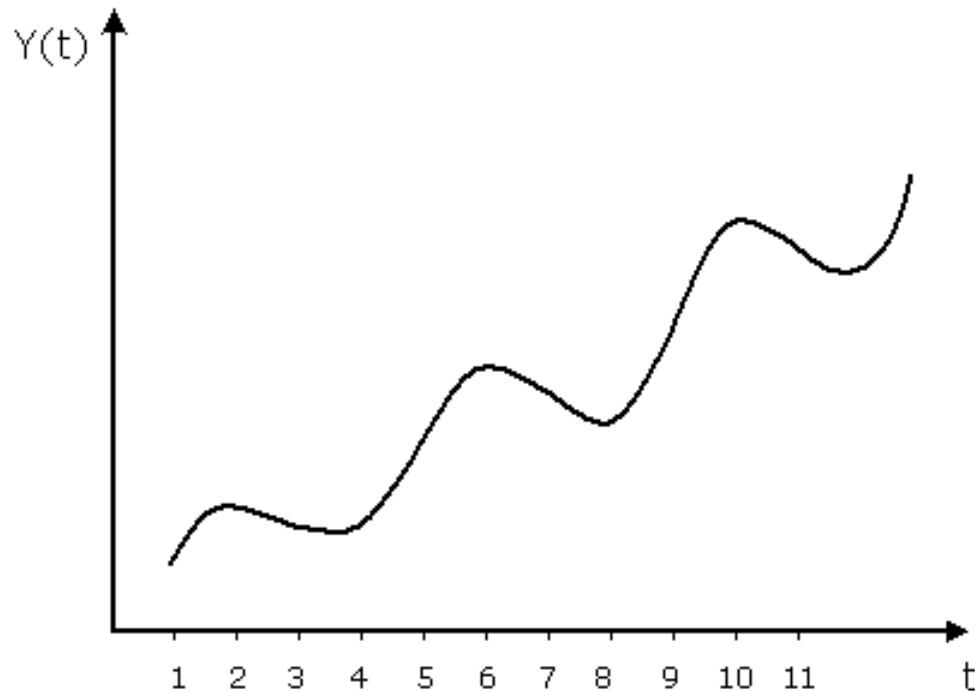
VARIACIONES ESTACIONALES (VE)

- Movimientos que se repiten de forma periódica, producida normalmente por el clima, producción, tiempo Lo usual es que tenga una frecuencia anual.



Predicción por descomposición de series

Las tendencias y estacionalidades pueden darse simultáneamente.



Predicción por descomposición de series

Modelos Clásicos

Un modelo clásico de series de tiempo, supone que la serie $Y(1), \dots, Y(n)$ puede ser expresada como suma o producto de tres componentes:

tendencia,

componente estacional,

término de error aleatorio.

Predicción por descomposición de series

Modelos Clásicos

1. $Y(t) = T(t) + E(t) + A(t)$ Modelo aditivo
2. $Y(t) = T(t) E(t) A(t)$ Modelo multiplicativo
3. $Y(t) = T(t) (1 - C(t))(1 - VE(t)) + A(t)$ Modelo mixto

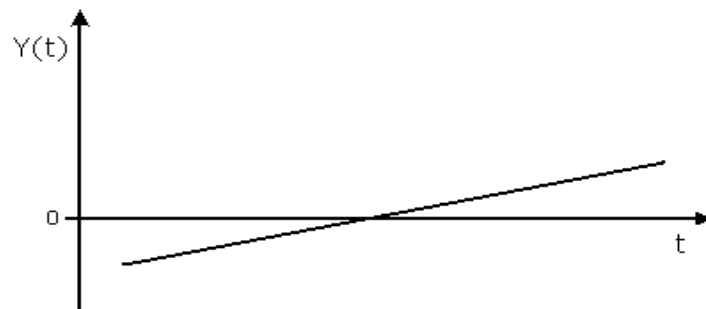
donde:

T: Tendencia de la serie.

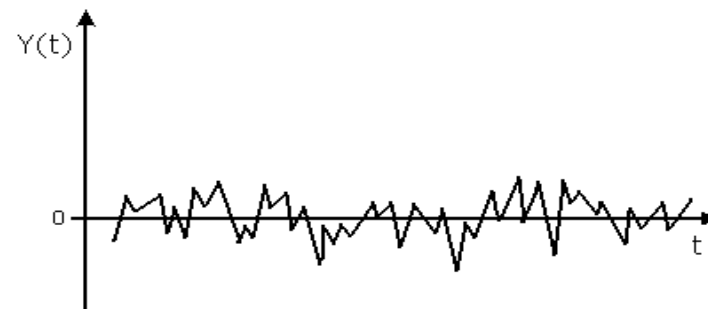
E: Variación Estacional.

A: Variaciones aleatorias.

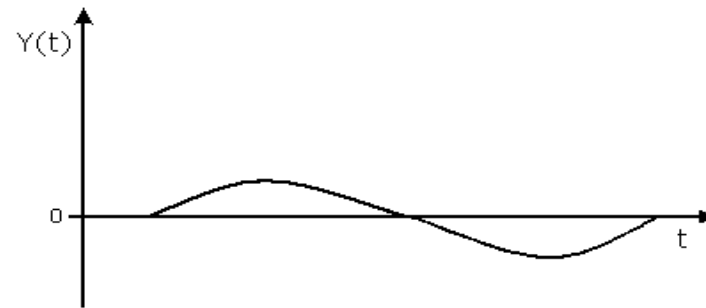
La serie y sus componentes, para el caso aditivo.



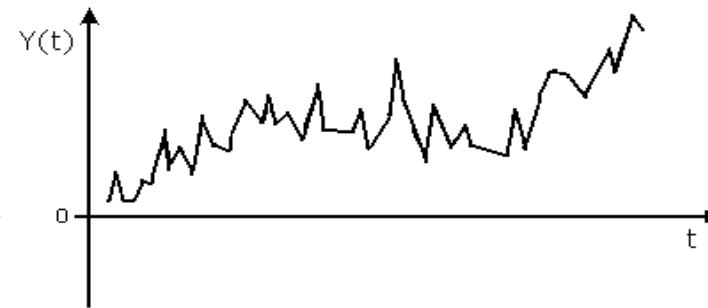
Tendencia



Componente aleatoria



Componente estacional



Serie de tiempo observada

El problema que se presenta es modelar adecuadamente las componentes de la serie.

Predicción por descomposición de series

ESTIMACIÓN DE LA TENDENCIA

Hay varios métodos para estimar la tendencia $T(t)$, uno de ellos es utilizar un modelo de regresión lineal.

Se pueden utilizar otros tipos de regresiones, como regresión cuadrática, logística, exponencial, entre otros.

ESTIMACIÓN DE LA TENDENCIA

EJEMPLO 1: La tabla presenta parte de los datos de una serie de energía eléctrica. Son 24 datos mensuales referentes a los años 1977 a 1978.

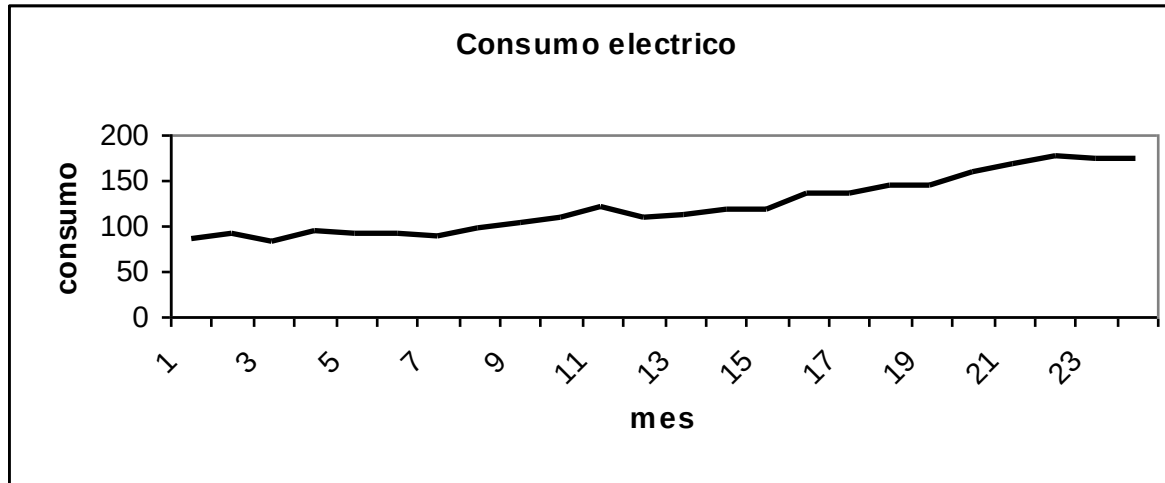
Consumo de Energía Eléctrica

t	Y(t)
1	84,6
2	89,9
3	81,9
4	95,4
5	91,2
6	89,8
7	89,7
8	97,9
9	103,4
10	107,6
11	120,4
12	109,6

t	Y(t)
13	110,3
14	118,1
15	116,5
16	134,2
17	134,7
18	144,8
19	144,4
20	159,2
21	168,2
22	175,2
23	174,5
24	173,7

ESTIMACIÓN DE LA TENDENCIA

Gráfico de la serie:



El modelo de tendencia propuesto es un modelo de regresión lineal:

$$Y(t) = \beta_0 + \beta_1 t + A(t)$$

Recurriendo al método de **mínimos cuadrados** se estiman los parámetros y se obtiene

$$T(t) = 68.45 + 4.24 * t$$

Predicción por descomposición de series

ESTIMACION DE LA COMPONENTE ESTACIONAL

Para estimarla, se debe conocer el período, y se deben tener datos de varios períodos consecutivos.

Por ejemplo, datos mensuales, estacionalidad de un año.

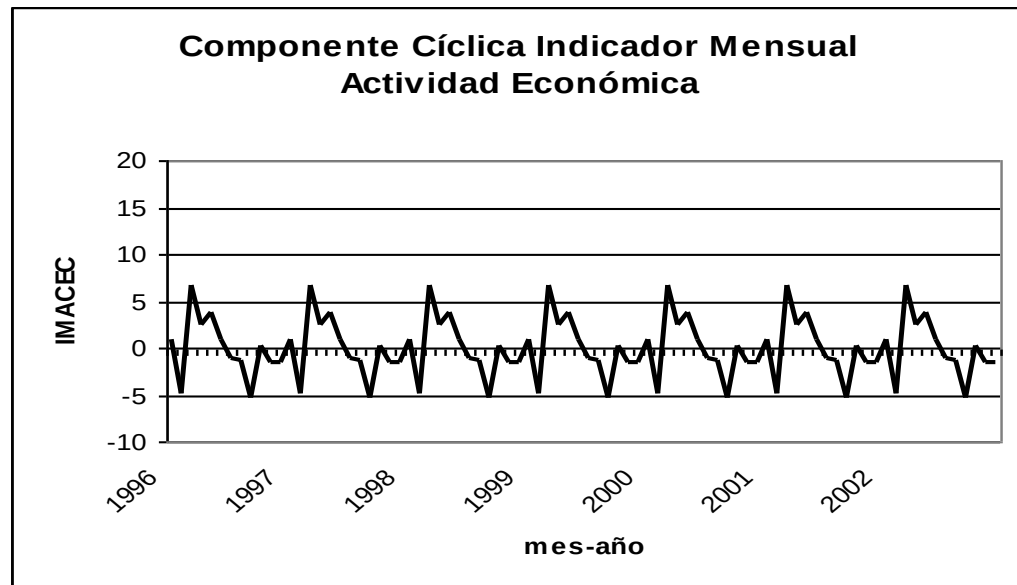
El ejemplo siguiente ilustra la forma de obtener la componente estacional.

ESTIMACION DE LA COMPONENTE ESTACIONAL

Asumiremos un modelo clásico aditivo.

- Entonces para obtener una estimación de la estacionalidad, restamos los valores ajustados de la tendencia a los datos, obteniendo una serie sin tendencia.
- Luego promediamos todos los valores de enero, los de febrero, los de marzo, etc., obteniendo doce valores mensuales promedio:

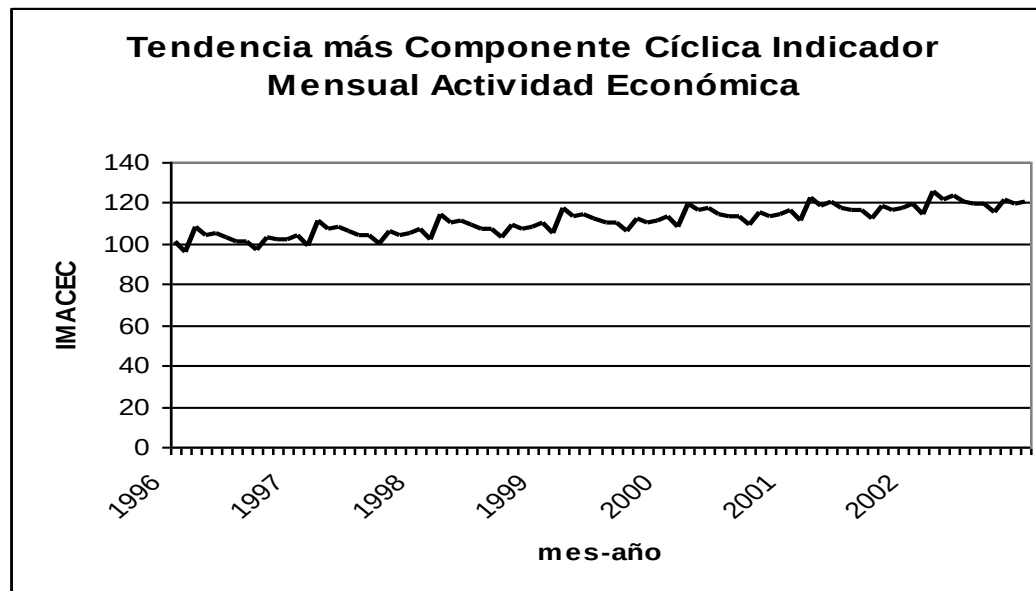
Mes	Prom.
Enero	0.8
Febrero	-4.9
Marzo	6.7
Abril	2.4
Mayo	3.8
Junio	0.8
Julio	-1.2
Agosto	-1.3
Septiembre	-5.4
Octubre	0.2
Noviembre	-1.7
Diciembre	-1.5



ESTIMACION DE LA COMPONENTE ESTACIONAL

Se observan valores altos a partir de marzo, y bajos en torno a septiembre.

Si recomponemos la serie con tendencia y componente cíclica, sin la componente aleatoria, tenemos la situación que se ilustra en el gráfico siguiente:



Con esto se
pueden hacer
predicciones
futuras,

Predicción Series Temporales

- Los modelos ARIMA responden al acrónimo de procesos **AutoR**egresivos, **I**ntegrados, y **M**edias móviles (**M**oving **A**verage),
- La idea subyacente fundamental consiste en admitir que las series temporales son generadas mediante un Proceso Generador de Datos que puede ser identificado y cuantificado y que, por tanto, pueden ser inferidos sus valores a futuro.
- Cuando realizamos una predicción de la evolución de una determinada serie temporal mediante la descomposición en los componentes estacional, tendencial, cíclico e irregular, el procedimiento consiste en identificar comportamientos regulares a lo largo de la serie (movimientos estacionales, tendenciales y cíclicos) y extrapolarlos a futuro, **asumiendo que los comportamientos irregulares tendrán un efecto promedio nulo.**

INTRODUCCIÓN A LOS MODELOS ARIMA

Fundamentos :

La modelización ARIMA asume que toda serie temporal está generada por un proceso estocástico (Proceso Generador de datos PGD) en la que los distintos valores observados Y_t responden a realizaciones (muestras) concretas de un conjunto de N variables aleatorias Z_t , que tienen unas determinadas probabilidades de ocurrencia asociadas a sus respectivas funciones de densidad $f(Z_t)$.

Estas funciones de densidad serán, en general, desconocidas y no pueden ser estimadas ya que sólo disponemos de una observación de cada una de ellas, por lo que se hace necesario asumir una serie de simplificaciones para poder realizar cualquier tipo de inferencia estadística.

INTRODUCCIÓN A LOS MODELOS ARIMA

Procesos estocásticos elementales: Camino aleatorio.

El *camino aleatorio* es un proceso tal que la diferencia entre dos valores consecutivos de la variable se comporta como un ruido blanco.

$$Z_t - Z_{t-1} = a_t \quad \text{o bien} \quad Z_t = Z_{t-1} + a_t$$

Si existe una tendencia sistemática en el cambio se denomina *camino aleatorio con deriva*.

$$Z_t - Z_{t-1} = m + a_t \quad \text{o bien} \quad Z_t = m + Z_{t-1} + a_t$$

INTRODUCCIÓN A LOS MODELOS ARIMA

Procesos estocásticos elementales: Proceso Autorregresivo.

Definimos un *proceso autorregresivo de primer orden* AR(1) como un proceso aleatorio que responde a una expresión del tipo

$$Z_t = \rho_0 + \rho_1 Z_{t-1} + a_t \quad \text{o bien} \quad \check{Z}_t = \rho_1 \check{Z}_{t-1} + a_t \quad \text{con } \check{Z}_t = Z_t - \rho_0$$

Para que el proceso AR(1) sea estacionario se debe cumplir que $-1 < \rho_1 < 1$, para que σ_z^2 sea finita y no negativa.

Los procesos autoregresivos pueden generalizarse al orden p AR(p) sin más que añadir términos retardados en la expresión general.

$$Z_t = \rho_0 + \rho_1 Z_{t-1} + \rho_2 Z_{t-2} + \dots + \rho_p Z_{t-p} + a_t$$

INTRODUCCIÓN A LOS MODELOS ARIMA

Procesos estocásticos elementales: Medias móviles.

Definimos una *media móvil de primer orden* MA(1) como un proceso aleatorio que responde a una expresión del tipo

$$Z_t = a_t + \theta_1 a_{t-1} \quad \text{con } Z_t \text{ en diferencias a la media}$$

Los procesos de medias móviles son estacionarios y, al igual que los autoregresivos pueden generalizarse al orden q MA(q) sin más que añadir términos retardados en la expresión general.

$$Z_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \dots + \theta_q a_{t-q}$$

INTRODUCCIÓN A LOS MODELOS ARIMA

Procesos estocásticos elementales: Procesos integrados.

Un proceso *integrado* es aquel que puede convertirse en estacionario aplicando diferencias.

Así, por ejemplo, un camino aleatorio sería un proceso integrado de orden 1 $I(1)$, ya que puede convertirse en estacionario tomando primeras diferencias.

Definimos el orden de integración de un proceso como el número de diferencias que debemos aplicarle para convertirlo en estacionario.

En algunas ocasiones las diferencias deben aplicarse sobre el valor estacional.

$$Z_t - Z_{t-s} = e_t \quad \text{con } s = 4 \text{ ó } 12 \quad e_t \text{ estacionario}$$

INTRODUCCIÓN A LOS MODELOS ARIMA

Proceso Generador de Datos.

Mediante la adecuada combinación de estos procesos elementales: integración, AR(p), y MA(q) podemos representar la evolución de cualquier serie temporal.

$$\nabla Y_t = \rho_1 \nabla Y_{t-1} + \rho_2 \nabla Y_{t-2} + \cdots + \rho_p \nabla Y_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q}$$

$$\nabla Y_t \rho_p(B) = \theta_q(B) a_t \Rightarrow \nabla Y_t = \frac{\theta_q(B)}{\rho_p(B)} a_t \quad \text{con} \quad \nabla Y_t = Y_t - Y_{t-1} = Y_t(1-B)$$