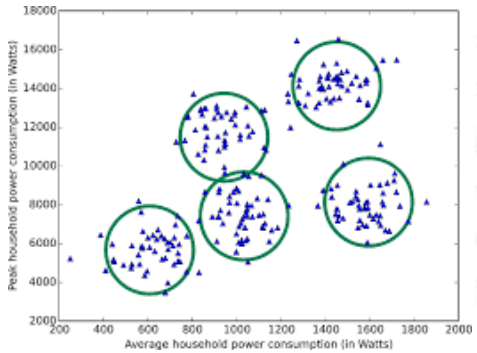
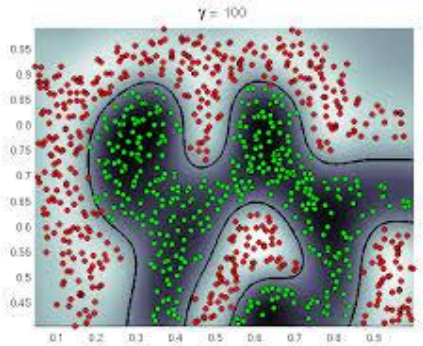
AGENDA

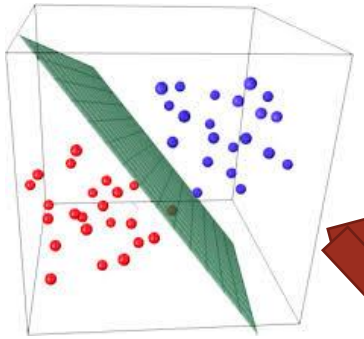
Aprendizaje automático



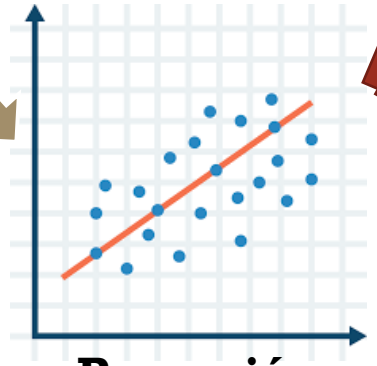
Aprendizaje no supervisado



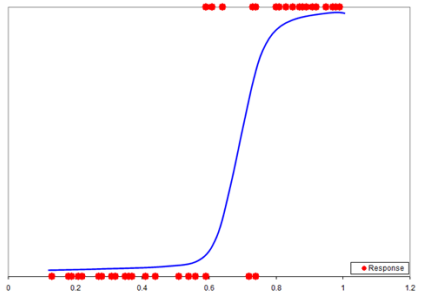
Aprendizaje supervisado



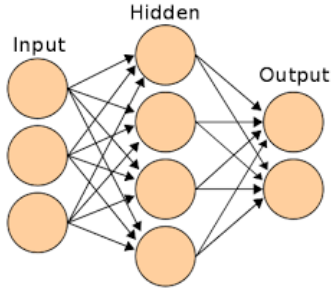
Clasificación



Regresión

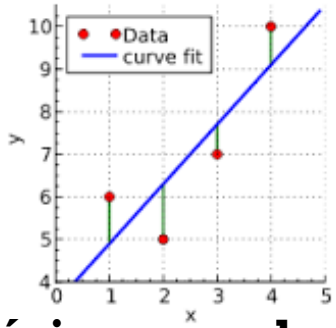


Regresión logística



Redes neuronales artificiales

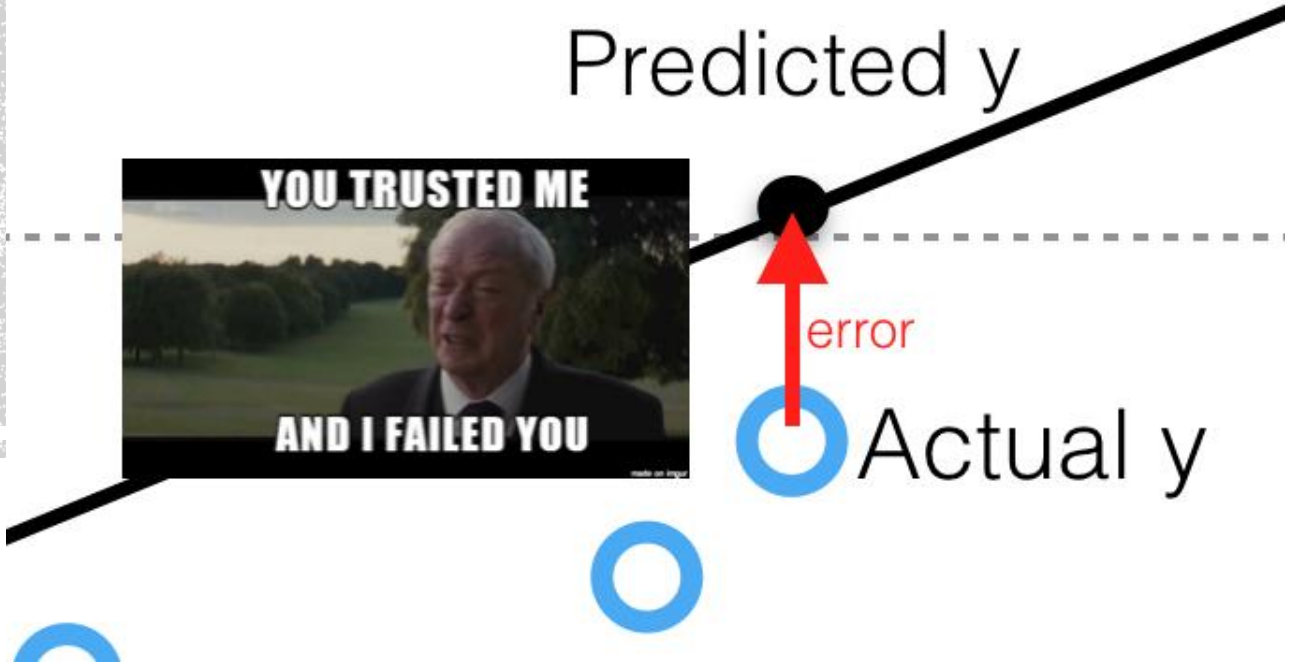
Métricas de Evaluación de la regresión



Mínimos cuadrados ordinarios

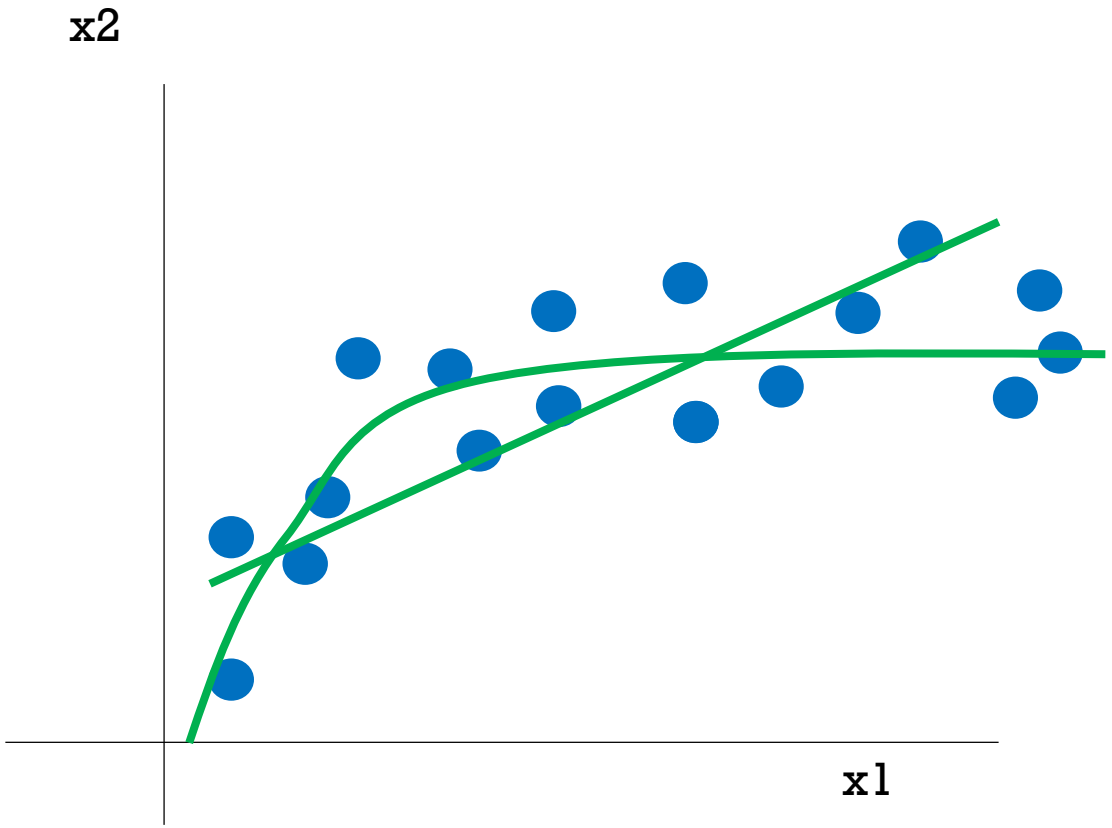
REGRESIÓN LINEAL

(Mínimos Cuadros Ordinarios)



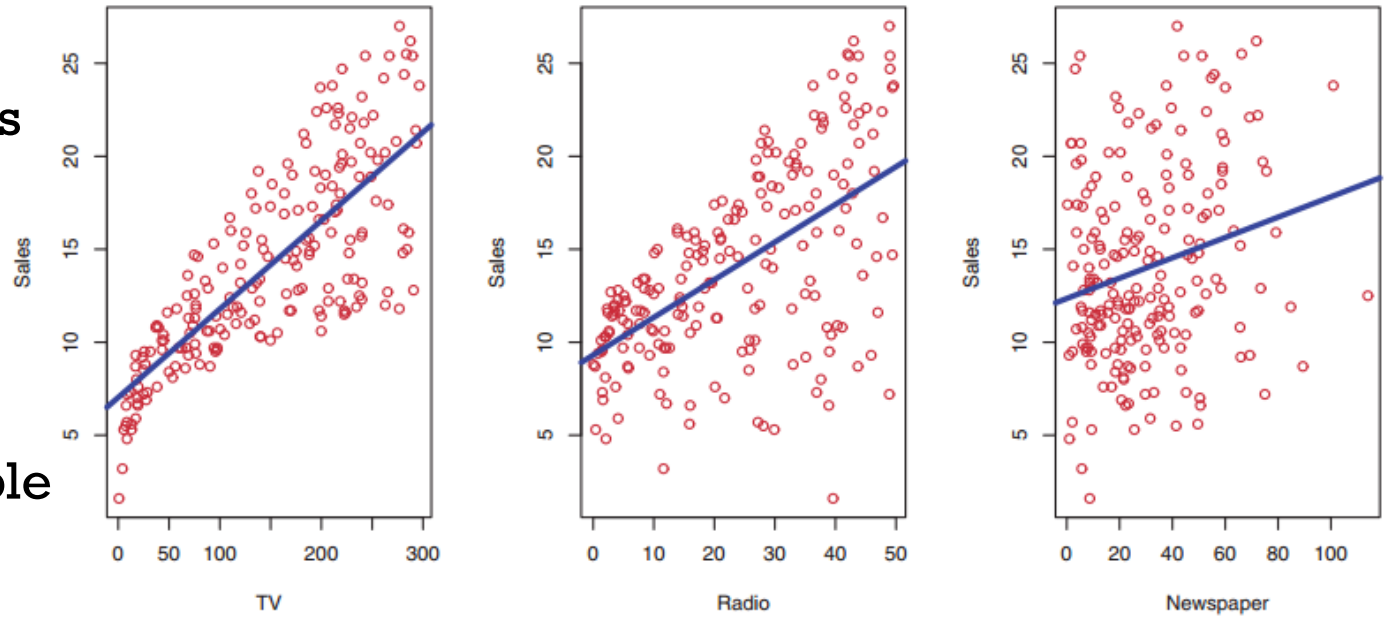
REGRESIÓN

- Encontrar modelos que permitan predecir valores continuos:
  - KNN
  - Regresión lineal
  - Regresión polinómica
  - Árboles de regresión
  - …
- Valores **continuos** de la variable objetivo
- **Baseline**: medida de evaluación dada por un modelo que predice una medida de tendencia central (e.g. el promedio)



REGRESIÓN

- **Predicción**:
  - Procesos de caja negra
  - Estimar el valor objetivo Y dado los valores de los predictores X
- **Inferencia**:
  - ¿Cuáles son los predictores asociados con la respuesta?
  - ¿Cuál es la relación entre la variable respuesta y cada uno de los predictores?
  - ¿Se puede resumir esa relación linealmente o se trata de una relación mas compleja?



$$Ventas = f(TV, Radio, Periódicos)$$

ISLR, 2013

REGRESIÓN LINEAL – MÍNIMOS CUADRADOS

- Se busca una relación lineal entre los atributos predictores (x_i) y el atributo objetivo (Y)

$$Y = h_\Theta(\boldsymbol{X}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n + \varepsilon$$

Residuos (a minimizar)

Aprendizaje: Como aprendemos los parámetros?

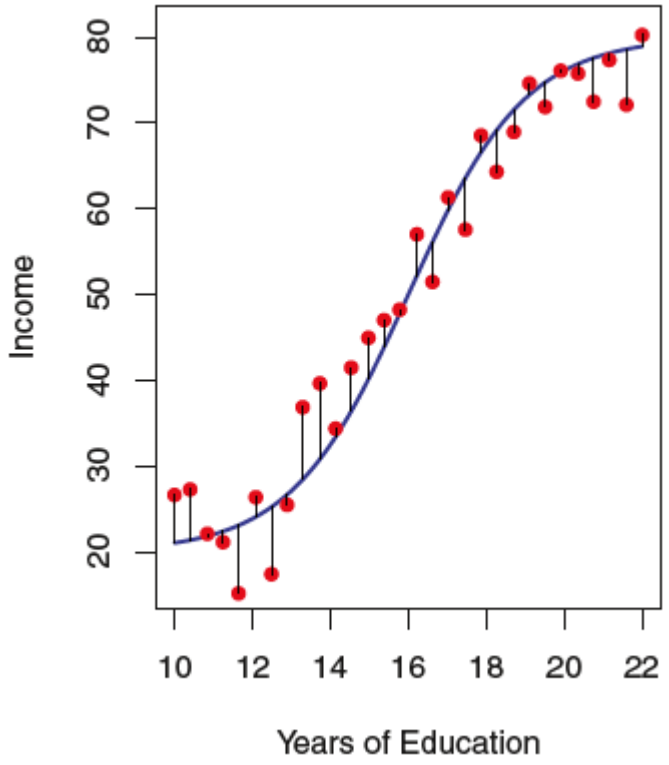
- Regresión lineal simple (un predictor), y múltiple (varios predictores)
- Los parámetros θ_i son estimados teniendo como objetivo la minimización de los residuos o diferencias cuadradas entre las predicciones ($\hat{Y}$) y los valores reales (Y):

$$argmin_\Theta \sum_1^m (\theta_0 + \theta_1 x_1^{(m)} + \cdots + \theta_n x_n^{(m)} - y^{(m)})^2 ,$$

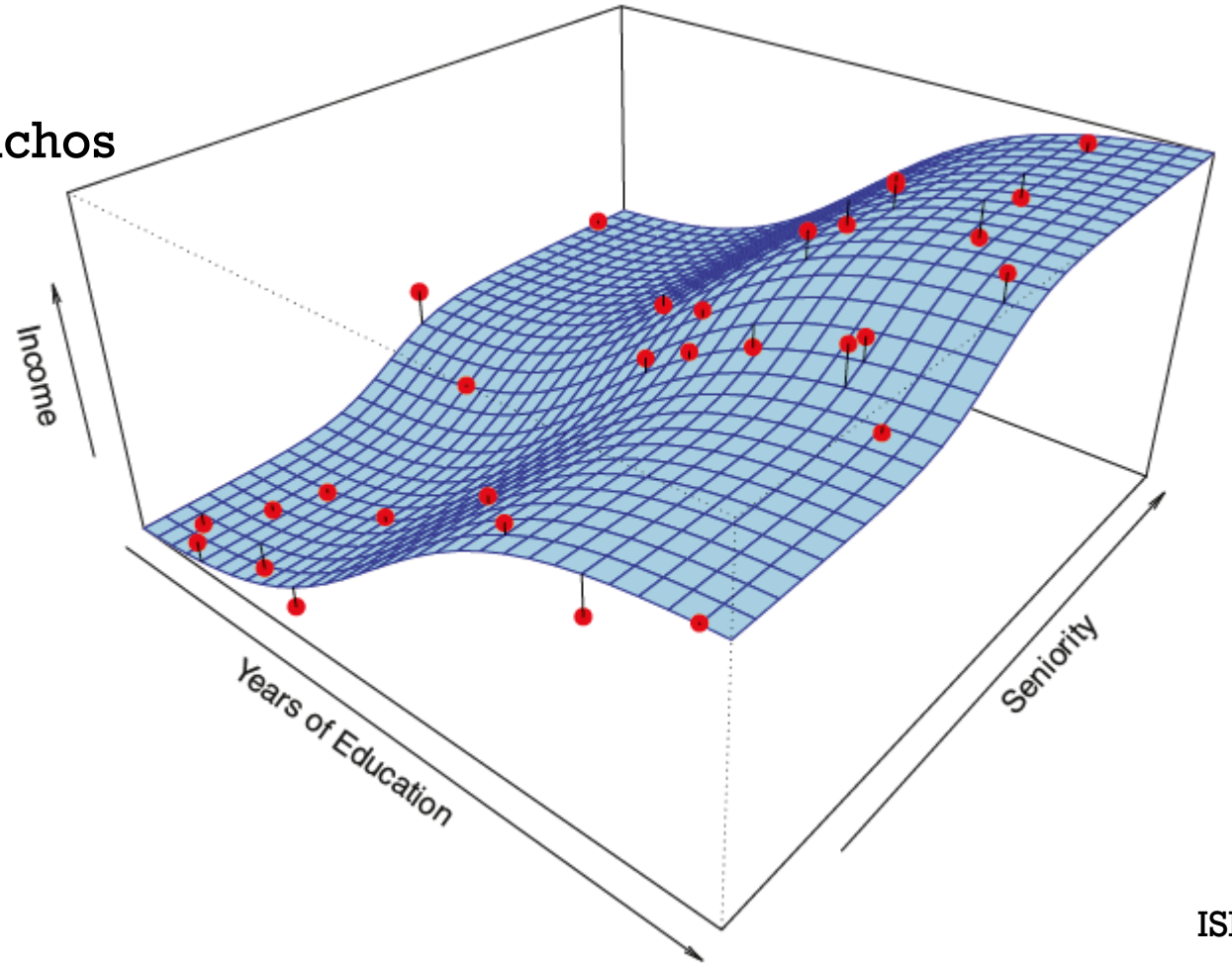
donde $x_i^{(m)}$ se interpreta como el valor de la i-esima variable independientes de la la m-ésima instancia de datos

RESIDUOS

Residuos: diferencia entre los valores reales y los valores predichos



Regresión polinómica

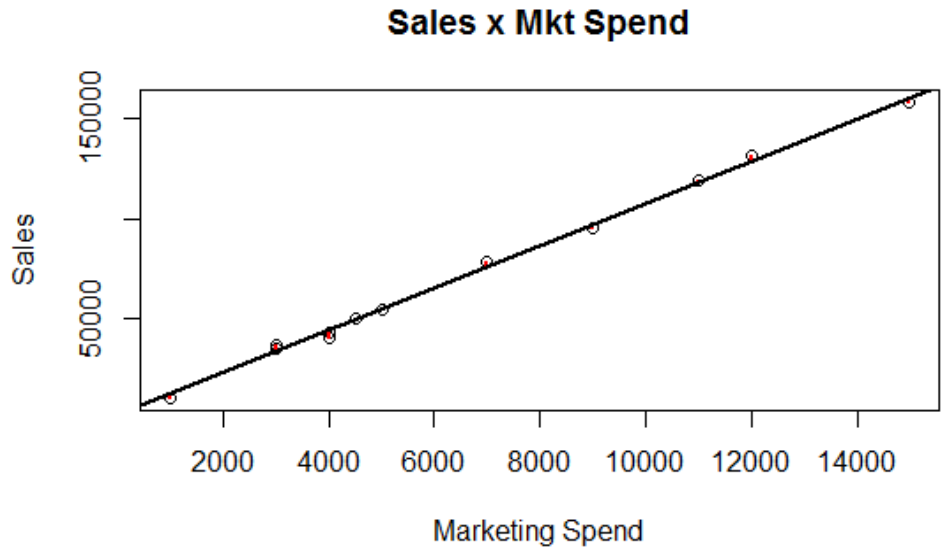


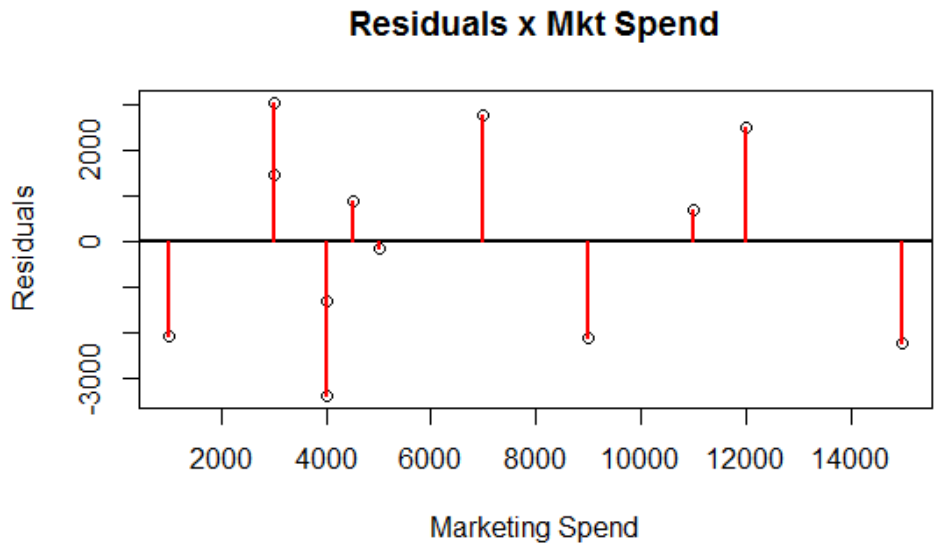
ISLR, 2013

Regresión no paramétrica: thin-plate spline

RESIDUOS

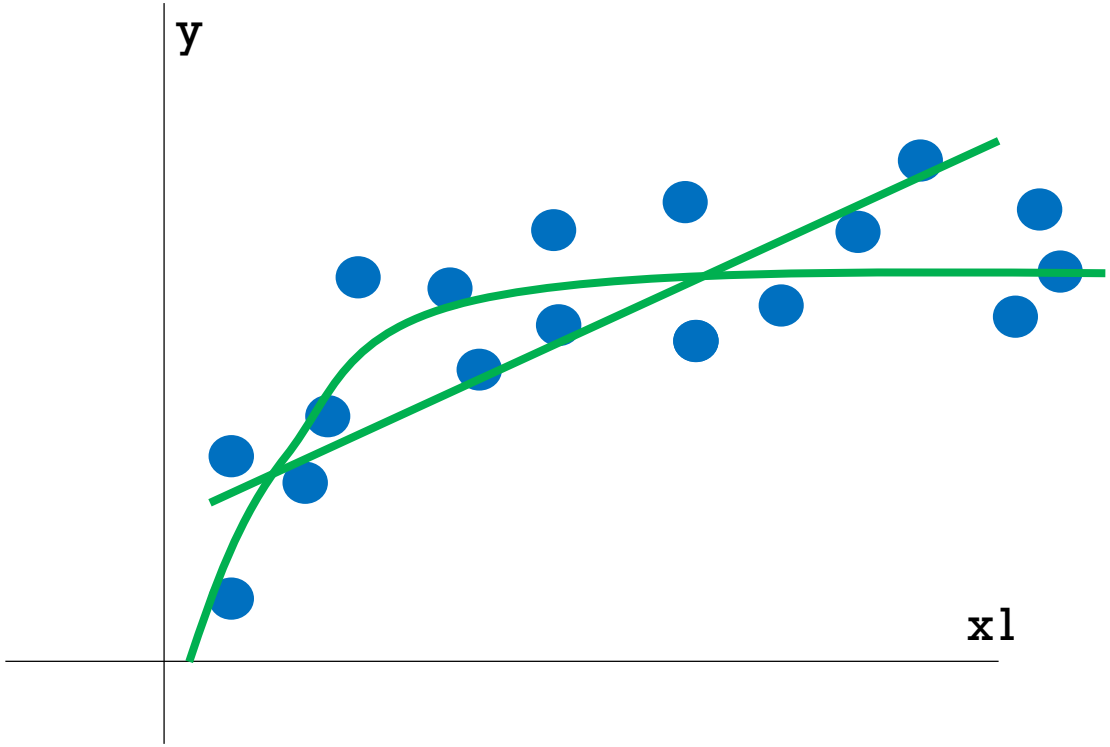
- Los **residuos** representan la variación que no se puede explicar por la regresión
  - Son estimadores de los errores
  - Pueden ser explicados en parte por la no consideración de otros predictores
  - $\sum e_i = 0$
  - $\sum x_i e_i = 0$
- El análisis de los **residuos** proporciona una manera de evaluar la calidad de la regresión:
  - Errores deben ser independientes
  - Su varianza debe ser constante
  - Deben estar normalmente distribuidos
  - No deben presentar patrones (e.g sinusoide)





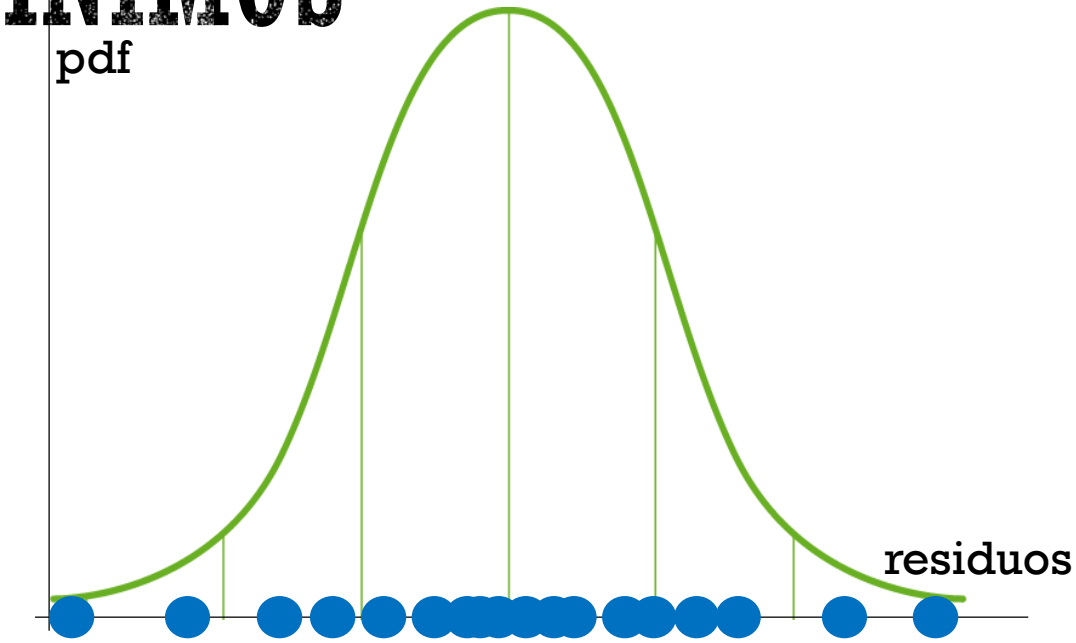
REGRESIÓN LINEAL – MÍNIMOS CUADRADOS

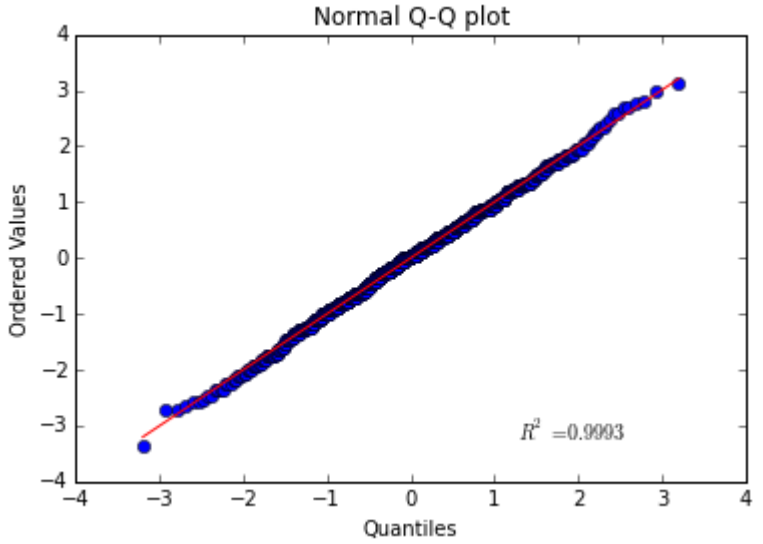
- Supuestos estadísticos para la utilización de la regresión lineal:
  - →**Linealidad y aditividad** en la relación entre variables independientes y variable dependiente
- Para validar:
  - Validación visual en plot de valores reales vs. valores ajustados (distribución simétrica con respecto a la diagonal)
  - Validación visual en plot de residuos vs. valores ajustados (distribución simétrica con respecto al eje de los valores ajustados)

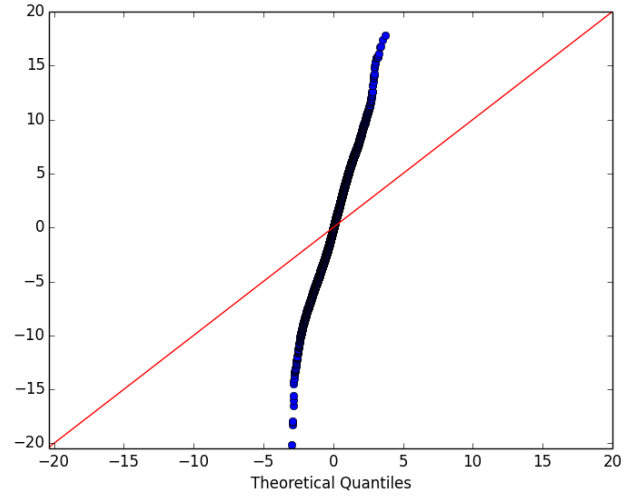


REGRESIÓN LINEAL – MÍNIMOS CUADRADOS

- Supuestos estadísticos para la utilización de la regresión lineal:
  - →**Normalidad** de la variable dependiente para valores fijos de las variables independientes
- Para validar:
  - Los residuos siguen una distribución normal
  - Utilizar un QQ-Plot
  - Prueba de Kolmogorov-Smirnov
  - Prueba de Shapiro-Wilk
  - Prueba de Jarque-Bera
  - Prueba de Anderson-Darling

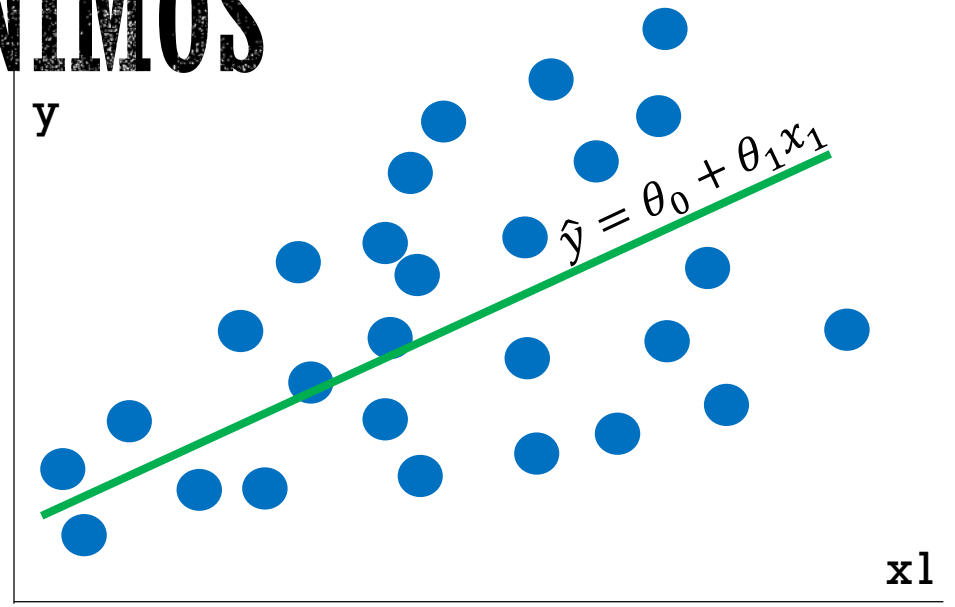


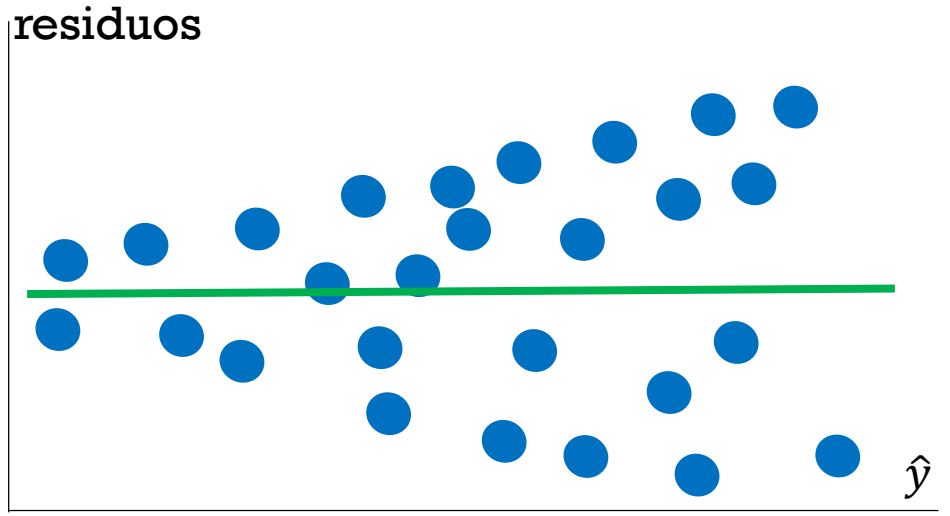




REGRESIÓN LINEAL – MÍNIMOS CUADRADOS

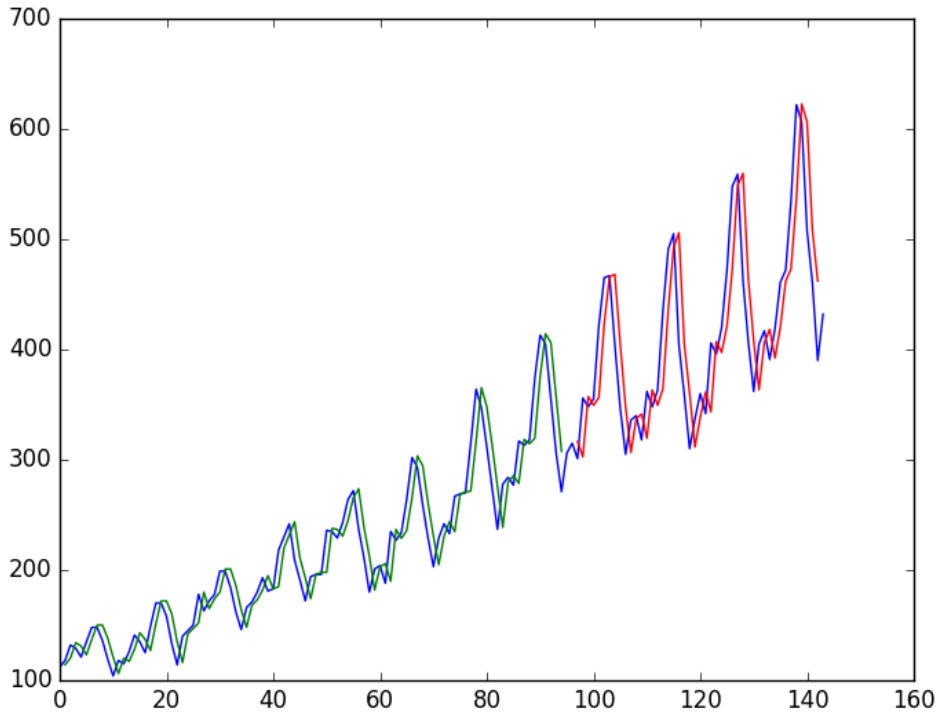
- Supuestos estadísticos para la utilización de la regresión lineal:
  - →**Homocedasticidad** de la variable dependiente no varia para diferentes valores de las variables independientes
- Para validar:
  - La dispersión de los residuos permanece constante con respecto a la variable ajustada
  - No hay aumento ni reducción de la varianza en los extremos del eje de los valores ajustados





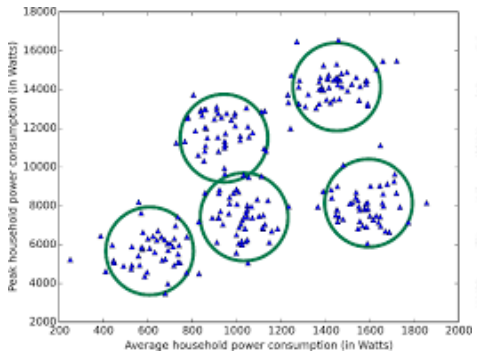
REGRESIÓN LINEAL – MÍNIMOS CUADRADOS

- Supuestos estadísticos para la utilización de la regresión lineal:
  - →**Independencia** de los valores de la variable dependiente (y de los residuos)
- Para validar:
  - Sentido común:
    - Analizar si hay alguna razón por la que un valor de la variables dependiente de una instancia influencie el de otra instancia.
    - Analizar la variable dependiente con respecto a una variable independiente temporal
  - Prueba Durbin-Watson de auto correlación en los residuos consecutivos (series de tiempo)

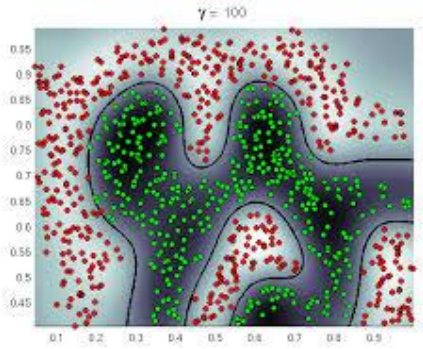


AGENDA

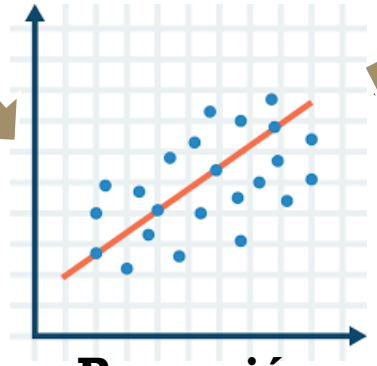
Aprendizaje automático



Aprendizaje no supervisado

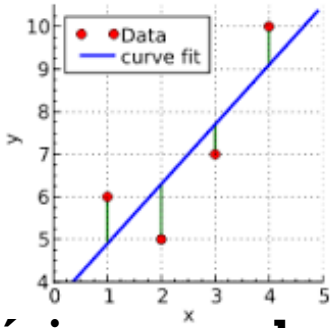


Aprendizaje supervisado



Regresión

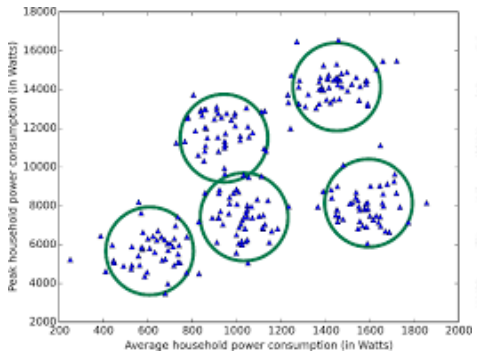
Métricas de Evaluación de la regresión



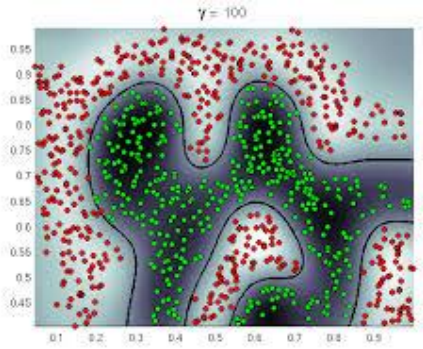
Mínimos cuadrados ordinarios

AGENDA

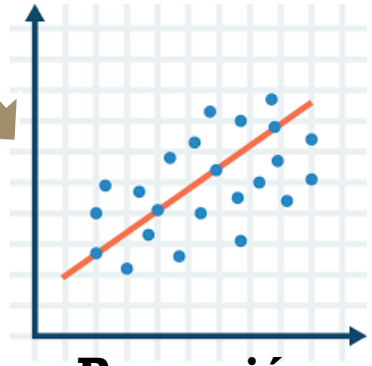
Aprendizaje automático



Aprendizaje no supervisado

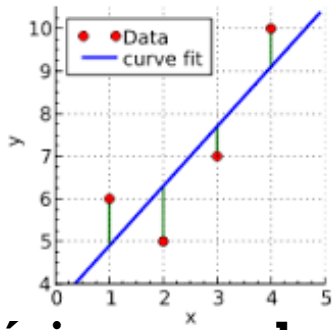


Aprendizaje supervisado



Regresión

Métricas de Evaluación de la regresión



Mínimos cuadrados ordinarios

MÉTRICAS DE REGRESIÓN

Coeficiente de correlación (Pearson $\rho \in [-1;1]$): indica la fuerza de la relación linear entre los predictores y la variables objetivo, que puede ser positive o negativa

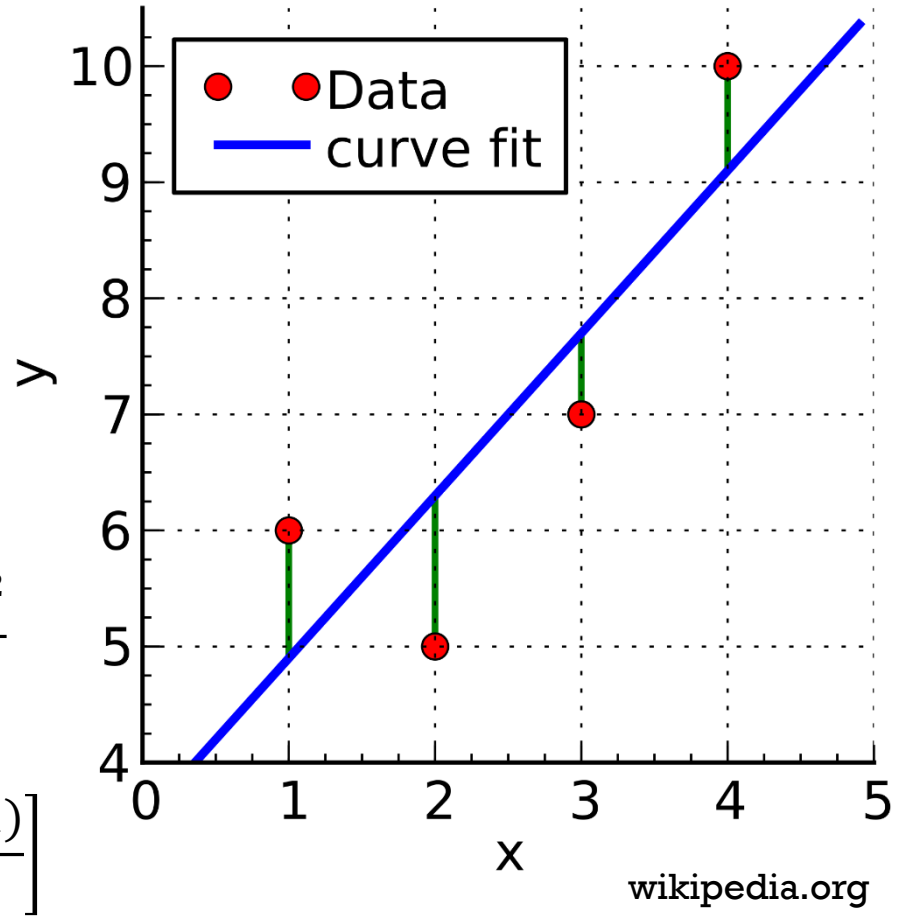
- $|\rho| = 0$ no hay correlación
- $|\rho| = 0.10$ correlación muy débil
- $|\rho| = 0.25$ correlación débil
- $|\rho| = 0.50$ correlación media
- $|\rho| = 0.75$ correlación fuerte
- $|\rho| = 0.90$ correlación muy fuerte
- $|\rho| = 1$ correlación perfecta

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Coeficiente de determinación ($R^2 = \rho^2$): indica el porcentaje de la varianza que pudo ser explicada por los predictores a partir de la relación linear

MÉTRICAS DE REGRESIÓN

- MAE (mean absolute error): $\frac{1}{m}\sum_{1}^{m}|h_\theta(x_i)-y_i|$
- MSE (mean square error): $\frac{1}{m}\sum_{1}^{m}(h_\theta(x_i)-y_i)^2$
- MSE (root mean square error): $\sqrt{\frac{1}{m}\sum_{1}^{m}(h_\theta(x_i)-y_i)^2}$
- R^2 (coeficiente de determinación): $1-\frac{\sum_1^m(h_\theta(x_i)-y_i)^2}{\sum_1^m(y_i-\bar{y})^2}$
- R^2 ajustado (penaliza el número k de variables independientes) $R^2_{adj}=1\left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$



VARIABLES CATEGÓRICAS

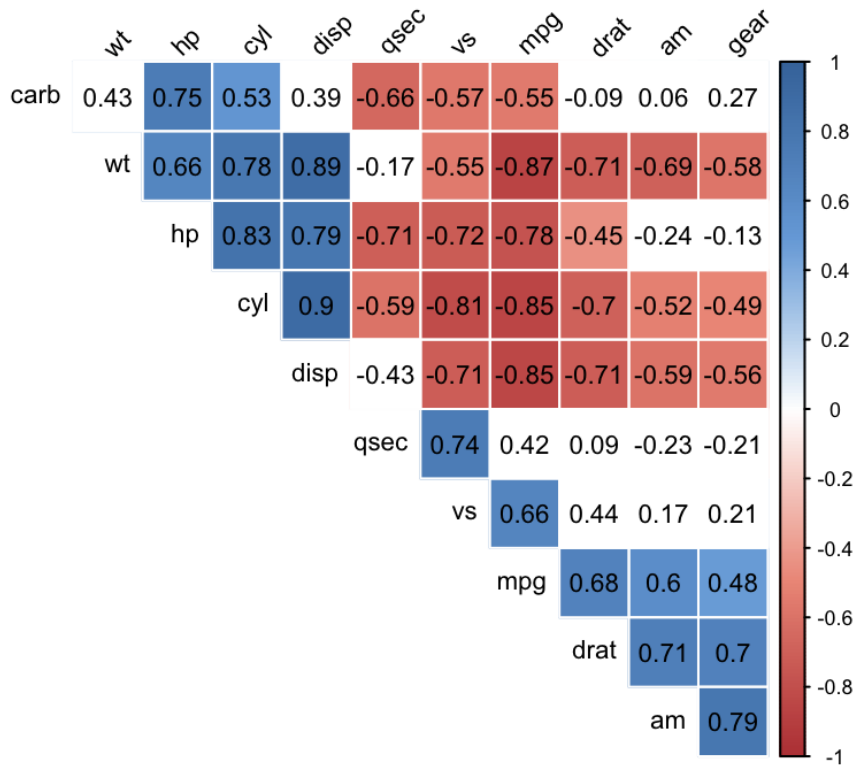
- Las variables predictoras deben ser numéricas.
- Las variables categóricas debes ser convertidas en numéricas:
  - One hot encoding: se crea una variable para cada valor posible de cada variable categórica
  - Contraste o "dummy": se crea una variable para cada valor posible menos 1 de cada variable categórica.

Ejemplo: variable estrato con 3 valores posibles (bajo, medio y alto)

	Estrato_bajo	Estrato_medio
Valor = bajo	1	0
Valor = medio	0	1
Valor = alto	0	0

REGRESIÓN LINEAL MÚLTIPLE – MÍNIMOS CUADRADOS

- Supuestos estadísticos para la utilización de la regresión lineal: múltiple
  - →**Variables predictoras linealmente independientes entre ellas**
  - →Evitar problema de la **multicolinealidad**
- Para validar:
  - Correlograma, filtrar variables con correlaciones altas (e.g. > |0.85|)



REGRESIÓN LINEAL MÚLTIPLE

Si se dispone de varias variables independientes, hay varias maneras de establecer el modelo de regresión múltiple a utilizar, dada una medida de bondad de ajuste:

- Completo: se evalúan todas las posibles combinaciones de variables independientes, y se escoge la mejor
- Tamaño fijo: se evalúan todas las posibles combinaciones de K variables independientes y se escoge la mejor
- Stepwise
  - Forward: se prueba una a una con las variables independientes aún no escogidas y se evalúa el modelo conjuntamente con las variables previamente escogidas, escogiendo la mejor. Se para cuando la medida de bondad de ajuste no mejore.
  - Backward: sigue un proceso contrario al forward, empezando con todas las variables y eliminando la variable que al no considerarla optimice la medida de bondad de ajuste.
- PCA: se transforman los datos en nuevas dimensiones no correlacionadas ordenadas con respecto a la cantidad de varianza que describen.

REGRESIÓN LINEAL – MÍNIMOS CUADRADOS

Los coeficientes que se obtienen para cada variable independiente deben ser analizados uno a uno, pues puede que no sean realmente significativos en la relación con la variable dependientes, dados las demás variables independientes.

Las herramientas de regresión lineal incluyen pruebas de hipótesis que comparan cada coeficiente contra el valor 0, y proveen el valor-p correspondiente. Si el valor-p es inferior a un alpha dado, se concluye entonces que el coeficiente es diferente de 0 y la relación entre variable independiente y dependiente es significativa.

REGRESIÓN LINEAL – MÍNIMOS CUADRADOS

$$h_{\boldsymbol{\Theta}}(x) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

- Si n=1, hay una solución simple

$$\theta_1 = \rho_{x_1,y} \frac{s_{x_1}}{s_y} \qquad \rho_{x_1,y} = \frac{\sum(x_1 - \overline{x_1})(y - \bar{y})}{s_{x_1} s_y}$$

$$\theta_0 = \bar{y} - \theta_1 \overline{x_1}$$

- Su n>1, la **ecuación normal** proporciona un solución analítica exacta para el problema de regresión linear, minimizando los mínimos cuadrados:

$$\Theta = (\boldsymbol{x}^T \boldsymbol{x})^{-1} \boldsymbol{x}^T \boldsymbol{y}$$

  - No se necesita normalizar los datos con la ecuación normal
  - Puede ser demasiado costosa cuando n es muy grande. La complejidad computacional de $(\boldsymbol{x}^T \boldsymbol{x})^{-1}$ es $O(n^3)$.
  - $\boldsymbol{x}^T \boldsymbol{x}$ debe tener inverso (no deben haber dependencias lineares entre las variables, y m<n).
- **Big data (n>10000)** → usar **descenso de gradiente**

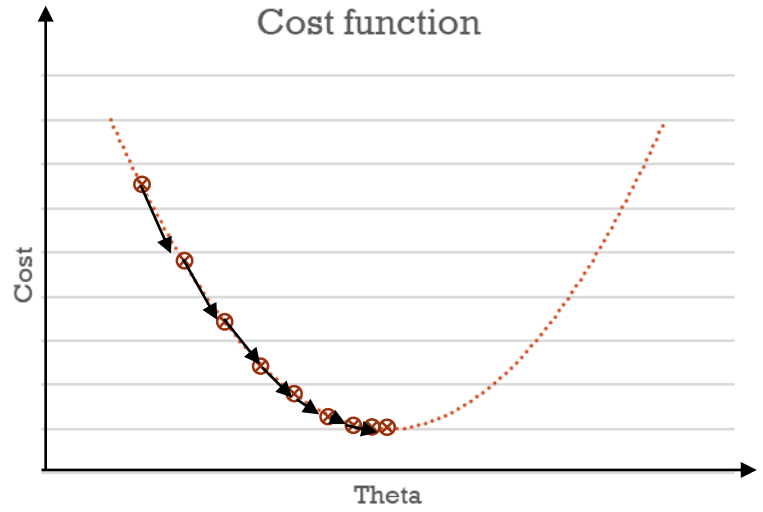
REGRESIÓN LINEAL – MÍNIMOS CUADRADOS

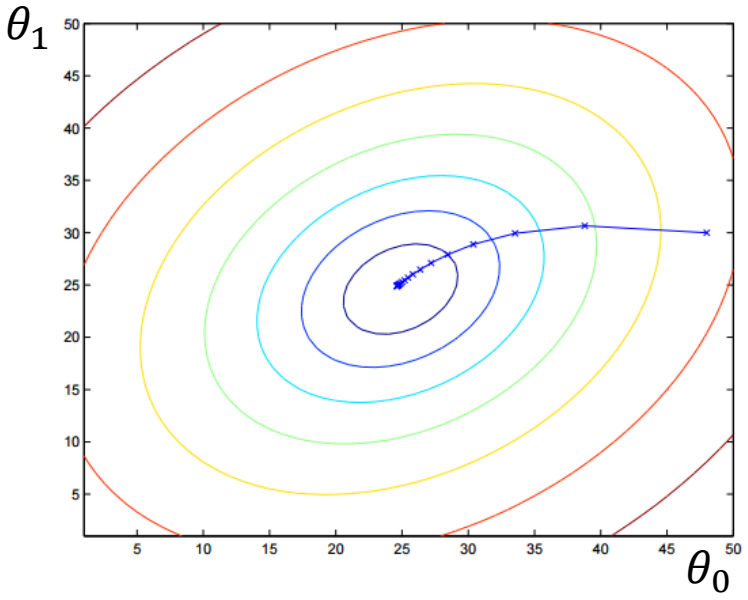
Descenso de gradiente

- Objetivo: encontrar los parámetros Θ que minimizan una función de costo (o loss) $J(\Theta)$.
- Para la regresión lineal múltiple: $J(\Theta) = \frac{1}{2m}\sum_{j=1}^{m}(\theta_0 + \theta_1 x_1^{(j)} + \cdots + \theta_n x_n^{(j)} - y^{(j)})^2$, dadas las m instancias de datos disponibles.
- Algoritmo:
  1. Escoger aleatoriamente valores para cada θ_i.
  2. Actualizar todos los θ_i simultáneamente: $\theta_i := \theta_i - \alpha\frac{\partial}{\partial x}J(\Theta)$
  3. Parar cuando se llegue a convergencia
- Solución:
  - Para la intercepción: $\theta_0 := \theta_0 - \alpha\frac{1}{m}\sum_{j=1}^{m}\left(\theta_0 + \theta_1 x_1^{(j)} + \cdots + \theta_n x_n^{(j)} - y^{(j)}\right)$
  - Para los coeficientes: $\theta_i := \theta_i - \alpha\frac{1}{m}\sum_{j=1}^{m}\left(\theta_0 + \theta_1 x_1^{(j)} + \cdots + \theta_n x_n^{(j)} - y^{(j)}\right) * x_i$

REGRESIÓN LINEAL – MÍNIMOS CUADRADOS

Descenso de gradiente

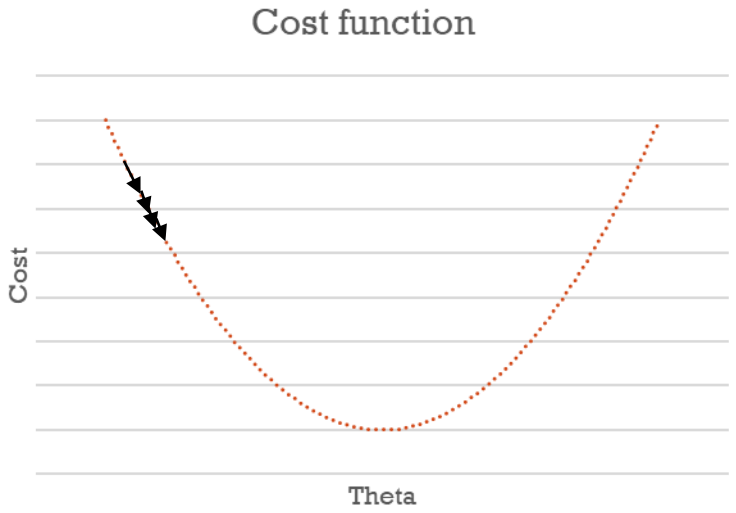




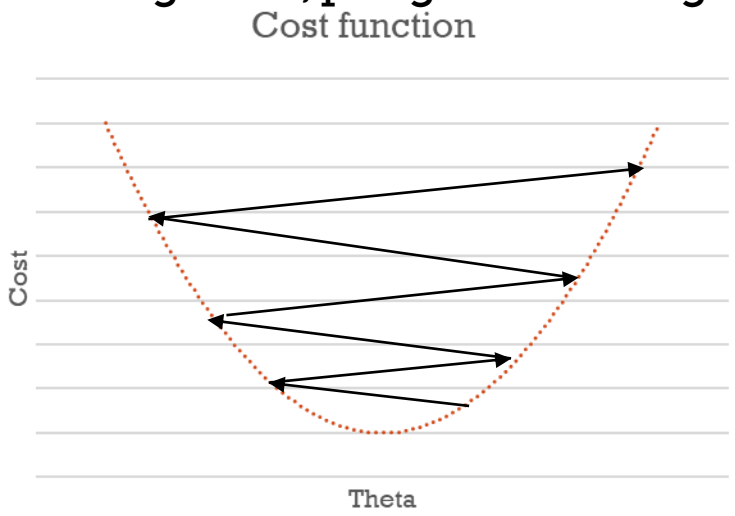
REGRESIÓN LINEAL – MÍNIMOS CUADRADOS

Descenso de gradiente

- Escogencia de la **taza de aprendizaje** α:
  - Si α muy pequeño: demorado llegar a convergencia



  - Si α muy grande: demorado llegar a convergencia, peligro de divergencia

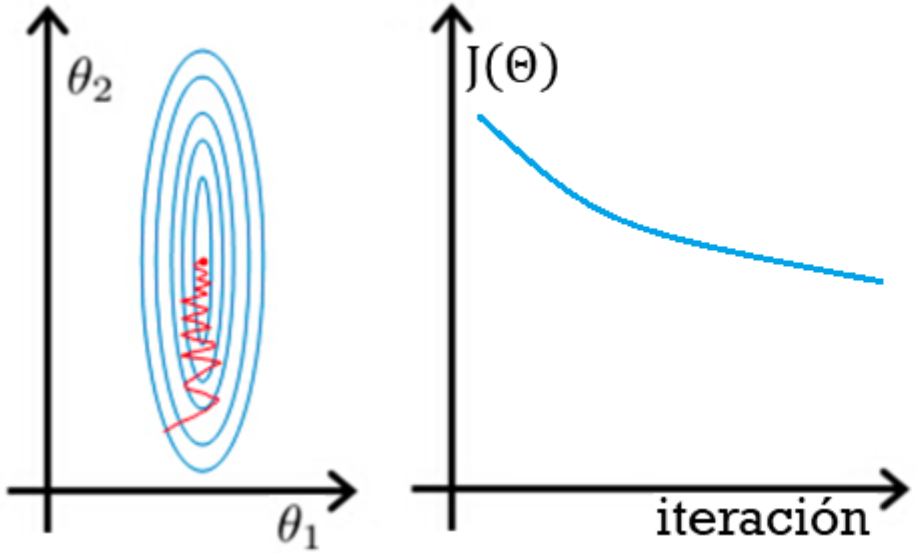


  - α debe decrecer siempre en cada iteración; en caso contrario, se debe reducir el valor de α
  - Se debe intentar con varios valores de α: 0.001, 0.01, 0.1, 1

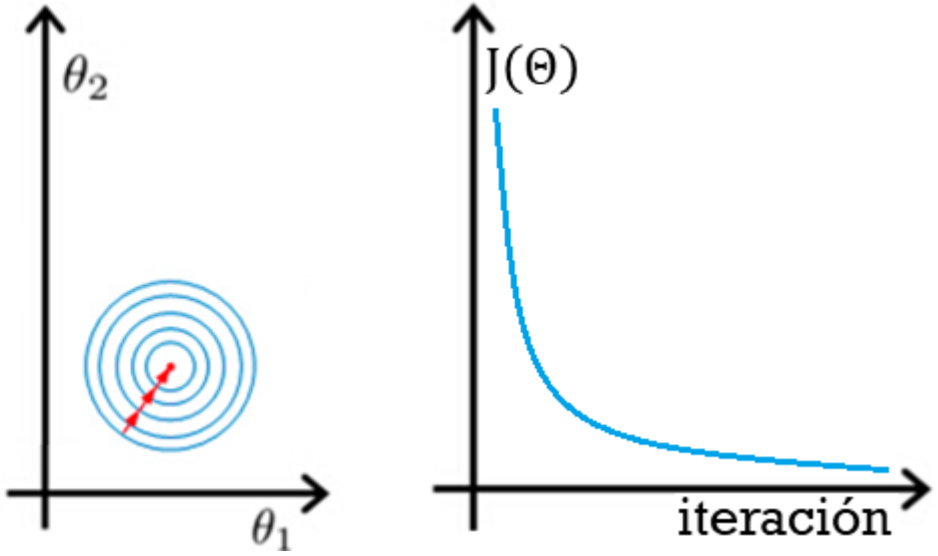
REGRESIÓN LINEAL – MÍNIMOS CUADRADOS

Feature Scaling

- Si escalas muy diferentes: demorado llegar a convergencia, sobre influencia de las derivadas parciales de las variables de mayor escala iteración

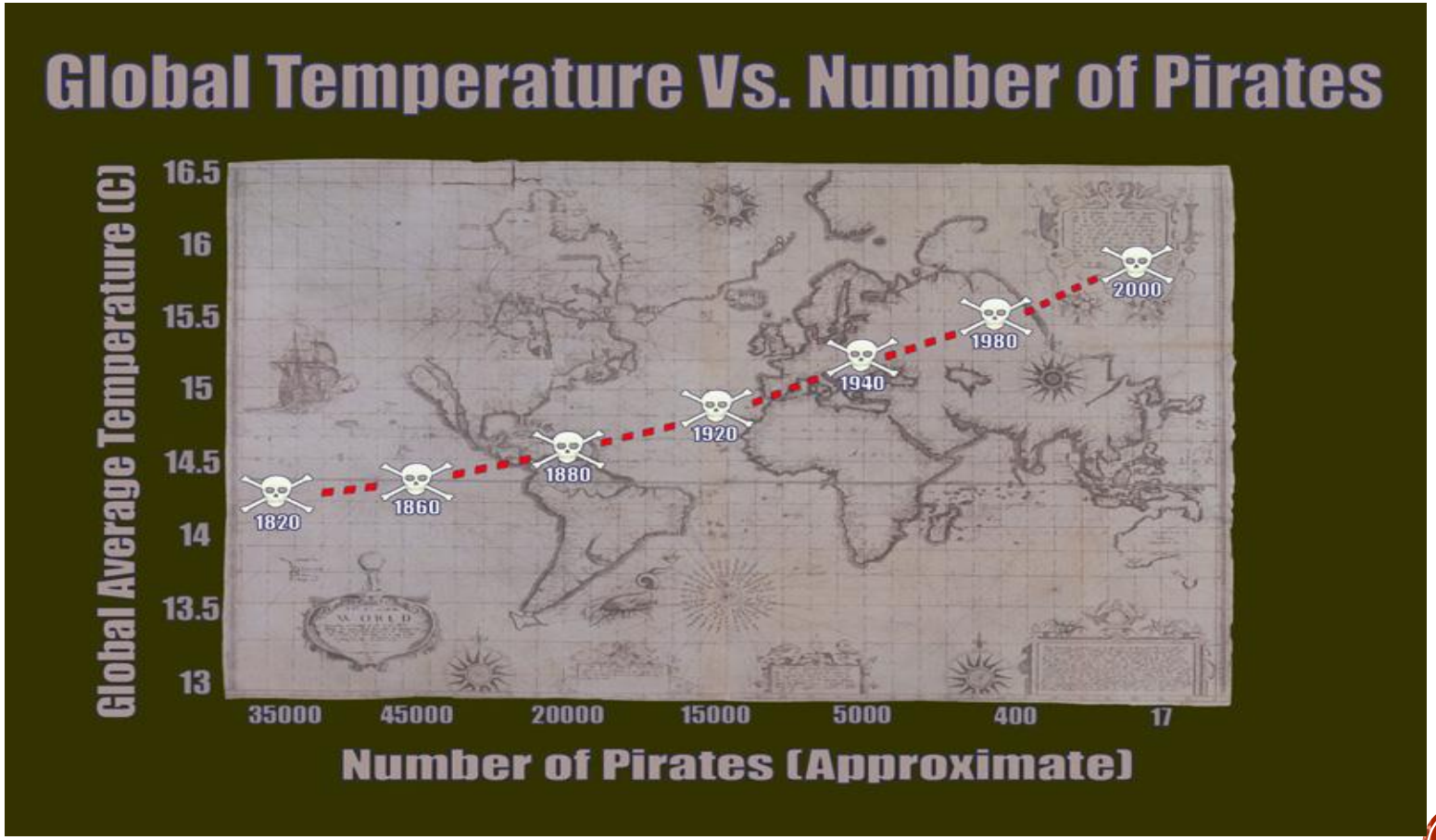


- Si misma escala: influencia igual de las derivadas parciales de todos los parámetros



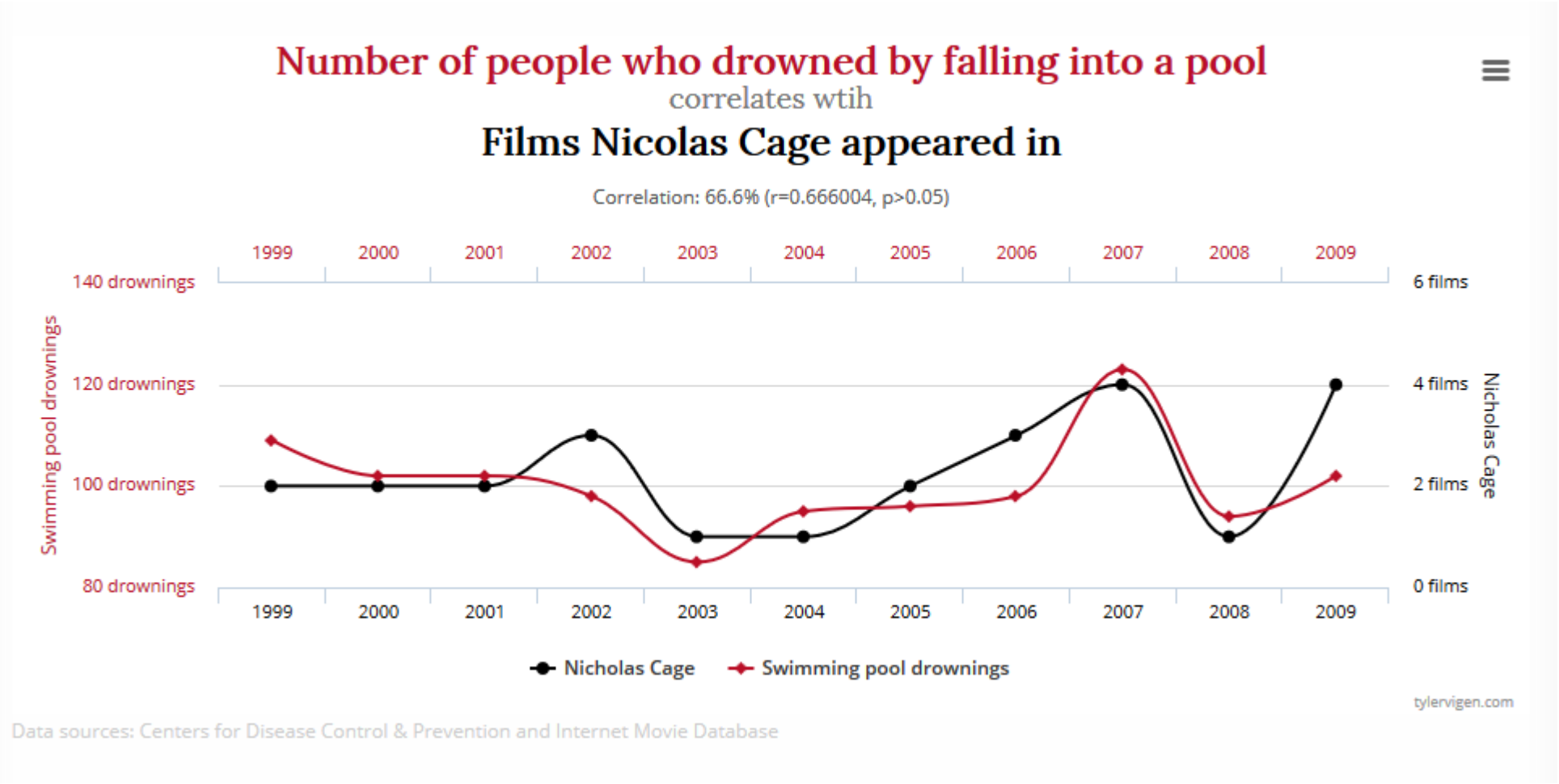
REGRESIÓN – CUIDADO!

Correlación y causalidad son dos cosas muy diferentes



Erika Andersen, Forbes, 2012

REGRESIÓN – CUIDADO!

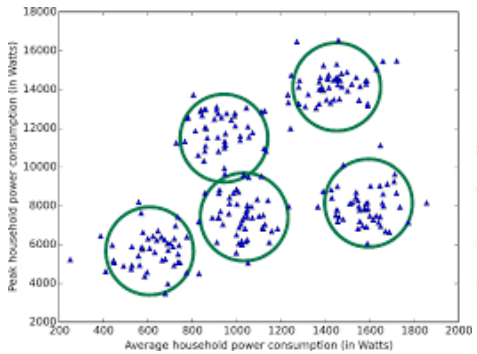


TALLER: REGRESIÓN MÍNIMOS CUADRADOS

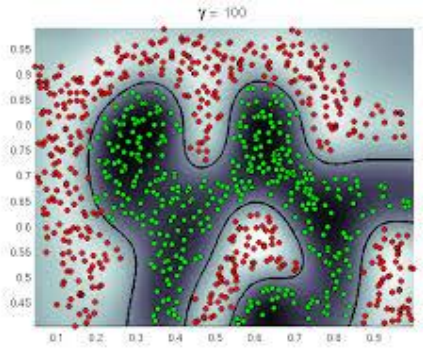
- Basarse en el notebook de regresión sobre el dataset "MPG Auto" para crear un modelo de regresión para el dataset "Residential Building".

AGENDA

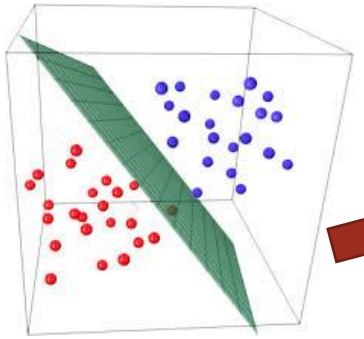
Aprendizaje automático



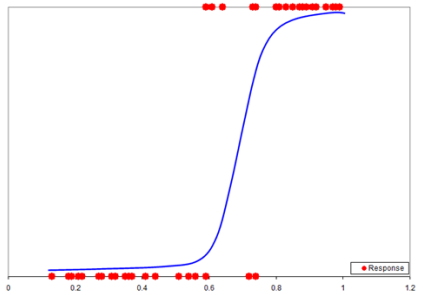
Aprendizaje no supervisado



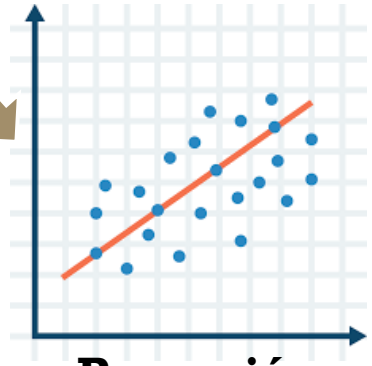
Aprendizaje supervisado



Clasificación

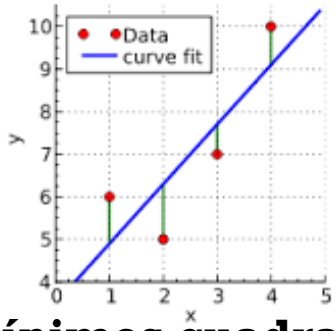


Regresión logística



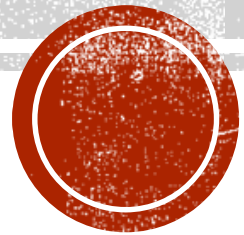
Regresión

Métricas de Evaluación de la regresión



Mínimos cuadrados ordinarios

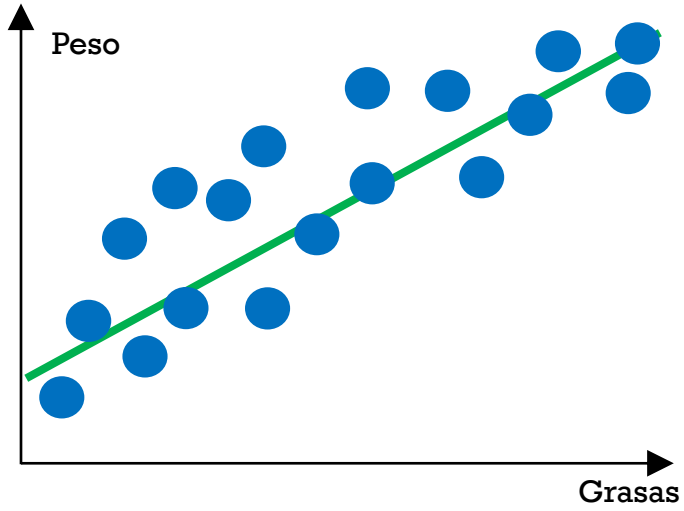
REGRESIÓN LOGÍSTICA

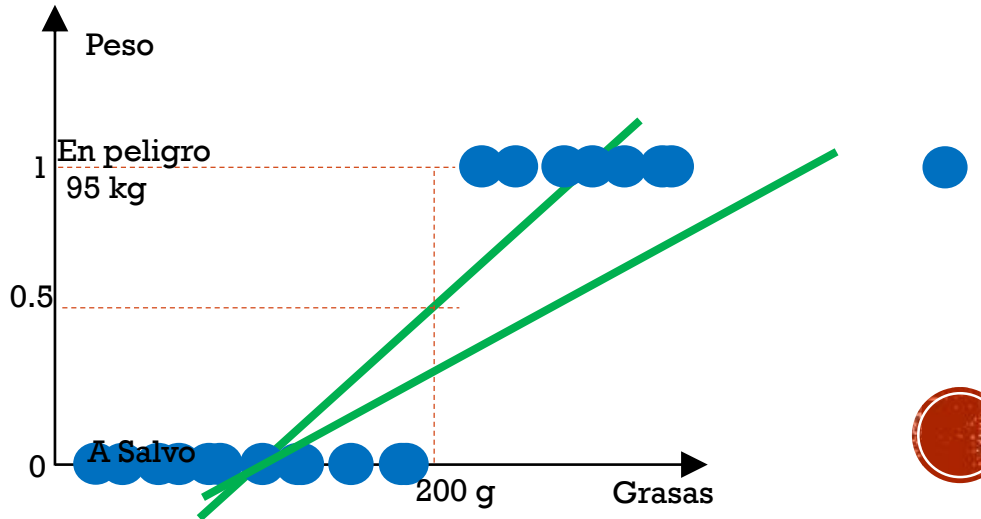


¿REGRESIÓN LINEAL PARA CLASIFICACIÓN?

Ejemplo: tenemos los datos que relacionan la cantidad de grasas consumidas y el peso de las personas → Regresión

- Si un doctor estima que mas de 95kg implica riesgo de diabetes, el problema se convierte en uno de clasificación: 0=a salvo, 1=en peligro
- Una regresión lineal podría ayudar a estimar el límite sobre el cual se estaría en peligro de diabetes
- Pero no se puede interpretar sus predicciones como probabilidades (valores no están en [0;1])
- Y no sería muy robusto…





REGRESIÓN LOGÍSTICA

- Algoritmo de **clasificación**, no de **regresión**
- Parte de la idea de la regresión lineal, cuyo resultado es modificado para poder obtener una salida **binaria**: sólo permite distinguir entre 2 clases.
  - Churn vs. Stay
  - Compra vs. No compra
  - Cliente valioso vs. Cliente no valioso
- Se agrega una transformación del resultado de la regresión lineal a partir de una función de distribución acumulativa logística, también conocida como función **logit** o **sigmoide**.

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

REGRESIÓN LOGÍSTICA

- El modelo pasa de:

$$h_\Theta(\boldsymbol{X}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

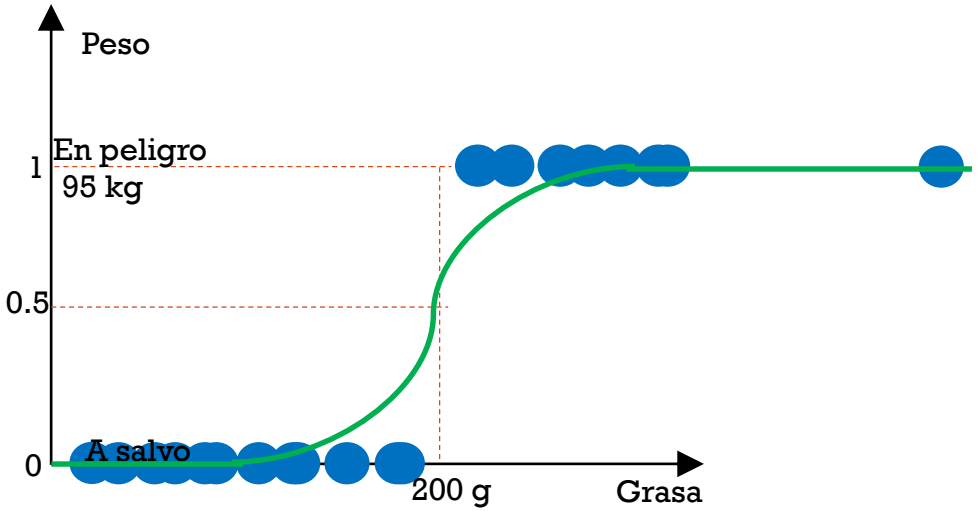
a

$$h_\Theta(\boldsymbol{X}) = \boldsymbol{f(z)} = \boldsymbol{f}(\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n),$$

con max(f(z))=1 y min(f(z))=0

- $\boldsymbol{f(z)}$ es la función **sigmoide** o **logística**
- Se pueden interpretar los valores de ***f(z)*** como **probabilidades** de que una instancia con atributos $\boldsymbol{X}$ pertenezca a la clase Y=1, $p_{1(\boldsymbol{X})}$
- $\log\left(\frac{p_{1(\boldsymbol{X})}}{1-p_{1(\boldsymbol{X})}}\right) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$

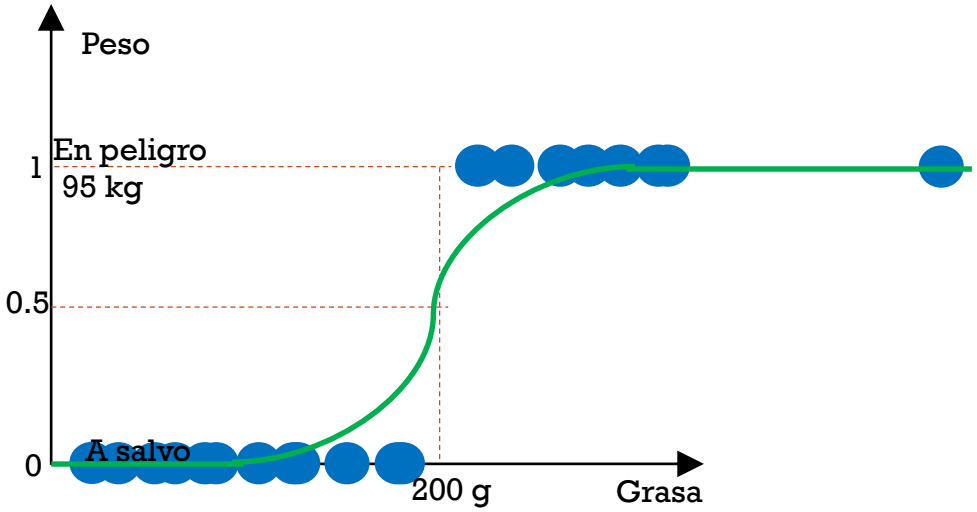
$$h_\Theta(\boldsymbol{X}) = p_{1(\boldsymbol{X})} = \frac{1}{1 + e^{-\boldsymbol{\Theta}^T \boldsymbol{X}}}$$



REGRESIÓN LOGÍSTICA

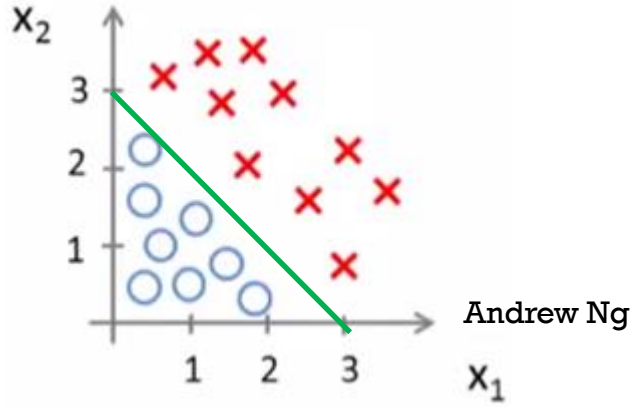
- Comportamiento:
  - Si y=1, queremos que $h_\Theta(X) \approx 1$, que $\Theta^T X \gg 0$
  - Si y=0, queremos que $h_\Theta(X) \approx 0$, que $\Theta^T X \ll 0$
- Predicción: se establece un valor de umbral, por ejemplo 0.5
  - Predecir clase 1 si $h_\Theta(X) \geq 0.5$, cuando $\Theta^T X \geq 0$
  - Predecir clase 0 de otra manera
- Se pueden establecer un umbral diferente si se quiere ser mas o menos robusto en la calificación
- Es necesario establecer si los θ_i encontrados por la regresión logística son o no significativos (valores p, lo veremos en R, en Weka no lo tenemos)

$$h_\Theta(X) = p_{1(X)} = \frac{1}{1 + e^{-\Theta^T X}}$$



REGRESIÓN LOGÍSTICA

- El algoritmo de regresión logística determina una frontera de decisión lineal



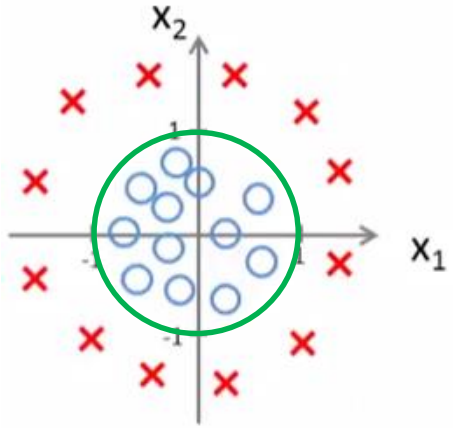
Andrew Ng

$$h_\Theta(\boldsymbol{X}) = f(-3 + x_1 + x_2)$$

Predecir la clase roja de cruz cuando:

- $h_\Theta(\boldsymbol{X}) \geq 0.5$
- $f(-3 + x_1 + x_2) \geq 0{,}5$

- Para fronteras de decisión no lineales: usar polinomios de un mayor orden



$$h_\Theta(\boldsymbol{X}) = f(-1 + {x_1}^2 + {x_2}^2)$$

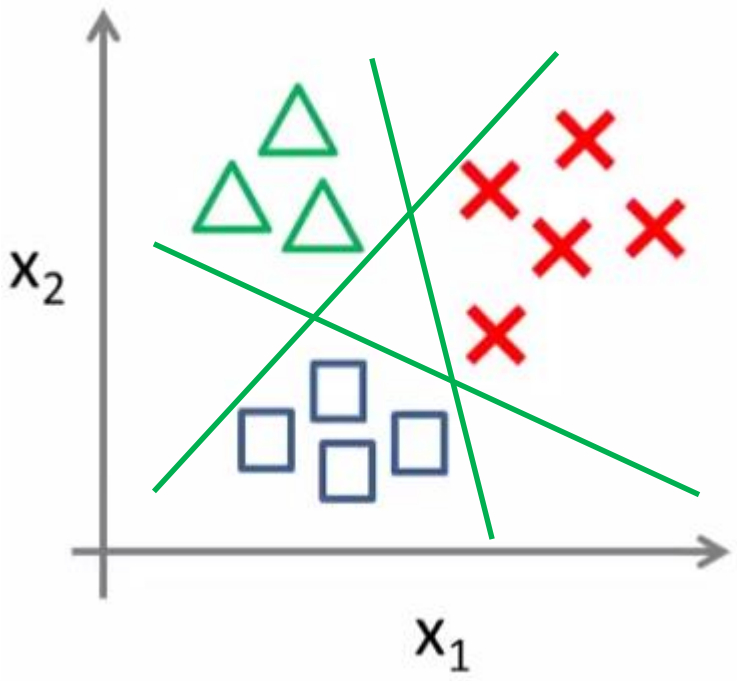
Predecir la clase roja de cruz cuando:

- $h_\Theta(\boldsymbol{X}) \geq 0.5$
- $f(-1 + {x_1}^2 + {x_2}^2) \geq 0{,}5$

REGRESIÓN LOGÍSTICA

¿Qué se puede hacer si se tienen más de 2 clases?

- Para problemas de clasificación con más de 2 clases, es necesario utilizar una aproximación de **1 vs. todos**
- Un clasificador por regresión logística es necesaria para cada clase
- Para una nueva instancia, la clase con la mayor probabilidad en su propio modelo es predicha



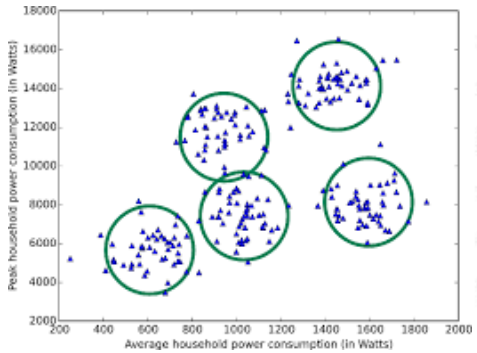
Andrew Ng

TALLER: REGRESIÓN LOGÍSTICA

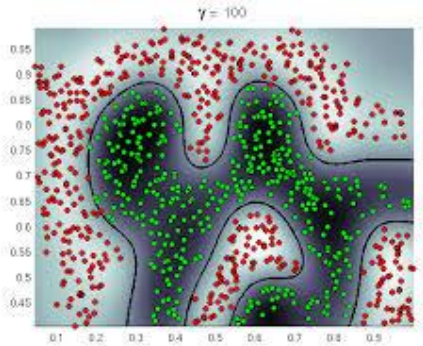
- Desarrollar el taller de regresión logística de predicción de créditos impagos

AGENDA

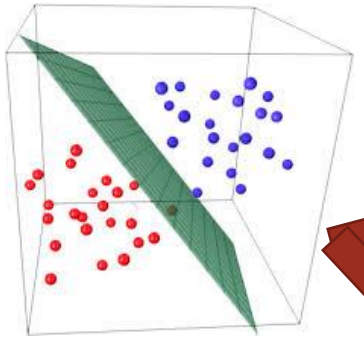
Aprendizaje automático



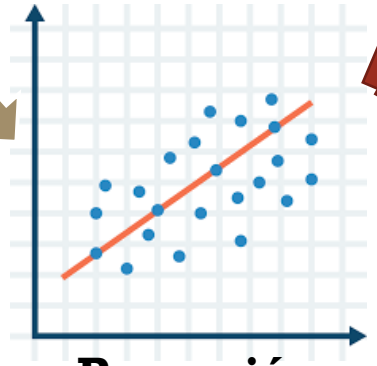
Aprendizaje no supervisado



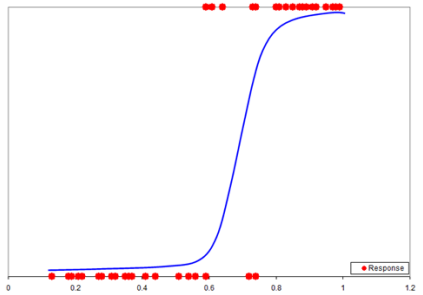
Aprendizaje supervisado



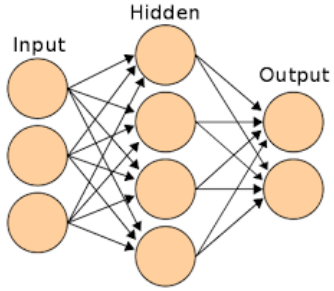
Clasificación



Regresión

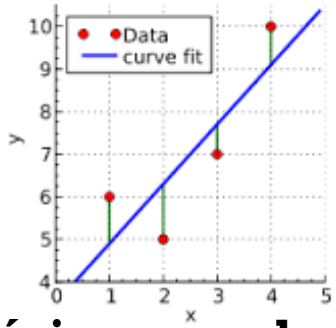


Regresión logística



Redes neuronales artificiales

Métricas de Evaluación de la regresión



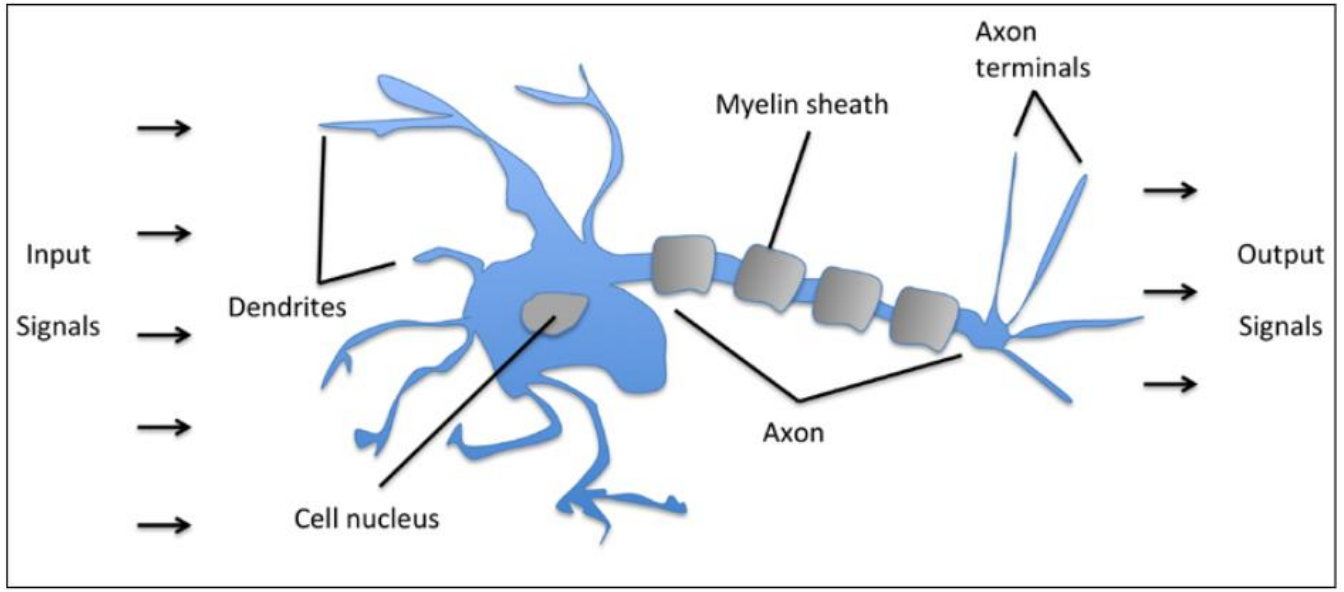
Mínimos cuadrados ordinarios

REDES NEURONALES

- Modela la relación entre un conjunto de señales de entrada y una señal de salida
- Son la base del **Deep Learning**, técnica muy actual que lleva el machine learning a solucionar problemas muy complejos, hasta hace 10 años imposibles para un computador: reconocimiento de imágenes, de audio, de escritura a mano, los carros o drones auto-pilotados, etc.
- Pueden ser utilizados para clasificación, regresión y últimamente aprendizaje no supervisado y por refuerzo
- Base del "Artificial Brain" de Google, y el "Watson" de IBM
- Son un modelo "**caja negra**", pues es imposible interpretarlo por su complejidad

REDES NEURONALES

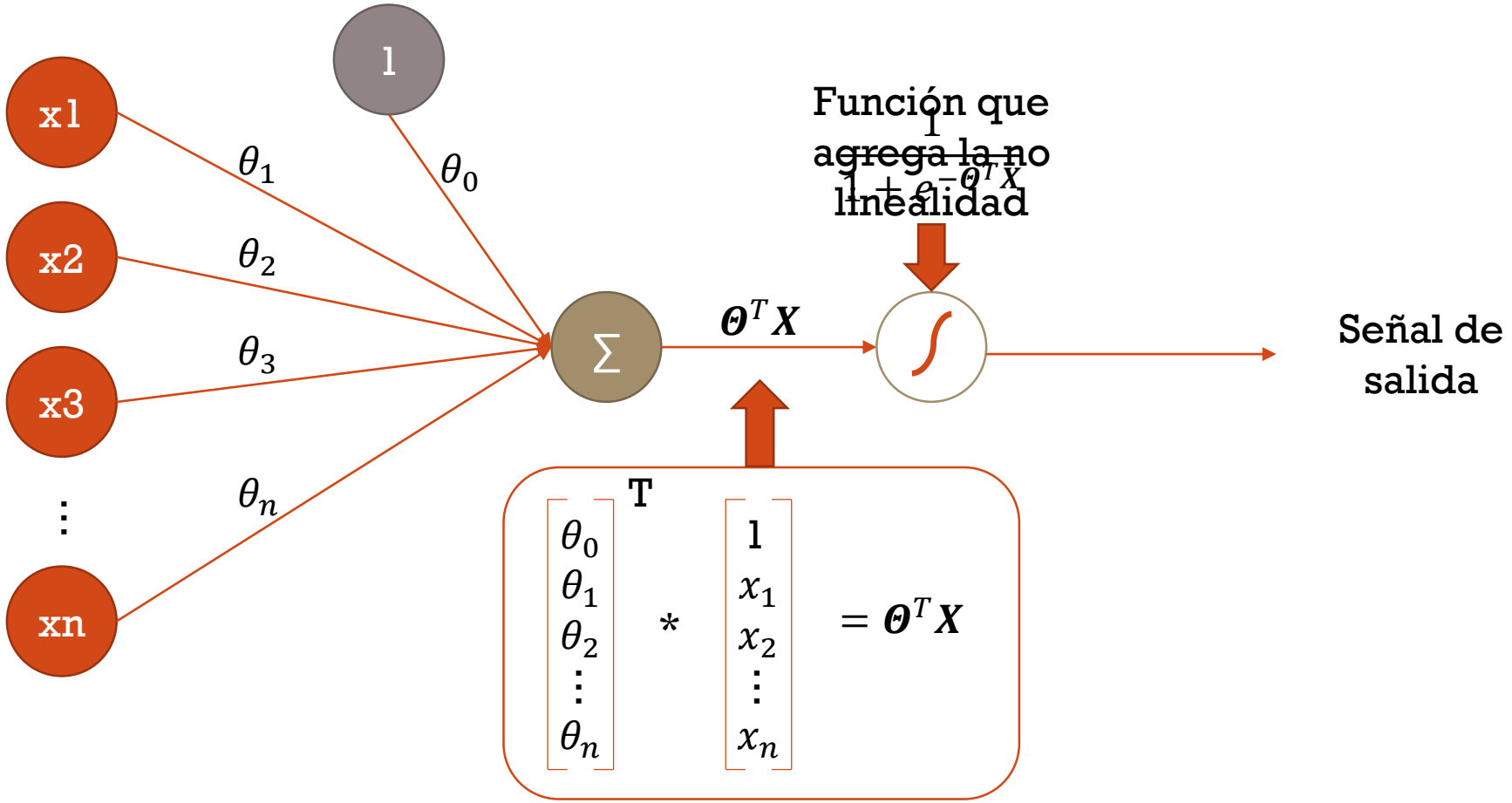
- Modelos de aprendizaje automático **bio-inspirados**: tratan de modelar cómo funciona el cerebro → 1943, transmisión de señales eléctricas y químicas
- Simplificación. Una neurona:
  - recibe **múltiples** señales de entrada,
  - que se **acumulan** en el cuerpo de la neurona
  - emiten una señal binaria que evalúa el sobrepaso de un **umbral**
  - Se conectan a otras neuronas a partir de sinapsis entre axones y dendritas



Python Machine Learning, 2015

¿cómo se parece esto a una regresión logística?

REGRESIÓN LOGÍSTICA



REDES NEURONALES

Una red neuronal se distingue por:

- La **topología de arquitectura de red**: describe el número de neuronas en cada una de las capas y la manera como se conectan entre ellas
- La **función de activación**: transforma la combinación de los inputs en una sola señal de salida a ser comunicada a las siguientes neuronas
- El **algoritmo de entrenamiento**: especifica como los pesos de las conexiones se establecen de tal manera que se cohíba o incite la activación de las neuronas en proporción de las señales de entrada.

REDES NEURONALES

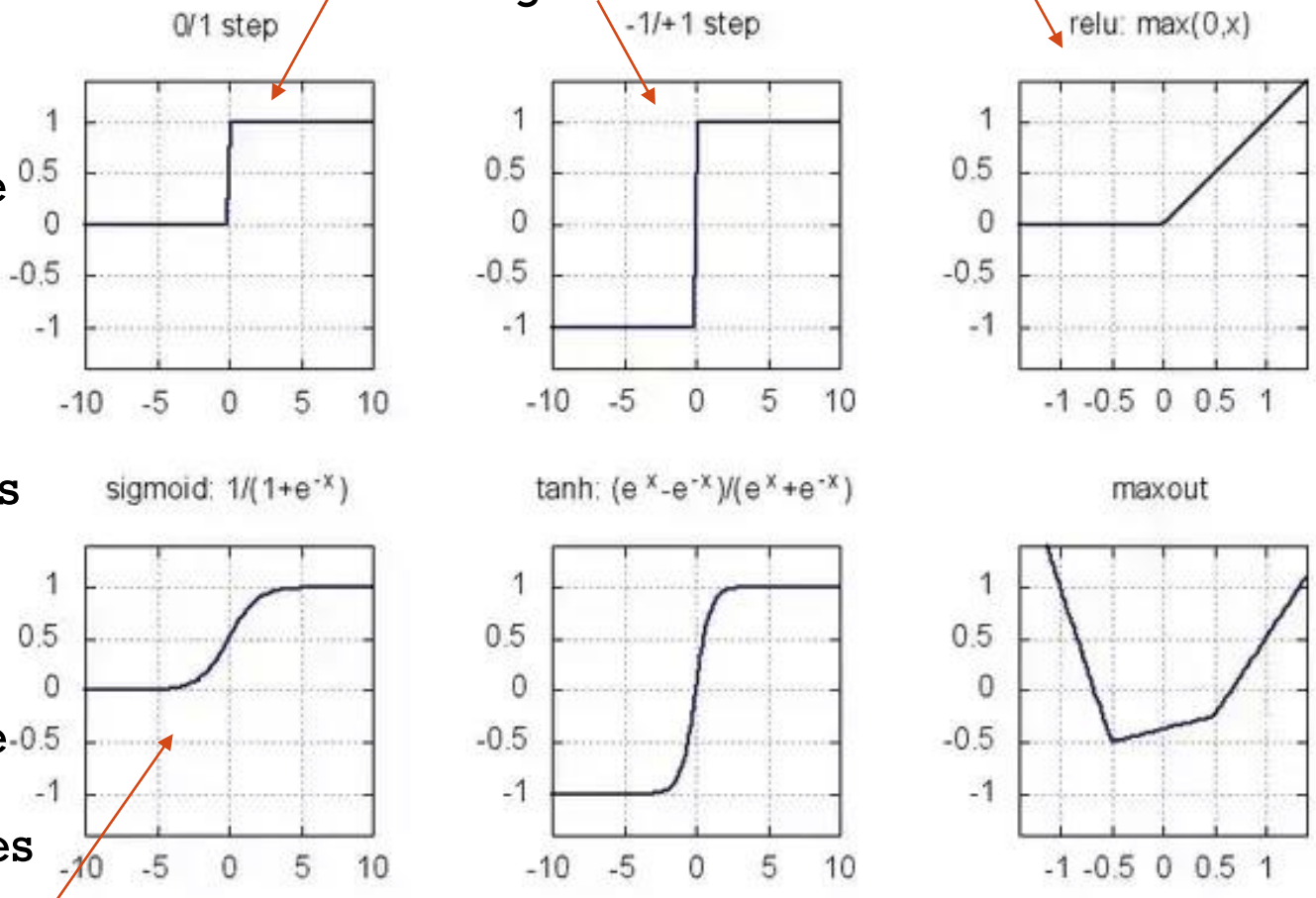
Funciones de activación: mecanismo de procesamiento de la información entrante que permite la propagación de la señal en la red

Se prefieren las que tengan buenas propiedades matemáticas (simples, derivables)

Para evitar problemas de aprendizaje, se acostumbra normalizar las entradas, para que los valores se encuentren en los rangos dinámicos de las funciones de activación.

Las más parecidas a la realidad biológica

Muy usada en Deep Learning

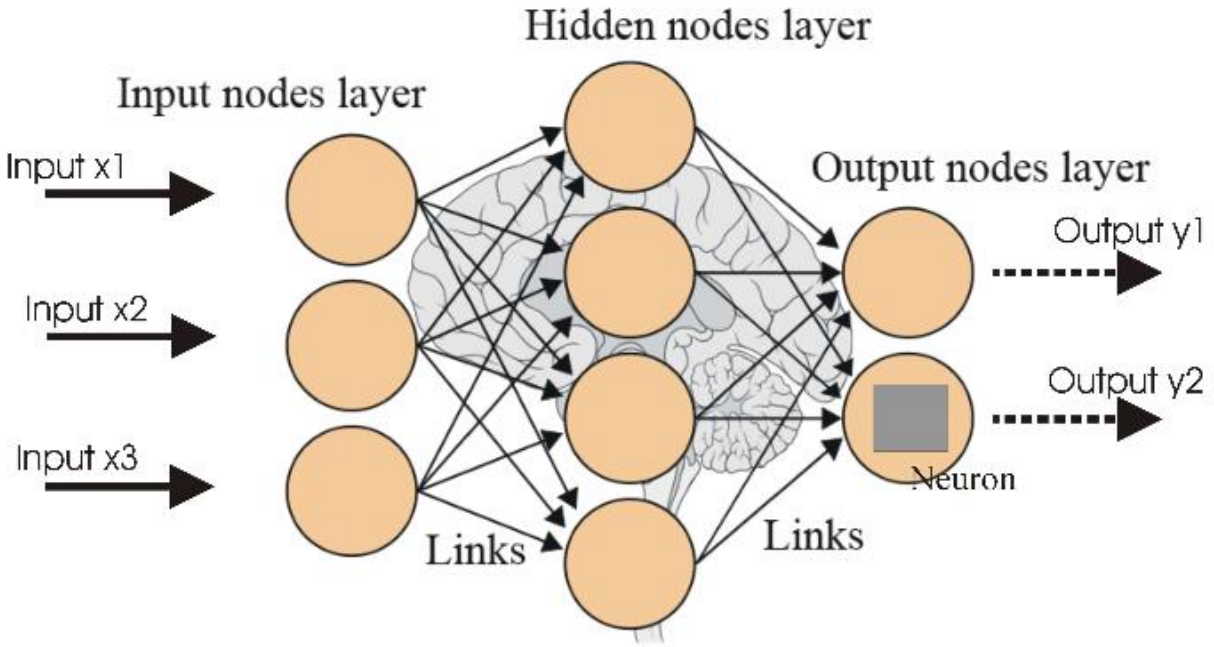


Regresión logística

REDES NEURONALES

Topología: determina la complejidad de las tareas que se pueden aprender

- Número de capas y como se conectan entre ellas. Deep Learning se refiere a la profundidad de las capas.
- **Número de neuronas en cada capa**: salvo por la capa de entrada y salida, depende de la complejidad del problema y calidad de los datos. Cuidado con **overfitting.**
- Dirección del envío de la información
  - Feedforward: hacia adelante
  - Recurrent: se permite retorno (short term memory)



www.analyticsvidhya.com/wp-content/uploads/2016/08/Artificial-Intelligence-Neural-Network-Nodes.jpg

REDES NEURONALES

Back-propagation: el algoritmo más común para entrenar una red neuronal feed-forward de múltiples capas (MLP – Multi Layer Perceptron, mediados de los 80's).

- 2 fases
  - **Forward**: para cada ejemplo propagar la información hasta llegar al final, calcular el error de predicción.
  - **Backward**: modificar los pesos de la capa inmediatamente anterior de tal manera que se reduzcan los errores. Continuar el proceso con las capas anteriores.
- Basado en el **descenso de gradiente** (derivadas parciales de las funciones de activación que van en la dirección de la reducción del error)
- Computacionalmente intensivo
- Influencia de la inicialización aleatoria de las neuronas
- Tasa de aprendizaje

TALLER: REDES NEURONALES

- Desarrollar el taller de redes neuronales para la detección del cancer

REFERENCIAS

- *Introduction to Statistical Learning with Applications in R (ISLR),* G. James, D. Witten, T. Hastie & R. Tibshirani, 2014
- *Machine Learning,* Tom M. Mitchell, McGraw-Hill, 1997
- *Python Machine Learning*, Sebastian Raschka, Packt, 2017