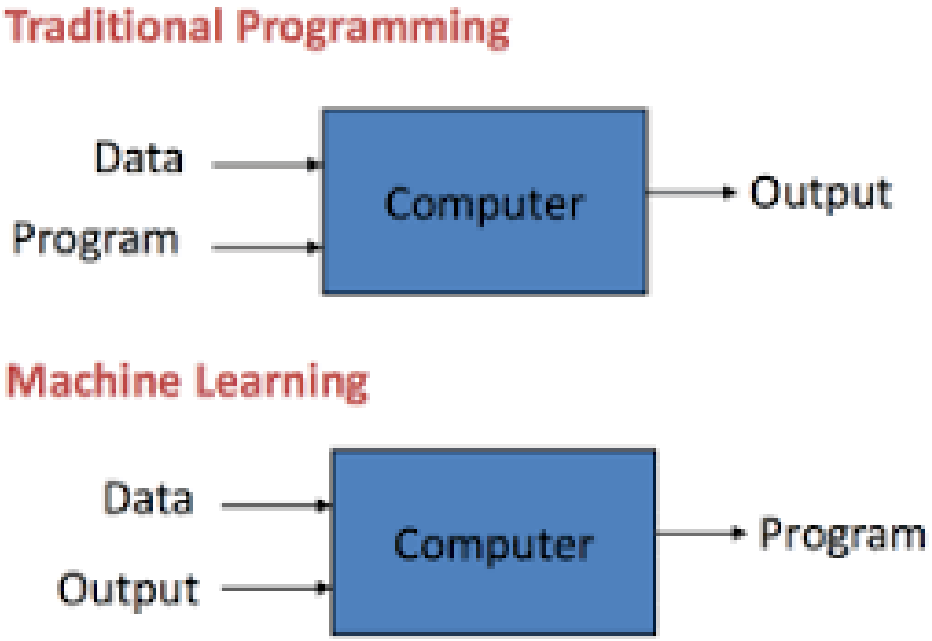
Aprendizaje de Máquina
Introducción

Aprendizaje de máquina (Machine Learning)

- Es una rama de la **inteligencia artificial** cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender algoritmos capaces de generalizar comportamientos a partir de información no estructurada suministrada en forma de ejemplos.

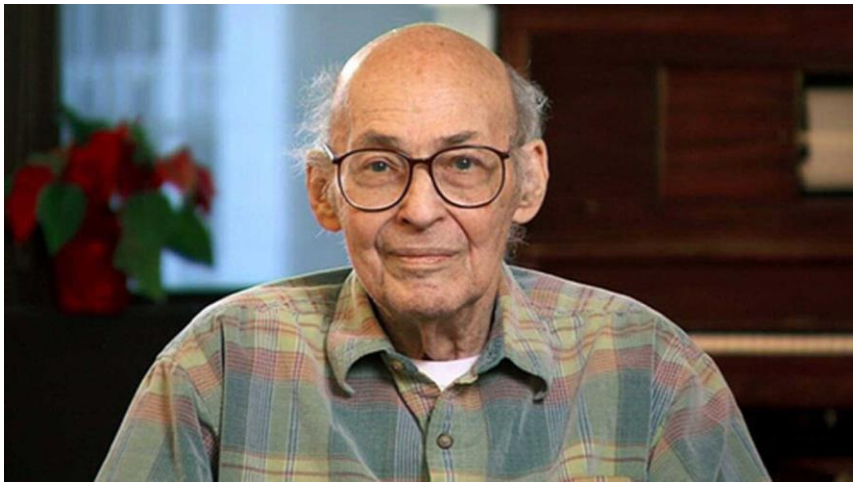
- El aprendizaje se cataloga como **supervisado**, **por refuerzo** o **no supervisado** dependiendo de si el algoritmo debe contar o no con información específica de datos satisfactorios para el objetivo del aprendizaje

Cambio de paradigma



Paradoja de Moravec

- Marvin Minsky escribió “En general, no somos conscientes de nuestras mejores habilidades”, añadiendo que “somos más conscientes de los pequeños procesos que nos cuestan que de los complejos que se realizan de forma fluida”



Paradoja de Moravec

Relativamente fácil de programar

- Tareas conscientes
  - Razonamiento de alto nivel
  - Jugar ajedrez
  - Resolver una integral
  - Demostrar un teorema
- Habilidades aprendidas recientemente en una escala evolutiva

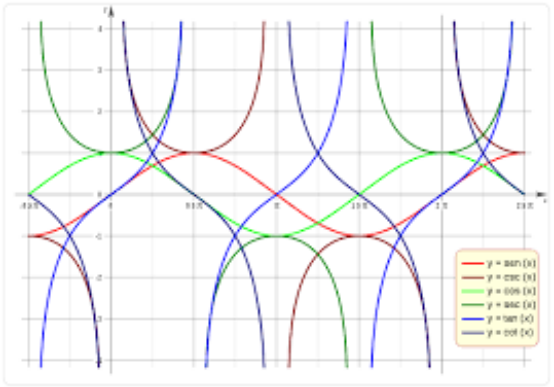
Difícil o extremadamente difícil de programar

- Tareas Inconscientes
  - Caminar o correr
  - Ver y reconocer
  - Oler
  - Escuchar y entender
- Habilidades perfeccionadas a través de millones de años de evolución

Algoritmo

- “Es un procedimiento correctamente definido que toma algún valor, o conjunto de valores, como entrada y produce algún valor, o conjunto de valores, como salida” Cormen et al.

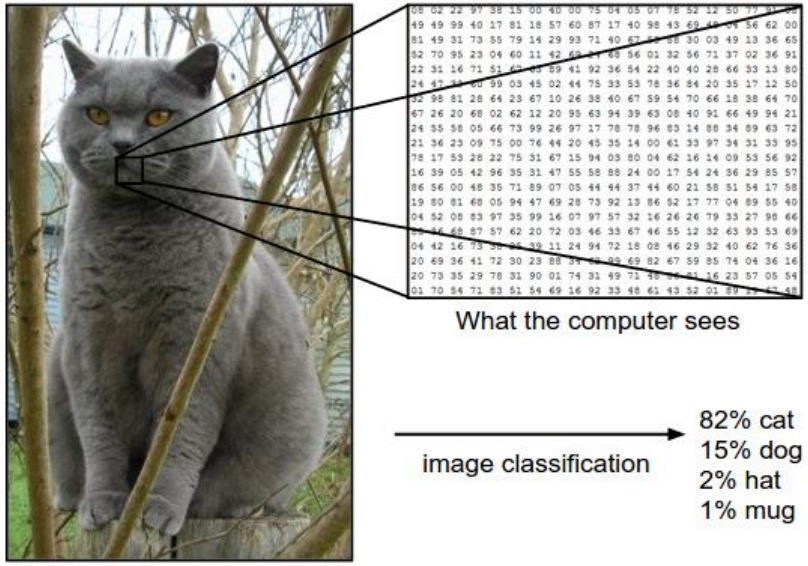
f(x)



Algoritmo para jugar gato (fácil de programar)

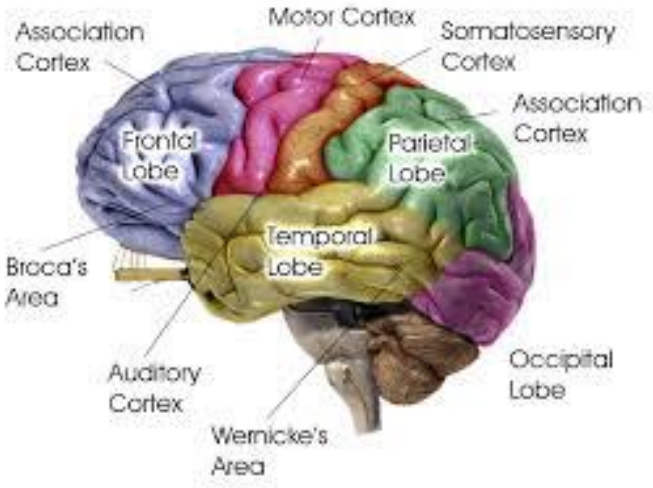
- Si tu oponente tiene dos fichas en una fila, tira en la casilla vacía de esa fila
- De otra forma, si hay una tirada que genere dos filas con dos fichas tuyas en ellas, tira ahí
- De otra forma, si el centro está libre, tira ahí
- De otra forma, si tu contrincante tiró en una esquina, tú tira en la esquina opuesta
- De otra forma, si hay una esquina libre, tira ahí
- De otra forma, tira en cualquier casilla vacía

Clasificación de imágenes (difícil de programar)



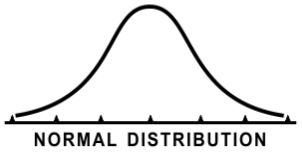
¿De dónde nos podemos inspirar?

- Unos de los proyectos de ingeniería inversa más retadores es entender cómo funciona el cerebro



Aprendizaje de máquina

- Es la ciencia que estudia como aprender a partir de datos.
- ¿Qué eso no es lo que hace la estadística?



Aprendizaje de máquina

- Empata en muchos aspectos con la estadística, pero el enfoque es distinto:
  - La estadística busca modelos simples que expliquen el porqué de los fenómenos
  - El aprendizaje de máquina busca que las predicciones sean lo más certeras posible
  - El aprendizaje de máquina se enfoca más en el aspecto computacional dada la complejidad de los algoritmos

Estadística Tradicional vs. Aprendizaje de Máquina

Estadística Tradicional

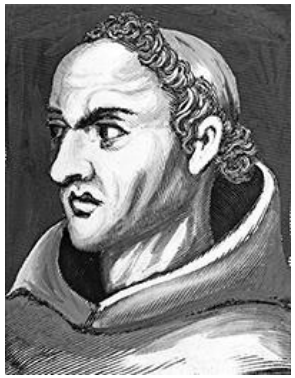
- Hacen hincapié en la **inferencia** de superpoblación
- Modelos más simples se prefieren a los complejos (parsimonia), aunque los modelos más complejos lo representen mejor
- Énfasis en la capacidad de **interpretar los parámetros**
- Modelado estadístico y los supuestos de muestreo conectan los datos a una población de interés
- Preocupación por los supuestos

Aprendizaje de Máquina

- Hace hincapié en las **predicciones**
- Preocupación por el sobre entrenamiento, pero no por la complejidad del modelo per sé
- Énfasis en el **rendimiento**
- La generalización se obtiene a través de la aplicación sobre nuevos conjuntos de datos. Por lo general no hay un modelo de superpoblación específica
- Preocupación por el rendimiento

Principio de parsimonia

- Conocido como navaja de Ockham, principio de parsimonia o de economía
  - “dadas las mismas condiciones, la explicación que suele resultar correcta es la más sencilla“.



Estadística tradicional vs. Aprendizaje de máquina

Aprendizaje de máquina

Más interpretables

Estadística tradicional

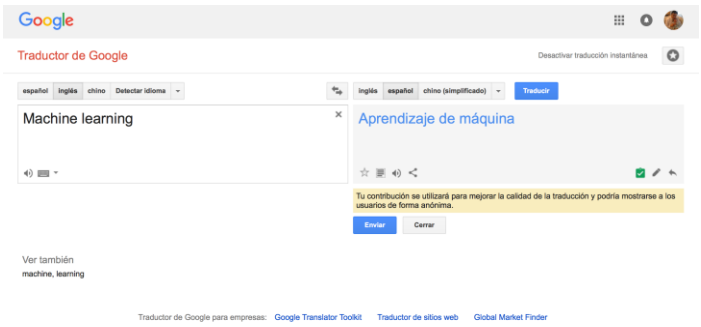
Mejores predicciones

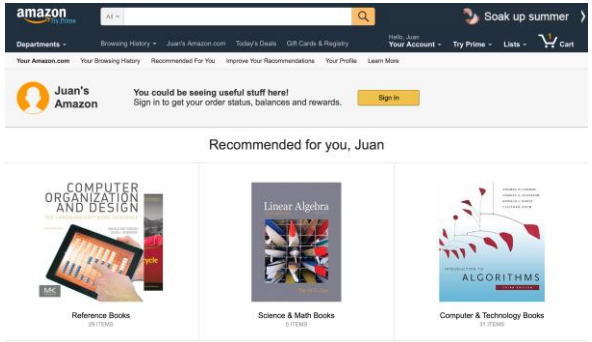
RISE OF THE MACHINES
Larry Wasserman

Ejemplos de programas que utilizan aprendizaje de máquina

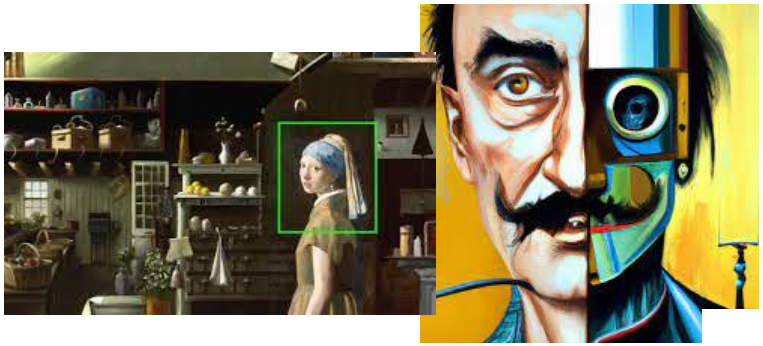
- Predecir:
  - si un paciente hospitalizado debido a un ataque al corazón, tendrá un segundo ataque al corazón utilizando información demográfica, la dieta y mediciones clínicas para ese paciente.
  - el precio de una acción en 6 meses con base en las medidas de rendimiento de la empresa y datos económicos.
- Identificar:
  - los números de un código postal escrito a mano, a partir de una imagen digitalizada.
  - los factores de riesgo para el cáncer de próstata, basado en variables clínicas y demográficas.

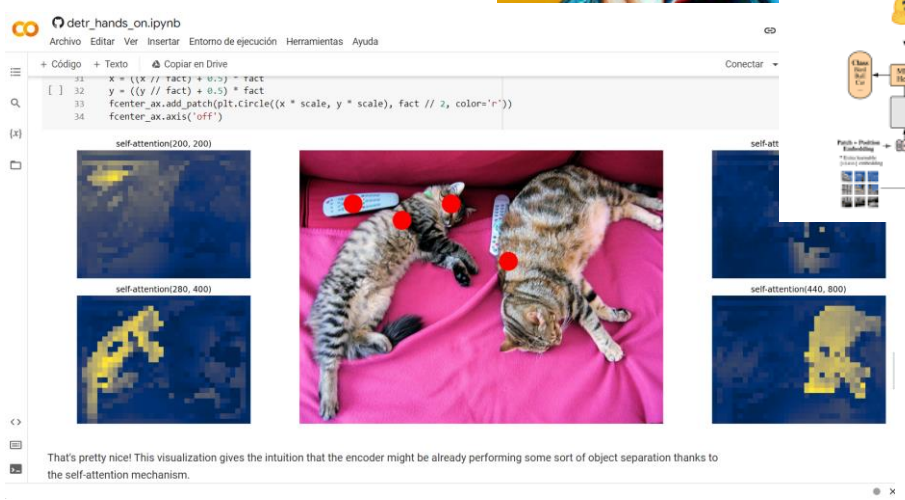
Ejemplos de programas que utilizan aprendizaje de máquina

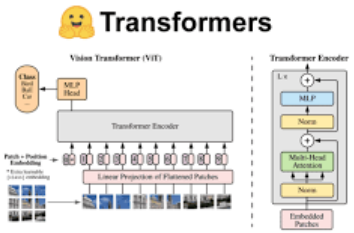


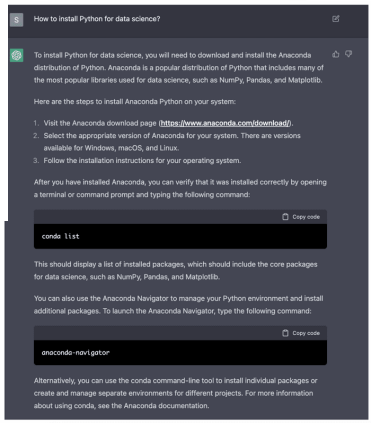


El presente y futuro cercano









UN POCO DE HISTORIA....

Neurona lógica
McCulloch-Pitts

1947

Inteligencia Artificial
A. Turing

1950

Perceptrón
Rosenblatt

1957

Perceptrón
Minsky

1969

Aprendizaje de Refuerzo
Sutton

1984

Back-propagation
Rumelhart

1986

Bosques aleatorios
Breiman

MVS
Vapnik

1995

Deep Learning
Hinton

2006

IA conexionista sub-simbólica

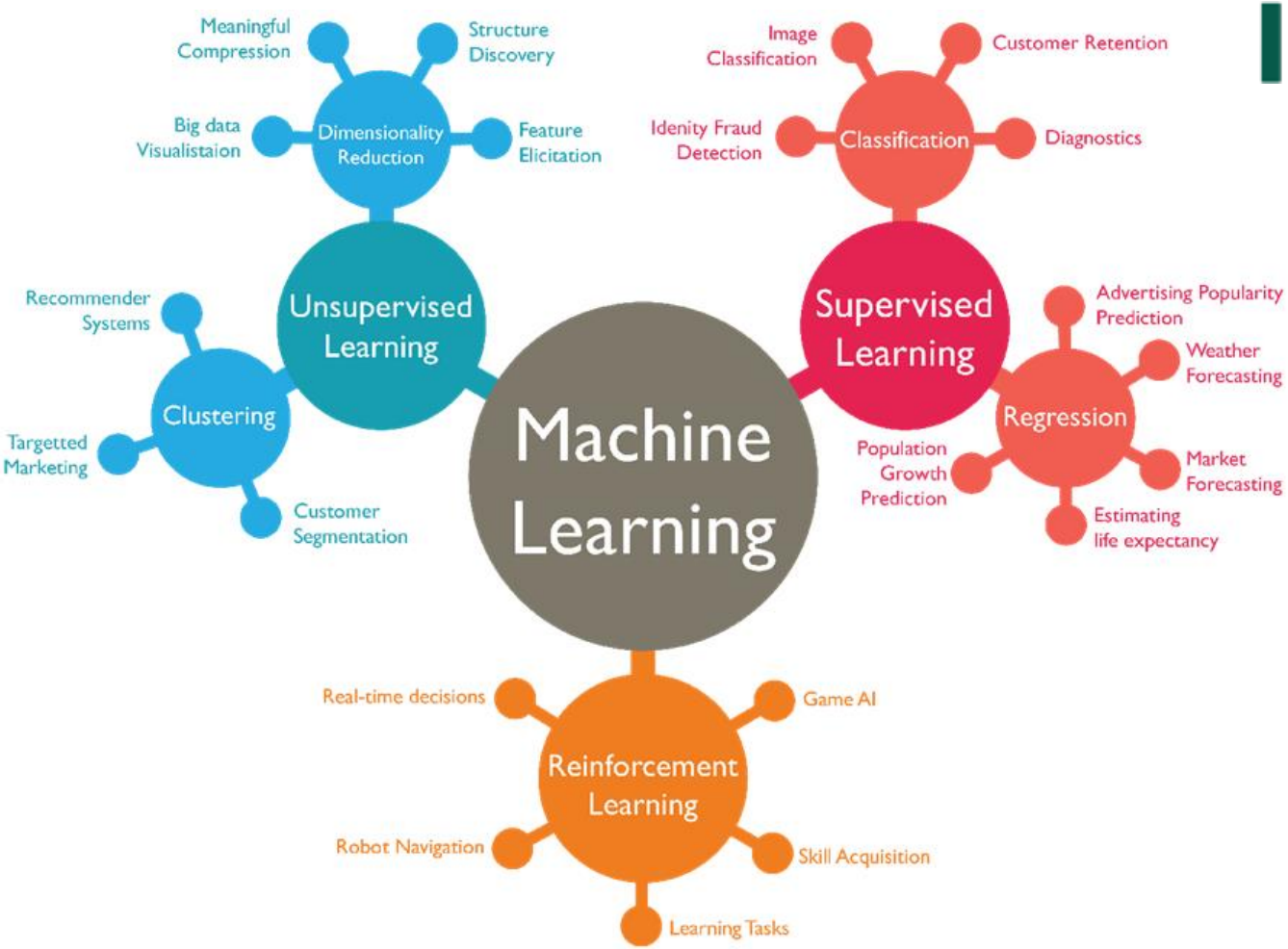
IA simbólica

Alpha GO
Deep Mind

2016

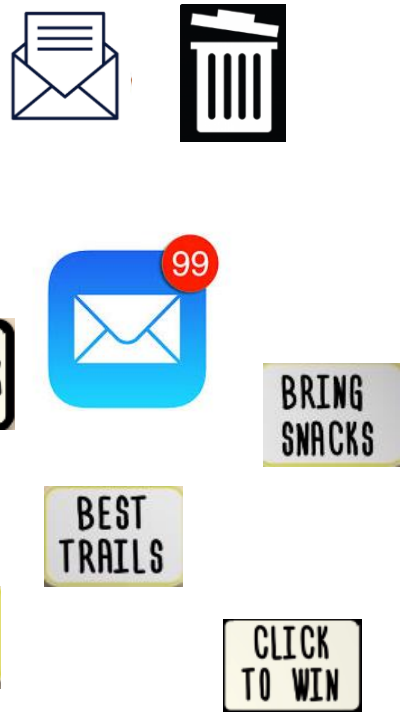
Transformers
Vaswani

2017


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

¿Qué significa “aprender” para una máquina?
Extraer patrones ocultos de un conjunto de datos

Supervisado:
aprender a partir de datos **etiquetados** a lo largo del tiempo
ejemplo: spam



No Supervisado:
aprender de datos que **no tienen etiquetas**
ejemplo: clusters

Aprendizaje supervisado

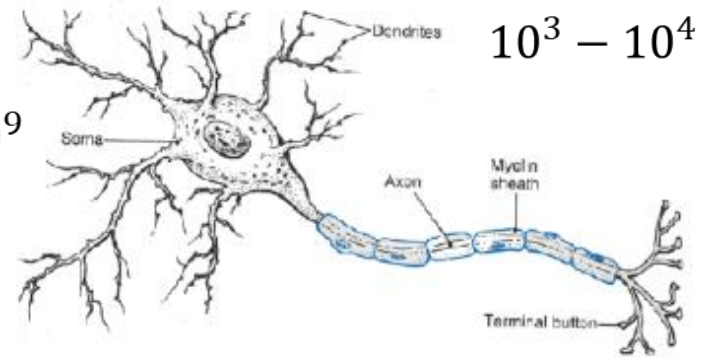
- Infiere una función a partir de datos de entrenamiento etiquetados.
- Cada dato de entrenamiento es un par que consta de un objeto de entrada (típicamente un vector) y un valor de salida deseado (también llamada la señal de supervisión).
- Dos tipos de salida
  - Numérico continuo: Regresión
  - Valores discretos (clases): Clasificación

Aprendizaje supervisado

- Se utilizan ampliamente en analítica predictiva
- Algoritmos
  - Regresiones lineales
  - Regresiones logísticas
  - Redes neuronales
  - Máquinas de vectores de soporte
  - Árboles de decisión

Aprendizaje supervisado: Redes neuronales

$86\ x\ 10^9$

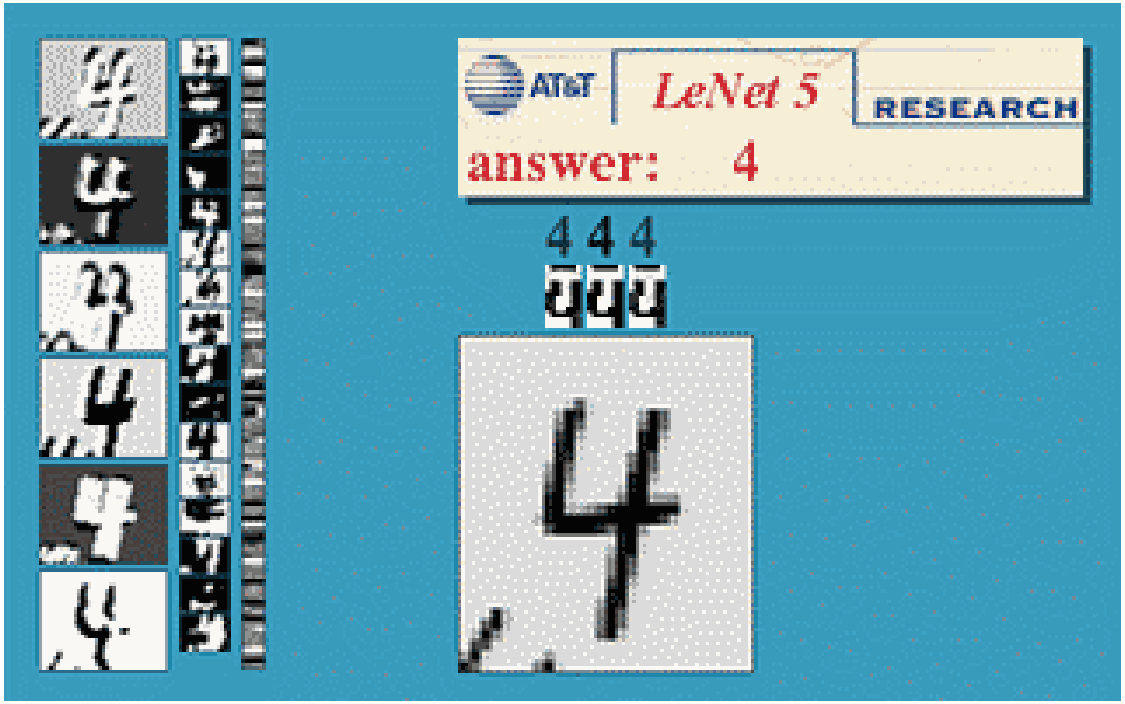


$10^3 - 10^4$

10^2

$6\ capas$

Ejemplo



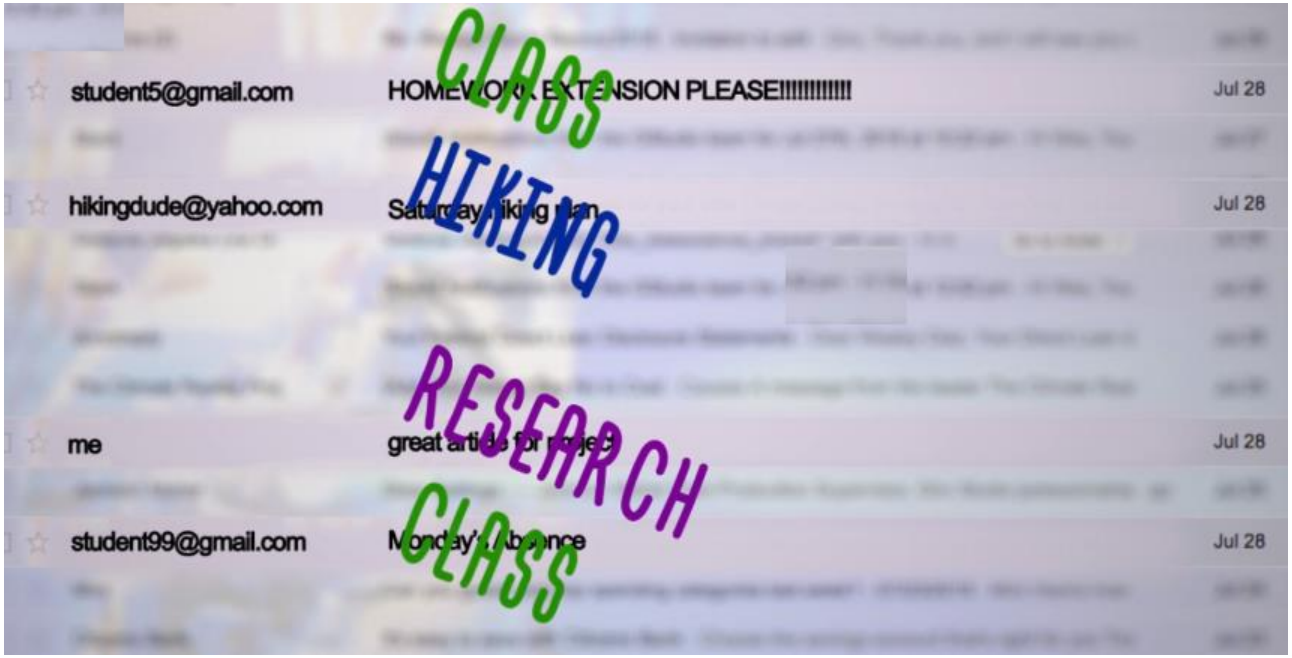
Aprendizaje no supervisado

- Infiere una función que describe la estructura de datos no etiquetados
- No hay señal de error ni de recompensa para evaluar una solución potencial
- Busca resumir y explicar la principales características de los datos

Aprendizaje no supervisado

- Encargado de detectar patrones o asociaciones, no fácilmente observables dentro de los datos.
- Se utiliza principalmente en minería de datos.
- Algoritmos
  - Agrupamiento (clustering)
  - Componentes principales
  - Modelo de mezclas (mixture models)

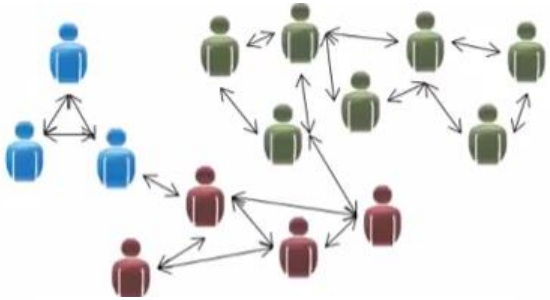
Aprendizaje no supervisado



Ejemplo: clasificación de correos electrónicos

Aprendizaje no supervisado

Segmentación de mercado

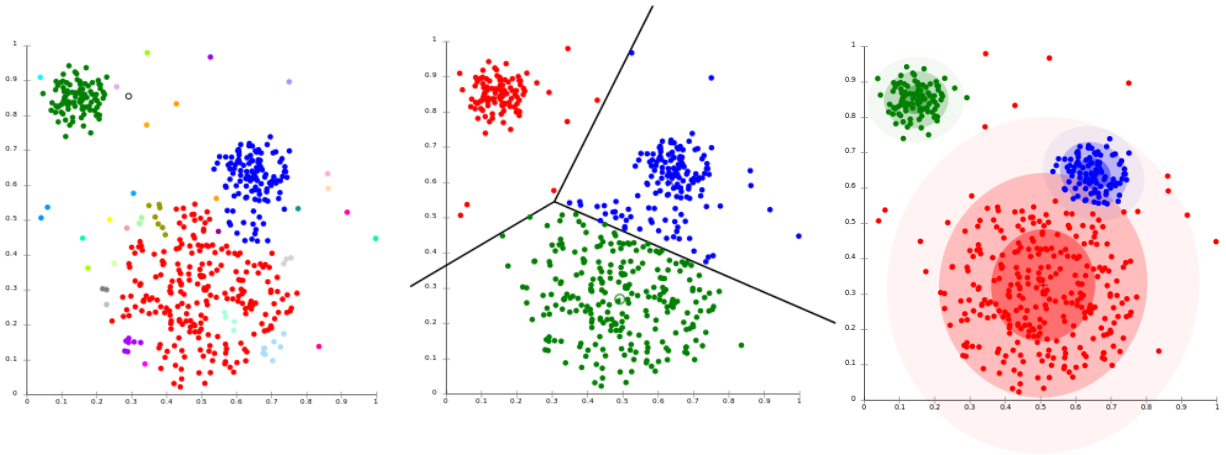


Análisis de redes sociales

Organizar centros de datos

Análisis de datos astronómicos

Aprendizaje no supervisado



Jerárquico K-medias Mezcla de modelos

Aprendizaje por refuerzo

- Es el aprendizaje de qué hacer - cómo mapear situaciones a acciones - con el fin de maximizar una señal de recompensa numérica.
- Al aprendiz no se le dice qué acciones tomar, como en la mayoría de las formas de aprendizaje de la máquina, sino que debe descubrir qué acciones producen la mayor recompensa probándolas.

Aprendizaje por refuerzo

- ¿Qué hace que el aprendizaje por refuerzo sea diferente de otros paradigmas de aprendizaje de máquina?
  - No hay ningún supervisor, solamente una señal de recompensa
  - La retroalimentación no es instantánea
  - El tiempo realmente importa (secuencial, observaciones no son independientes e idénticamente distribuidas)
    - Las acciones del agente afectan a los datos posteriores que recibe

Deep reinforcement learning

Ejemplo de deep learning

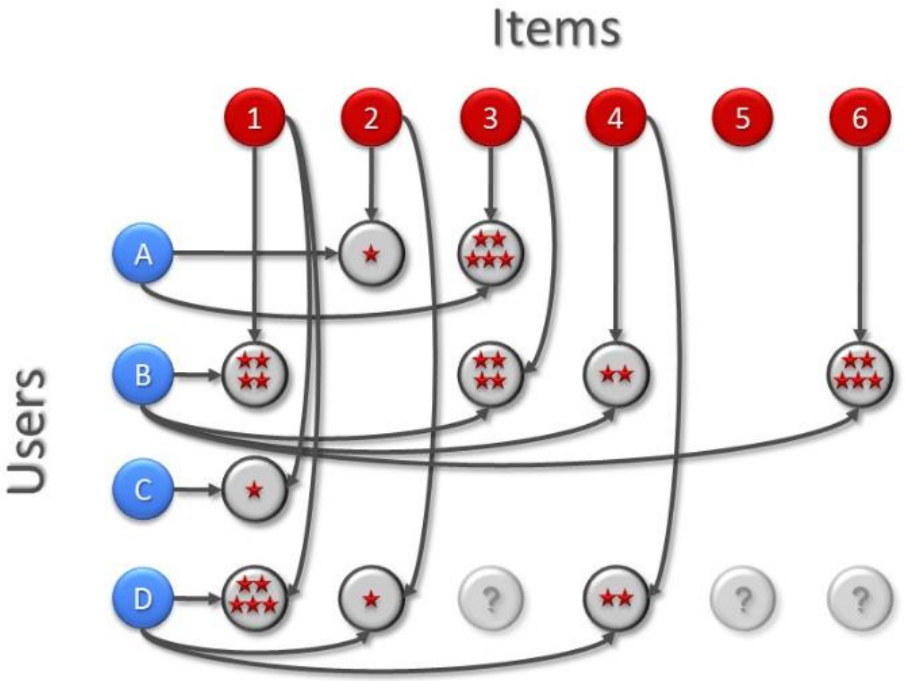
https://www.youtube.com/watch?v=V1eYniJ0Rnk

Sistemas de recomendación

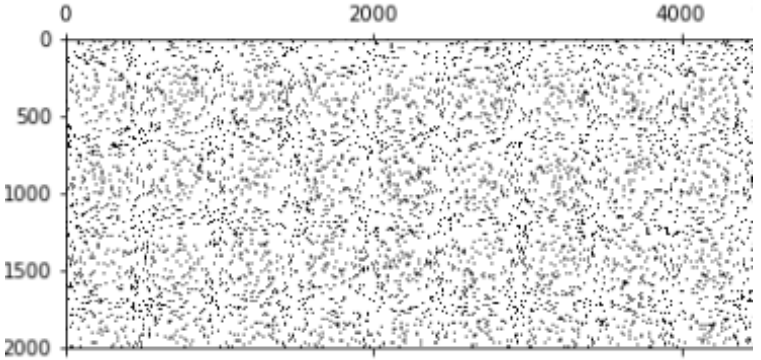
- Son útiles para entender mejor la demanda y planificar adecuadamente la oferta con el fin de optimizar el proceso de producción.
- Muchos sectores económicos (Transporte, retail, entretenimiento, finanzas) dependen de las preferencias de las personas involucradas en las transacciones.

La **teoría de la elección social** es un marco teórico para el análisis de la combinación de opiniones individuales, preferencias, intereses o bienestar para llegar a una decisión colectiva o bienestar social.

¿Cómo predecir ratings?



- Es necesario construir una matriz con los usuarios como renglones, los productos como columnas y las calificaciones en las entradas.
- Hay un gran porcentaje de datos faltantes porque típicamente los usuarios sólo califican un pequeño subconjunto de productos.



¿Supervisado o No Supervisado?

Datos:
Mediciones de ***concentración de contaminantes*** en diferentes localidades, bajo diferentes condiciones climáticas y días de la semana.

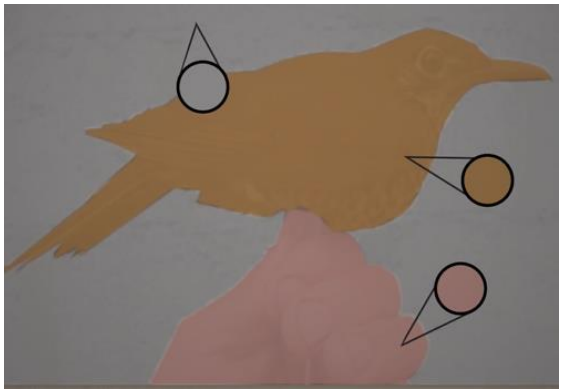
Pregunta:
¿Cuál será la concentración de contaminantes en una nueva localidad dado el clima y el día de la semana?

MIT 2016

¿Supervisado o No Supervisado?

Comprimir de videos para envío por email

Datos:
Pixeles por segundo agrupados por semejanzas en colores en determinado tiempo



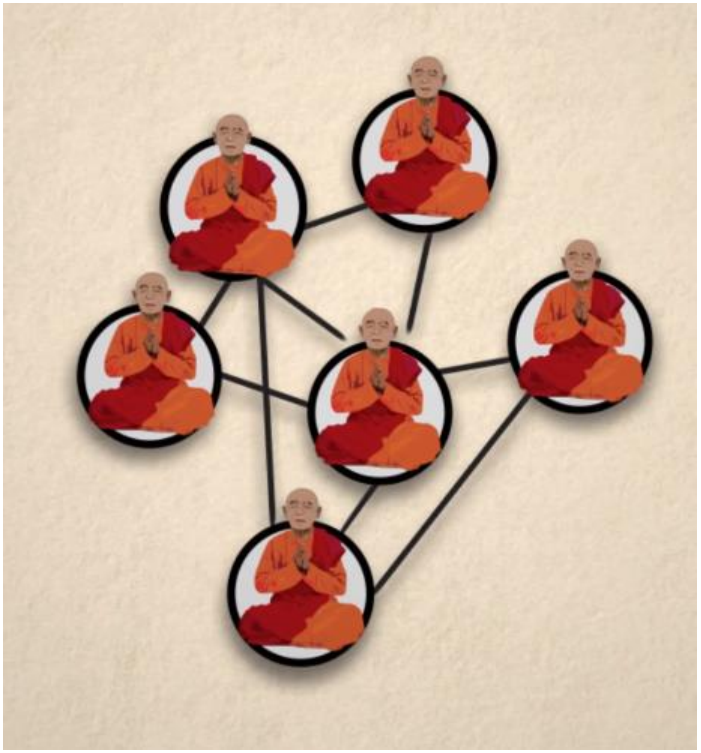
MIT 2016

¿Supervisado o No Supervisado?

Frank Sampson,
estudiante de PhD estuvo un mes en un monasterio durante los años 60.

Datos:
Las interacciones entre los monjes

Pregunta:
¿Cuáles serán las relaciones de amistad o grupos sociales en el monasterio?



MIT 2016

Aprendizaje Supervisado

Infiere una función a partir de **datos** de entrenamiento **etiquetados**.

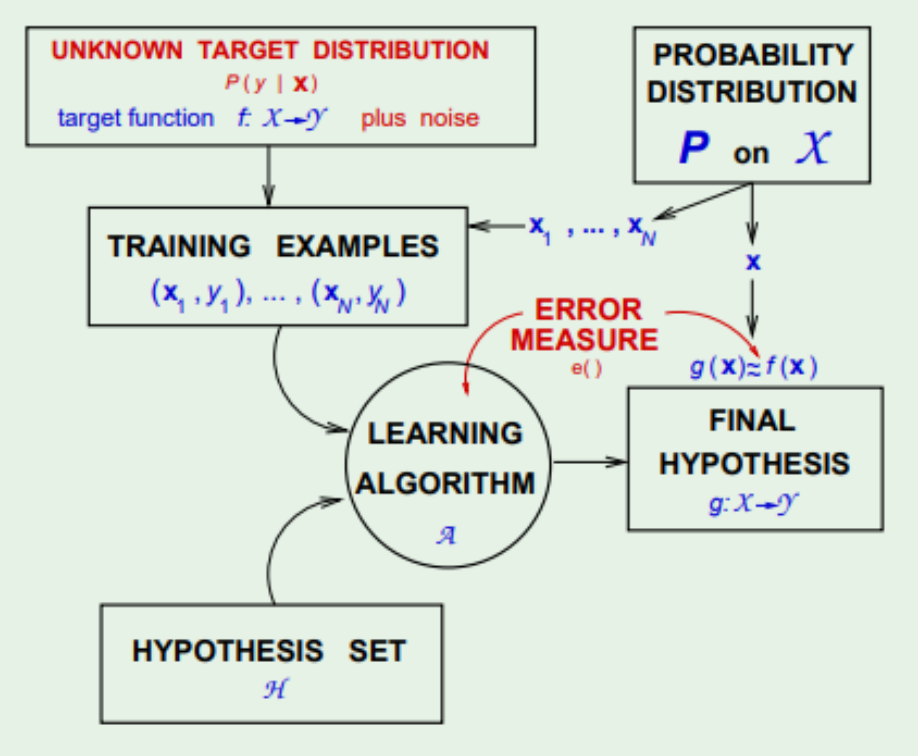
Dos tipos de salida **(variable dependiente)**

- Numérico **continuo: Regresión**
- Valores **discretos (clases): Clasificación**

Algoritmos

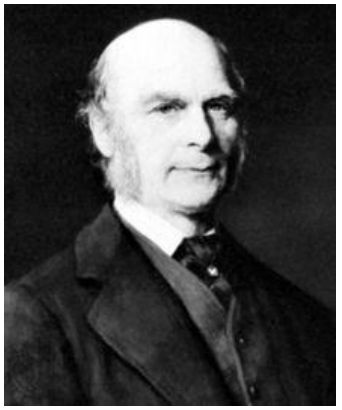
- Regresiones lineales
- Regresiones logísticas
- Redes neuronales
- Máquinas de soporte vectorial
- Árboles de decisión

El diagrama de aprendizaje



Regresión hacia la mediocridad

- Galton acuñó el término regresión para describir un hecho observable en la herencia de rasgos genéticos cuantitativos multifactoriales
  - el descendiente de los padres que se encuentran en las colas de la distribución tiende a estar más cerca del centro, la media de la distribución.
  - Él cuantificó esta tendencia, y al hacerlo inventó el análisis de regresión lineal (el método de mínimos cuadrados se acredita a Gauss, pero lo publicó primero Legendre)

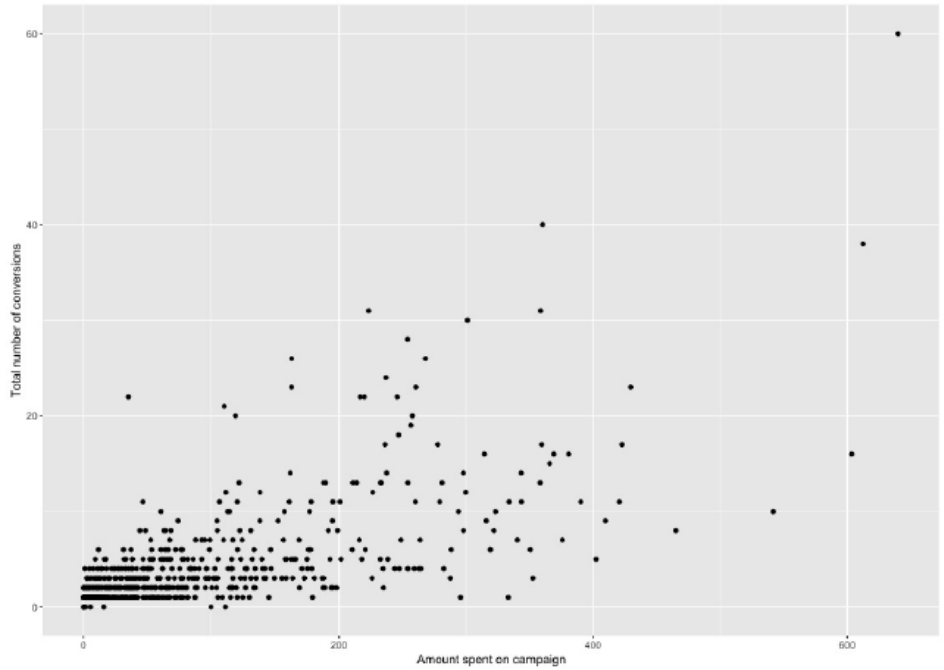


Objetivos

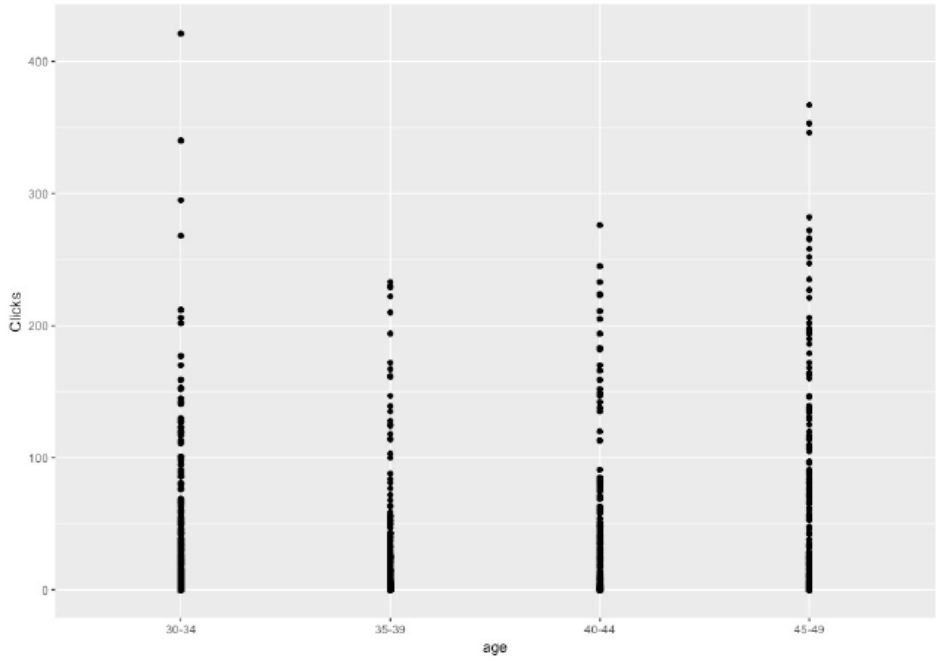
- Considere dos variables aleatorias X e Y. Por ejemplo,
  1. X es la cantidad de $ gastado en anuncios de Facebook e Y es la tasa de conversión total
  2. X es la edad de la persona e Y es el número de clics

- Dadas dos variables aleatorias (X, Y), podemos preguntar lo siguiente:
  - ¿Cómo predecir Y a partir de X?
  - ¿Cuántas conversiones Y más por un dólar adicional?
  - ¿El número de clics depende de la edad?
  - ¿Qué pasa si X es un vector aleatorio? Por ejemplo, $X = (X_1, X_2)$ donde X_1 es la cantidad de $ gastados en anuncios de Facebook y X_2 es la duración en días de la campaña.

Conversiones vs. monto gastado



Clics vs. edad



Distribución conjunta

- Podríamos responder estas preguntas si contáramos con la distribución conjunta de X y Y
- (X_i, Y_i), i = 1,. . . , n son i.i.d de alguna distribución conjunta desconocida **P**

- **P** puede describirse mediante:
  1. Un PDF conjunto $h(x, y)$
  2. La densidad marginal de X

     $$h(x) = \int h(x, y)dy$$

     y la densidad condicional

     $$h(y|x) = \frac{h(x,y)}{h(x)}$$

- $h(y|x)$ responde a todas nuestras preguntas.
  - Contiene toda la información sobre Y dado X

Usos de la distribución conjunta

- Predecir
- Determinar causalidad
- Entender de mejor manera el mundo

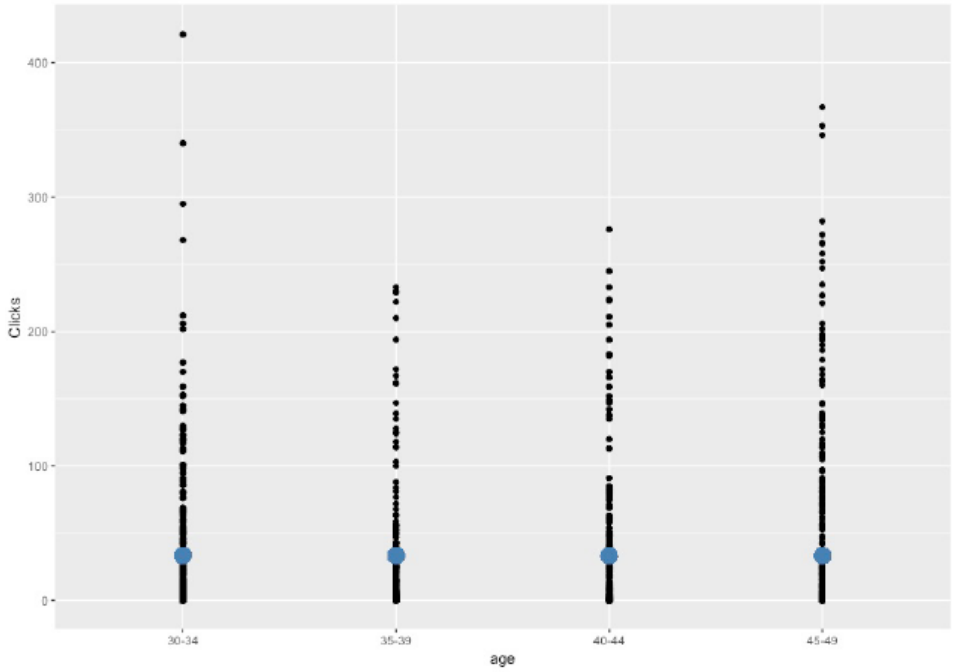
- Nosotros nos enfocaremos en predecir

Modelado parcial

- También podemos describir la distribución solo parcialmente, por ejemplo, usando
- La esperanza de Y: $\mathrm{E}[Y]$
- La esperanza condicional de Y dado $X = x$: $\mathrm{E}[Y|X = x]$
  - La función

    $x \to \mathrm{f}(x) := \mathrm{E}[Y|X = x] = \int y\, h(y|x)\, dy$
  - se le conoce como **función de regresión**
- Otras posibilidades:
  - La mediana condicional: $\mathrm{m}(x)$ tal que $\int_{-\infty}^{m(x)} h(y|x)\, dy = \frac{1}{2}$
  - Cuantiles condicionales
  - Varianza condicional (no informativa sobre la ubicación)

Esperanza condicional



Densidad condicional y cuantiles condicionales

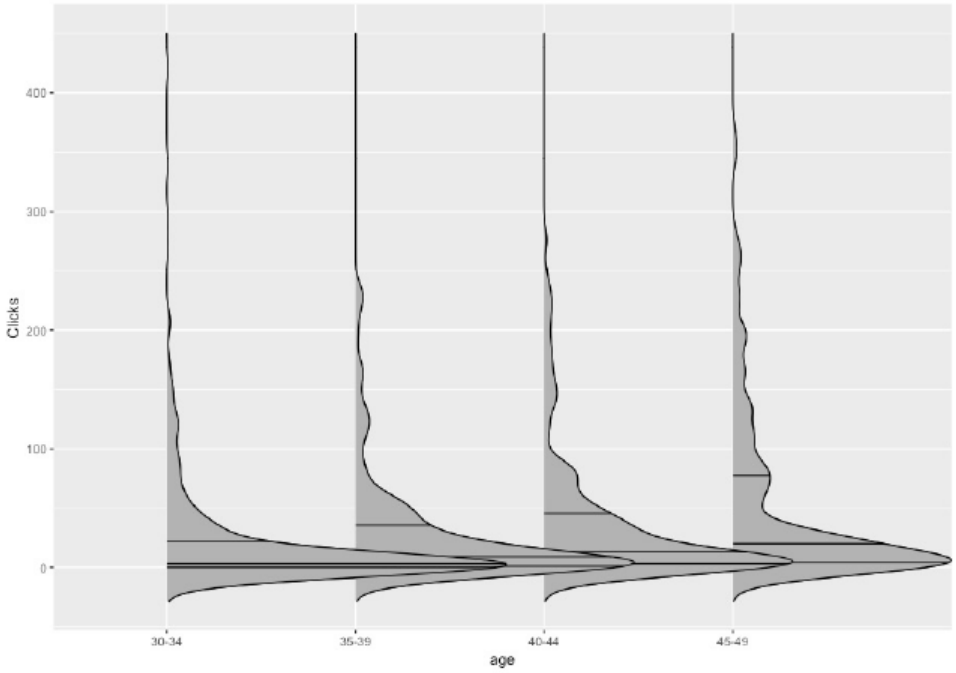
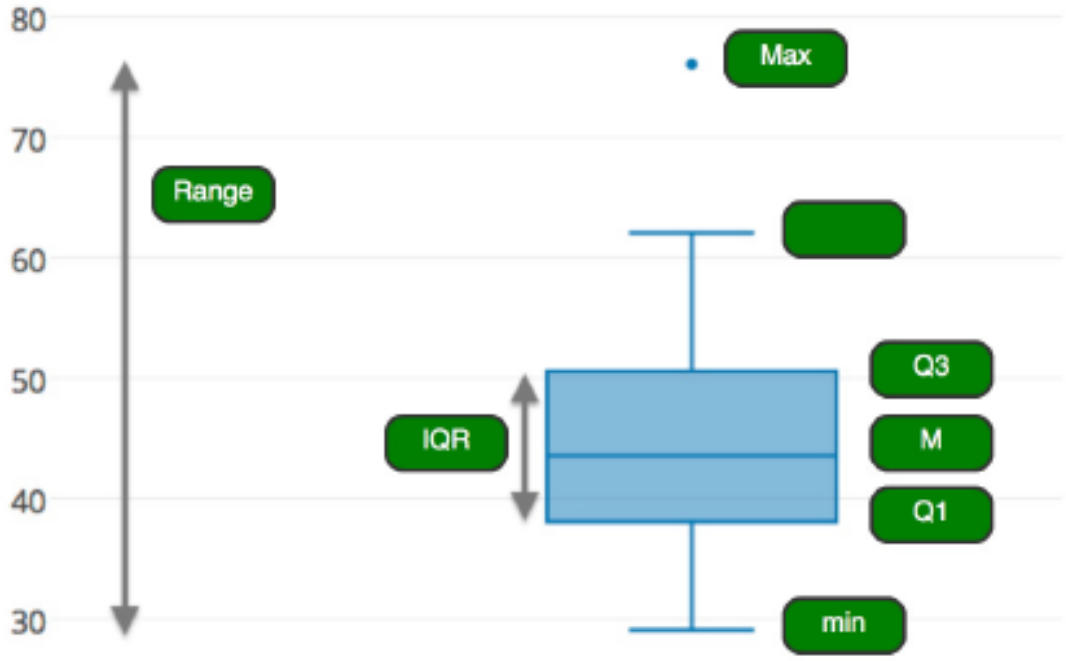
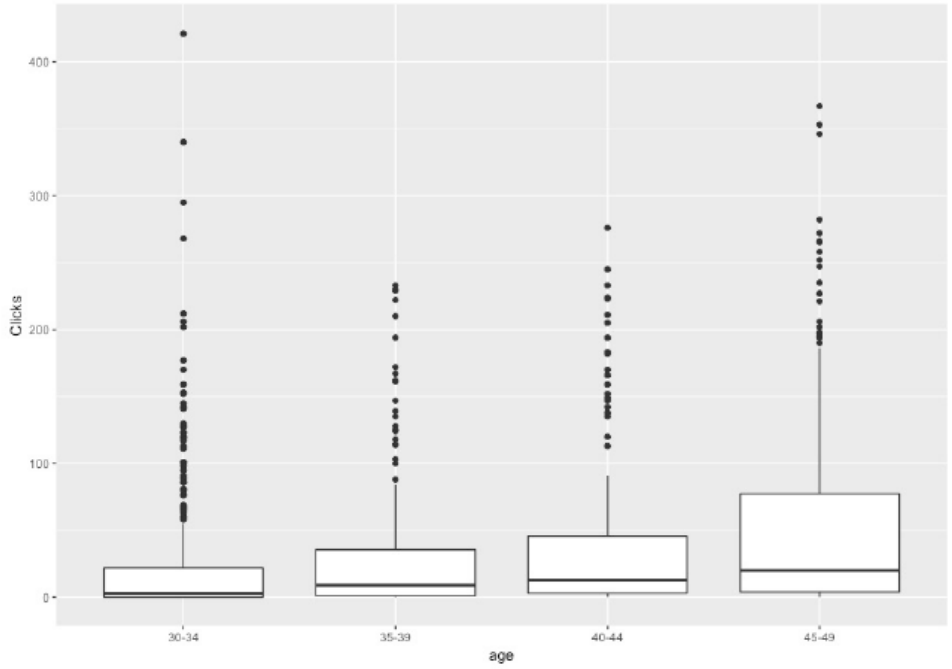


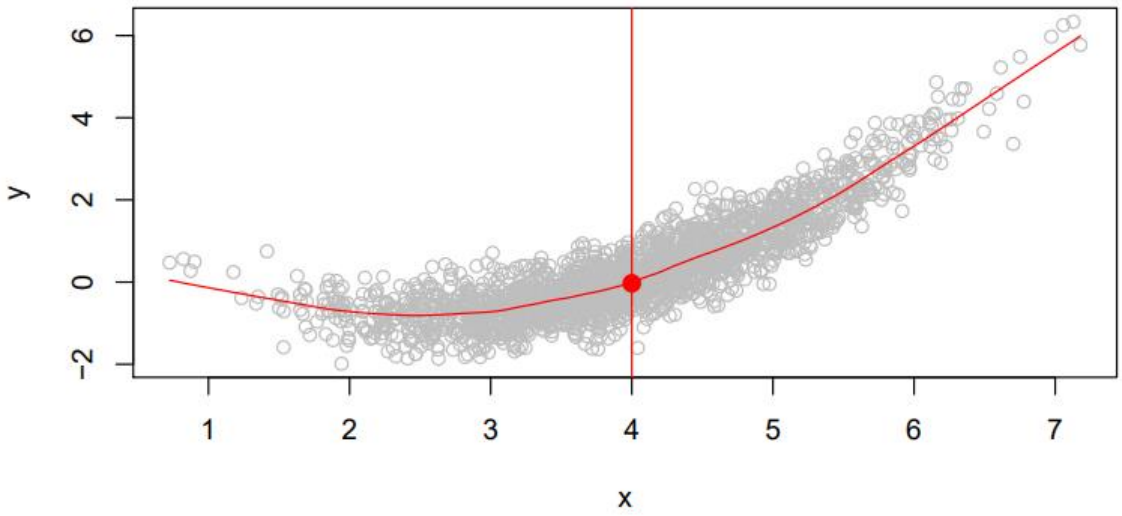
Diagrama de caja y brazos



Distribución condicional: diagrama de caja y brazos



¿Hay una $f(x)$ ideal?
La función de regresión



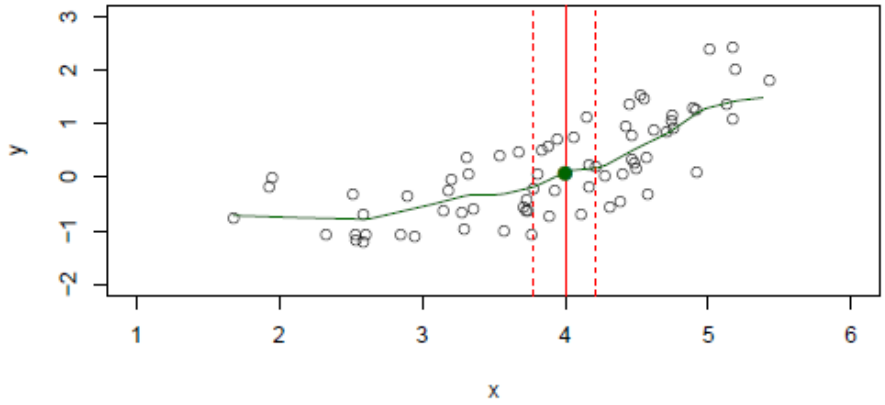
La función de regresión $\mathrm{E}[Y|X = x]$ minimiza $\mathrm{E}[(Y - g(x))^2|X = x]$ sobre todas las funciones $g(x)$ en los puntos $X = x$

¿Cómo estimamos $f(x)$?

- Por lo general, tenemos pocos o ningún punto con $X = x$ exactamente.
- ¡Entonces no podemos calcular $\mathrm{E}[Y|X = x]$!
- Solución: relajamos la definición y dejamos

$$\hat{f}(x) = Mean(Y|X \in N(x))$$

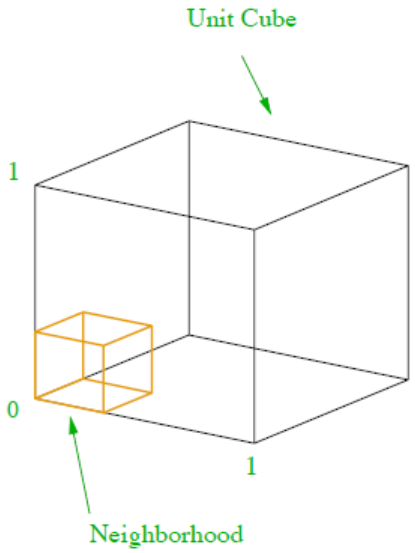
- donde $N(x)$ es una vecindad de x.

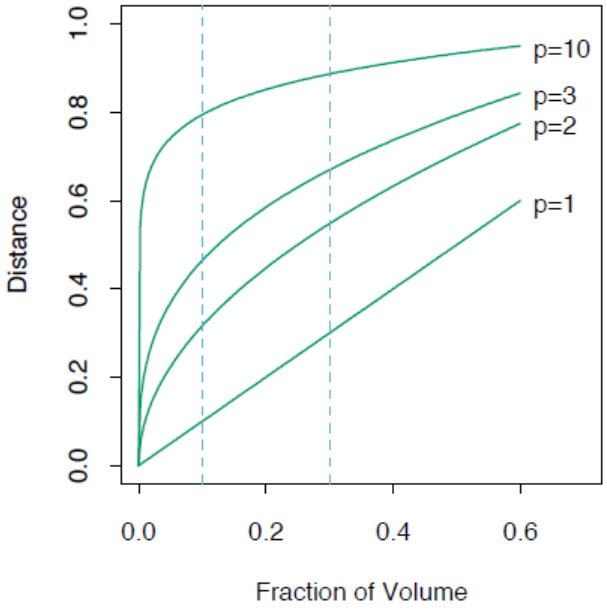


Vecinos más cercanos

- El promedio de vecinos más cercanos puede ser bastante bueno para dimensiones pequeñas
  - es decir, $d \leq 4$ y n grande.
- Los métodos de vecino más cercano pueden ser malos cuando d es grande.
- Motivo: **la maldición de la dimensionalidad**. Vecinos más cercanos tienden a estar muy lejos en altas dimensiones.
  - Necesitamos obtener una fracción razonable de los N valores de y_i para promediar y así reducir la varianza | p. ej. 10%
  - Un vecindario del 10% en grandes dimensiones ya no necesita ser local, entonces perdemos el espíritu de estimar $\mathrm{E}[Y|X = x]$ por un promedio local.

La maldición de la dimensionalidad





$$e_p(r) = r^{1/p}$$

Regresión lineal

- Primero nos enfocamos en modelar la función de regresión

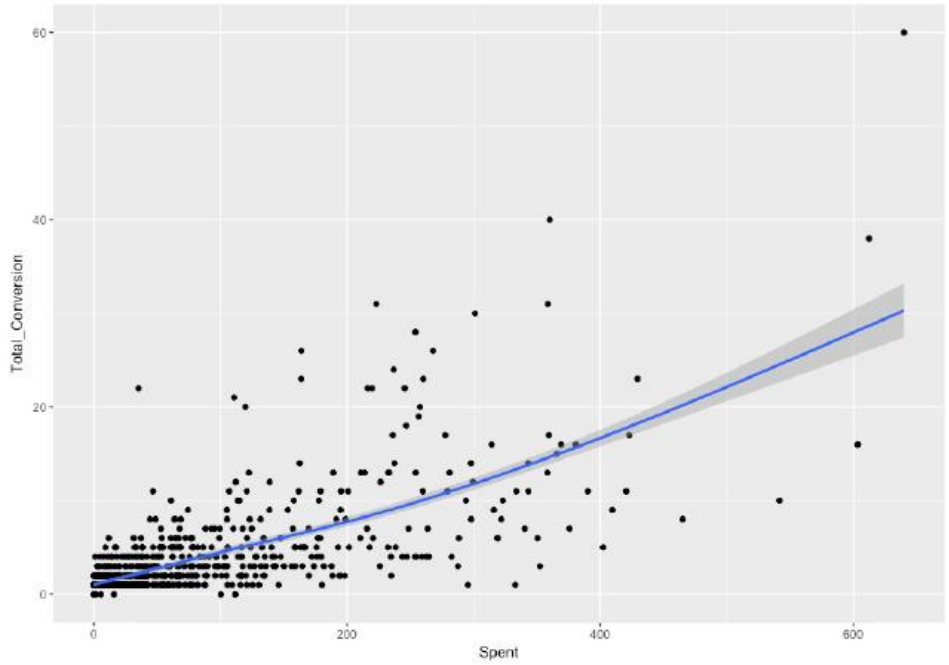
$$f(x) = \mathrm{E}[Y|X = x]$$

  - Demasiadas funciones f de regresión posibles (no paramétricas)
  - Es útil restringir a funciones simples que son descritas por pocos parámetros
  - La más simple:

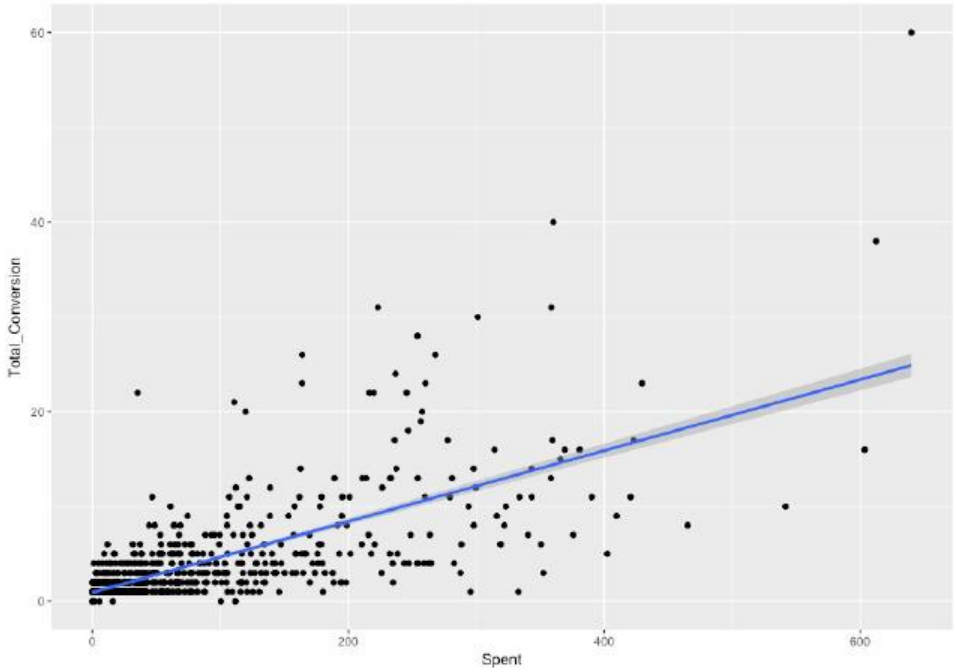
$$f(x) = a + bx$$

  - Bajo esta suposición, hablamos de regresión lineal

Regresión no paramétrica (LOESS)



Regresión lineal



Análisis probabilístico

- Sean X y Y dos variables aleatorias reales (no necesariamente independientes) con dos momentos y tal que $\mathrm{var}(X) > 0$
- La regresión lineal teórica de Y en X es la línea $x \to a^* + b^* x$ donde

$$(a^*, b^*) = \underset{(a,b)\in R^2}{argmin} E[(Y - a - bX)^2]$$

- Igualando las derivadas parciales a cero obtenemos

$$b^* = \frac{cov(X,Y)}{var(X)}$$

$$a^* = E[Y] - b^* E[X] = E[Y] - \frac{cov(X,Y)}{var(X)} E[X]$$

Modelos paramétricos y estructurados

- Un modelo lineal se especifica en términos de d + 1 parámetros

$$f_L(x) = \beta_0 + \beta_1 X + \beta_2 X + \cdots \beta_d X$$

- Estimamos los parámetros ajustando el modelo a datos de entrenamiento.
- Aunque casi nunca es correcto, un modelo lineal a menudo sirve como una aproximación buena e interpretable a la función verdadera desconocida $f(x)$.

Ruido

- A veces hay aleatoriedad real (estudiante aleatorio, moneda sesgada, error de medición, etc.)
- A veces es un fenómeno determinista pero demasiado complejo:
- Modelado estadístico
  - Proceso complicado "=" Proceso simple + ruido aleatorio
- (buen) modelado consiste en elegir un proceso simple (plausible) y la distribución de ruido.

Ruido

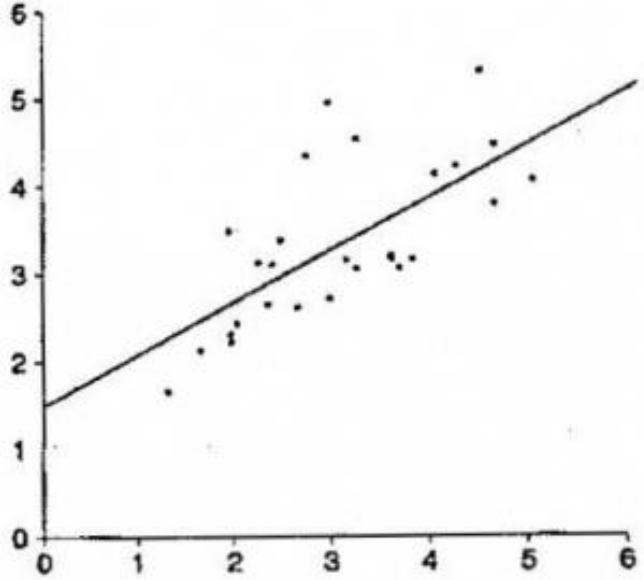
- Claramente los puntos no están exactamente en la línea $x \rightarrow a^* + b^* x$ si la $\text{var}(Y|X = x)$>0
- La variable aleatoria $\varepsilon = Y - (a^* + b^* x)$ es llamada ruido y satisface

$$Y = a^* + b^* x + \varepsilon$$

- Con

$$E[\varepsilon] = 0$$
$$cov(\varepsilon, X) = 0$$

All this data, and statisticians still miss every point.



Error irreducible

- $\varepsilon = Y - f(x)$ es el error irreducible, aun si conociéramos $f(x)$, todavía tendríamos errores en la predicción, ya que en cada $X = x$ normalmente hay una distribución de posibles valores de Y.
- Para cualquier estimado $\hat{f}(x)$ de $f(x)$, tenemos

$$\mathrm{E}\left[(Y - \hat{f}(x))^2 | X = x\right] = \left[f(x) - \hat{f}(x)\right]^2 + var(\varepsilon)$$

Reducible Irreducible

Problema estadístico

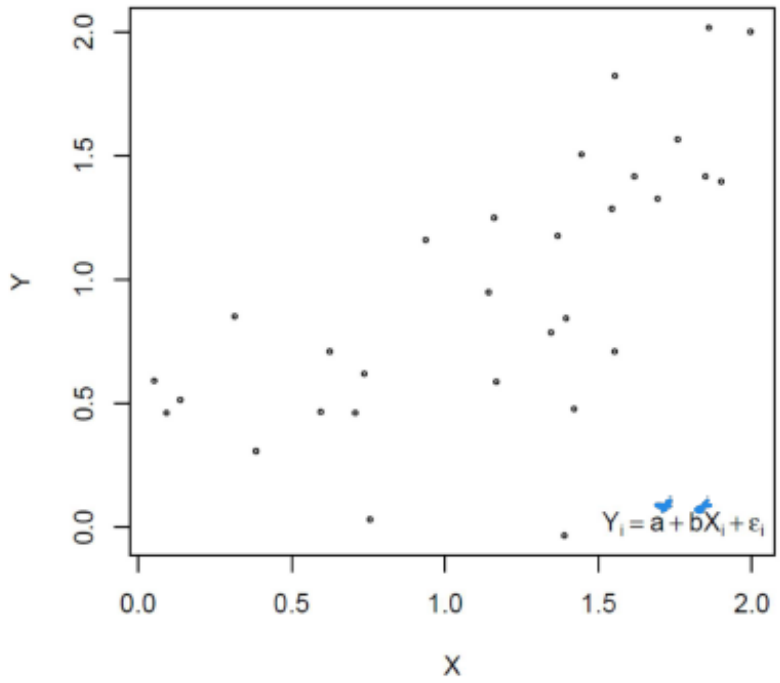
- En la práctica, a^* y b^* deben estimarse a partir de los datos.
- Supongamos que observamos n i.i.d. pares aleatorios
- $(X_1, Y_1), \ldots, (X_n, Y_n)$, con la misma distribución que (X, Y):

$$Y_i = a^* + b^* X_i + \varepsilon_i$$

- Queremos estimar a^* y b^*

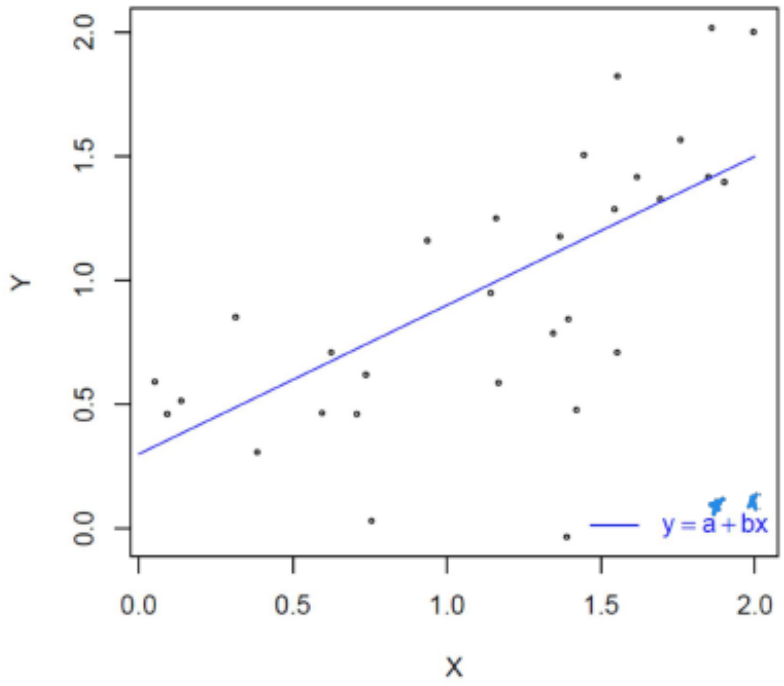
Datos

$$Y_i = a^* + b^* X_i + \varepsilon_i$$



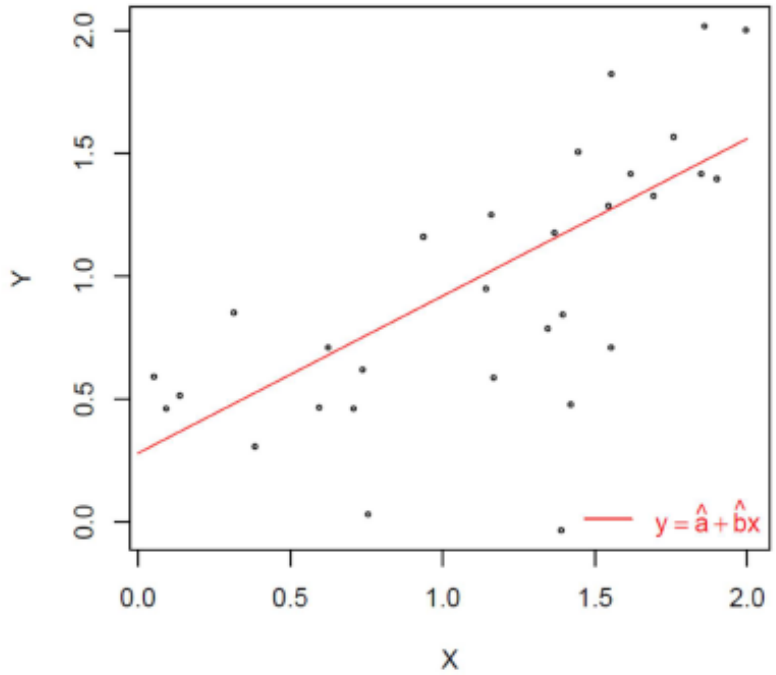
Parámetros reales

$$Y_i = a^* + b^* X_i + \varepsilon_i$$



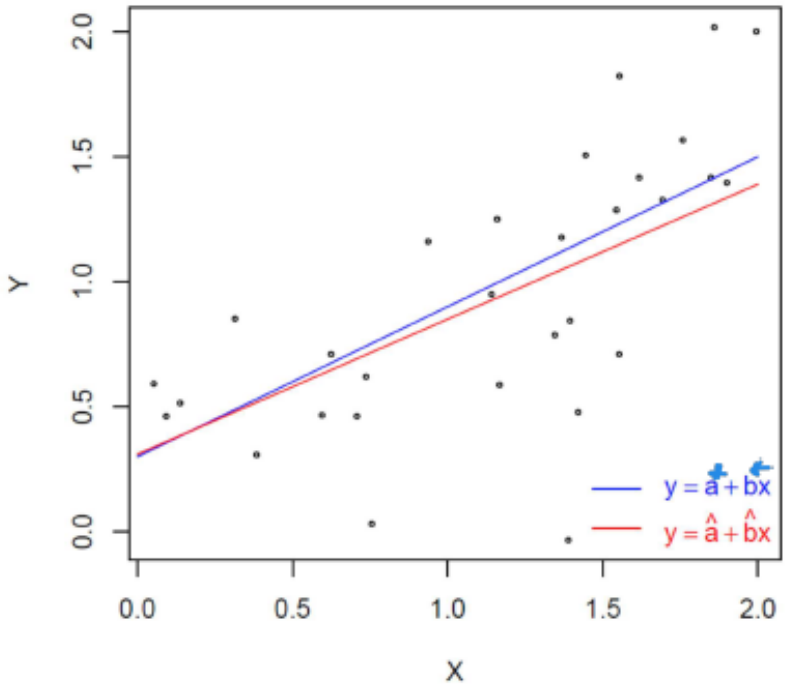
Parámetros estimados

$$Y_i = a^* + b^* X_i + \varepsilon_i$$



Comparación

$$Y_i = a^* + b^* X_i + \varepsilon_i$$



Mínimos cuadrados

- Definición
- El estimador de mínimos cuadrados (LSE) de (a, b) es el minimizador de la suma de los errores al cuadrado:

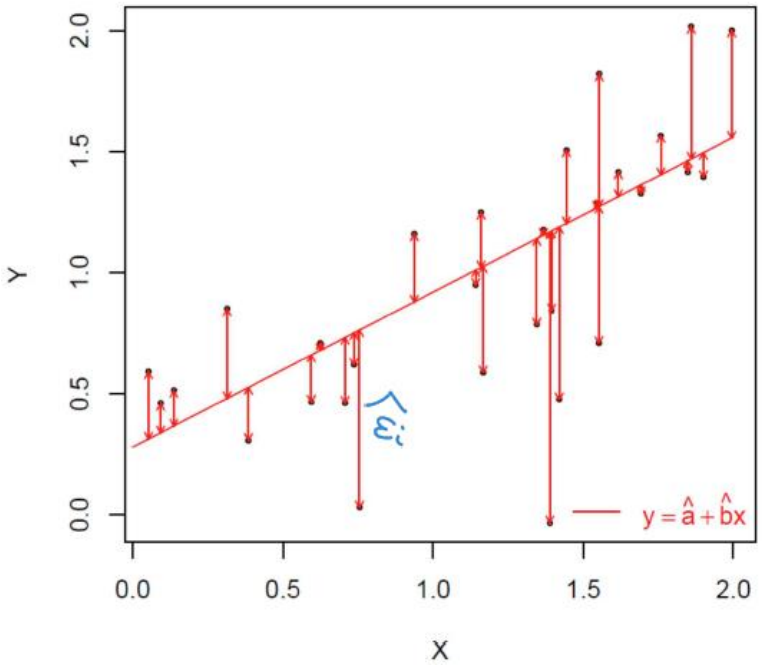
$$\sum_{i=1}^{n}(Y_i - a - bX_i)^2$$

- $(\hat{a}, \hat{b})$ *está dado por*

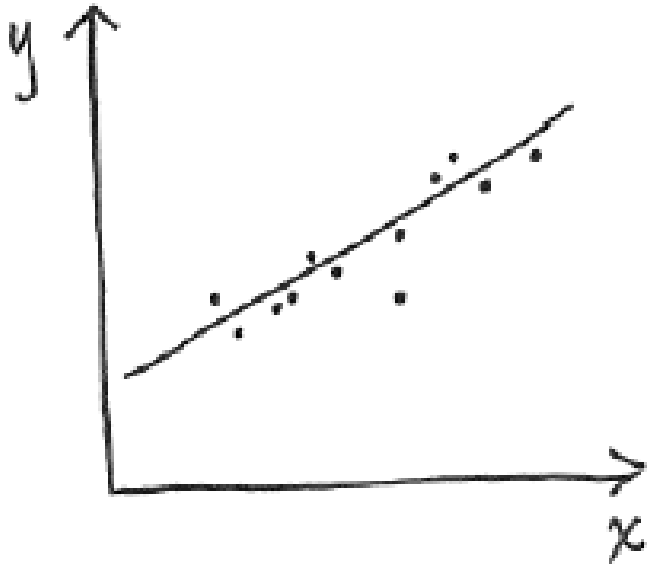
$$\hat{b} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \bar{X}^2}$$

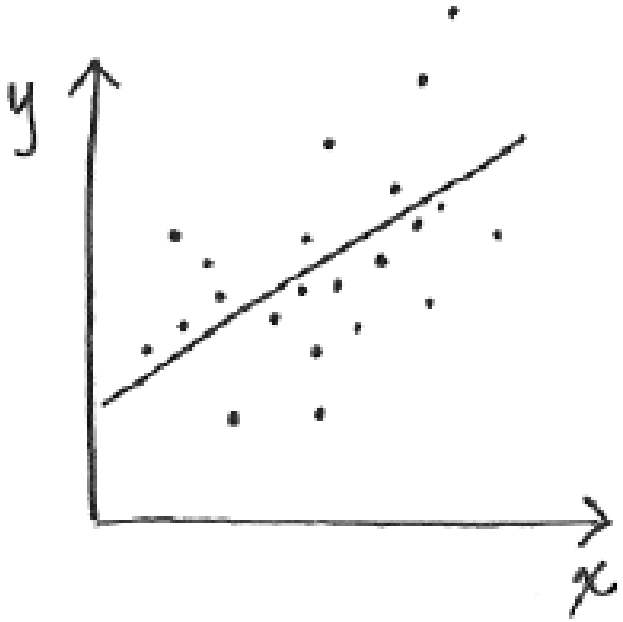
$$\hat{a} = \bar{Y} - \hat{b}\bar{X}$$

Residuales

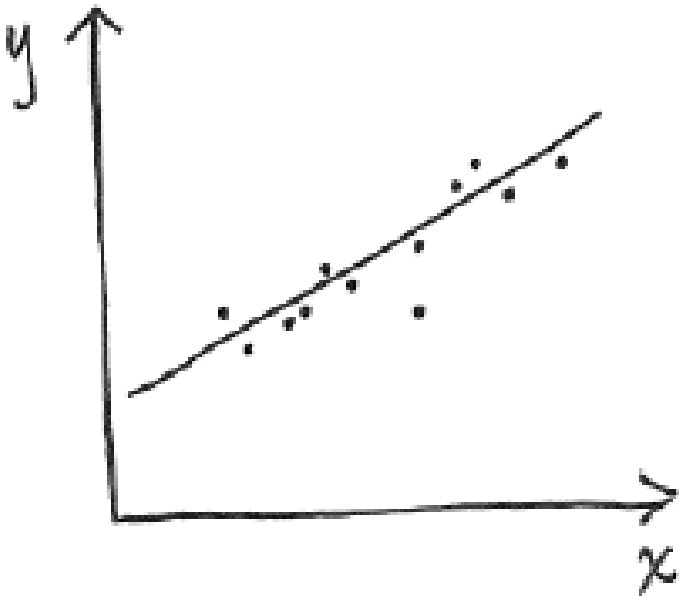


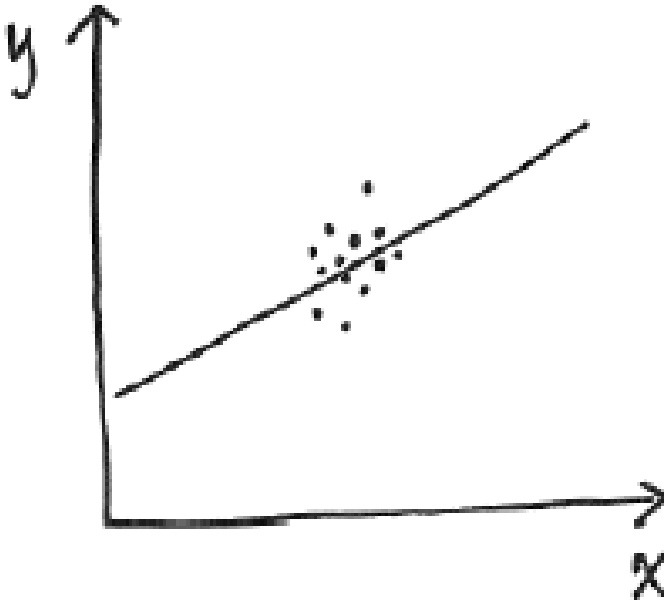
Mayor varianza de los errores





Mayor varianza de X





Regresión multivariada

$$Y_i = X_i{}^T \beta^* + \varepsilon_i, \; i = 1,2, \dots, n$$

- Vector de variables explicativas o covariables: $X_i \in R^d$ (suponga que su primera coordenada es 1).
- Respuesta / Variable dependiente: Y_i
- $\beta_i^* = a^*$ se conoce como el intercepto

Definición

- El estimador de mínimos cuadrados (LSE) de β^* es el minimizador de la suma de los errores al cuadrado:

$$\beta^* = \underset{\beta \in R^d}{argmin} \sum_{i=1}^{n} (Y_i - X_i{}^T \beta^*)^2$$

Solución analítica

- Obteniendo el gradiente e igualándolo a cero obtenemos

$$\hat{\beta} = \underset{\beta \in R^d}{argmin} \sum_{i=1}^{n} (Y_i - X_i{}^T \beta)^2$$

$$\hat{\beta} = \underset{\beta \in R^d}{argmin} \|\boldsymbol{Y} - \mathbb{X}\beta\|^2$$

$$\nabla_\beta \|\boldsymbol{Y} - \mathbb{X}\beta^*\|^2 = \mathbf{0}$$

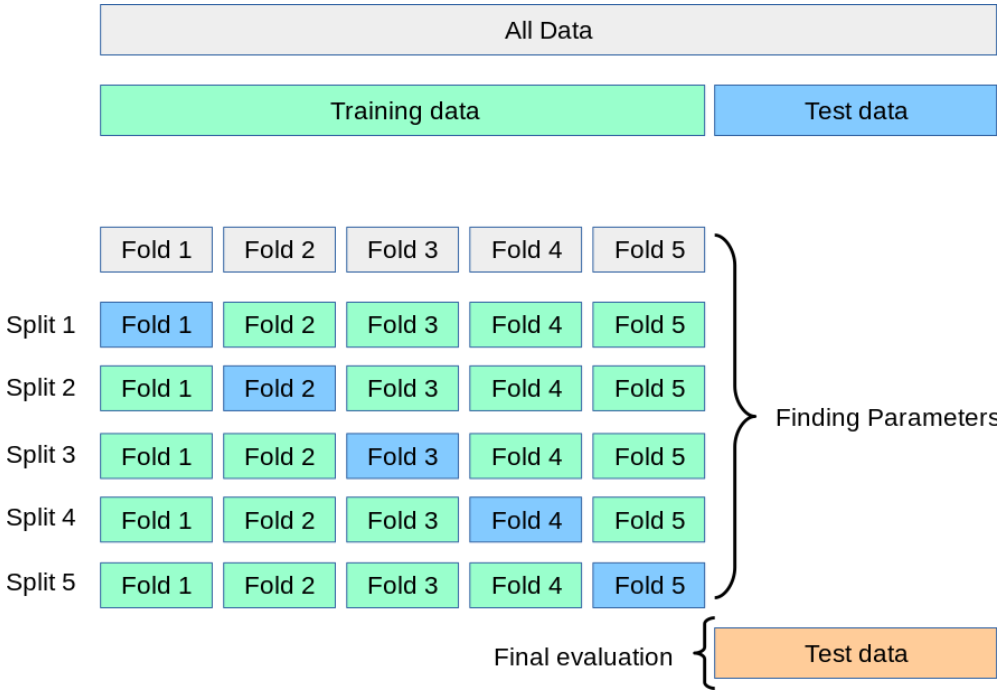
$$-\mathbf{2}\mathbb{X}^{\boldsymbol{T}}(\boldsymbol{Y} - \mathbb{X}\beta^*) = \mathbf{0}$$

$$\mathbb{X}^{\boldsymbol{T}}\boldsymbol{Y} - \mathbb{X}^{\boldsymbol{T}}\mathbb{X}\beta^* = \mathbf{0}$$

$$\mathbb{X}^{\boldsymbol{T}}\mathbb{X}\beta^* = \mathbb{X}^{\boldsymbol{T}}\boldsymbol{Y}$$

$$\beta^* = \left(\mathbb{X}^{\boldsymbol{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\boldsymbol{T}}\boldsymbol{Y}$$

Prueba y Validación cruzada



Validación cruzada

Validación-cruzada($\mathcal{S}, k$)

dividimos $\mathcal{S}$ en k pedazos $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k$ ($\approx$ *mismo tamaño*)

for $i = 1$ *to* k

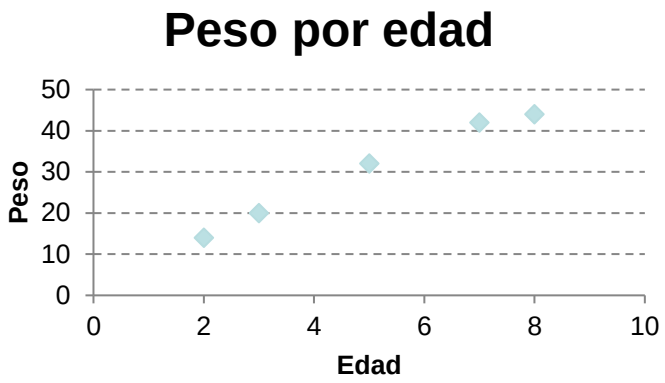
entrena h_i *con* $\mathcal{S} \setminus \mathcal{S}_i$ (sin $\mathcal{S}_i$)

calcula el error $\mathcal{E}_i(h_i)$ con $\mathcal{S}_i$

return $\frac{1}{k}\sum_{i=1}^{n} \mathcal{E}_i(h_i)$

Regresión lineal

Edad	Peso
2	14
3	20
5	32
7	42
8	44



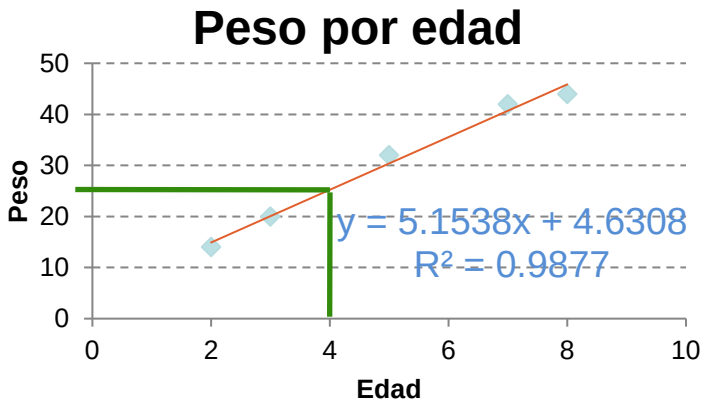
¿Cuál será el peso de alguien de 4 años de edad?

Regresión lineal

5 valores de entrenamiento (m)

Entrada (x) Salida (y)

Edad	Peso
2	14
3	20
5	32
7	42
8	44

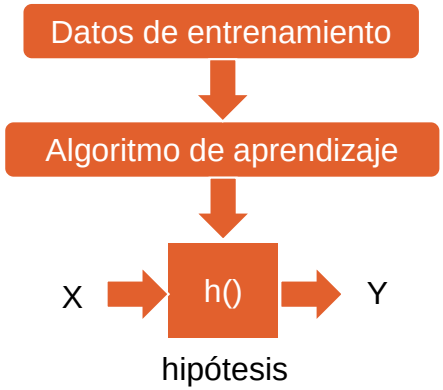


peso = 5.1538 * edad + 4.6308

edad=4 => peso=25.25

Regresión lineal

- Para obtener la ecuación de la recta puedo hacerlo de dos formas:
  - Descenso por gradiente (algoritmo de aprendizaje)
  - Analítica



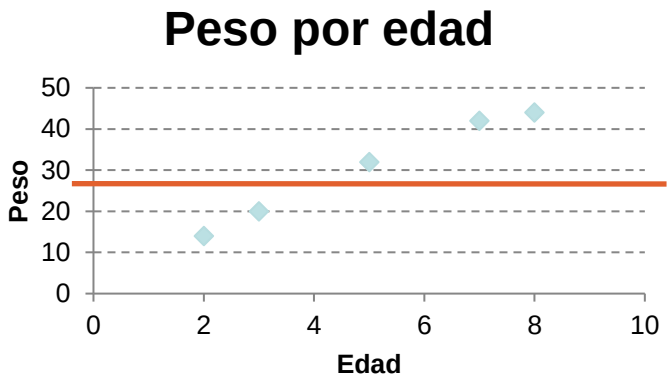
¿Cómo representamos h()?

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

Regresión lineal univariada

Regresión lineal

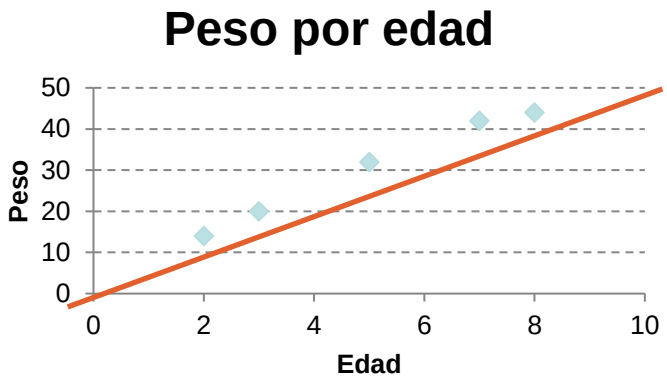
- ¿Cómo obtenemos los valores de los parámetros θ_0 y θ_1?



$\theta_0 = 25$ y $\theta_1 = 0$

Regresión lineal

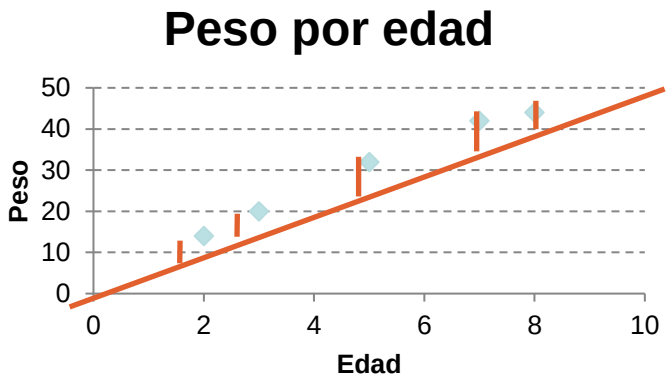
- ¿Cómo obtenemos los valores de los parámetros θ_0 y θ_1?



$\theta_0 = 0$ y $\theta_1 = 5$

Regresión lineal

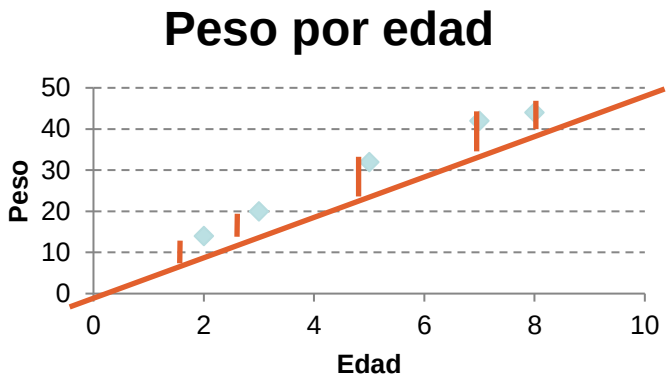
- ¿Cómo obtenemos los valores de los parámetros θ_0 y θ_1?



Debemos reducir la distancia entre los datos de entrenamiento y el valor de la recta, es decir, minimizar $\Sigma(h_\theta(x) - y)^2$ para todos los datos de entrenamiento

Regresión lineal

- ¿Cómo obtenemos los valores de los parámetros θ_0 y θ_1?

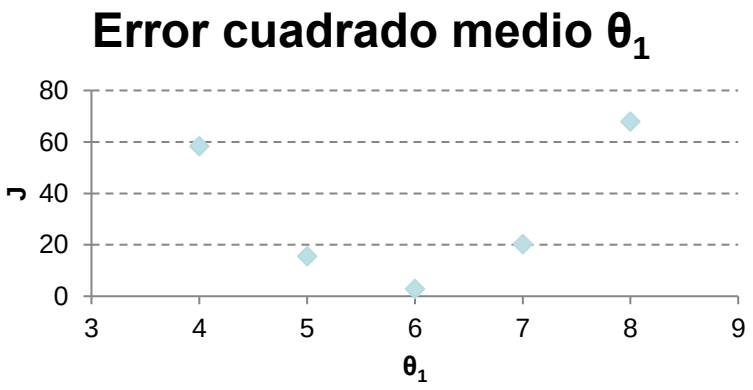


Minimizar $J(\theta_0, \theta_1) = 1/2m \bullet \Sigma(h_\theta(x) - y)^2$ (función de costo, error cuadrado medio)

Error cuadrado medio

- Error cuadrado medio (sólo θ_1)

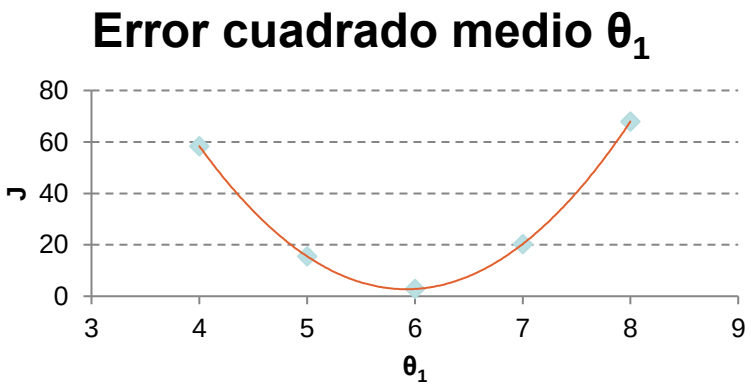
J(4)	J(5)	J(6)	J(7)	J(8)
36	16	4	0	4
64	25	4	1	16
144	49	4	9	64
196	49	0	49	196
144	16	16	144	400
58.4	15.5	2.8	20.3	68



Error cuadrado medio

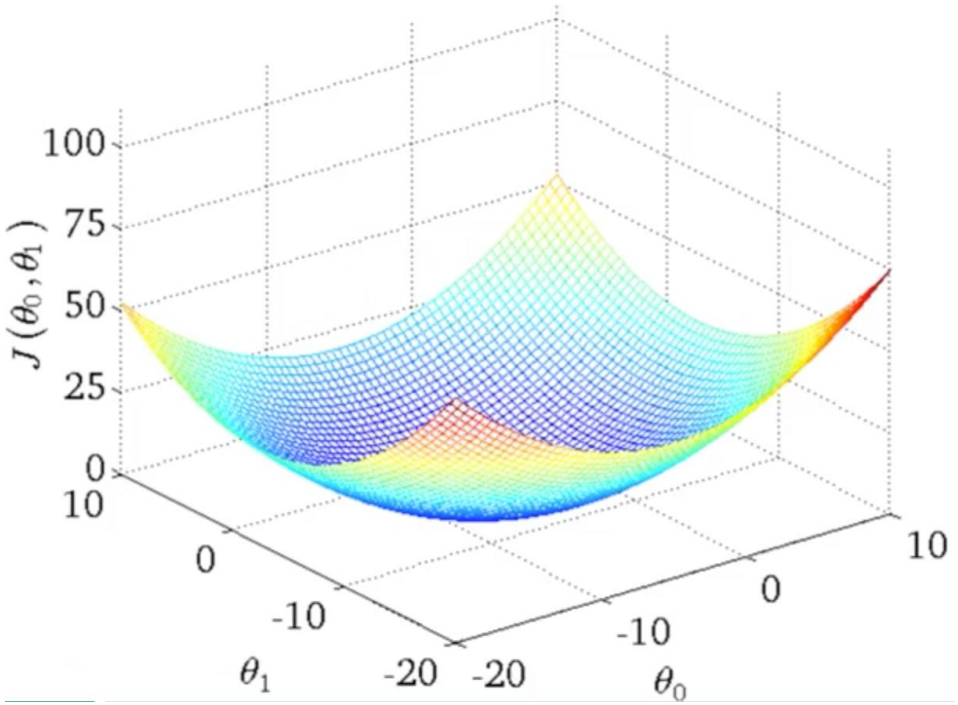
- Error cuadrado medio (sólo θ_1)

J(4)	J(5)	J(6)	J(7)	J(8)
36	16	4	0	4
64	25	4	1	16
144	49	4	9	64
196	49	0	49	196
144	16	16	144	400
58.4	15.5	2.8	20.3	68

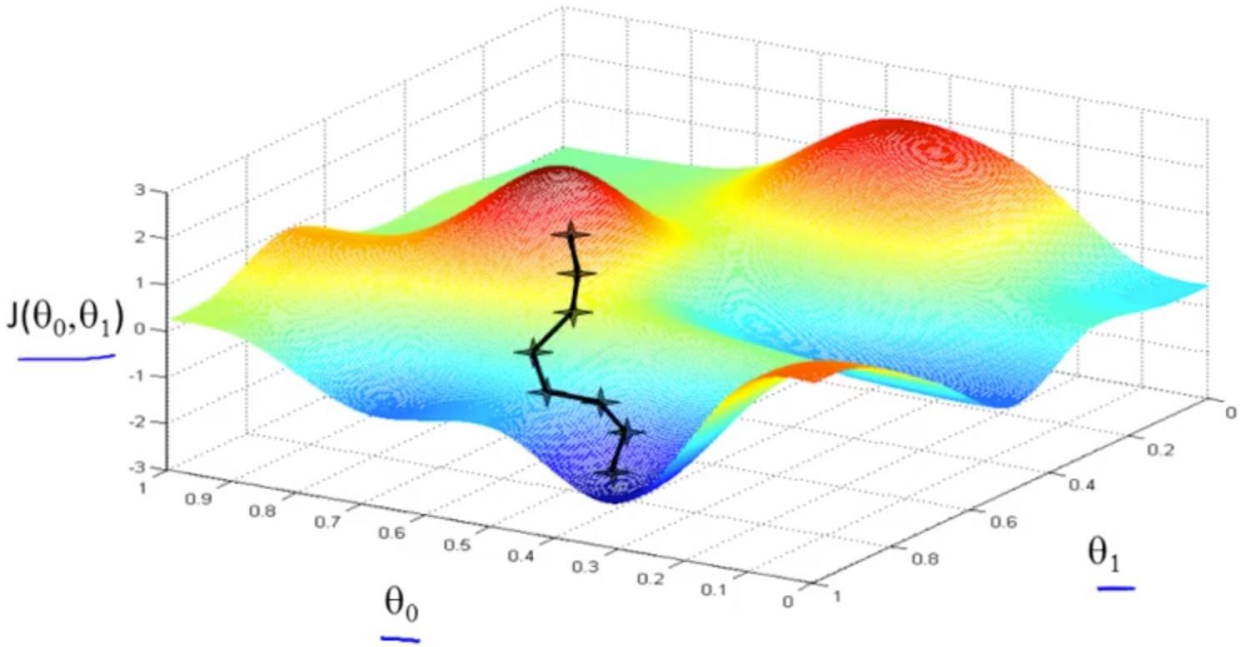


El valor mínimo del ECM se obtiene con θ_1=5.1538 y θ_0=4.6308

Error cuadrado medio



Descenso por gradiente



Descenso por gradiente

- Algoritmo de descenso por gradiente

  Repetir hasta que converja {

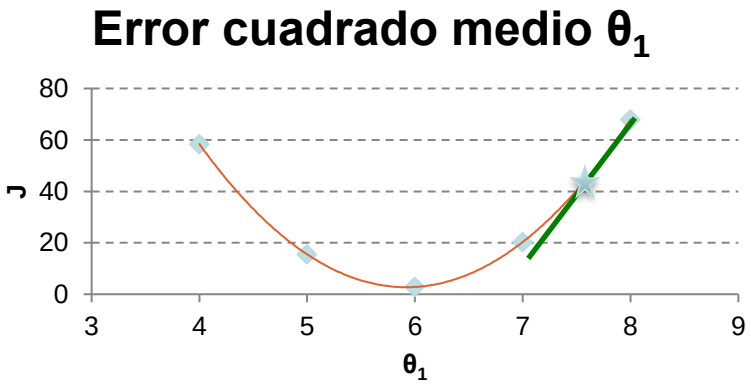
$$\theta_j := \theta_j - \alpha \cdot \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$ Para j=1 y j=2

  }

  α: tasa de aprendizaje

Descenso por gradiente

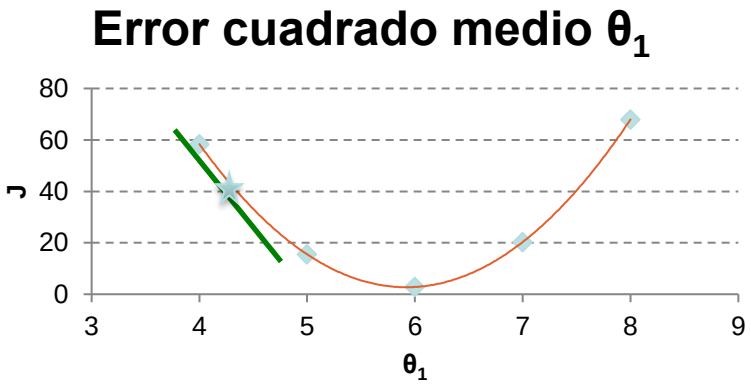
- Error cuadrado medio (sólo θ_1)



$$\theta_j := \theta_j - \alpha \cdot \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

Descenso por gradiente

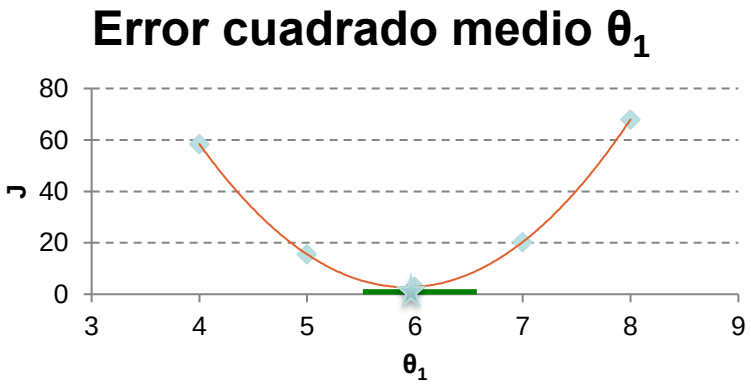
- Error cuadrado medio (sólo θ_1)



$$\theta_j := \theta_j - \alpha \cdot \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

Descenso por gradiente

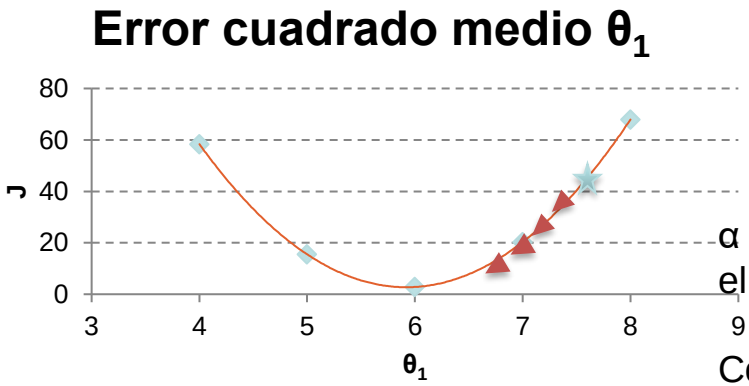
- Error cuadrado medio (sólo θ_1)



$$\theta_j := \theta_j - \alpha \cdot \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

Descenso por gradiente

- Error cuadrado medio (sólo θ_1)



α determina el paso

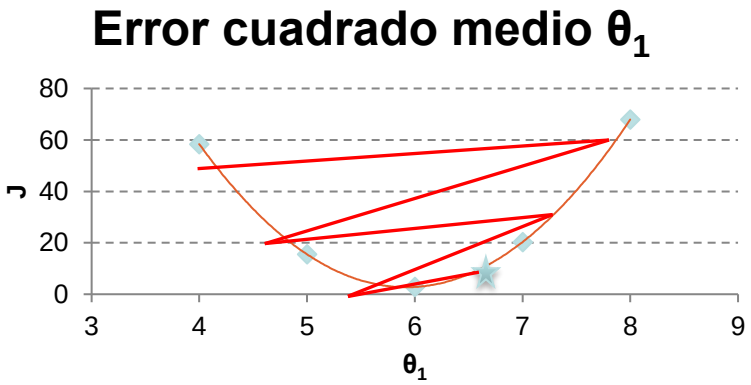
Conforme te acercas al mínimo disminuye el cambio (sin modificar α)

Si la variación de θ_j entre un paso y otro es $< 10^{-3}$ declaramos que el algoritmo convergió

$$\theta_j := \theta_j - \alpha \cdot \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

Descenso por gradiente

- Error cuadrado medio (sólo θ_1)



$$\theta_j := \theta_j - \alpha \cdot \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

Si α es muy grande no converge el algoritmo

Descenso por gradiente

- ¿Cuál es la derivada parcial de J con respecto a θ?

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

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

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

Descenso por gradiente

- Algoritmo de descenso por gradiente

  Repetir hasta que converja {

$$\theta_0 := \theta_0 - \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})$$

$$\theta_1 := \theta_1 - \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x^{(i)}$$

  }

  – A este algoritmo se le conoce como descenso por gradiente por lotes, ya que en cada paso utiliza todo el conjunto de entrenamiento

Descenso por gradiente

- Algoritmo de descenso por gradiente multivariado

  Repetir hasta que converja {

$$\theta_j := \theta_j - \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x^{(i)}$$ Para j=1 hasta n

  }

Coeficiente de determinación R^2

- ¿Qué tanto es mejor nuestro modelo que si sólo tomáramos la media?
- Siempre tiene valores entre

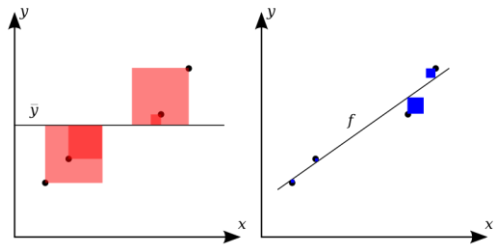
  0: no lo hacemos mejor que la media

  1: predecimos cada dato perfectamente
- Se puede ver como el porcentaje de la variación en los resultados que se explica con el modelo

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

$$SS_{res} = \sum_i (y_i - f_i)^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$



Escalamiento de características

- Con el fin de que el algoritmo de descenso por gradiente funcione adecuadamente es necesario que las variables se encuentren en la misma escala:
  - Edad: 18 a 65, Sueldo: 10,000 a 100,000

  Transformarlas para que estén en el rango ≈ -1 a 1:

$$x_j := \frac{x_j - \mu_{x_j}}{\max(x_j) - \min(x_j)}$$

$$x_j := \frac{x_j - \mu_{x_j}}{\sigma_{x_j}}$$

-3 a 3

normalización

Obtención de parámetros sin escalar

$$z_j = \frac{x_j - \mu_{x_j}}{\sigma_{x_j}}$$

$$\hat{y} = \hat{\beta}_o + \sum_{j=1}^{k} \hat{\beta}_j z_j$$

$$\hat{y} = \hat{\beta}_o + \sum_{j=1}^{k} \hat{\beta}_j \left(\frac{x_j - \mu_{x_j}}{\sigma} \right)$$

$$\hat{y} = \left(\hat{\beta}_o - \sum_{j=1}^{k} \hat{\beta}_j \frac{\mu_{x_j}}{\sigma_{x_j}} \right) + \sum_{j=1}^{k} \left(\frac{\hat{\beta}_j}{\sigma_{x_j}} \right) x_j$$

Tasa de aprendizaje

- Si la tasa de aprendizaje es muy pequeña, el algoritmo se tardará mucho en converger
- Si la tasa de aprendizaje es muy grande, el algoritmo puede no converger

- Valores comunes para α son .001, .01, .1, 1, …

Regularización

- Regresión Ridge y Lasso
  - Podemos ajustar un modelo que contenga todos los p predictores utilizando una técnica que restrinja o regularice las estimaciones de los coeficientes, o de manera equivalente, que reduzca las estimaciones de los coeficientes hacia cero.
  - Puede que no sea inmediatamente obvio por qué tal restricción debería mejorar el ajuste, pero resulta que la reducción de las estimaciones de los coeficientes puede reducir significativamente su varianza.

Regresión Ridge

- Recordemos que el procedimiento de cálculo de mínimos cuadrados estimamos β_0, β_1, β_2, …, β_p utilizando los valores que minimizan

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_j\right)^2$$

- Por el contrario, las estimaciones del coeficiente de regresión Ridge $\hat{\beta}^R$ son los valores que minimizan

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$$

- donde $\lambda \geq 0$ es un parámetro de ajuste, que se determinará por separado.

Escalamiento de predictores

- Las estimaciones de los coeficientes de mínimos cuadrados estándar son equivariantes de escala: multiplicar X_j por una constante c simplemente conduce a escalar las estimaciones de coeficientes de mínimos cuadrados por un factor de $1/c$. En otras palabras, independientemente de cómo se escala el j-ésimo predictor, X_j $\hat{\beta}_j$ seguirá siendo el mismo.
- Por el contrario, las estimaciones del coeficiente de regresión de Ridge puede cambiar sustancialmente al multiplicar un predictor dado por una constante, debido a la suma del término de los coeficientes al cuadrado en la parte de penalización de la función objetivo de regresión de la cresta.
- Por lo tanto, es mejor aplicar la regresión de crestas después de estandarizar los predictores, usando la fórmula

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}}$$

Regresión Lasso

- La regresión Ridge tiene una desventaja obvia: a diferencia de la selección de subconjuntos, que generalmente seleccionará modelos que involucran solo un subconjunto de las variables, la regresión Ridge incluirá todos los predictores p en el modelo final
- El Lasso es una alternativa relativamente reciente a la regresión Ridge que supera esta desventaja. Los coeficientes de lasso, $\widehat{\beta}_{\lambda}^{L}$, minimizan la cantidad

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|=RSS+\lambda\sum_{j=1}^{p}|\beta_j|$$

- En el lenguaje estadístico, el lasso usa un ℓ_1 (pronunciado penalización "ele 1") en lugar de penalización de un ℓ_2. La norma ℓ_1 de un vector de coeficientes β está dada por $\|\beta\|_1=\sum|\beta_j|$

Solución analítica con regularización Ridge

$$\beta = (\mathbb{X}^T\mathbb{X} + \lambda I_n)^{-1} \cdot \mathbb{X}^T \boldsymbol{Y}$$

Agregando este término además la hacemos
la matriz invertible si era singular

Descenso por gradiente vs. Solución analítica

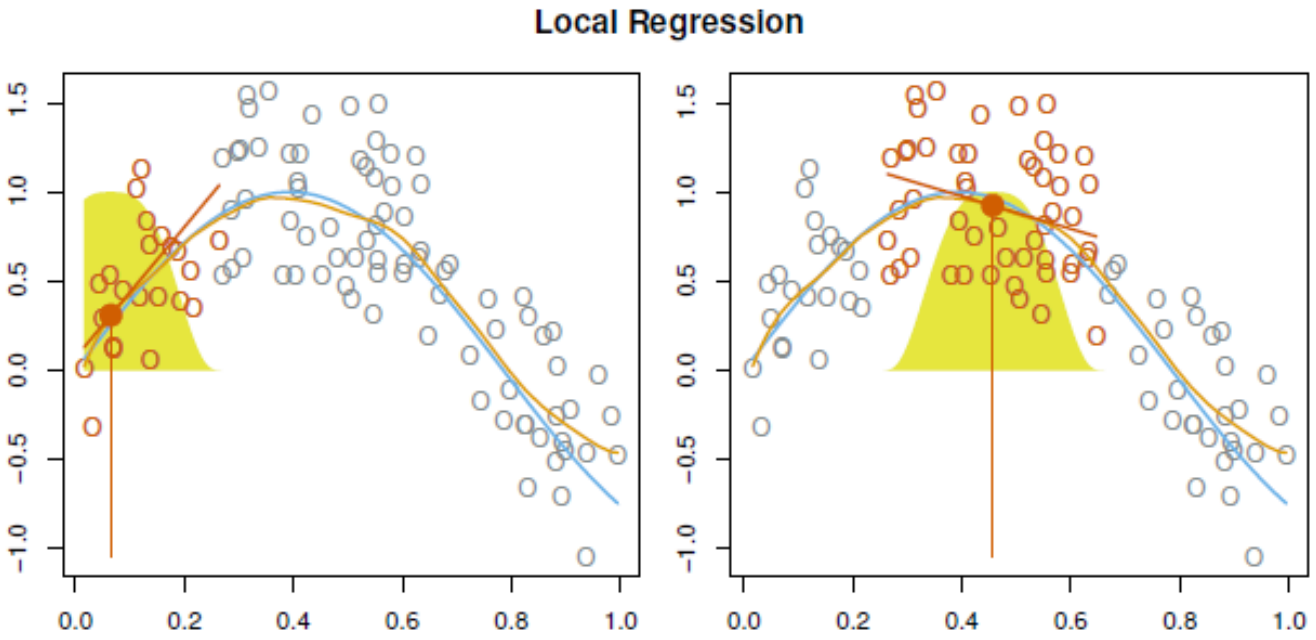
Descenso por gradiente

- Se necesita escoger α
- Se necesitan muchas iteraciones
- Funciona bien aún para un número grande de características

Solución analítica

- No se necesita escoger α
- No se necesita iterar
- Se necesita calcular $(X^TX)^{-1}$. La inversión de matrices $O(n^3)$
- Muy lento cuando hay un número muy grande de características (>10,000)

Locally estimated scatterplot smoothing (LOESS)



Supuestos básicos en regresión lineal

i. $X_i \; y \; \varepsilon_i$ no correlacionados

ii. $var(X) > 0$

iii. $\mathbb{E}(\varepsilon_i) = 0$

iv. homocedasticidad $\mathbb{E}({\varepsilon_i}^2) = \sigma^2 \; para \; toda \; i$

v. sin correlación serial $\mathbb{E}\left(\varepsilon_i \varepsilon_j\right) = 0 \; si \; i \neq j$

Por conveniencia podemos imponer que las X_i no son estocásticas como alternativa a i)

Los supuestos del iii) al v) los podemos cumplir con un supuesto más fuerte donde las $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

Generalización

- Los dos componentes del modelo (que vamos a relajar) son

1. **Componente aleatorio**: la variable de respuesta Y es continua y $Y|X = x$ es gaussiana con media $\mu(x)$

2. **Función de regresión**: $\mu(x) = x^T\beta$ (lineal)

Modelo lineal

- Un modelo lineal gaussiano asume

$$Y|X = x \sim \mathcal{N}_d(\mu(x), \sigma^2 I_d)$$

- Y

$$\mathbb{E}[Y|X = x] = \mu(x) = x^T\beta \in \mathbb{R}$$

Componentes de un modelo lineal

- Un modelo lineal generalizado (GLM) generaliza modelos lineales gaussianos en las siguiente forma

1. **Componente aleatorio**:

$$Y|X = x \sim \text{ es alguna distribución}$$

(por ejemplo Bernoulli, exponencial, Poisson)

1. **Función de regresión**: $g(\mu(x)) = x^T\beta$ (lineal)

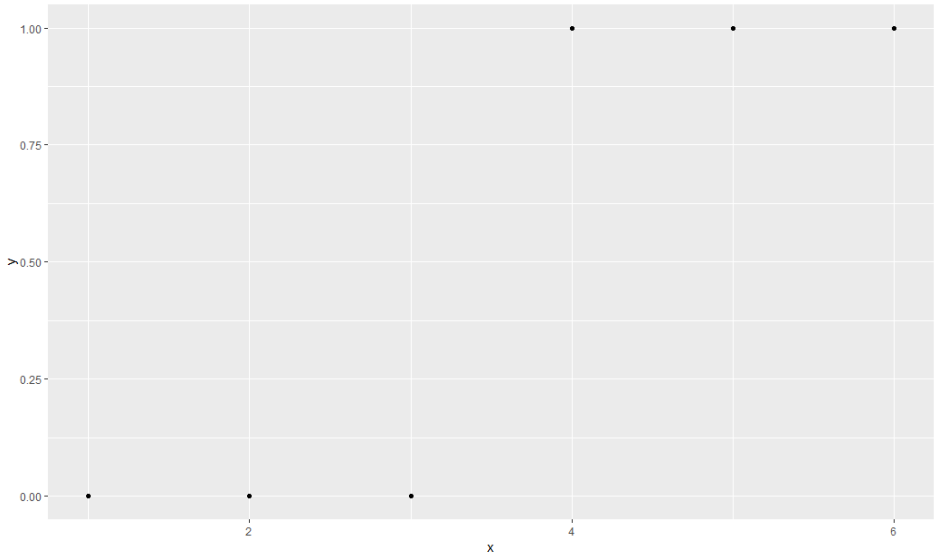
Donde g se le conoce como función liga y $\mu(x) = \mathbb{E}[Y|X = x]$ es la función de regresión

Familia Exponencial

- La familia exponencial comprende un conjunto de distribuciones flexibles que abarcan tanto variables aleatorias continuas como discretas.
  - Gaussiana: $\mathbb{R}^q$
  - Bernoulli: binaria{0, 1}
  - Multinomial: categórica
  - Binomial: conteos de éxitos/fracasos
  - Poisson: $\mathbb{N}^+$
  - Exponencial: $\mathbb{R}^+$

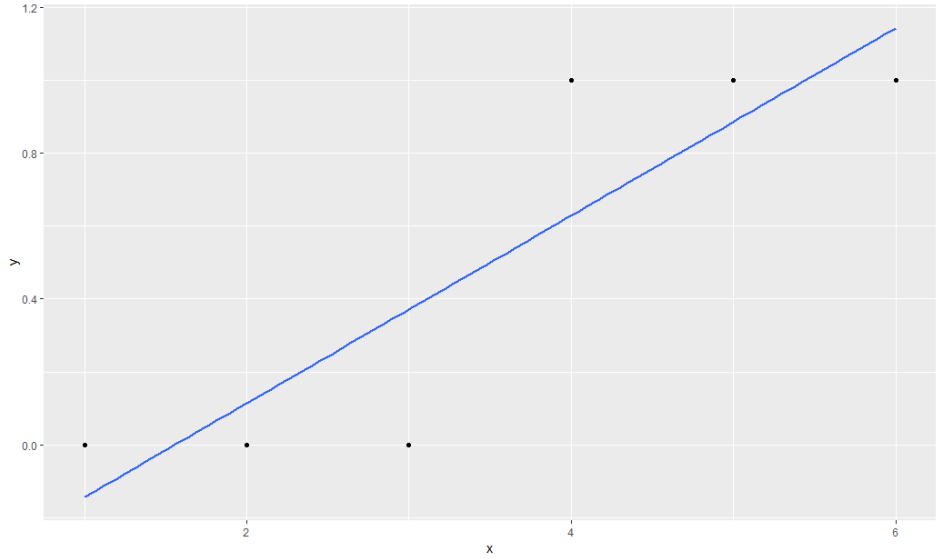
Clasificación

- Debemos poder separar de alguna forma



Clasificación

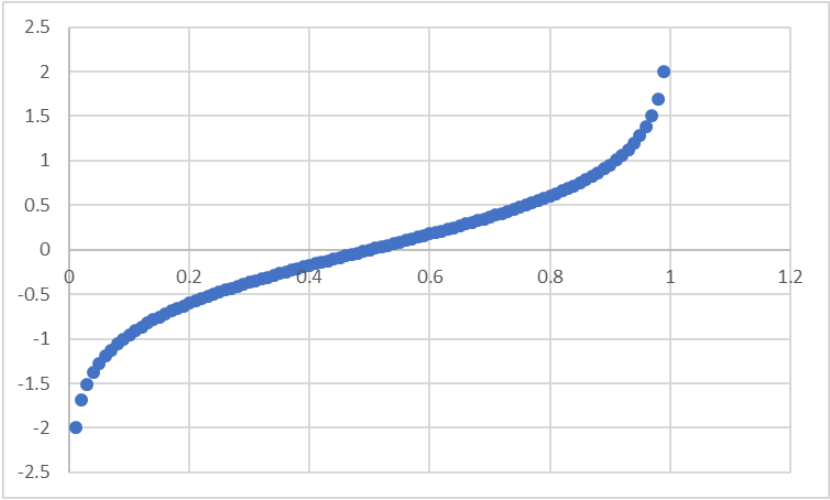
- Podríamos tratar de utilizar la regresión lineal
  - En muchos casos, como en este, una línea no separa de forma correcta los valores



Liga Logit

$$b'^{-1}(\mu(x)) = -\ln\left(\frac{1-\mu(x)}{\mu(x)}\right)$$

$$= \ln\left(\frac{\mu(x)}{1-\mu(x)}\right)$$



Regresión logística

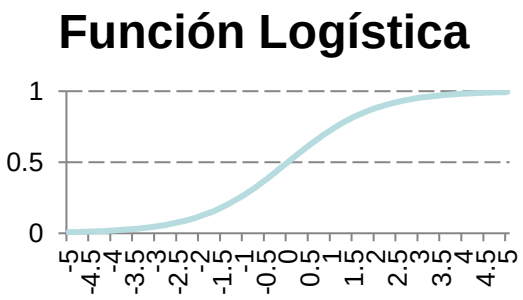
- Las predicciones de este método siempre están entre 0 y 1:

$$0 \leq h(x) \leq 1$$

- Por motivos históricos se le llama regresión, pero es un método para clasificación.

$$h_\theta(X) = g(\theta^T X)$$

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

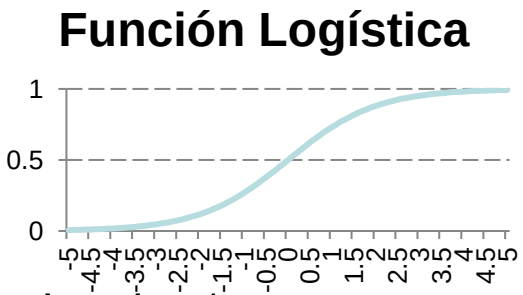


Regresión logística

- El resultado lo tomaré como la probabilidad de que la clasificación sea 1.
  - 0.7 indicaría que con 70% de probabilidad las características se clasificarían como 1

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

$$h_\theta(X) = p(y=1 \mid x;\theta)$$

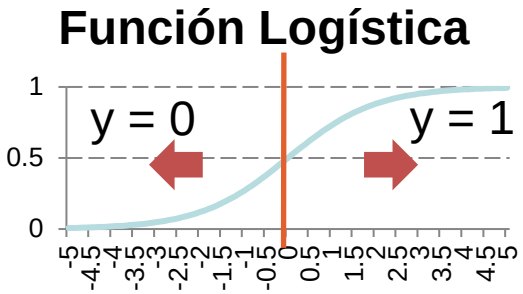


Probabilidad de que "y" sea igual a 1 dado x, parametrizado por θ

Límite de decisión

- $\theta^T X$ define el límite de decisión con el cual clasificamos a nuestra entrada
  - Dependiendo de cómo se concentren los datos de entrada, se determina qué tipo de límite utilizo

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



Si $\theta^T X \geq 0$ entonces lo clasificamos como 1
Si $\theta^T X < 0$ entonces lo clasificamos como 0

Función de costo

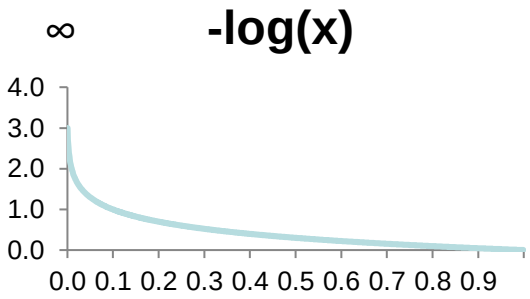
- Si usáramos la misma función de costo que la rergesión obtendríamos una función no convexa

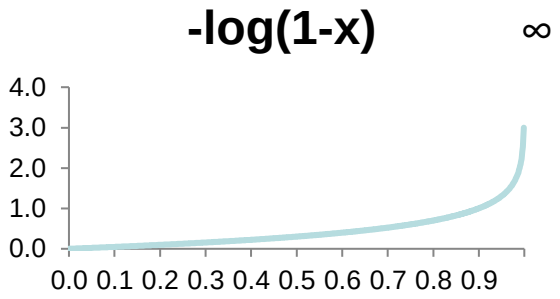
$$J(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2$$

- Para asegurar que sea convexa usamos:

$$J(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{m}\sum_{i=1}^{m}\text{costo}(h_\theta(x^{(i)}), y^{(i)})$$

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





Función de costo

- Reescribimos la función de costo

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

$$\text{costo}(h_\theta(x), y) = -y \cdot \log(h_\theta(x)) - (1 - y)\log(1 - h_\theta(x))$$

- Derivamos la función

$$\theta_j := \theta_j - \alpha \cdot \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1, \ldots, \theta_n)$$

$$J(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{m} \sum_{i=1}^{m} \text{costo}(h_\theta(x^{(i)}), y^{(i)})$$

$$\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}$$

Descenso por gradiente

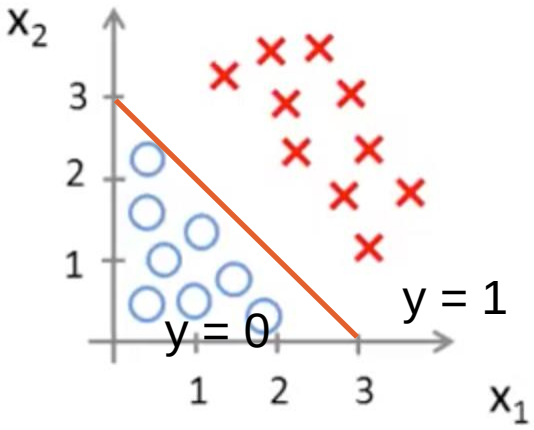
- Algoritmo de descenso por gradiente para regresión lineal

  Repetir hasta que converja {

$$\theta_j := \theta_j - \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}$$ Para j=0 hasta n

  }

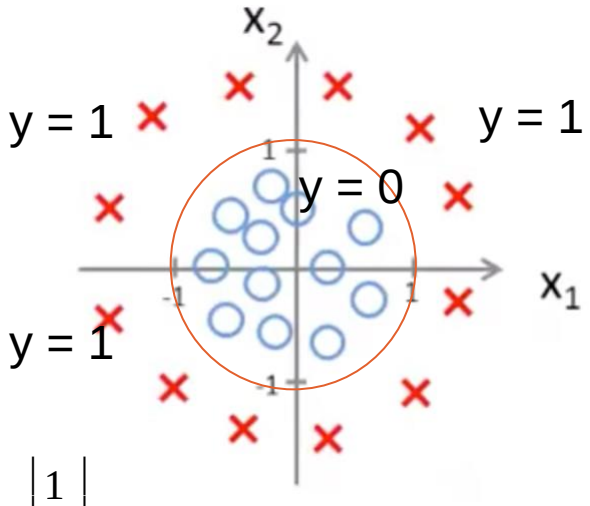
Límite de decisión



$$\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$$

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

$x_1 + x_2 \geq 3$



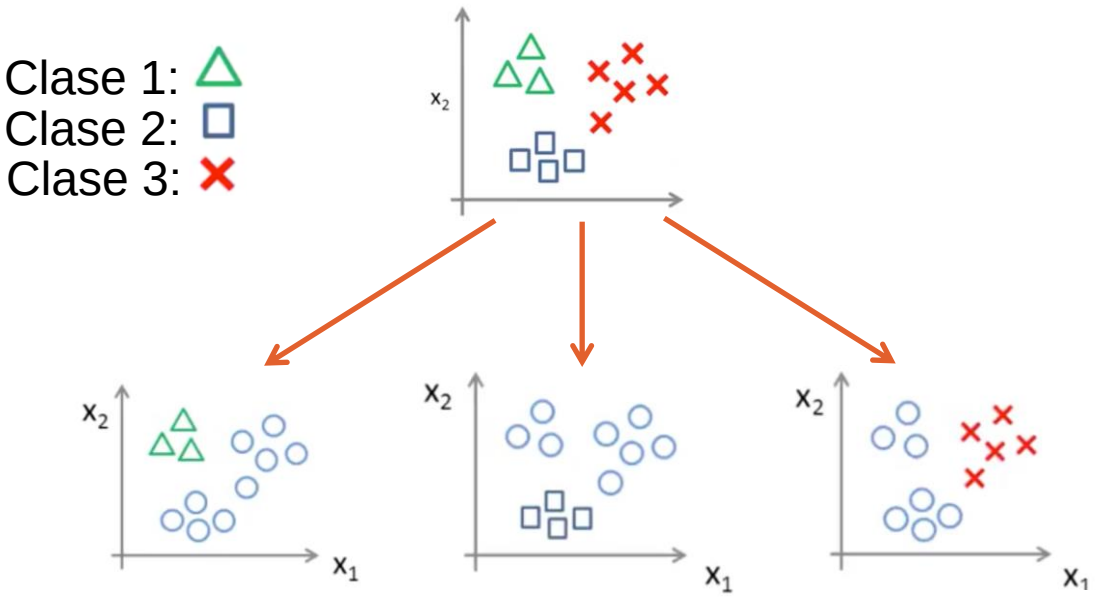
$$\theta = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

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

$x_1^2 + x_2^2 \geq 1$

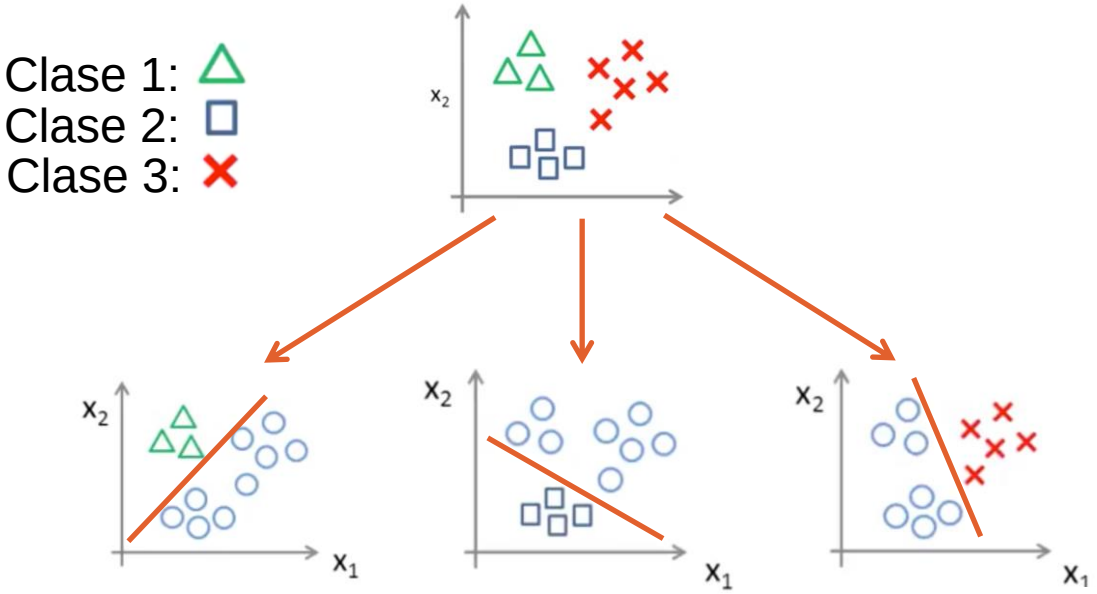
Clasificación multiclase

- Uno vs. Resto



Clasificación multiclase

- Uno vs. Resto



$$h_\theta^{(i)}(X) = p(y = i \mid x;\theta)$$

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

Evaluación de modelos de clasificación

- Exactitud
  - ¿Qué tan frecuentemente el clasificador realiza la predicción correcta?
  - Objetivo de negocio: Necesitamos que la mayoría de las decisiones sean las correctas.

$$exactitud = \frac{\#\ predicciones\ correctas}{\#\ total\ de\ observaciones}$$

$$= \frac{TP + TN}{TP + FP + TN + FN}$$

Exactitud

- La exactitud es un ejemplo de micro-average
- No sirve para conjuntos de datos no balanceados (por ejemplo en detección de fraude).
  - En este caso el modelo nulo es muy preciso, pero obviamente esto no lo hace el mejor modelo.
  - Hay que considerar una función de costo.

Matriz de confusión

- Una de las desventajas de la métrica de exactitud es que no hace distinción entre las clases, y en muchas aplicaciones los falsos positivos no cuestan lo mismo que los falsos negativos.

- Una matriz de confusión muestra un resumen más detallado de las clasificaciones correctas o incorrectas para cada clase.

Matriz de confusión

Predicción		Condición positiva	Condición negativa
	Condición positiva	Verdaderos positivos	Falsos positivos (Error tipo I)
	Condición negativa	Falsos negativos (Error tipo II)	Verdaderos negativos

$$\text{sensibilidad} = \frac{\text{verdaderos positivos}}{\text{verdaderos positivos} + \text{falsos negativos}}$$

$$\text{especificidad} = \frac{\text{verdaderos negativos}}{\text{verdaderos negativos} + \text{falsos positivos}}$$

$$\text{precisión} = \frac{\text{verdaderos positivos}}{\text{verdaderos positivos} + \text{falsos positivos}}$$

$$\text{F1 score} = 2 \cdot \frac{1}{\frac{1}{\text{sensibilidad}} + \frac{1}{\text{precisión}}} = 2 \cdot \frac{\text{precisión} \cdot \text{sensibilidad}}{\text{precisión} + \text{sensibilidad}}$$

Precisión

- Objetivo de negocio: Lo que predigamos como falso o verdadero, más vale que lo sea.
- ¿Qué fracción clasificada por el modelo están en la clase a la que pertenece?
- Cuando el modelo dice que la observación pertenece a la clase, que tan frecuentemente le atina.
- La proporción de observaciones clasificadas como C correctamente de todas las que se clasificaron como C
- Mide la capacidad del sistema de rechazar aquellas observaciones no relevantes en el conjunto clasificado
- La precisión se ve afectada por la cantidad de falsos positivos (el sistema clasificó una observación erróneamente)

Sensibilidad (Recall)

- Objetivo de negocio: Queremos reducir x en un tanto por ciento.
- ¿Qué fracción que están en la clase fueron detectadas por el modelo?
- Qué tan frecuentemente el clasificador encuentra lo que debe de encontrar.
- La proporción de observaciones clasificadas como C de todas las posibles que podían ser C.
- Mide la capacidad del sistema de encontrar todas las observaciones relevantes
- Se ve afectado por la cantidad de falsos negativos (el sistema falló al clasificar una observación relevante)

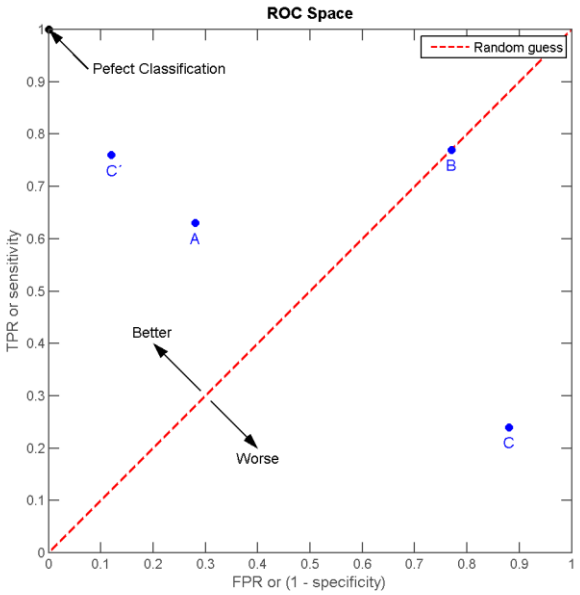
F1 Score

- Se usa en conjunto con precisión y sensibilidad.
- Mide el sacrificio de sensibilidad y/o precisión uno respecto al otro.
- Promedio armónico de las métricas de precisión y sensibilidad.

Especificidad

- Pregunta de negocio: No podemos equivocarnos en x, el sistema debe de proporcionar este servicio con altos niveles de certeza.
- Conocida también como tasa de verdaderos negativos

Curva característica operativa del receptor (ROC)



A

VP=63	FP=28	91
FN=37	VN=72	109
100	100	200

VPR = 0.63
FPR = 0.28
ACC = 0.68

B

VP=77	FP=77	154
FN=23	VN=23	46
100	100	200

VPR = 0.77
FPR = 0.77
ACC = 0.50

C

VP=24	FP=88	112
FN=76	VN=12	88
100	100	200

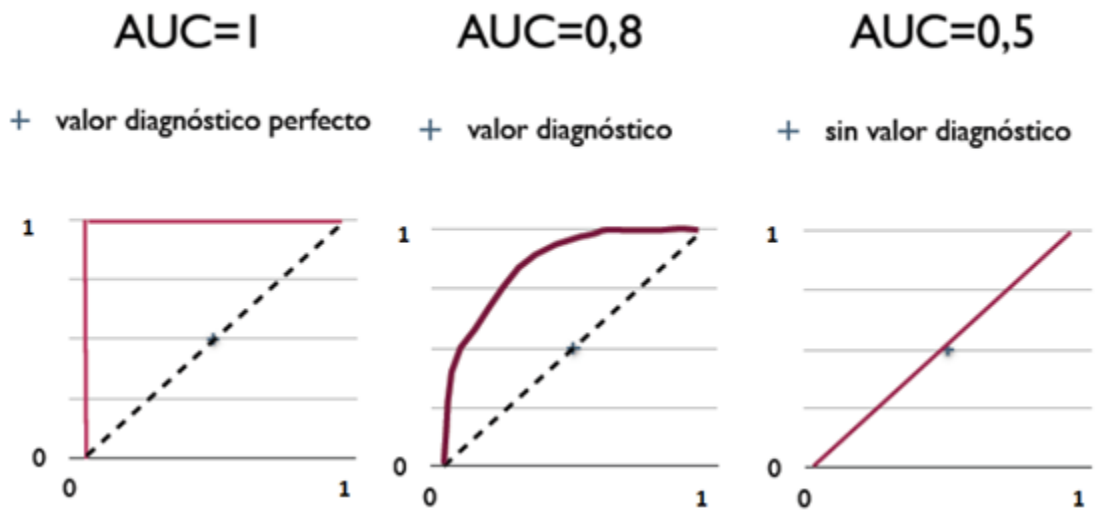
VPR = 0.24
FPR = 0.88
ACC = 0.18

C′

VP=76	FP=12	88
FN=24	VN=88	112
100	100	200

VPR = 0.76
FPR = 0.12
ACC = 0.82

Área bajo la curva



Área bajo la curva

[0.5, 0.6): Test malo.

[0.6, 0.75): Test regular.

[0.75, 0.9): Test bueno.

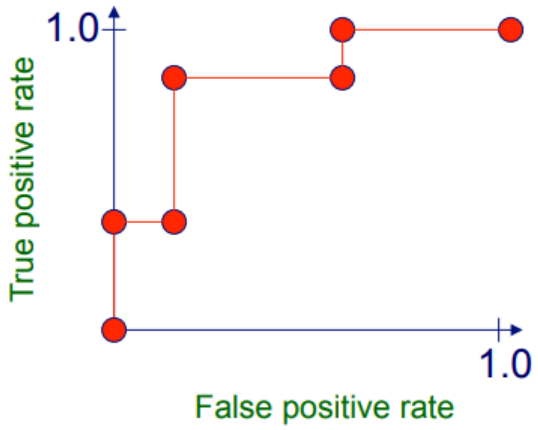
[0.9, 0.97): Test muy bueno.

[0.97, 1): Test excelente.

El área bajo la curva es la probabilidad de que una muestra positiva aleatoria tenga una puntuación más alta que una muestra negativa aleatoria.

Graficando la curva ROC

instance	confidence positive		correct class
Ex 9	.99		+
Ex 7	.98	TPR= 2/5, FPR= 0/5	+
Ex 1	.72	TPR= 2/5, FPR= 1/5	-
Ex 2	.70		+
Ex 6	.65	TPR= 4/5, FPR= 1/5	+
Ex 10	.51		-
Ex 3	.39	TPR= 4/5, FPR= 3/5	-
Ex 5	.24	TPR= 5/5, FPR= 3/5	+
Ex 4	.11		-
Ex 8	.01	TPR= 5/5, FPR= 5/5	-



Graficando la curva ROC

http://www.navan.name/roc/

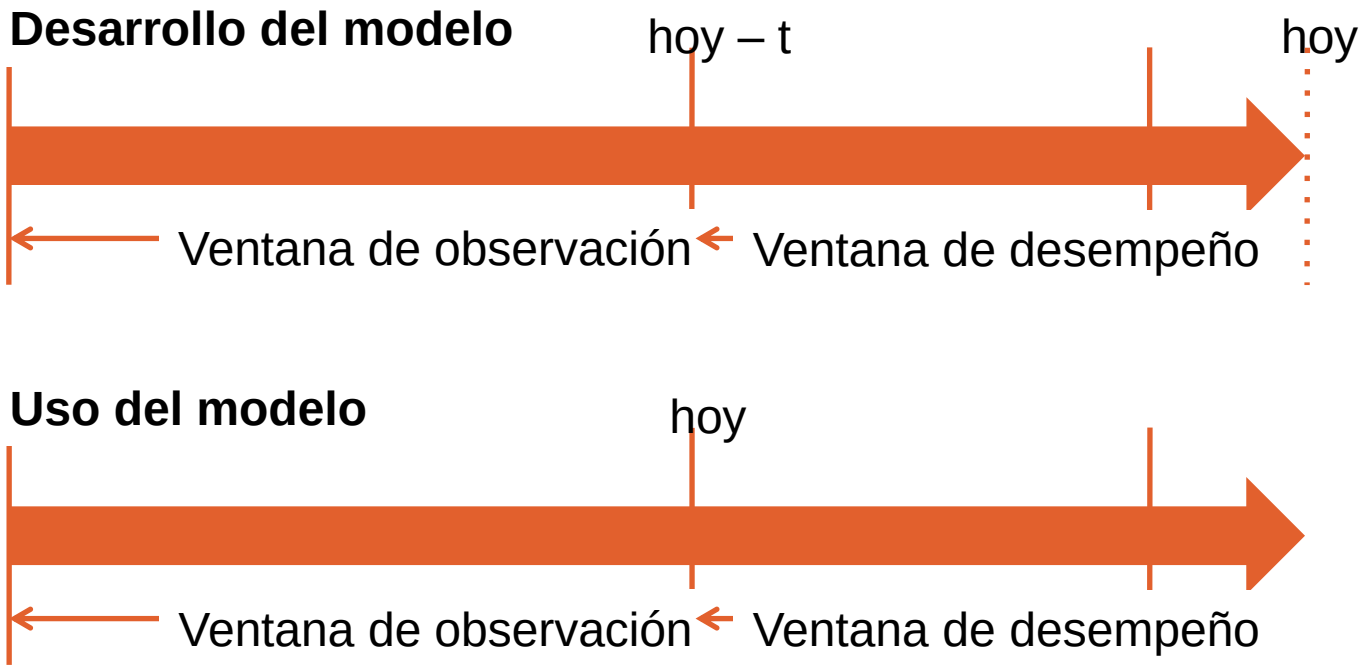
Ventanas de observación y desempeño

- Para realizar predicciones debemos:
  - Identificar a la población objetivo
    - Clientes nuevos, leales, etc.
  - Determinar el tiempo que se necesita para generar la variable a predecir
    - Clientes que abandonarán en los siguientes 6 meses, requeriremos un periodo de observación de 6 meses.
  - Eliminar los datos que no deban incluirse en la generación del modelo
    - Clientes que antes del cumplimiento del periodo de observación pertenezcan a la categoría que queremos predecir

Ventanas de observación y desempeño

- Para realizar predicciones debemos:
  - Determinar si se observará un periodo o un punto fijo en el tiempo
    - Promedio de los últimos 6 meses o valor de hace 6 meses de una variable en particular
  - Definir si dependiendo de las características de los datos se tomarán o no múltiples puntos de observación en el tiempo.
    - 4 submuestras con distintos puntos de observación
  - Determinar si los datos se generaron en momentos económicos similares
    - No usar datos de 2008 o 2009 para predecir comportamientos de pago de hipotecas

Ventanas de observación y desempeño



Créditos

- Parte de este material está basado en cursos de estadística, aprendizaje de máquina y aprendizaje estadístico del MIT, Stanford y Caltech