Regresión

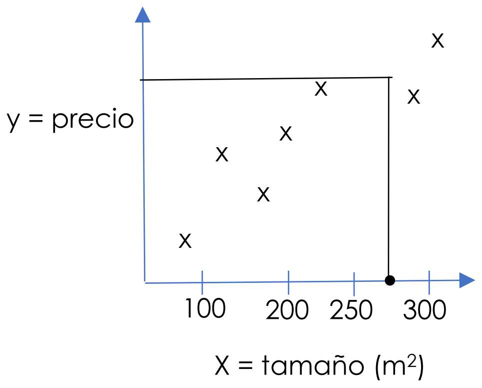
Dra. Consuelo Varinia García Mendoza

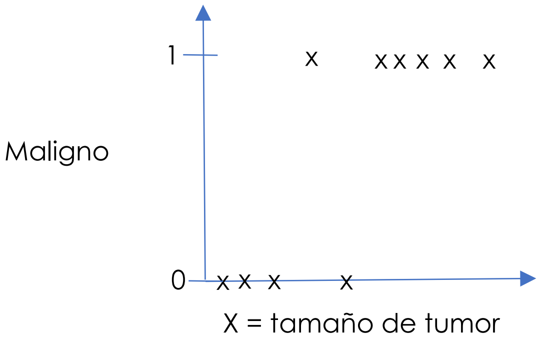
Regresión y correlación

- En la práctica a menudo se requiere resolver problemas que implican conjuntos de variables que están correlacionadas.
  - El rendimiento del combustible de un motor está relacionado con su volumen
  - ¿Qué variables están relacionadas con el precio de una casa?
- Es de interés un método de pronóstico
  - El rendimiento de cualquier motor dado su volumen
  - El precio de cualquier casa dado tu tamaño

Aprendizaje supervisado

- En el aprendizaje supervisado se tiene un conjunto de datos *x* y a partir de ellos se intenta predecir un conjunto de valores *y*
- El objetivo es crear un modelo que aprenda de los datos *x* y los relacione con los valores de *y*
- Esta relación (mapping) entre los datos de x y los valores y se puede determinar mediante una función *h(x)*
- Los valores a predecir pueden ser continuos o discretos





Relación no determinista

Ejemplos:

- Autos con motores del mismo volumen pueden tener distinto rendimiento de combustible
- Casas con la misma superficie de construcción distintos precios

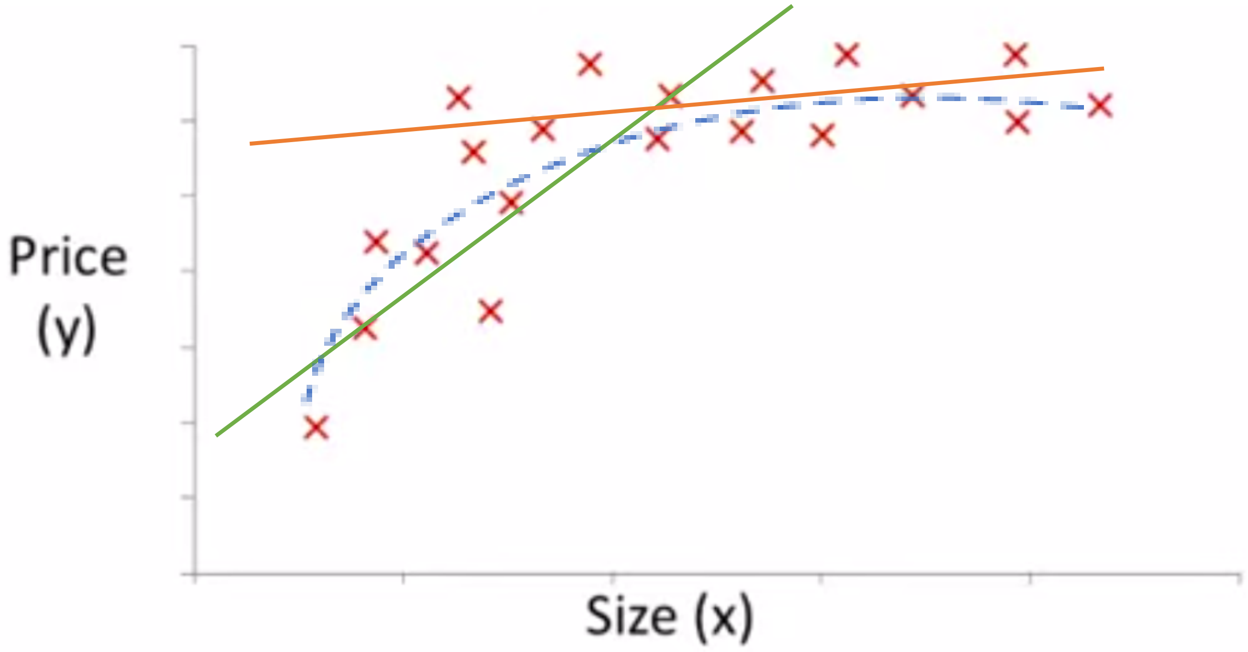
Componente aleatorio relacionado no otras variables independientes o incluso elementos desconocidos.

- A pesar de que no se pude hacer un pronóstico exacto si es posible hacer un pronóstico estimado o ajustado

Regresión

- Regresión lineal
- Regresión no lineal

H → y



Pronóstico estimado o ajustado $\hat{y}$

Relación lineal

$$\hat{y} = b_0 + b_1 x$$

donde

b_0 : intersección

b_1 : pendiente

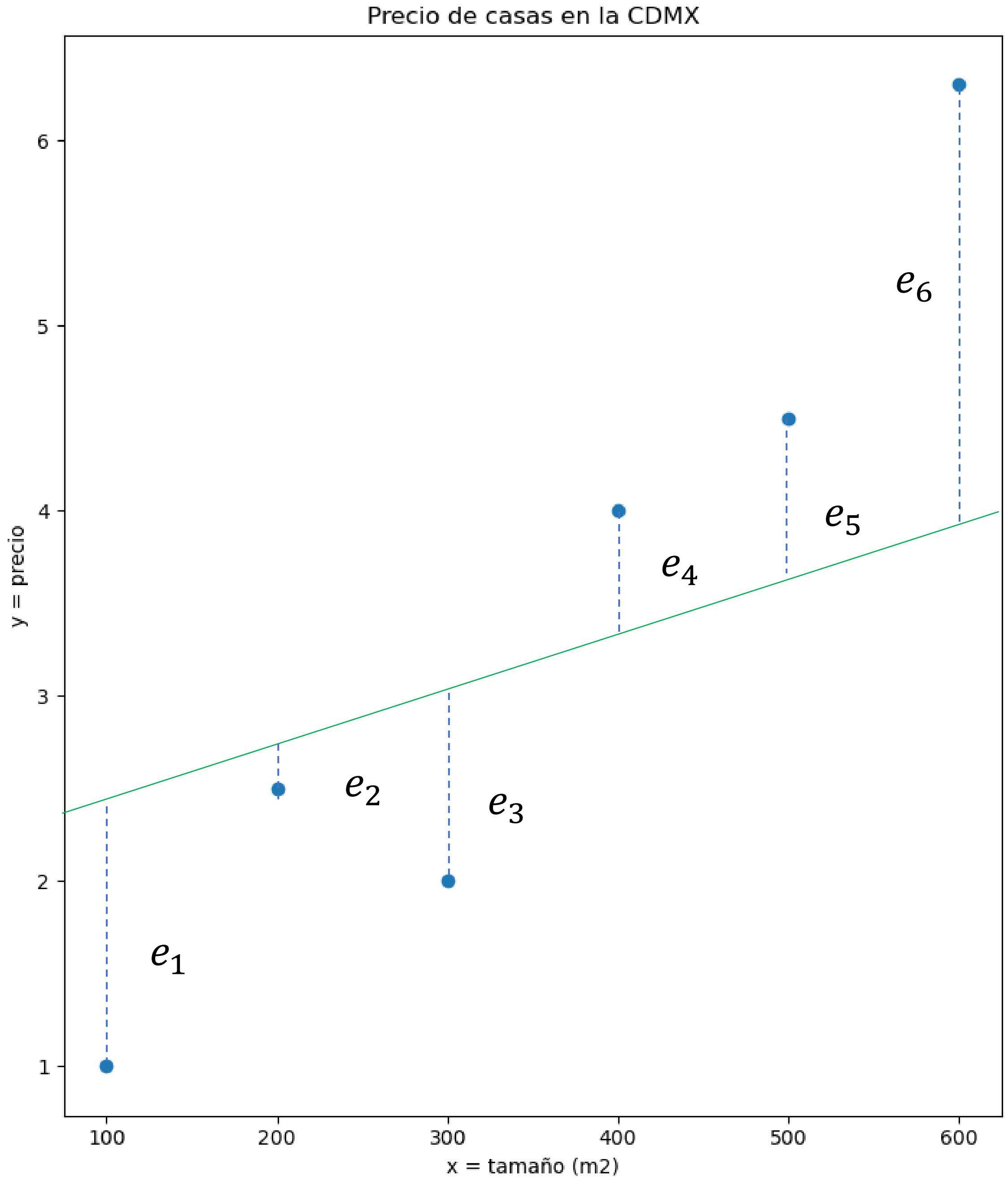
x : variable independiente o regresor (tamaño de la casa, volumen del motor)

$\hat{y}$: variable dependiente o respuesta estimada (precio, rendimiento del combustible)

Error de estimación

Cada recta de regresión que elijamos para modelar los datos tendrá asociada una suma del error respecto a la recta de regresión estimada

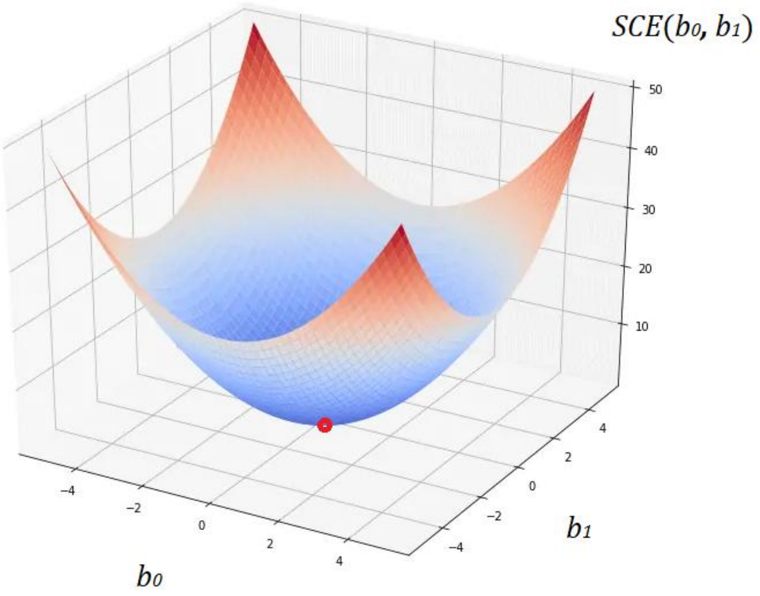
$$\sum_{I=1}^{n} |y_i - \widehat{y}_i|$$



Función de perdida

Suma de los cuadrados del error respecto a la recta de regresión estimada

$$\mathrm{SCE}(b_{0,}\, b_1) = \sum_{I=1}^{n} {e_i}^2 = \sum_{I=1}^{n} (y_i - \widehat{y_i})^2$$



- **Problema de optimización**

$$\frac{\partial \mathrm{SCE}(b_{0,}\, b_1)}{\partial b_0} = 0$$

$$\frac{\partial \mathrm{SCE}(b_{0,}\, b_1)}{\partial b_1} = 0$$

Mínimos cuadrados ordinarios (OLS)

- Método analítico

$$\frac{\partial \mathrm{SCE}\left(b_{0,}\, b_1\right)}{\partial b_0} = \frac{\partial \sum_{I=1}^{n}(y_i - b_0 - b_1 x_i)^2}{\partial b_0} = 0$$

$$\frac{\partial \mathrm{SCE}\left(b_{0,}\, b_1\right)}{\partial b_1} = \frac{\partial \sum_{I=1}^{n}(y_i - b_0 - b_1 x_i)^2}{\partial b_1} = 0$$

- $b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n}$

- $b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} {x_i}^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$

Mínimos cuadrados ordinarios (OLS)

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n}$$

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n \sum_{i=1}^{n} {x_i}^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\rightarrow$$

$$\hat{y} = b_0 + b_1 x$$

Desventajas de OLS

Con este método no hay aprendizaje, no se aprende del error

Tiene dificultades cuando las variables independientes presentan multicolinealidad entre ellas, esto es que están correlacionadas

Las variables independientes (características) deben tener correlación con la salida (target) pero no entre ellas.

Desventajas de OLS

Otra desventaja de este método es que su costo computacional el cuál aumenta considerablemente al aplicarlo a conjuntos de datos grandes con muchas instancias y variables (características)

- $\hat{y} = b_0 + b_1 x$

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_n x_n$$

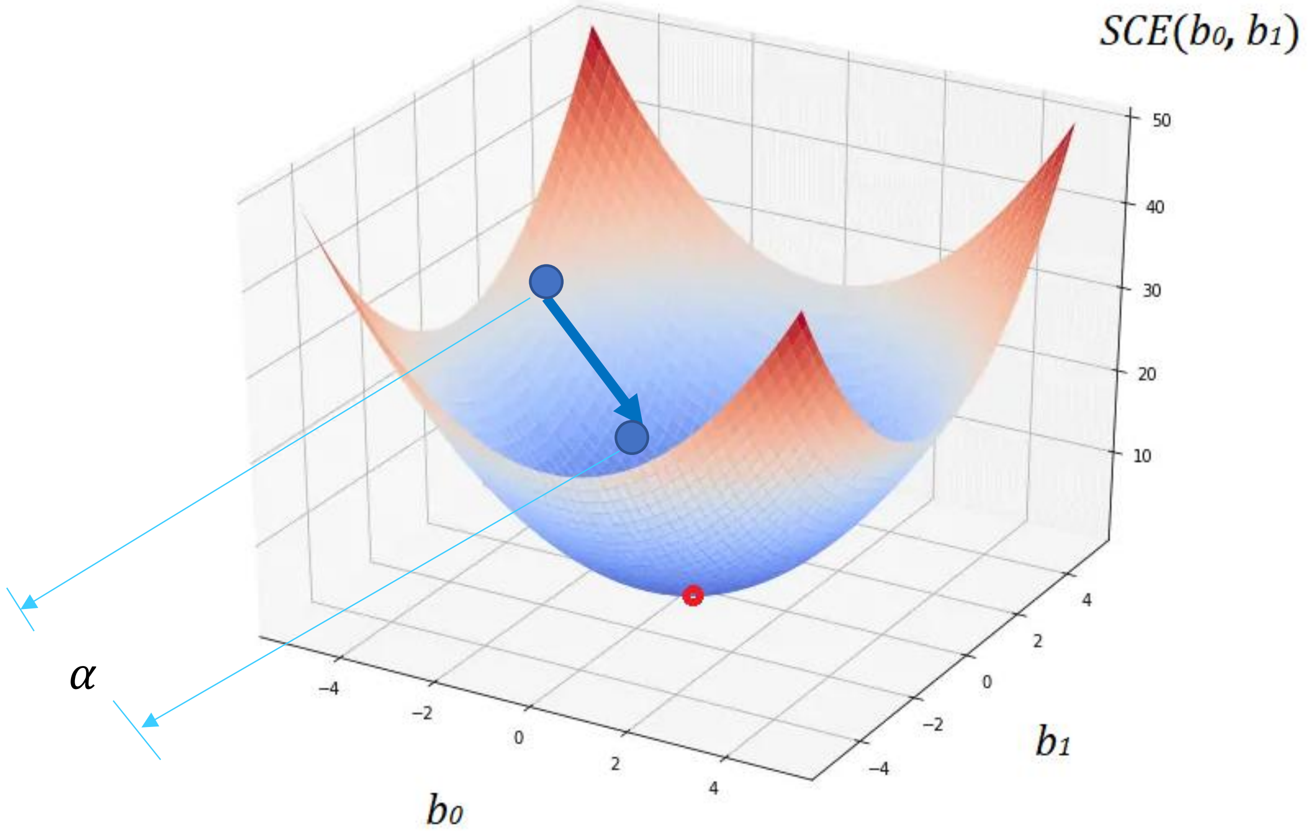
$$\frac{\partial \mathrm{SCE}(b_0, b_1)}{\partial b_0} = \frac{\partial \sum_{I=1}^{n} (y_i - b_0 - b_1 x)^2}{\partial b_0} = 0$$

$$\frac{\partial \mathrm{SCE}(b_0, b_1)}{\partial b_1} = \frac{\partial \sum_{I=1}^{n} (y_i - b_0 - b_1 x)^2}{\partial b_1} = 0$$

- Requiere calcular $n + 1$ derivadas
- Resolver un sistema de $n + 1$ ecuaciones
- Hacer sumarias de m muestras, ejemplos o instancias

Gradiente descendiente por lotes (BGD)

- Tamaño de paso = $\alpha \in (0,1)$
- Dirección -> derivada
- No. de pasos (iteraciones)



Limitaciones del algoritmo de BGD

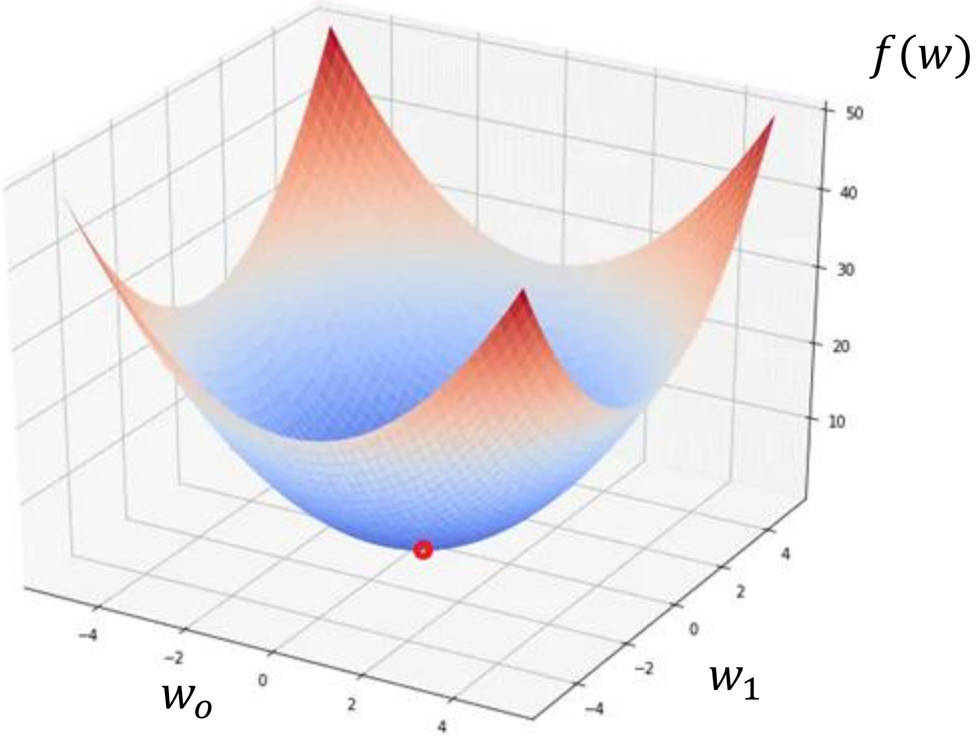
- Permite ajustar los pesos de forma iterativa
- En cada iteración se calculan los pesos de cada característica y del error
- Considera todos los ejemplos en su cálculo
- Para conjuntos de datos grandes esto puede ser muy costoso y BGD se vuelve lento

$$w_k = w_k - 2\alpha \sum_{r=0}^{e-1} \left(w_k x_{r,k} - y_r\right) \cdot x_{r,k}$$

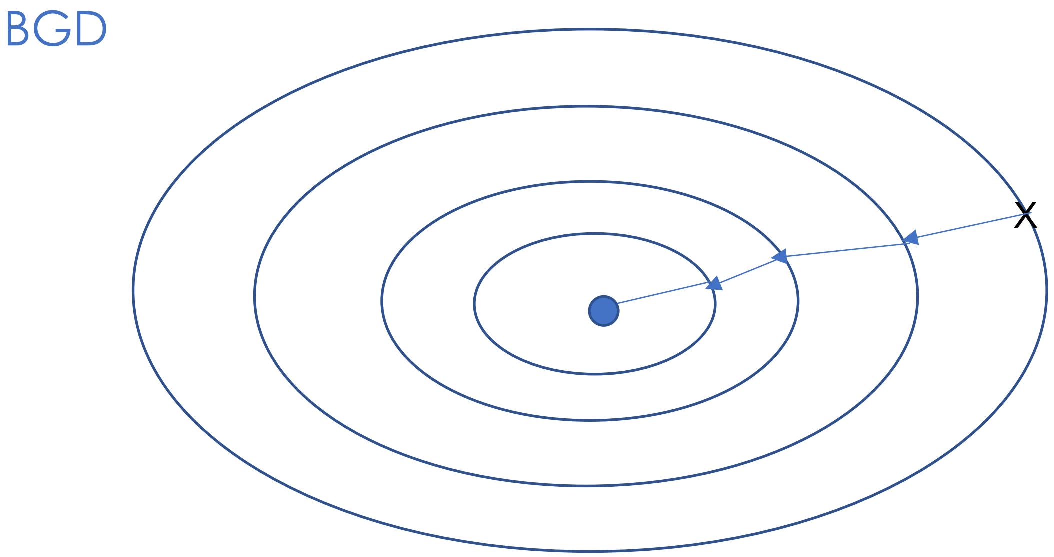
Gradiente descendente estocástico (SGD)

- Existe una variante del algoritmo BGD que considera sólo un ejemplo para el ajuste de pesos en cada iteración. Eligiendo de manera aleatoria este ejemplo es decir se establece r de manera aleatoria en cada iteración
- Con este único ejemplo se ajustan los pesos y se intenta una aproximación al óptimo
- Este algoritmo reduce el número de operaciones que se realizan en cada iteración por lo que puede manejar grandes cantidades de datos
- Sin embargo, se debe considerar que no siempre se encontrarán los pesos óptimos, pero si una buena aproximación

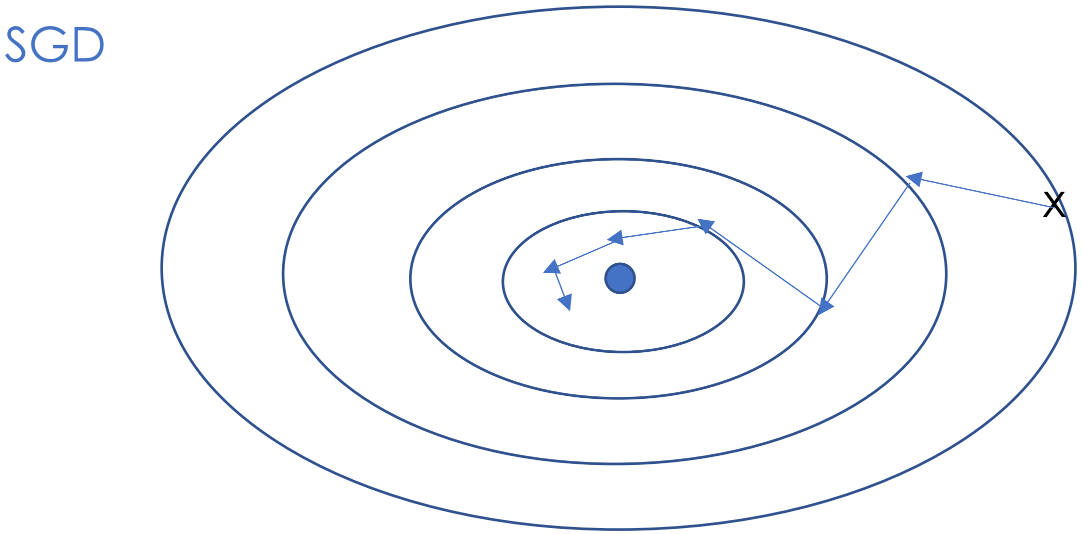
$$w_k = w_k - 2\alpha\left(w_k x_{r,k} - y_r\right) \cdot x_{r,k}$$



$f(w)$
50
40
30
20
10
-4
-2
0
2
4
w_o
4
2
0
-2
-4
w_1



BGD
X



SGD
X

Implementación de BGD o SGD

- Funciones de perdida
  - Suma de errores al cuadrado

$$f(w_k) = \sum_{r=0}^{e-1}\left(w_k x_{r,k} - y_r\right)^2$$

  - Error cuadrático medio

$$f(w_k) = \frac{1}{e}\sum_{r=0}^{e-1}\left(w_k x_{r,k} - y_r\right)^2$$