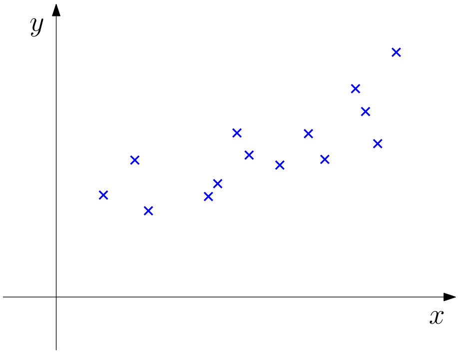
Regresión lineal

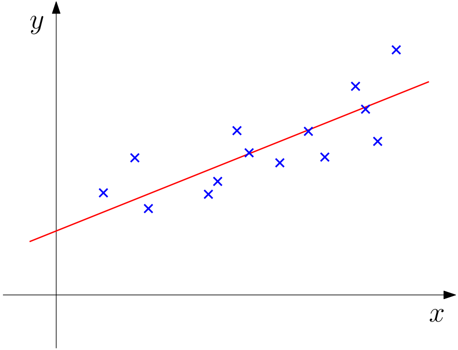
Fernando Lozano

Universidad de los Andes

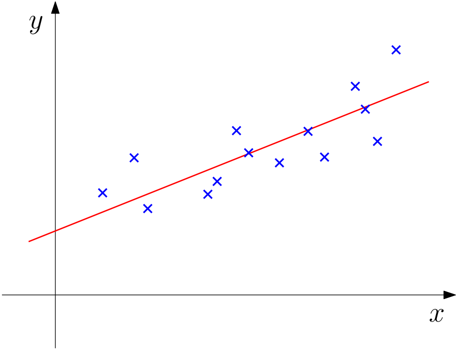
5 de agosto de 2015



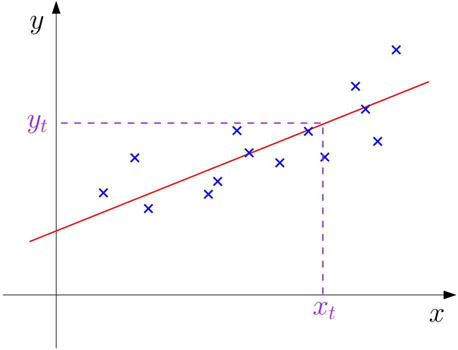
- n datos $\{x_i, y_i\}_{i=1}^{n}$.



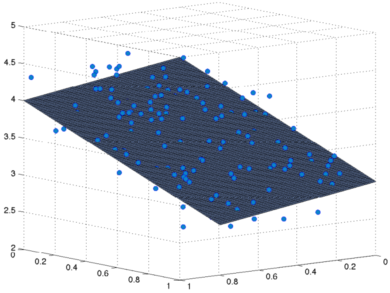
- n datos $\{x_i, y_i\}_{i=1}^n$.
- Dependencia aproximadamente lineal (afín) entre x y y.



- n datos $\{x_i, y_i\}_{i=1}^{n}$.
- Dependencia aproximadamente lineal (afín) entre x y y.
- Queremos modelar dependencia: $y = wx + b$

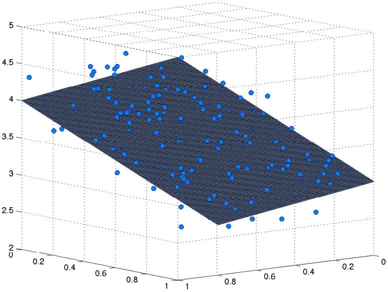


- n datos $\{x_i, y_i\}_{i=1}^{n}$.
- Dependencia aproximadamente lineal (afín) entre x y y.
- Queremos modelar dependencia: $y = wx + b$
- Usar modelo para conocer respuesta y_t para valores nuevos de x_t



- Similarmente en 2 dimensiones:

$$y = w_1 x_1 + w_2 x_2 + b$$

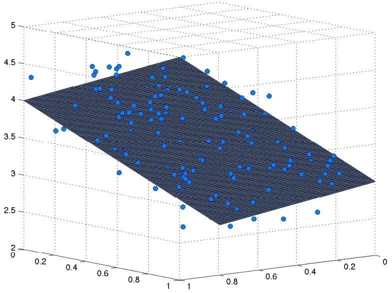


- Similarmente en 2 dimensiones:

$$y = w_1 x_1 + w_2 x_2 + b$$

- En general:

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b$$



- Similarmente en 2 dimensiones:

$$y = w_1 x_1 + w_2 x_2 + b$$

- En general:

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b = \mathbf{w}^T \mathbf{x} + b$$

- Con la convención $x_0 = 1$, $w_0 = b$, tenemos $y = \mathbf{w}^T\mathbf{x}$

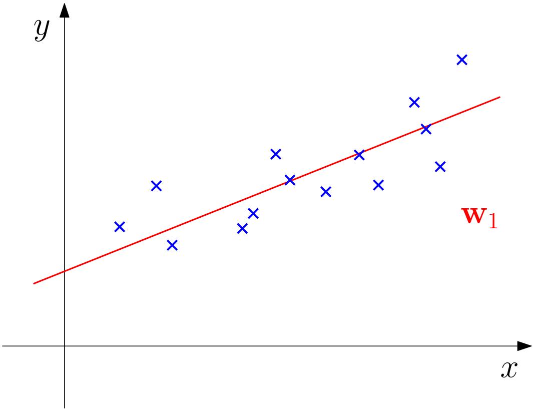
- Con la convención $x_0 = 1$, $w_0 = b$, tenemos $y = \mathbf{w}^T \mathbf{x}$
- Queremos encontrar $\mathbf{w}$ que modele apropiadamente la dependencia entre $\mathbf{x}$ y y:

- Con la convención $x_0 = 1$, $w_0 = b$, tenemos $y = \mathbf{w}^T \mathbf{x}$
- Queremos encontrar $\mathbf{w}$ que modele apropiadamente la dependencia entre $\mathbf{x}$ y y:
  - Qué es un buen modelo?

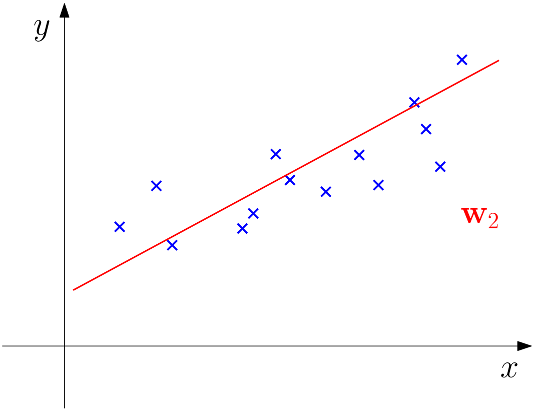
- Con la convención $x_0 = 1$, $w_0 = b$, tenemos $y = \mathbf{w}^T \mathbf{x}$
- Queremos encontrar $\mathbf{w}$ que modele apropiadamente la dependencia entre $\mathbf{x}$ y y:
  - Qué es un buen modelo?
  - Intuitivamente $\mathbf{w}$ debe corresponder a un modelo que se ajuste bien a los datos $\{x_i, y_i\}_{i=1}^{n}$

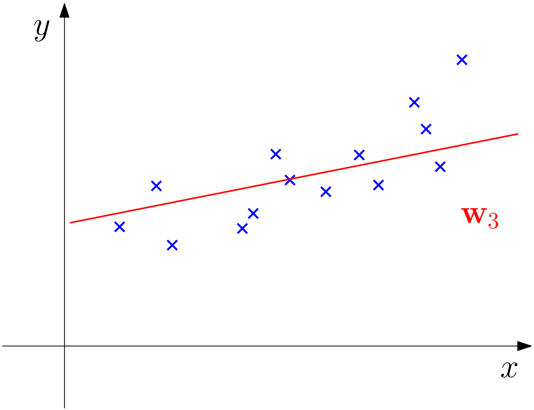
- Con la convención $x_0 = 1$, $w_0 = b$, tenemos $y = \mathbf{w}^T \mathbf{x}$
- Queremos encontrar $\mathbf{w}$ que modele apropiadamente la dependencia entre $\mathbf{x}$ y y:
  - Qué es un buen modelo?
  - Intuitivamente $\mathbf{w}$ debe corresponder a un modelo que se ajuste bien a los datos $\{x_i, y_i\}_{i=1}^n \Rightarrow$ Optimización.

- Con la convención $x_0 = 1$, $w_0 = b$, tenemos $y = \mathbf{w}^T \mathbf{x}$
- Queremos encontrar $\mathbf{w}$ que modele apropiadamente la dependencia entre $\mathbf{x}$ y y:
  - Qué es un buen modelo?
  - Intuitivamente $\mathbf{w}$ debe corresponder a un modelo que se ajuste bien a los datos $\{x_i, y_i\}_{i=1}^n \Rightarrow$ Optimización.

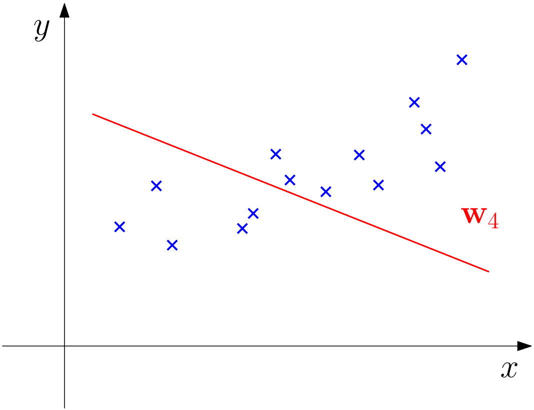


y
$\mathbf{w}_1$
x

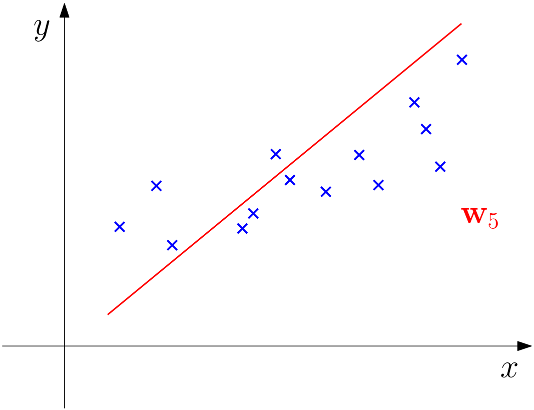

y
x
$\mathbf{w}_2$



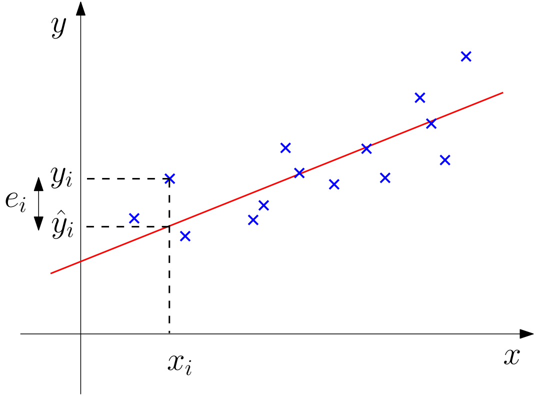
y
x
$\mathbf{w}_3$



y
x
$\mathbf{w}_4$

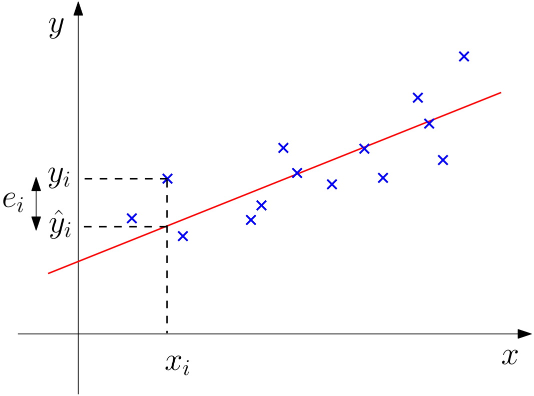


y
x
$\mathbf{w}_5$



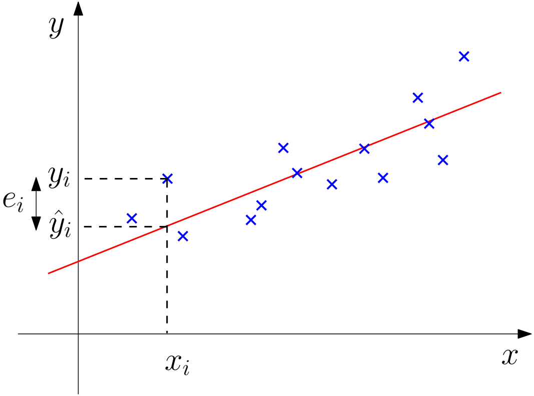
- Función de error:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 =$$



- Función de error:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i))^2$$



- Función de error:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i))^2$$

- Problema de optimización:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} E(\mathbf{w})$$

Examinando la función de error

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2$$

Examinando la función de error

$$
\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2 \\
&= \frac{1}{2}\sum_{i=1}^{n}((\mathbf{w}^T\mathbf{x}_i)^2 - 2y_i\mathbf{w}^T\mathbf{x}_i + y_i^2)
\end{aligned}
$$

Examinando la función de error

$$
\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2 \\
&= \frac{1}{2}\sum_{i=1}^{n}((\mathbf{w}^T\mathbf{x}_i)^2 - 2y_i\mathbf{w}^T\mathbf{x}_i + y_i^2) \\
&= \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{w} - 2y_i\mathbf{w}^T\mathbf{x}_i + y_i^2)
\end{aligned}
$$

Examinando la función de error

$$
\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2 \\
&= \frac{1}{2}\sum_{i=1}^{n}((\mathbf{w}^T\mathbf{x}_i)^2 - 2y_i\mathbf{w}^T\mathbf{x}_i + y_i^2) \\
&= \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{w} - 2y_i\mathbf{w}^T\mathbf{x}_i + y_i^2) \\
&= \frac{1}{2}\mathbf{w}^T\mathbf{H}\mathbf{w} - \mathbf{b}^T\mathbf{w} + c
\end{aligned}
$$

Examinando la función de error

$$
\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2 \\
&= \frac{1}{2}\sum_{i=1}^{n}((\mathbf{w}^T\mathbf{x}_i)^2 - 2y_i\mathbf{w}^T\mathbf{x}_i + y_i^2) \\
&= \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{w} - 2y_i\mathbf{w}^T\mathbf{x}_i + y_i^2) \\
&= \frac{1}{2}\mathbf{w}^T\mathbf{H}\mathbf{w} - \mathbf{b}^T\mathbf{w} + c
\end{aligned}
$$

donde:

$$\mathbf{H} = \textstyle\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T, \quad \mathbf{b} = \sum_{i=1}^{n}\mathbf{x}_i y_i, \quad c = \frac{1}{2}\sum_{i=1}^{n} y_i$$

Derivadas de primer y segundo orden

Gradiente:

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

Derivadas de primer y segundo orden

Gradiente:

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

Hessiana:

$$\nabla^2_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix}$$

- $E(\mathbf{w})$ es una función cuadrática de $\mathbf{w}$.

- $E(\mathbf{w})$ es una función cuadrática de $\mathbf{w}$.
- Derivando:

- $E(\mathbf{w})$ es una función cuadrática de $\mathbf{w}$.
- Derivando:
  - $\nabla_{\mathbf{w}} E =$

- $E(\mathbf{w})$ es una función cuadrática de $\mathbf{w}$.
- Derivando:
  - $\nabla_{\mathbf{w}} E = \mathbf{H}\mathbf{w} - \mathbf{b}$.

- $E(\mathbf{w})$ es una función cuadrática de $\mathbf{w}$.
- Derivando:
  - $\nabla_{\mathbf{w}} E = \mathbf{H}\mathbf{w} - \mathbf{b}$.
  - $\nabla^2_{\mathbf{w}} E =$

- $E(\mathbf{w})$ es una función cuadrática de $\mathbf{w}$.
- Derivando:
  - $\nabla_{\mathbf{w}} E = \mathbf{H}\mathbf{w} - \mathbf{b}$.
  - $\nabla^2_{\mathbf{w}} E = \mathbf{H}$

- $E(\mathbf{w})$ es una función cuadrática de $\mathbf{w}$.
- Derivando:
  - $\nabla_{\mathbf{w}} E = \mathbf{H}\mathbf{w} - \mathbf{b}$.
  - $\nabla^2_{\mathbf{w}} E = \mathbf{H}$
- Note que para cualquier vector $\mathbf{z} \in \mathbb{R}^{d+1}$,

$$\mathbf{z}^T \mathbf{H}\mathbf{z}$$

- $E(\mathbf{w})$ es una función cuadrática de $\mathbf{w}$.
- Derivando:
    - $\nabla_{\mathbf{w}} E = \mathbf{H}\mathbf{w} - \mathbf{b}$.
    - $\nabla^2_{\mathbf{w}} E = \mathbf{H}$
- Note que para cualquier vector $\mathbf{z} \in \mathbb{R}^{d+1}$,

$$\mathbf{z}^T \mathbf{H} \mathbf{z} = \mathbf{z}^T \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{z}$$

- $E(\mathbf{w})$ es una función cuadrática de $\mathbf{w}$.
- Derivando:
  - $\nabla_{\mathbf{w}} E = \mathbf{H}\mathbf{w} - \mathbf{b}$.
  - $\nabla^2_{\mathbf{w}} E = \mathbf{H}$
- Note que para cualquier vector $\mathbf{z} \in \mathbb{R}^{d+1}$,

$$\mathbf{z}^T \mathbf{H} \mathbf{z} = \mathbf{z}^T \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{z} = \sum_{i=1}^{n} (\mathbf{z}^T \mathbf{x}_i)(\mathbf{x}_i^T \mathbf{z}) = \sum_{i=1}^{n} (\mathbf{z}^T \mathbf{x}_i)^2 \geq 0$$

  es decir $\mathbf{H}$ es

- $E(\mathbf{w})$ es una función cuadrática de $\mathbf{w}$.
- Derivando:
  - $\nabla_{\mathbf{w}} E = \mathbf{H}\mathbf{w} - \mathbf{b}$.
  - $\nabla^2_{\mathbf{w}} E = \mathbf{H}$
- Note que para cualquier vector $\mathbf{z} \in \mathbb{R}^{d+1}$,

$$\mathbf{z}^T \mathbf{H} \mathbf{z} = \mathbf{z}^T \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{z} = \sum_{i=1}^{n} (\mathbf{z}^T \mathbf{x}_i)(\mathbf{x}_i^T \mathbf{z}) = \sum_{i=1}^{n} (\mathbf{z}^T \mathbf{x}_i)^2 \geq 0$$

  es decir $\mathbf{H}$ es positiva semidefinida, luego $E(\mathbf{w})$ es

- $E(\mathbf{w})$ es una función cuadrática de $\mathbf{w}$.
- Derivando:
  - $\nabla_{\mathbf{w}} E = \mathbf{H}\mathbf{w} - \mathbf{b}$.
  - $\nabla^2_{\mathbf{w}} E = \mathbf{H}$
- Note que para cualquier vector $\mathbf{z} \in \mathbb{R}^{d+1}$,

$$\mathbf{z}^T \mathbf{H} \mathbf{z} = \mathbf{z}^T \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{z} = \sum_{i=1}^{n} (\mathbf{z}^T \mathbf{x}_i)(\mathbf{x}_i^T \mathbf{z}) = \sum_{i=1}^{n} (\mathbf{z}^T \mathbf{x}_i)^2 \geq 0$$

es decir $\mathbf{H}$ es positiva semidefinida, luego $E(\mathbf{w})$ es convexa.

- $E(\mathbf{w})$ es una función cuadrática de $\mathbf{w}$.
- Derivando:
  - $\nabla_{\mathbf{w}} E = \mathbf{H}\mathbf{w} - \mathbf{b}$.
  - $\nabla^2_{\mathbf{w}} E = \mathbf{H}$
- Note que para cualquier vector $\mathbf{z} \in \mathbb{R}^{d+1}$,

$$\mathbf{z}^T \mathbf{H} \mathbf{z} = \mathbf{z}^T \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{z} = \sum_{i=1}^{n} (\mathbf{z}^T \mathbf{x}_i)(\mathbf{x}_i^T \mathbf{z}) = \sum_{i=1}^{n} (\mathbf{z}^T \mathbf{x}_i)^2 \geq 0$$

  es decir $\mathbf{H}$ es positiva semidefinida, luego $E(\mathbf{w})$ es convexa.
- Si $\mathbf{H} > 0$, $E(\mathbf{w})$ tiene un único mínimo global:

- $E(\mathbf{w})$ es una función cuadrática de $\mathbf{w}$.
- Derivando:
  - $\nabla_{\mathbf{w}} E = \mathbf{H}\mathbf{w} - \mathbf{b}$.
  - $\nabla^2_{\mathbf{w}} E = \mathbf{H}$
- Note que para cualquier vector $\mathbf{z} \in \mathbb{R}^{d+1}$,

$$\mathbf{z}^T \mathbf{H} \mathbf{z} = \mathbf{z}^T \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{z} = \sum_{i=1}^{n} (\mathbf{z}^T \mathbf{x}_i)(\mathbf{x}_i^T \mathbf{z}) = \sum_{i=1}^{n} (\mathbf{z}^T \mathbf{x}_i)^2 \geq 0$$

  es decir $\mathbf{H}$ es positiva semidefinida, luego $E(\mathbf{w})$ es convexa.
- Si $\mathbf{H} > 0$, $E(\mathbf{w})$ tiene un único mínimo global:

$$\nabla_{\mathbf{w}} E =$$

- $E(\mathbf{w})$ es una función cuadrática de $\mathbf{w}$.
- Derivando:
  - $\nabla_{\mathbf{w}} E = \mathbf{H}\mathbf{w} - \mathbf{b}$.
  - $\nabla^2_{\mathbf{w}} E = \mathbf{H}$
- Note que para cualquier vector $\mathbf{z} \in \mathbb{R}^{d+1}$,

$$\mathbf{z}^T \mathbf{H} \mathbf{z} = \mathbf{z}^T \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{z} = \sum_{i=1}^{n} (\mathbf{z}^T \mathbf{x}_i)(\mathbf{x}_i^T \mathbf{z}) = \sum_{i=1}^{n} (\mathbf{z}^T \mathbf{x}_i)^2 \geq 0$$

  es decir $\mathbf{H}$ es positiva semidefinida, luego $E(\mathbf{w})$ es convexa.
- Si $\mathbf{H} > 0$, $E(\mathbf{w})$ tiene un único mínimo global:

$$\nabla_{\mathbf{w}} E = 0$$

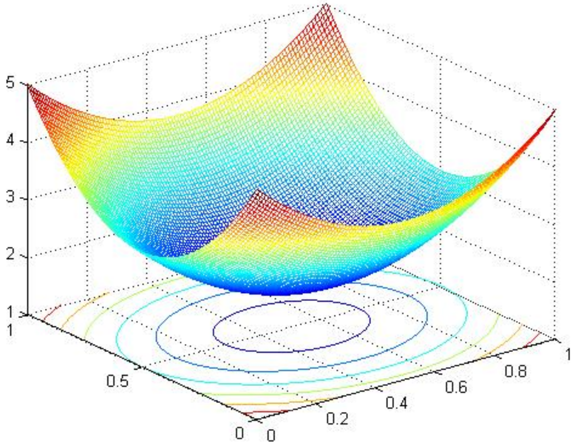
- $E(\mathbf{w})$ es una función cuadrática de $\mathbf{w}$.
- Derivando:
  - $\nabla_{\mathbf{w}} E = \mathbf{H}\mathbf{w} - \mathbf{b}$.
  - $\nabla^2_{\mathbf{w}} E = \mathbf{H}$
- Note que para cualquier vector $\mathbf{z} \in \mathbb{R}^{d+1}$,

$$\mathbf{z}^T\mathbf{H}\mathbf{z} = \mathbf{z}^T\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{z} = \sum_{i=1}^{n}(\mathbf{z}^T\mathbf{x}_i)(\mathbf{x}_i^T\mathbf{z}) = \sum_{i=1}^{n}(\mathbf{z}^T\mathbf{x}_i)^2 \geq 0$$

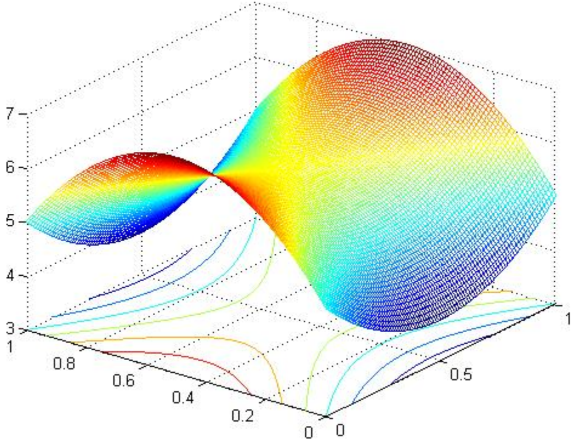
  es decir $\mathbf{H}$ es positiva semidefinida, luego $E(\mathbf{w})$ es convexa.
- Si $\mathbf{H} > 0$, $E(\mathbf{w})$ tiene un único mínimo global:

$$\nabla_{\mathbf{w}} E = 0 \Rightarrow \mathbf{w}^* = \mathbf{H}^{-1}\mathbf{b}$$

Función cuadrática convexa



Función cuadrática no convexa



Algoritmo iterativo de descenso

Inicialice $\mathbf{w}_0$

Algoritmo iterativo de descenso

Inciale $\mathbf{w}_0$
repeat

Algoritmo iterativo de descenso

Inicialice $\mathbf{w}_0$
repeat
    Escoja dirección $\mathbf{d}$

Algoritmo iterativo de descenso

Inicialice $\mathbf{w}_0$

repeat

Escoja dirección $\mathbf{d}$

$\mathbf{w}_{k+1} = \mathbf{w}_k + \eta_k \mathbf{d}$

Algoritmo iterativo de descenso

Inicialice $\mathbf{w}_0$
repeat
    Escoja dirección $\mathbf{d}$
    $\mathbf{w}_{k+1} = \mathbf{w}_k + \eta_k \mathbf{d}$
until Condición de terminación.

Algoritmo iterativo de descenso

Inicialice $\mathbf{w}_0$

repeat

  Escoja dirección $\mathbf{d}$

  $\mathbf{w}_{k+1} = \mathbf{w}_k + \eta_k \mathbf{d}$

until Condición de terminación.

Escogencia de η_k

Escogencia de η_k

- Constante

Escogencia de η_k

- Constante (tasa de aprendizaje)

Escogencia de η_k

- Constante (tasa de aprendizaje)
- Idealmente $\eta_k = \arg\min_{\eta} f\left(\mathbf{w}_k + \eta\mathbf{d}\right)$

Escogencia de η_k

- Constante (tasa de aprendizaje)
- Idealmente $\eta_k = \arg\min_{\eta} f(\mathbf{w}_k + \eta \mathbf{d})$ $\Rightarrow$ Algoritmos de búsqueda de línea.

Escogencia de η_k

- Constante (tasa de aprendizaje)
- Idealmente $\eta_k = \arg\min_\eta f(\mathbf{w}_k + \eta \mathbf{d}) \Rightarrow$ Algoritmos de búsqueda de línea.
    1. Golden Search, Fibonacci.
    2. Newton, ajuste cúbico.
    3. Backtracking.
    4. $\vdots$

Escogencia de η_k

- Constante (tasa de aprendizaje)
- Idealmente $\eta_k = \arg\min_\eta f(\mathbf{w}_k + \eta\mathbf{d})$ $\Rightarrow$ Algoritmos de búsqueda de línea.
  1. Golden Search, Fibonacci.
  2. Newton, ajuste cúbico.
  3. Backtracking.
  4. $\vdots$
- Variación dinámica de η

Escogencia de **d**

Escogencia de d

- Función a partir de un punto fijo **w**, a lo largo de una dirección **d**:

$$g(\eta) = E(\mathbf{w} + \eta \mathbf{d})$$

Escogencia de $\mathbf{d}$

- Función a partir de un punto fijo $\mathbf{w}$, a lo largo de una dirección $\mathbf{d}$:

$$g(\eta) = E(\mathbf{w} + \eta\mathbf{d})$$

- Derivada direccional:

Escogencia de **d**

- Función a partir de un punto fijo $\mathbf{w}$, a lo largo de una dirección $\mathbf{d}$:

$$g(\eta) = E(\mathbf{w} + \eta\mathbf{d})$$

- Derivada direccional:

$$g'(\eta) =$$

Escogencia de **d**

- Función a partir de un punto fijo $\mathbf{w}$, a lo largo de una dirección $\mathbf{d}$:

$$g(\eta) = E(\mathbf{w} + \eta\mathbf{d})$$

- Derivada direccional:

$$g'(\eta) = \mathbf{d}^T$$

Escogencia de **d**

- Función a partir de un punto fijo $\mathbf{w}$, a lo largo de una dirección $\mathbf{d}$:

$$g(\eta) = E(\mathbf{w} + \eta\mathbf{d})$$

- Derivada direccional:

$$g'(\eta) = \mathbf{d}^T \nabla_{\mathbf{w}} E(\mathbf{w} + \eta\mathbf{d})$$

Escogencia de **d**

- Función a partir de un punto fijo $\mathbf{w}$, a lo largo de una dirección $\mathbf{d}$:

$$g(\eta) = E(\mathbf{w} + \eta\mathbf{d})$$

- Derivada direccional:

$$g'(\eta) = \mathbf{d}^T \nabla_{\mathbf{w}} E(\mathbf{w} + \eta\mathbf{d}) \Rightarrow g'(0) = \mathbf{d}^T \nabla_{\mathbf{w}} E(\mathbf{w})$$

Escogencia de **d**

- Función a partir de un punto fijo $\mathbf{w}$, a lo largo de una dirección $\mathbf{d}$:

$$g(\eta) = E(\mathbf{w} + \eta\mathbf{d})$$

- Derivada direccional:

$$g'(\eta) = \mathbf{d}^T \nabla_{\mathbf{w}} E(\mathbf{w} + \eta\mathbf{d}) \Rightarrow g'(0) = \mathbf{d}^T \nabla_{\mathbf{w}} E(\mathbf{w}) = \langle \mathbf{d}, \nabla_{\mathbf{w}} E(\mathbf{w}) \rangle$$

Escogencia de **d**

- Función a partir de un punto fijo $\mathbf{w}$, a lo largo de una dirección $\mathbf{d}$:

$$g(\eta) = E(\mathbf{w} + \eta\mathbf{d})$$

- Derivada direccional:

$$g'(\eta) = \mathbf{d}^T \nabla_{\mathbf{w}} E(\mathbf{w} + \eta\mathbf{d}) \Rightarrow g'(0) = \mathbf{d}^T \nabla_{\mathbf{w}} E(\mathbf{w}) = \langle \mathbf{d}, \nabla_{\mathbf{w}} E(\mathbf{w}) \rangle$$

- Para $\|\mathbf{d}\|$ constante la derivada direccional es máximamente negativa cuando

Escogencia de **d**

- Función a partir de un punto fijo $\mathbf{w}$, a lo largo de una dirección $\mathbf{d}$:

$$g(\eta) = E(\mathbf{w} + \eta\mathbf{d})$$

- Derivada direccional:

$$g'(\eta) = \mathbf{d}^T \nabla_{\mathbf{w}} E(\mathbf{w} + \eta\mathbf{d}) \Rightarrow g'(0) = \mathbf{d}^T \nabla_{\mathbf{w}} E(\mathbf{w}) = \langle \mathbf{d}, \nabla_{\mathbf{w}} E(\mathbf{w}) \rangle$$

- Para $\|\mathbf{d}\|$ constante la derivada direccional es máximamente negativa cuando $\mathbf{d} =$

Escogencia de $\mathbf{d}$

- Función a partir de un punto fijo $\mathbf{w}$, a lo largo de una dirección $\mathbf{d}$:

$$g(\eta) = E(\mathbf{w} + \eta \mathbf{d})$$

- Derivada direccional:

$$g'(\eta) = \mathbf{d}^T \nabla_{\mathbf{w}} E(\mathbf{w} + \eta \mathbf{d}) \Rightarrow g'(0) = \mathbf{d}^T \nabla_{\mathbf{w}} E(\mathbf{w}) = \langle \mathbf{d}, \nabla_{\mathbf{w}} E(\mathbf{w}) \rangle$$

- Para $\|\mathbf{d}\|$ constante la derivada direccional es máximamente negativa cuando $\mathbf{d} = -\nabla_{\mathbf{w}} E(\mathbf{w})$

Escogencia de $\mathbf{d}$

- Función a partir de un punto fijo $\mathbf{w}$, a lo largo de una dirección $\mathbf{d}$:

$$g(\eta) = E(\mathbf{w} + \eta\mathbf{d})$$

- Derivada direccional:

$$g'(\eta) = \mathbf{d}^T \nabla_{\mathbf{w}} E(\mathbf{w} + \eta\mathbf{d}) \Rightarrow g'(0) = \mathbf{d}^T \nabla_{\mathbf{w}} E(\mathbf{w}) = \langle \mathbf{d}, \nabla_{\mathbf{w}} E(\mathbf{w}) \rangle$$

- Para $\|\mathbf{d}\|$ constante la derivada direccional es máximamente negativa cuando $\mathbf{d} = -\nabla_{\mathbf{w}} E(\mathbf{w})$
- El gradiente negativo es la dirección de máximo descenso.

Descenso de Gradiente (GD)

Inicialice $\mathbf{w}_0$

Descenso de Gradiente (GD)

Inicialice $\mathbf{w}_0$
repeat

Descenso de Gradiente (GD)

Inicialice $\mathbf{w}_0$

repeat

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta_k \nabla_{\mathbf{w}} E(\mathbf{w}_k)$$

Descenso de Gradiente (GD)

Inicialice $\mathbf{w}_0$

repeat

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta_k \nabla_{\mathbf{w}} E(\mathbf{w}_k)$$

until Condición de terminación.

Descenso de Gradiente (GD)

Inicialice $\mathbf{w}_0$

repeat

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta_k \nabla_{\mathbf{w}} E(\mathbf{w}_k)$$

until Condición de terminación.

Análisis

- Consideramos el caso en el que $\eta_k = \eta$ es constante.

Análisis

- Consideramos el caso en el que $\eta_k = \eta$ es constante.
- Consideramos la función:

$$E_c(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^T \mathbf{H} (\mathbf{w} - \mathbf{w}^*), \qquad \mathbf{H} > 0$$

Análisis

- Consideramos el caso en el que $\eta_k = \eta$ es constante.
- Consideramos la función:

$$E_c(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^T \mathbf{H}(\mathbf{w} - \mathbf{w}^*), \qquad \mathbf{H} > 0$$

- Es fácil ver que $E_c(\mathbf{w}) = E(\mathbf{w}) + \frac{1}{2}\mathbf{w}^{*T}\mathbf{H}\mathbf{w}^*$.

Análisis

- Consideramos el caso en el que $\eta_k = \eta$ es constante.
- Consideramos la función:

$$E_c(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^T \mathbf{H}(\mathbf{w} - \mathbf{w}^*), \qquad \mathbf{H} > 0$$

- Es fácil ver que $E_c(\mathbf{w}) = E(\mathbf{w}) + \frac{1}{2}\mathbf{w}^{*T}\mathbf{H}\mathbf{w}^*$.
- Iteración de Descenso de Gradiente:

$$\Delta \mathbf{w}_k = \mathbf{w}_k - \mathbf{w}_{k-1}$$

Análisis

- Consideramos el caso en el que $\eta_k = \eta$ es constante.
- Consideramos la función:

$$E_c(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^T \mathbf{H}(\mathbf{w} - \mathbf{w}^*), \qquad \mathbf{H} > 0$$

- Es fácil ver que $E_c(\mathbf{w}) = E(\mathbf{w}) + \frac{1}{2}\mathbf{w}^{*T}\mathbf{H}\mathbf{w}^*$.
- Iteración de Descenso de Gradiente:

$$\begin{aligned}\Delta\mathbf{w}_k &= \mathbf{w}_k - \mathbf{w}_{k-1} \\ &= -\eta\nabla E(\mathbf{w}_{k-1})\end{aligned}$$

Análisis

- Consideramos el caso en el que $\eta_k = \eta$ es constante.
- Consideramos la función:

$$E_c(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^T \mathbf{H}(\mathbf{w} - \mathbf{w}^*), \qquad \mathbf{H} > 0$$

- Es fácil ver que $E_c(\mathbf{w}) = E(\mathbf{w}) + \frac{1}{2}\mathbf{w}^{*T}\mathbf{H}\mathbf{w}^*$.
- Iteración de Descenso de Gradiente:

$$\begin{aligned} \Delta \mathbf{w}_k &= \mathbf{w}_k - \mathbf{w}_{k-1} \\ &= -\eta \nabla E(\mathbf{w}_{k-1}) \\ &= -\eta \mathbf{H}(\mathbf{w}_{k-1} - \mathbf{w}^*) \end{aligned}$$

El viejo truco...

El viejo truco...

- Suponga que $\mathbf{H}$ tiene valores propios λ_i y vectores propios correspondientes $\mathbf{u}_i$:

El viejo truco...

- Suponga que $\mathbf{H}$ tiene valores propios λ_i y vectores propios correspondientes $\mathbf{u}_i$:

$$\mathbf{w} - \mathbf{w}^* = \sum_i \alpha_i \mathbf{u}_i$$

El viejo truco...

- Suponga que $\mathbf{H}$ tiene valores propios λ_i y vectores propios correspondientes $\mathbf{u}_i$:

$$\mathbf{w} - \mathbf{w}^* = \sum_i \alpha_i \mathbf{u}_i$$

donde $\alpha_i =$

El viejo truco...

- Suponga que $\mathbf{H}$ tiene valores propios λ_i y vectores propios correspondientes $\mathbf{u}_i$:

$$\mathbf{w} - \mathbf{w}^* = \sum_i \alpha_i \mathbf{u}_i$$

donde $\alpha_i = \langle \mathbf{w} - \mathbf{w}^*, \mathbf{u}_i \rangle$.

El viejo truco...

- Suponga que $\mathbf{H}$ tiene valores propios λ_i y vectores propios correspondientes $\mathbf{u}_i$:

$$\mathbf{w} - \mathbf{w}^* = \sum_i \alpha_i \mathbf{u}_i$$

  donde $\alpha_i = \langle \mathbf{w} - \mathbf{w}^*, \mathbf{u}_i \rangle$.

- Luego

$$\Delta \mathbf{w}_k = (\mathbf{w}_k - \mathbf{w}^*) - (\mathbf{w}_{k-1} - \mathbf{w}^*)$$

El viejo truco...

- Suponga que $\mathbf{H}$ tiene valores propios λ_i y vectores propios correspondientes $\mathbf{u}_i$:

$$\mathbf{w} - \mathbf{w}^* = \sum_i \alpha_i \mathbf{u}_i$$

donde $\alpha_i = \langle \mathbf{w} - \mathbf{w}^*, \mathbf{u}_i \rangle$.

- Luego

$$\Delta \mathbf{w}_k = (\mathbf{w}_k - \mathbf{w}^*) - (\mathbf{w}_{k-1} - \mathbf{w}^*) = \sum_i (\alpha_i^{(k)} - \alpha_i^{(k-1)}) \mathbf{u}_i$$

El viejo truco...

- Suponga que $\mathbf{H}$ tiene valores propios λ_i y vectores propios correspondientes $\mathbf{u}_i$:

$$\mathbf{w} - \mathbf{w}^* = \sum_i \alpha_i \mathbf{u}_i$$

donde $\alpha_i = \langle \mathbf{w} - \mathbf{w}^*, \mathbf{u}_i \rangle$.

- Luego

$$\Delta \mathbf{w}_k = (\mathbf{w}_k - \mathbf{w}^*) - (\mathbf{w}_{k-1} - \mathbf{w}^*) = \sum_i (\alpha_i^{(k)} - \alpha_i^{(k-1)}) \mathbf{u}_i$$

- o en términos del gradiente:

$$\Delta \mathbf{w}_k = -\eta \mathbf{H} (\mathbf{w}_{k-1} - \mathbf{w}^*)$$

El viejo truco...

- Suponga que $\mathbf{H}$ tiene valores propios λ_i y vectores propios correspondientes $\mathbf{u}_i$:

$$\mathbf{w} - \mathbf{w}^* = \sum_i \alpha_i \mathbf{u}_i$$

donde $\alpha_i = \langle \mathbf{w} - \mathbf{w}^*, \mathbf{u}_i \rangle$.

- Luego

$$\Delta \mathbf{w}_k = (\mathbf{w}_k - \mathbf{w}^*) - (\mathbf{w}_{k-1} - \mathbf{w}^*) = \sum_i (\alpha_i^{(k)} - \alpha_i^{(k-1)}) \mathbf{u}_i$$

- o en términos del gradiente:

$$\Delta \mathbf{w}_k = -\eta \mathbf{H}(\mathbf{w}_{k-1} - \mathbf{w}^*) = \sum_i -\eta \lambda_i \alpha_i^{(k-1)} \mathbf{u}_i$$

El viejo truco...

- Suponga que $\mathbf{H}$ tiene valores propios λ_i y vectores propios correspondientes $\mathbf{u}_i$:

$$\mathbf{w} - \mathbf{w}^* = \sum_i \alpha_i \mathbf{u}_i$$

donde $\alpha_i = \langle \mathbf{w} - \mathbf{w}^*, \mathbf{u}_i \rangle$.

- Luego

$$\Delta \mathbf{w}_k = (\mathbf{w}_k - \mathbf{w}^*) - (\mathbf{w}_{k-1} - \mathbf{w}^*) = \sum_i (\alpha_i^{(k)} - \alpha_i^{(k-1)})\mathbf{u}_i$$

- o en términos del gradiente:

$$\Delta \mathbf{w}_k = -\eta \mathbf{H}(\mathbf{w}_{k-1} - \mathbf{w}^*) = \sum_i -\eta \lambda_i \alpha_i^{(k-1)} \mathbf{u}_i$$

- Comparando:

$$\alpha_i^{(k)} = (1 - \eta\lambda_i)\alpha_i^{(k-1)}$$

- En T iteraciones:

$$\alpha_i^{(T)} = (1 - \eta\lambda_i)^T \alpha_i^{(0)}$$

- En T iteraciones:

$$\alpha_i^{(T)} = (1 - \eta\lambda_i)^T \alpha_i^{(0)}$$

- Si garantizamos $|1 - \eta\lambda_i| < 1$, tenemos que cuando $T \to \infty$,

- En T iteraciones:

$$\alpha_i^{(T)} = (1 - \eta\lambda_i)^T \alpha_i^{(0)}$$

- Si garantizamos $|1 - \eta\lambda_i| < 1$, tenemos que cuando $T \to \infty$,

$$\alpha_i^{(T)} \to 0 \Rightarrow$$

- En T iteraciones:

$$\alpha_i^{(T)} = (1 - \eta\lambda_i)^T \alpha_i^{(0)}$$

- Si garantizamos $|1 - \eta\lambda_i| < 1$, tenemos que cuando $T \to \infty$,

$$\alpha_i^{(T)} \to 0 \Rightarrow \mathbf{w}^T \to$$

- En T iteraciones:

$$\alpha_i^{(T)} = (1 - \eta\lambda_i)^T \alpha_i^{(0)}$$

- Si garantizamos $|1 - \eta\lambda_i| < 1$, tenemos que cuando $T \to \infty$,

$$\alpha_i^{(T)} \to 0 \Rightarrow \mathbf{w}^T \to \mathbf{w}^*$$

- En T iteraciones:
$$\alpha_i^{(T)} = (1 - \eta\lambda_i)^T \alpha_i^{(0)}$$

- Si garantizamos $|1 - \eta\lambda_i| < 1$, tenemos que cuando $T \to \infty$,
$$\alpha_i^{(T)} \to 0 \Rightarrow \mathbf{w}^T \to \mathbf{w}^*$$

- $\eta \upuparrows \Rightarrow$

- En T iteraciones:
$$\alpha_i^{(T)} = (1 - \eta\lambda_i)^T \alpha_i^{(0)}$$

- Si garantizamos $|1 - \eta\lambda_i| < 1$, tenemos que cuando $T \to \infty$,
$$\alpha_i^{(T)} \to 0 \Rightarrow \mathbf{w}^T \to \mathbf{w}^*$$

- $\eta \upuparrows \Rightarrow$ convergencia más rápida.

- En T iteraciones:
$$\alpha_i^{(T)} = (1 - \eta\lambda_i)^T \alpha_i^{(0)}$$
- Si garantizamos $|1 - \eta\lambda_i| < 1$, tenemos que cuando $T \to \infty$,
$$\alpha_i^{(T)} \to 0 \Rightarrow \mathbf{w}^T \to \mathbf{w}^*$$
- $\eta \upuparrows \Rightarrow$ convergencia más rápida.
- Valor máximo?

- En T iteraciones:

$$\alpha_i^{(T)} = (1 - \eta\lambda_i)^T \alpha_i^{(0)}$$

- Si garantizamos $|1 - \eta\lambda_i| < 1$, tenemos que cuando $T \to \infty$,

$$\alpha_i^{(T)} \to 0 \Rightarrow \mathbf{w}^T \to \mathbf{w}^*$$

- $\eta \upuparrows \Rightarrow$ convergencia más rápida.
- Valor máximo?

$$|1 - \eta\lambda_i| < 1$$

- En T iteraciones:
$$\alpha_i^{(T)} = (1 - \eta\lambda_i)^T \alpha_i^{(0)}$$
- Si garantizamos $|1 - \eta\lambda_i| < 1$, tenemos que cuando $T \to \infty$,
$$\alpha_i^{(T)} \to 0 \Rightarrow \mathbf{w}^T \to \mathbf{w}^*$$
- $\eta \upuparrows \Rightarrow$ convergencia más rápida.
- Valor máximo?
$$|1 - \eta\lambda_i| < 1 \Rightarrow \eta < \frac{2}{\lambda_{\text{máx}}}$$

- En T iteraciones:

$$\alpha_i^{(T)} = (1 - \eta\lambda_i)^T \alpha_i^{(0)}$$

- Si garantizamos $|1 - \eta\lambda_i| < 1$, tenemos que cuando $T \to \infty$,

$$\alpha_i^{(T)} \to 0 \Rightarrow \mathbf{w}^T \to \mathbf{w}^*$$

- $\eta \upuparrows \Rightarrow$ convergencia más rápida.
- Valor máximo?

$$|1 - \eta\lambda_i| < 1 \Rightarrow \eta < \frac{2}{\lambda_{\text{máx}}}$$

- Con $\eta \approx \frac{2}{\lambda_{\text{máx}}}$, convergencia es governada por:

$$\left(1 - 2\frac{\lambda_{\text{mín}}}{\lambda_{\text{máx}}}\right) = \left(1 - \frac{2}{\kappa}\right)$$

- En T iteraciones:

$$\alpha_i^{(T)} = (1 - \eta\lambda_i)^T \alpha_i^{(0)}$$

- Si garantizamos $|1 - \eta\lambda_i| < 1$, tenemos que cuando $T \to \infty$,

$$\alpha_i^{(T)} \to 0 \Rightarrow \mathbf{w}^T \to \mathbf{w}^*$$

- $\eta \upuparrows \Rightarrow$ convergencia más rápida.
- Valor máximo?

$$|1 - \eta\lambda_i| < 1 \Rightarrow \eta < \frac{2}{\lambda_{\text{máx}}}$$

- Con $\eta \approx \frac{2}{\lambda_{\text{máx}}}$, convergencia es governada por:

$$\left(1 - 2\frac{\lambda_{\text{mín}}}{\lambda_{\text{máx}}}\right) = \left(1 - \frac{2}{\kappa}\right)$$

- $\kappa \upuparrows \Rightarrow$

- En T iteraciones:

$$\alpha_i^{(T)} = (1 - \eta\lambda_i)^T \alpha_i^{(0)}$$

- Si garantizamos $|1 - \eta\lambda_i| < 1$, tenemos que cuando $T \to \infty$,

$$\alpha_i^{(T)} \to 0 \Rightarrow \mathbf{w}^T \to \mathbf{w}^*$$

- $\eta \upuparrows \Rightarrow$ convergencia más rápida.
- Valor máximo?

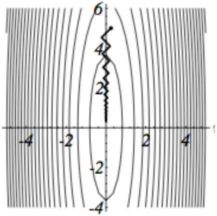
$$|1 - \eta\lambda_i| < 1 \Rightarrow \eta < \frac{2}{\lambda_{\text{máx}}}$$

- Con $\eta \approx \frac{2}{\lambda_{\text{máx}}}$, convergencia es governada por:

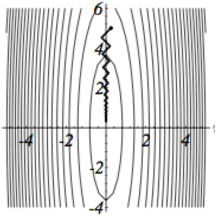
$$\left(1 - 2\frac{\lambda_{\text{mín}}}{\lambda_{\text{máx}}}\right) = \left(1 - \frac{2}{\kappa}\right)$$

- $\kappa \upuparrows \Rightarrow$convergencia puede ser muy lenta!

- Convergencia lenta.

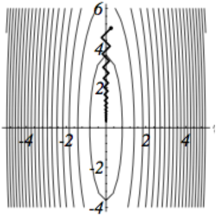


- Convergencia lenta.



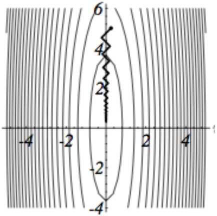
- Simple: sólo primeras derivadas

- Convergencia lenta.



- Simple: sólo primeras derivadas
- Con búsqueda de línea, tasa de convergencia es

- Convergencia lenta.



- Simple: sólo primeras derivadas
- Con búsqueda de línea, tasa de convergencia es $\frac{\kappa-1}{\kappa+1}$

Descenso de Gradiente Estocástico/ en línea

$$
\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i))^2 \\
\nabla_{\mathbf{w}} E &= \sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)\mathbf{x}_i
\end{aligned}
$$

Descenso de Gradiente Estocástico/ en línea

$$
\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i))^2 \\
\nabla_{\mathbf{w}} E &= \sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)\mathbf{x}_i
\end{aligned}
$$

- $\nabla_{\mathbf{w}} E|_{\mathbf{w}_k}$ se estima a partir de un minibatch de los datos

Descenso de Gradiente Estocástico/ en línea

$$
\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i))^2 \\
\nabla_{\mathbf{w}} E &= \sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)\mathbf{x}_i
\end{aligned}
$$

- $\nabla_{\mathbf{w}} E|_{\mathbf{w}_k}$ se estima a partir de un minibatch de los datos
- Varias pasadas por los datos.

Descenso de Gradiente Estocástico/ en línea

$$\begin{aligned} E(\mathbf{w}) &= \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i))^2 \\ \nabla_{\mathbf{w}} E &= \sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)\mathbf{x}_i \end{aligned}$$

- $\nabla_{\mathbf{w}} E|_{\mathbf{w}_k}$ se estima a partir de un minibatch de los datos
- Varias pasadas por los datos.
- En el caso extremo se usa un solo dato

Descenso de Gradiente Estocástico/ en línea

$$
\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i))^2 \\
\nabla_{\mathbf{w}} E &= \sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)\mathbf{x}_i
\end{aligned}
$$

- $\nabla_{\mathbf{w}} E|_{\mathbf{w}_k}$ se estima a partir de un minibatch de los datos
- Varias pasadas por los datos.
- En el caso extremo se usa un solo dato:

$$
\begin{aligned}
\nabla_{\mathbf{w}} E &\approx (\mathbf{w}^T\mathbf{x}_i - y_i)\mathbf{x}_i \\
&= e_i\mathbf{x}_i
\end{aligned}
$$

Algoritmo LMS

Incialize $\mathbf{w}_0$ a valores pequeños.

Algoritmo LMS

Incialize $\mathbf{w}_0$ a valores pequeños.
repeat
    Escoja $(\mathbf{x}_i, y_i)$

Algoritmo LMS

Incialize $\mathbf{w}_0$ a valores pequeños.

repeat

Escoja $(\mathbf{x}_i, y_i)$

$g = \mathbf{w}_k^T \mathbf{x}_i$

Algoritmo LMS

Incialize $\mathbf{w}_0$ a valores pequeños.

repeat

  Escoja $(\mathbf{x}_i, y_i)$

  $g = \mathbf{w}_k^T \mathbf{x}_i$

  $e = g - y_i$

Algoritmo LMS

Incialize $\mathbf{w}_0$ a valores pequeños.

repeat

Escoja $(\mathbf{x}_i, y_i)$

$g = \mathbf{w}_k^T \mathbf{x}_i$

$e = g - y_i$

$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta_k e \mathbf{x}_i$

Algoritmo LMS

Incialize $\mathbf{w}_0$ a valores pequeños.
repeat
    Escoja $(\mathbf{x}_i, y_i)$
    $g = \mathbf{w}_k^T \mathbf{x}_i$
    $e = g - y_i$
    $\mathbf{w}_{k+1} = \mathbf{w}_k - \eta_k e \mathbf{x}_i$
until Condición de terminación.

Ecuaciones Normales

- Defina:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T; & \mathbf{x}_2^T; & \cdots & \mathbf{x}_n^T \end{bmatrix}$$

Ecuaciones Normales

- Defina:

$$
\begin{aligned}
\mathbf{X} &= \begin{bmatrix} \mathbf{x}_1^T; & \mathbf{x}_2^T; & \cdots & \mathbf{x}_n^T \end{bmatrix} \\
\mathbf{y} &= \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^T
\end{aligned}
$$

Ecuaciones Normales

- Defina:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T; & \mathbf{x}_2^T; & \cdots & \mathbf{x}_n^T \end{bmatrix}$$
$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^T$$

- Entonces, para una solución con $E(\mathbf{w}) = 0$ se requiere:

Ecuaciones Normales

- Defina:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T; & \mathbf{x}_2^T; & \cdots & \mathbf{x}_n^T \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^T$$

- Entonces, para una solución con $E(\mathbf{w}) = 0$ se requiere:

$$\mathbf{X}\mathbf{w} = \mathbf{y}$$

Ecuaciones Normales

- Defina:

$$\mathbf{X} = \begin{bmatrix}\mathbf{x}_1^T; & \mathbf{x}_2^T; & \cdots & \mathbf{x}_n^T\end{bmatrix}$$
$$\mathbf{y} = \begin{bmatrix}y_1 & y_2 & \cdots & y_n\end{bmatrix}^T$$

- Entonces, para una solución con $E(\mathbf{w}) = 0$ se requiere:

$$\mathbf{X}\mathbf{w} = \mathbf{y}$$

- Es decir, $\mathbf{y}$ debe ser una combinación lineal de las columnas de $\mathbf{X}$.

Ecuaciones Normales

- Defina:

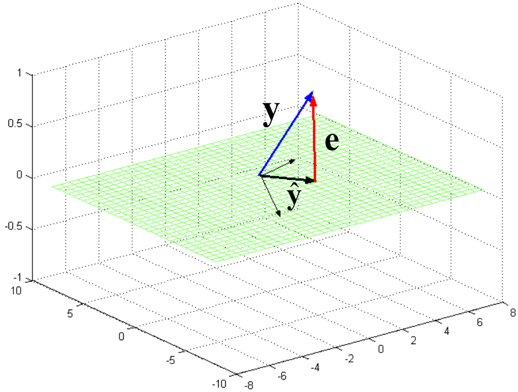
$$\mathbf{X} = \begin{bmatrix}\mathbf{x}_1^T; & \mathbf{x}_2^T; & \cdots & \mathbf{x}_n^T\end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix}y_1 & y_2 & \cdots & y_n\end{bmatrix}^T$$

- Entonces, para una solución con $E(\mathbf{w}) = 0$ se requiere:

$$\mathbf{X}\mathbf{w} = \mathbf{y}$$

- Es decir, $\mathbf{y}$ debe ser una combinación lineal de las columnas de $\mathbf{X}$.
- En general, no existe $\mathbf{w}$ que cumpla esta condición.


y
e
ŷ
1
0.5
0
-0.5
-1
10
5
0
-5
-10
-8
-6
-4
-2
0
2
4
6
8

Solución

$$\mathbf{e} \perp \hat{\mathbf{y}}, \hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}}$$

Solución

$$\mathbf{e} \perp \hat{\mathbf{y}}, \hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} \Rightarrow (\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}) \perp \mathbf{X}(:, i)$$

Solución

$$\mathbf{e} \perp \hat{\mathbf{y}}, \hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} \Rightarrow (\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}) \perp \mathbf{X}(:, i)$$

$$\mathbf{X}^T(\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}) \quad = \quad 0$$

Solución

$$\mathbf{e} \perp \hat{\mathbf{y}}, \hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} \Rightarrow (\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}) \perp \mathbf{X}(:, i)$$

$$\begin{aligned} \mathbf{X}^T(\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}) &= 0 \\ (\mathbf{X}^T\mathbf{X})\hat{\mathbf{w}} &= \mathbf{X}^T\mathbf{y} \end{aligned}$$

Solución

$$\mathbf{e} \perp \hat{\mathbf{y}}, \hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} \Rightarrow (\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}) \perp \mathbf{X}(:, i)$$

$$\begin{aligned} \mathbf{X}^T(\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}) &= 0 \\ (\mathbf{X}^T\mathbf{X})\hat{\mathbf{w}} &= \mathbf{X}^T\mathbf{y} \\ \hat{\mathbf{w}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned}$$

Solución

$$\mathbf{e} \perp \hat{\mathbf{y}}, \hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} \Rightarrow (\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}) \perp \mathbf{X}(:, i)$$

$$\begin{aligned}
\mathbf{X}^T(\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}) &= 0 \\
(\mathbf{X}^T\mathbf{X})\hat{\mathbf{w}} &= \mathbf{X}^T\mathbf{y} \\
\hat{\mathbf{w}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
\hat{\mathbf{w}} &= \mathbf{X}^+\mathbf{y}
\end{aligned}$$

Interpretación probabilística

- Suposiciones:
  - y_i y $\mathbf{x}_i$ están relacionadas por:

$$y_i = \check{\mathbf{w}}^T \mathbf{x}_i + \epsilon_i$$

Interpretación probabilística

- Suposiciones:
  - y_i y $\mathbf{x}_i$ están relacionadas por:

$$y_i = \check{\mathbf{w}}^T \mathbf{x}_i + \epsilon_i$$

  - $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, y son independientes:

Interpretación probabilística

- Suposiciones:
  - y_i y $\mathbf{x}_i$ están relacionadas por:

$$y_i = \check{\mathbf{w}}^T \mathbf{x}_i + \epsilon_i$$

  - $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, y son independientes:

$$p(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

Interpretación probabilística

- Suposiciones:
  - y_i y $\mathbf{x}_i$ están relacionadas por:

$$y_i = \check{\mathbf{w}}^T \mathbf{x}_i + \epsilon_i$$

  - $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, y son independientes:

$$p(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

- y_i es una variable aleatoria, con densidad:

$$p(y_i \mid x_i; \check{\mathbf{w}}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \check{\mathbf{w}}^T \mathbf{x}_i)^2}{2\sigma^2}\right)$$

Interpretación probabilística

- Suposiciones:
  - y_i y $\mathbf{x}_i$ están relacionadas por:

$$y_i = \check{\mathbf{w}}^T \mathbf{x}_i + \epsilon_i$$

  - $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, y son independientes:

$$p(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

- y_i es una variable aleatoria, con densidad:

$$p(y_i \mid x_i; \check{\mathbf{w}}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \check{\mathbf{w}}^T \mathbf{x}_i)^2}{2\sigma^2}\right)$$

- Pensamos en $\check{\mathbf{w}}$ como un parámetro de la densidad.

- Función de verosimilitud:

$$L(\check{\mathbf{w}}) = p(\mathbf{y} \mid \mathbf{X}; \check{\mathbf{w}})$$

- Función de verosimilitud:

$$L(\check{\mathbf{w}}) = p(\mathbf{y} \mid \mathbf{X}; \check{\mathbf{w}})$$

- Por independencia de los ϵ_i:

$$L(\check{\mathbf{w}}) = \prod_{i=1}^{n} p(y_i \mid x_i; \check{\mathbf{w}}) =$$

- Función de verosimilitud:

$$L(\check{\mathbf{w}}) = p(\mathbf{y} \mid \mathbf{X}; \check{\mathbf{w}})$$

- Por independencia de los ϵ_i:

$$L(\check{\mathbf{w}}) = \prod_{i=1}^{n} p(y_i \mid x_i; \check{\mathbf{w}}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \check{\mathbf{w}}^T \mathbf{x}_i)^2}{2\sigma^2}\right)$$

- Función de verosimilitud:

$$L(\check{\mathbf{w}}) = p(\mathbf{y} \mid \mathbf{X}; \check{\mathbf{w}})$$

- Por independencia de los ϵ_i:

$$L(\check{\mathbf{w}}) = \prod_{i=1}^{n} p(y_i \mid x_i; \check{\mathbf{w}}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \check{\mathbf{w}}^T \mathbf{x}_i)^2}{2\sigma^2}\right)$$

- Cuál es una buena opción para $\check{\mathbf{w}}$?

- Función de verosimilitud:

$$L(\check{\mathbf{w}}) = p(\mathbf{y} \mid \mathbf{X}; \check{\mathbf{w}})$$

- Por independencia de los ϵ_i:

$$L(\check{\mathbf{w}}) = \prod_{i=1}^{n} p(y_i \mid x_i; \check{\mathbf{w}}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \check{\mathbf{w}}^T \mathbf{x}_i)^2}{2\sigma^2}\right)$$

- Cuál es una buena opción para $\check{\mathbf{w}}$?
- Principio de máxima verosimilitud: escoger $\check{\mathbf{w}}$ de manera que la probabilidad de los datos sea máxima.

- Queremos encontrar el $\check{\mathbf{w}}$ que maximiza la verosimilitud logarítimica:

$$\log(L(\check{\mathbf{w}})) = -\log\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-\check{\mathbf{w}}^T\mathbf{x}_i)^2}{2\sigma^2}\right)\right)$$

- Queremos encontrar el $\check{\mathbf{w}}$ que maximiza la verosimilitud logarítimica:

$$\log(L(\check{\mathbf{w}})) = -\log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \check{\mathbf{w}}^T\mathbf{x}_i)^2}{2\sigma^2}\right)\right)$$
$$= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \check{\mathbf{w}}^T\mathbf{x}_i)^2}{2\sigma^2}\right)\right)$$

- Queremos encontrar el $\check{\mathbf{w}}$ que maximiza la verosimilitud logarítimica:

$$
\begin{aligned}
\log(L(\check{\mathbf{w}})) &= -\log\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-\check{\mathbf{w}}^T\mathbf{x}_i)^2}{2\sigma^2}\right)\right)\\
&= \sum_{i=1}^{n}\log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-\check{\mathbf{w}}^T\mathbf{x}_i)^2}{2\sigma^2}\right)\right)\\
&= n\log\frac{1}{\sqrt{2\pi}\sigma}-\frac{1}{\sigma^2}\times\frac{1}{2}\sum_{i=1}^{n}(y_i-\check{\mathbf{w}}^T\mathbf{x}_i)^2
\end{aligned}
$$

- Queremos encontrar el $\check{\mathbf{w}}$ que maximiza la verosimilitud logarítimica:

$$\log(L(\check{\mathbf{w}})) = -\log\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-\check{\mathbf{w}}^T\mathbf{x}_i)^2}{2\sigma^2}\right)\right)$$
$$= \sum_{i=1}^{n}\log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-\check{\mathbf{w}}^T\mathbf{x}_i)^2}{2\sigma^2}\right)\right)$$
$$= n\log\frac{1}{\sqrt{2\pi}\sigma}-\frac{1}{\sigma^2}\times\frac{1}{2}\sum_{i=1}^{n}(y_i-\check{\mathbf{w}}^T\mathbf{x}_i)^2$$

- Maximizar verosimilitud es equivalente a minimizar el error cuadrático en los datos.

Función de error?

- Análisis para el problema de regresión.

Función de error?

- Análisis para el problema de regresión.
- Suponga que h es una función de $\mathbf{x}$ que depende de parámetros $\mathbf{w}$

Función de error?

- Análisis para el problema de regresión.
- Suponga que h es una función de $\mathbf{x}$ que depende de parámetros $\mathbf{w}$

$$\hat{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(h(\mathbf{x}_i, \mathbf{w}) - y_i)^2$$

Función de error?

- Análisis para el problema de regresión.
- Suponga que h es una función de $\mathbf{x}$ que depende de parámetros $\mathbf{w}$

$$\hat{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(h(\mathbf{x}_i, \mathbf{w}) - y_i)^2$$

- En el límite $n \to \infty$:

$$E(\mathbf{w}) = \lim_{n\to\infty} \frac{\hat{E}(\mathbf{w})}{n}$$

Función de error?

- Análisis para el problema de regresión.
- Suponga que h es una función de $\mathbf{x}$ que depende de parámetros $\mathbf{w}$

$$\hat{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(h(\mathbf{x}_i, \mathbf{w}) - y_i)^2$$

- En el límite $n \to \infty$:

$$E(\mathbf{w}) = \lim_{n\to\infty} \frac{\hat{E}(\mathbf{w})}{n} = \frac{1}{2}\iint (h(\mathbf{x}, \mathbf{w}) - y)^2 p(\mathbf{x}, y)\, dy\, d\mathbf{x}$$

Función de error?

- Análisis para el problema de regresión.
- Suponga que h es una función de $\mathbf{x}$ que depende de parámetros $\mathbf{w}$

$$\hat{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(h(\mathbf{x}_i, \mathbf{w}) - y_i)^2$$

- En el límite $n \to \infty$:

$$E(\mathbf{w}) = \lim_{n\to\infty} \frac{\hat{E}(\mathbf{w})}{n} = \frac{1}{2}\iint (h(\mathbf{x}, \mathbf{w}) - y)^2 p(\mathbf{x}, y) dy d\mathbf{x}$$

- Reemplazando $p(\mathbf{x}, y) = p(y|\mathbf{x})p(\mathbf{x})$:

$$E(\mathbf{w}) = \frac{1}{2}\iint (h(\mathbf{x}, \mathbf{w}) - y)^2 p(y|\mathbf{x})p(\mathbf{x}) dy d\mathbf{x}$$

- Truco: $(h(\mathbf{x},\mathbf{w}) - y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}] + \mathbb{E}[y|\mathbf{x}] - y)^2$

- Truco: $(h(\mathbf{x},\mathbf{w}) - y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}] + \mathbb{E}[y|\mathbf{x}] - y)^2$

$$\begin{aligned}(h(\mathbf{x},\mathbf{w}) - y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 + (\mathbb{E}[y|\mathbf{x}] - y)^2 \\ + 2(h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])(\mathbb{E}[y|\mathbf{x}] - y)\end{aligned}$$

- Truco: $(h(\mathbf{x},\mathbf{w}) - y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}\left[y|\mathbf{x}\right] + \mathbb{E}\left[y|\mathbf{x}\right] - y)^2$

$$\begin{aligned}(h(\mathbf{x},\mathbf{w}) - y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}\left[y|\mathbf{x}\right])^2 + (\mathbb{E}\left[y|\mathbf{x}\right] - y)^2 \\ + 2(h(\mathbf{x},\mathbf{w}) - \mathbb{E}\left[y|\mathbf{x}\right])(\mathbb{E}\left[y|\mathbf{x}\right] - y)\end{aligned}$$

- integrando cada término con respecto a y:

- Truco: $(h(\mathbf{x},\mathbf{w}) - y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}] + \mathbb{E}[y|\mathbf{x}] - y)^2$

$$
\begin{aligned}
(h(\mathbf{x},\mathbf{w}) - y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 + (\mathbb{E}[y|\mathbf{x}] - y)^2 \\
+ 2(h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])(\mathbb{E}[y|\mathbf{x}] - y)
\end{aligned}
$$

- integrando cada término con respecto a y:

$$
\iint (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 p(y|\mathbf{x}) p(\mathbf{x}) dy d\mathbf{x}
$$

$$
=
$$

- Truco: $(h(\mathbf{x},\mathbf{w})-y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}] + \mathbb{E}[y|\mathbf{x}] - y)^2$

$$
\begin{aligned}
(h(\mathbf{x},\mathbf{w})-y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 + (\mathbb{E}[y|\mathbf{x}] - y)^2 \\
+ 2(h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])(\mathbb{E}[y|\mathbf{x}] - y)
\end{aligned}
$$

- integrando cada término con respecto a y:

$$
\begin{aligned}
&\iint (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 p(y|\mathbf{x}) p(\mathbf{x}) dy d\mathbf{x} \\
&\qquad = \int (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 p(\mathbf{x}) d\mathbf{x}
\end{aligned}
$$

- Truco: $(h(\mathbf{x},\mathbf{w})-y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}] + \mathbb{E}[y|\mathbf{x}] - y)^2$

$$
\begin{aligned}
(h(\mathbf{x},\mathbf{w})-y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 + (\mathbb{E}[y|\mathbf{x}] - y)^2 \\
+ 2(h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])(\mathbb{E}[y|\mathbf{x}] - y)
\end{aligned}
$$

- integrando cada término con respecto a y:

$$
\begin{aligned}
\iint (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 p(y|\mathbf{x})p(\mathbf{x})dyd\mathbf{x} \\
= \int (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 p(\mathbf{x})d\mathbf{x}
\end{aligned}
$$

$$
\int (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])(\mathbb{E}[y|\mathbf{x}] - y)p(y|\mathbf{x})dy =
$$

- Truco: $(h(\mathbf{x},\mathbf{w}) - y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}] + \mathbb{E}[y|\mathbf{x}] - y)^2$

$$\begin{aligned}(h(\mathbf{x},\mathbf{w}) - y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 + (\mathbb{E}[y|\mathbf{x}] - y)^2 \\ + 2(h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])(\mathbb{E}[y|\mathbf{x}] - y)\end{aligned}$$

- integrando cada término con respecto a y:

$$\begin{aligned}\iint (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 p(y|\mathbf{x})p(\mathbf{x})dyd\mathbf{x} \\ = \int (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 p(\mathbf{x})d\mathbf{x}\end{aligned}$$

$$\int (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])(\mathbb{E}[y|\mathbf{x}] - y)p(y|\mathbf{x})dy = 0$$

- Truco: $(h(\mathbf{x},\mathbf{w}) - y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}] + \mathbb{E}[y|\mathbf{x}] - y)^2$

$$\begin{aligned}(h(\mathbf{x},\mathbf{w}) - y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 + (\mathbb{E}[y|\mathbf{x}] - y)^2 \\ + 2(h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])(\mathbb{E}[y|\mathbf{x}] - y)\end{aligned}$$

- integrando cada término con respecto a y:

$$\begin{aligned}\iint (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 p(y|\mathbf{x})p(\mathbf{x})dyd\mathbf{x} \\ = \int (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 p(\mathbf{x})d\mathbf{x}\end{aligned}$$

$$(h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])(\mathbb{E}[y|\mathbf{x}] - \int y)p(y|\mathbf{x})dy = 0$$

- Truco: $(h(\mathbf{x},\mathbf{w})-y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}\left[y|\mathbf{x}\right] + \mathbb{E}\left[y|\mathbf{x}\right] - y)^2$

$$
\begin{aligned}
(h(\mathbf{x},\mathbf{w})-y)^2 = (h(\mathbf{x},\mathbf{w}) - \mathbb{E}\left[y|\mathbf{x}\right])^2 + (\mathbb{E}\left[y|\mathbf{x}\right] - y)^2 \\
+ 2(h(\mathbf{x},\mathbf{w}) - \mathbb{E}\left[y|\mathbf{x}\right])(\mathbb{E}\left[y|\mathbf{x}\right] - y)
\end{aligned}
$$

- integrando cada término con respecto a y:

$$
\begin{aligned}
&\iint (h(\mathbf{x},\mathbf{w}) - \mathbb{E}\left[y|\mathbf{x}\right])^2 p(y|\mathbf{x})p(\mathbf{x})dyd\mathbf{x} \\
&\qquad = \int (h(\mathbf{x},\mathbf{w}) - \mathbb{E}\left[y|\mathbf{x}\right])^2 p(\mathbf{x})d\mathbf{x}
\end{aligned}
$$

$$
\int (h(\mathbf{x},\mathbf{w}) - \mathbb{E}\left[y|\mathbf{x}\right])(\mathbb{E}\left[y|\mathbf{x}\right] - y)p(y|\mathbf{x})dy = 0
$$

$$\int (\mathbb{E}\left[y|\mathbf{x}\right] - y)^2 p(y|\mathbf{x}) dy =$$

$$\int (\mathbb{E}\left[y|\mathbf{x}\right] - y)^2 p(y|\mathbf{x}) dy = \mathbb{E}\left[y^2|\mathbf{x}\right] - \mathbb{E}\left[y|\mathbf{x}\right]^2$$

$$\int (\mathbb{E}\left[y|\mathbf{x}\right] - y)^2 p(y|\mathbf{x}) dy = \mathbb{E}\left[y^2|\mathbf{x}\right] - \mathbb{E}\left[y|\mathbf{x}\right]^2$$

- Todo junto:

$$\begin{aligned} E(\mathbf{w}) = \frac{1}{2} \int (h(\mathbf{x}, \mathbf{w}) - \mathbb{E}\left[y|\mathbf{x}\right])^2 p(\mathbf{x}) d\mathbf{x} \\ + \frac{1}{2} \int \mathbb{E}\left[y^2|\mathbf{x}\right] - \mathbb{E}\left[y|\mathbf{x}\right]^2 p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

$$\int (\mathbb{E}[y|\mathbf{x}] - y)^2 p(y|\mathbf{x}) dy = \mathbb{E}\left[y^2|\mathbf{x}\right] - \mathbb{E}[y|\mathbf{x}]^2$$

- Todo junto:

$$E(\mathbf{w}) = \frac{1}{2}\int (h(\mathbf{x},\mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 p(\mathbf{x}) d\mathbf{x} \\ + \frac{1}{2}\int \mathbb{E}\left[y^2|\mathbf{x}\right] - \mathbb{E}[y|\mathbf{x}]^2 p(\mathbf{x}) d\mathbf{x}$$

- Segundo término no depende de $\mathbf{w}$!

$$\int (\mathbb{E}\left[y|\mathbf{x}\right] - y)^2 p(y|\mathbf{x})dy = \mathbb{E}\left[y^2|\mathbf{x}\right] - \mathbb{E}\left[y|\mathbf{x}\right]^2$$

- Todo junto:

$$E(\mathbf{w}) = \frac{1}{2}\int (h(\mathbf{x},\mathbf{w}) - \mathbb{E}\left[y|\mathbf{x}\right])^2 p(\mathbf{x})d\mathbf{x} + \frac{1}{2}\int \mathbb{E}\left[y^2|\mathbf{x}\right] - \mathbb{E}\left[y|\mathbf{x}\right]^2 p(\mathbf{x})d\mathbf{x}$$

- Segundo término no depende de $\mathbf{w}$!
- $h(\mathbf{x},\mathbf{w})$ que minimiza $E(\mathbf{w})$ es

$$\int (\mathbb{E}\left[y|\mathbf{x}\right] - y)^2 p(y|\mathbf{x})dy = \mathbb{E}\left[y^2|\mathbf{x}\right] - \mathbb{E}\left[y|\mathbf{x}\right]^2$$

- Todo junto:

$$E(\mathbf{w}) = \frac{1}{2}\int (h(\mathbf{x},\mathbf{w}) - \mathbb{E}\left[y|\mathbf{x}\right])^2 p(\mathbf{x})d\mathbf{x} + \frac{1}{2}\int \mathbb{E}\left[y^2|\mathbf{x}\right] - \mathbb{E}\left[y|\mathbf{x}\right]^2 p(\mathbf{x})d\mathbf{x}$$

- Segundo término no depende de $\mathbf{w}$!
- $h(\mathbf{x},\mathbf{w})$ que minimiza $E(\mathbf{w})$ es $\mathbb{E}\left[y|\mathbf{x}\right]$.

$$\int (\mathbb{E}\left[y|\mathbf{x}\right] - y)^2 p(y|\mathbf{x})dy = \mathbb{E}\left[y^2|\mathbf{x}\right] - \mathbb{E}\left[y|\mathbf{x}\right]^2$$

- Todo junto:

$$E(\mathbf{w}) = \frac{1}{2}\int (h(\mathbf{x},\mathbf{w}) - \mathbb{E}\left[y|\mathbf{x}\right])^2 p(\mathbf{x})d\mathbf{x} + \frac{1}{2}\int \mathbb{E}\left[y^2|\mathbf{x}\right] - \mathbb{E}\left[y|\mathbf{x}\right]^2 p(\mathbf{x})d\mathbf{x}$$

- Segundo término no depende de $\mathbf{w}$!
- $h(\mathbf{x},\mathbf{w})$ que minimiza $E(\mathbf{w})$ es $\mathbb{E}\left[y|\mathbf{x}\right]$.
- $\mathbb{E}\left[y|\mathbf{x}\right]$ se llama función de regresión.

El dilema sesgo-varianza

- En la práctica n es finito. Considere conjuntos con n datos $D_1, D_2, \cdots \sim p(\mathbf{x}, y)$.

El dilema sesgo-varianza

- En la práctica n es finito. Considere conjuntos con n datos $D_1, D_2, \cdots \sim p(\mathbf{x}, y)$.
- Para un $\mathbf{x}$ fijo $(h(\mathbf{x}, \mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2$ depende del D_i particular.

El dilema sesgo-varianza

- En la práctica n es finito. Considere conjuntos con n datos $D_1, D_2, \cdots \sim p(\mathbf{x}, y)$.
- Para un $\mathbf{x}$ fijo $(h(\mathbf{x}, \mathbf{w}) - \mathbb{E}\left[y|\mathbf{x}\right])^2$ depende del D_i particular.
- Promediando sobre todos los D posibles:

El dilema sesgo-varianza

- En la práctica n es finito. Considere conjuntos con n datos $D_1, D_2, \cdots \sim p(\mathbf{x}, y)$.
- Para un $\mathbf{x}$ fijo $(h(\mathbf{x}, \mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2$ depende del D_i particular.
- Promediando sobre todos los D posibles:

$$\mathbb{E}[E(\mathbf{x})] = \mathbb{E}_D(h(\mathbf{x}, \mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2$$

El dilema sesgo-varianza

- En la práctica n es finito. Considere conjuntos con n datos $D_1, D_2, \cdots \sim p(\mathbf{x}, y)$.
- Para un $\mathbf{x}$ fijo $(h(\mathbf{x}, \mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2$ depende del D_i particular.
- Promediando sobre todos los D posibles:

$$\begin{aligned}\mathbb{E}[E(\mathbf{x})] = \mathbb{E}_D(h(\mathbf{x}, \mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 = \\ \mathbb{E}_D\left(h(\mathbf{x}, \mathbf{w}) - \mathbb{E}_D h(\mathbf{x}, \mathbf{w}) + \mathbb{E}_D h(\mathbf{x}, \mathbf{w}) - \mathbb{E}[y|\mathbf{x}]\right)^2\end{aligned}$$

El dilema sesgo-varianza

- En la práctica n es finito. Considere conjuntos con n datos $D_1, D_2, \cdots \sim p(\mathbf{x}, y)$.
- Para un $\mathbf{x}$ fijo $(h(\mathbf{x}, \mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2$ depende del D_i particular.
- Promediando sobre todos los D posibles:

$$
\begin{aligned}
\mathbb{E}[E(\mathbf{x})] = \mathbb{E}_D(h(\mathbf{x}, \mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 = \\
\mathbb{E}_D\left(h(\mathbf{x}, \mathbf{w}) - \mathbb{E}_D h(\mathbf{x}, \mathbf{w}) + \mathbb{E}_D h(\mathbf{x}, \mathbf{w}) - \mathbb{E}[y|\mathbf{x}]\right)^2
\end{aligned}
$$

- De nuevo la integral del término cruzado es cero, y tenemos:

El dilema sesgo-varianza

- En la práctica n es finito. Considere conjuntos con n datos $D_1, D_2, \cdots \sim p(\mathbf{x}, y)$.
- Para un $\mathbf{x}$ fijo $(h(\mathbf{x}, \mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2$ depende del D_i particular.
- Promediando sobre todos los D posibles:

$$\mathbb{E}[E(\mathbf{x})] = \mathbb{E}_D(h(\mathbf{x}, \mathbf{w}) - \mathbb{E}[y|\mathbf{x}])^2 = \\ \mathbb{E}_D\left(h(\mathbf{x}, \mathbf{w}) - \mathbb{E}_D h(\mathbf{x}, \mathbf{w}) + \mathbb{E}_D h(\mathbf{x}, \mathbf{w}) - \mathbb{E}[y|\mathbf{x}]\right)^2$$

- De nuevo la integral del término cruzado es cero, y tenemos:

$$\mathbb{E}[E(\mathbf{x})] = \mathbb{E}(h(\mathbf{x}, \mathbf{w}) - \mathbb{E}_D(h(\mathbf{x}, \mathbf{w})))^2 \\ + (\mathbb{E}_D(h(\mathbf{x}, \mathbf{w})) - \mathbb{E}[y|\mathbf{x}])^2$$

- $\left(\mathbb{E}_D(h(\mathbf{x},\mathbf{w})) - \mathbb{E}\left[y|\mathbf{x}\right]\right)^2$ es el sesgo de h en $\mathbf{x}$.

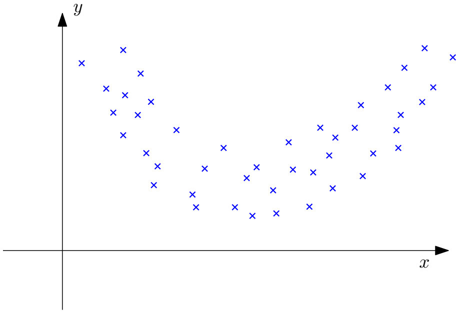
- $(\mathbb{E}_D(h(\mathbf{x}, \mathbf{w})) - \mathbb{E}\left[y|\mathbf{x}\right])^2$ es el sesgo de h en $\mathbf{x}$.
- $\mathbb{E}(h(\mathbf{x}, \mathbf{w}) - \mathbb{E}_D(h(\mathbf{x}, \mathbf{w})))^2$ es la varianza de h en $\mathbf{x}$.

- $(\mathbb{E}_D(h(\mathbf{x},\mathbf{w})) - \mathbb{E}\left[y|\mathbf{x}\right])^2$ es el sesgo de h en $\mathbf{x}$.
- $\mathbb{E}(h(\mathbf{x},\mathbf{w}) - \mathbb{E}_D(h(\mathbf{x},\mathbf{w})))^2$ es la varianza de h en $\mathbf{x}$.
- Integrando sobre $\mathbf{x}$:

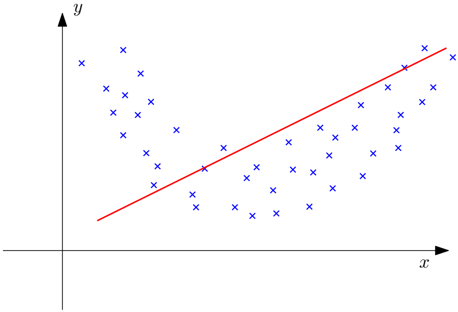
$$\text{sesgo}^2 = \frac{1}{2}\int (\mathbb{E}_D(h(\mathbf{x},\mathbf{w})) - \mathbb{E}\left[y|\mathbf{x}\right])^2 p(\mathbf{x})d\mathbf{x}$$

$$\text{varianza} = \frac{1}{2}\int \mathbb{E}(h(\mathbf{x},\mathbf{w}) - \mathbb{E}_D(h(\mathbf{x},\mathbf{w})))^2 p(\mathbf{x})d\mathbf{x}$$

Características/Preprocesamiento

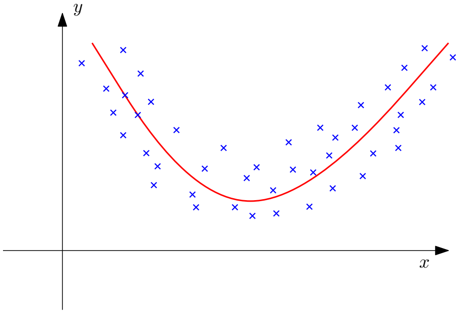


Características/Preprocesamiento



- Modelo no es apropiado.

Características/Preprocesamiento



- Modelo no es apropiado.
- Mejor opción:

$$y = w_2x^2 + w_1x + w_0$$

- En 2 dimensiones:

- En 2 dimensiones:

$$y = w_{11}x_1^2 + w_{22}x_2^2$$

- En 2 dimensiones:

$$y = w_{11}x_1^2 + w_{22}x_2^2 + w_{12}x_1x_2$$

- En 2 dimensiones:

$$y = w_{11}x_1^2 + w_{22}x_2^2 + w_{12}x_1x_2 + w_1x_1 + w_2x_2$$

- En 2 dimensiones:

$$y = w_{11}x_1^2 + w_{22}x_2^2 + w_{12}x_1x_2 + w_1x_1 + w_2x_2 + w_0$$

- En 2 dimensiones:

$$y = w_{11}x_1^2 + w_{22}x_2^2 + w_{12}x_1x_2 + w_1x_1 + w_2x_2 + w_0$$

- En d dimensiones:

- En 2 dimensiones:

$$y = w_{11}x_1^2 + w_{22}x_2^2 + w_{12}x_1x_2 + w_1x_1 + w_2x_2 + w_0$$

- En d dimensiones:

$$y = w_{11}x_1^2 + w_{22}x_2^2 + \cdots + w_{dd}x_d^2$$

- En 2 dimensiones:

$$y = w_{11}x_1^2 + w_{22}x_2^2 + w_{12}x_1x_2 + w_1x_1 + w_2x_2 + w_0$$

- En d dimensiones:

$$\begin{aligned} y =& w_{11}x_1^2 + w_{22}x_2^2 + \cdots + w_{dd}x_d^2 \\ & + w_{12}x_1x_2 + w_{13}x_1x_3 + \cdots + w_{23}x_2x_3 + \cdots + w_{(d-1)d}x_{d-1}x_d \end{aligned}$$

- En 2 dimensiones:

$$y = w_{11}x_1^2 + w_{22}x_2^2 + w_{12}x_1x_2 + w_1x_1 + w_2x_2 + w_0$$

- En d dimensiones:

$$\begin{aligned} y =& w_{11}x_1^2 + w_{22}x_2^2 + \cdots + w_{dd}x_d^2 \\ &+ w_{12}x_1x_2 + w_{13}x_1x_3 + \cdots + w_{23}x_2x_3 + \cdots + w_{(d-1)d}x_{d-1}x_d \\ &+ w_1x_1 + w_2x_2 + \cdots + w_dx_d + \end{aligned}$$

- En 2 dimensiones:

$$y = w_{11}x_1^2 + w_{22}x_2^2 + w_{12}x_1x_2 + w_1x_1 + w_2x_2 + w_0$$

- En d dimensiones:

$$\begin{aligned} y = & w_{11}x_1^2 + w_{22}x_2^2 + \cdots + w_{dd}x_d^2 \\ & + w_{12}x_1x_2 + w_{13}x_1x_3 + \cdots + w_{23}x_2x_3 + \cdots + w_{(d-1)d}x_{d-1}x_d \\ & + w_1x_1 + w_2x_2 + \cdots + w_dx_d + w_0 \end{aligned}$$

- En 2 dimensiones:

$$y = w_{11}x_1^2 + w_{22}x_2^2 + w_{12}x_1x_2 + w_1x_1 + w_2x_2 + w_0$$

- En d dimensiones:

$$\begin{aligned} y =& w_{11}x_1^2 + w_{22}x_2^2 + \cdots + w_{dd}x_d^2 \\ &+ w_{12}x_1x_2 + w_{13}x_1x_3 + \cdots + w_{23}x_2x_3 + \cdots + w_{(d-1)d}x_{d-1}x_d \\ &+ w_1x_1 + w_2x_2 + \cdots + w_dx_d + w_0 \end{aligned}$$

$$\binom{d+2}{2} \quad \text{términos}$$

$$y = w_1 x_1^2 + w_2 x_2^2 + w_3 x_1 x_2 + w_4 x_1 + w_5 x_2 + w_0$$

$$
\begin{aligned}
y &= w_1 x_1^2 + w_2 x_2^2 + w_3 x_1 x_2 + w_4 x_1 + w_5 x_2 + w_0 \\
&= w_1 z_1 + w_2 z_2 + w_3 z_3 + w_4 z_4 + w_5 z_5 + w_0
\end{aligned}
$$

$$
\begin{aligned}
y &= w_1 x_1^2 + w_2 x_2^2 + w_3 x_1 x_2 + w_4 x_1 + w_5 x_2 + w_0 \\
&= w_1 z_1 + w_2 z_2 + w_3 z_3 + w_4 z_4 + w_5 z_5 + w_0 \\
&= \mathbf{w}^T \mathbf{z} + w_0
\end{aligned}
$$

$$
\begin{aligned}
y &= w_1 x_1^2 + w_2 x_2^2 + w_3 x_1 x_2 + w_4 x_1 + w_5 x_2 + w_0 \\
&= w_1 z_1 + w_2 z_2 + w_3 z_3 + w_4 z_4 + w_5 z_5 + w_0 \\
&= \mathbf{w}^T \mathbf{z} + w_0
\end{aligned}
$$