

Redes Neuronales

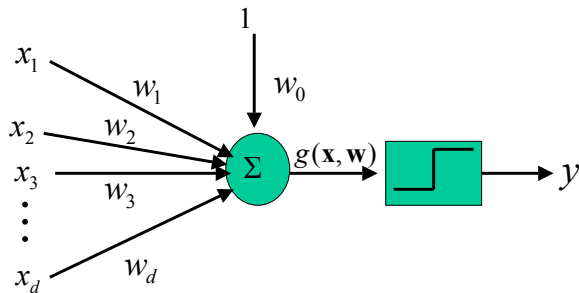
Fernando Lozano

Universidad de los Andes

19 de agosto de 2014

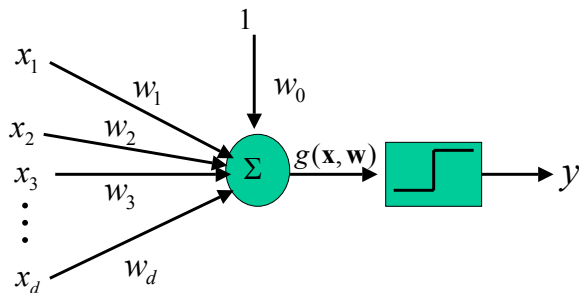


Modelo de una neurona



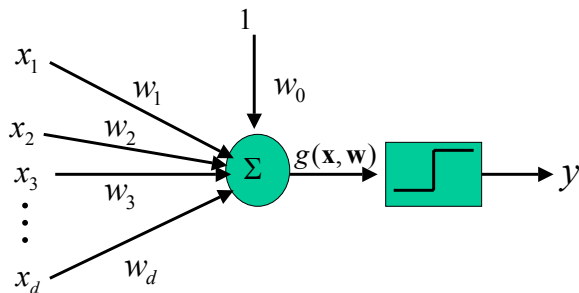
$$\mathbf{x} = [x_1 \ x_2 \ \dots \ x_d]^T$$

Modelo de una neurona



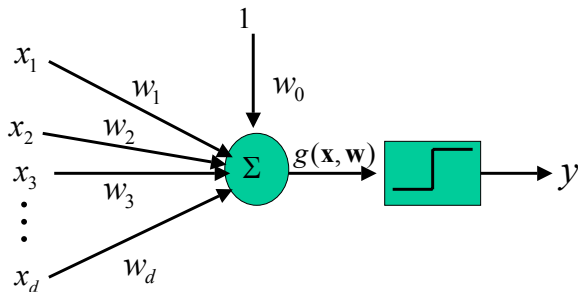
$$\mathbf{x} = [x_1 \quad x_2 \quad \dots \quad x_d]^T$$
$$\mathbf{w} = [w_1 \quad w_2 \quad \dots \quad w_d]^T$$

Modelo de una neurona



$$\begin{aligned}\mathbf{x} &= [x_1 \quad x_2 \quad \dots \quad x_d]^T \\ \mathbf{w} &= [w_1 \quad w_2 \quad \dots \quad w_d]^T \\ g(\mathbf{x}, \mathbf{w}) &= \mathbf{w}^T \mathbf{x} + w_0\end{aligned}$$

Modelo de una neurona



$$\begin{aligned}\mathbf{x} &= [x_1 \ x_2 \ \dots \ x_d]^T \\ \mathbf{w} &= [w_1 \ w_2 \ \dots \ w_d]^T \\ g(\mathbf{x}, \mathbf{w}) &= \mathbf{w}^T \mathbf{x} + w_0 \\ y &= f_{LD}(g(\mathbf{x})) = f_{LD}(\mathbf{w}^T \mathbf{x} + w_0)\end{aligned}$$

Vectores extendidos

$$\tilde{\mathbf{x}} = \begin{bmatrix} 1 & x_1 & x_2 & \dots & x_d \end{bmatrix}^T$$

Vectores extendidos

$$\begin{aligned}\tilde{\mathbf{x}} &= \begin{bmatrix} 1 & x_1 & x_2 & \dots & x_d \end{bmatrix}^T \\ \tilde{\mathbf{w}} &= \begin{bmatrix} w_0 & w_1 & w_2 & \dots & w_d \end{bmatrix}^T\end{aligned}$$

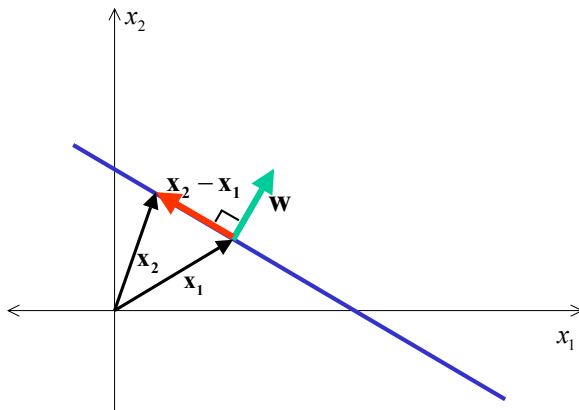
Vectores extendidos

$$\begin{aligned}\tilde{\mathbf{x}} &= \begin{bmatrix} 1 & x_1 & x_2 & \dots & x_d \end{bmatrix}^T \\ \tilde{\mathbf{w}} &= \begin{bmatrix} w_0 & w_1 & w_2 & \dots & w_d \end{bmatrix}^T \\ g(\mathbf{x}) &= \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}\end{aligned}$$

Vectores extendidos

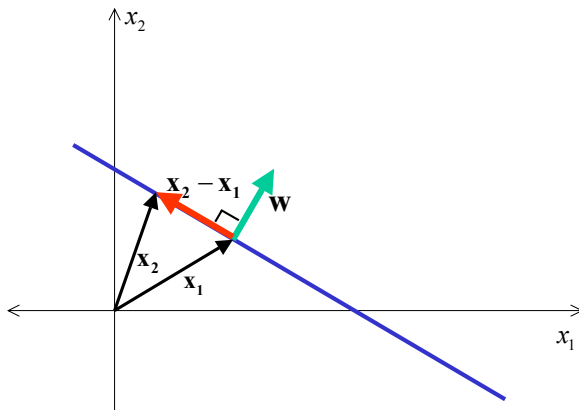
$$\begin{aligned}\tilde{\mathbf{x}} &= \begin{bmatrix} 1 & x_1 & x_2 & \dots & x_d \end{bmatrix}^T \\ \tilde{\mathbf{w}} &= \begin{bmatrix} w_0 & w_1 & w_2 & \dots & w_d \end{bmatrix}^T \\ g(\mathbf{x}) &= \tilde{\mathbf{w}}^T \tilde{\mathbf{x}} \\ u(\mathbf{x}) &= f_{LD}(g(\mathbf{x})) = f_{LD}(\tilde{\mathbf{w}}^T \tilde{\mathbf{x}})\end{aligned}$$

Interpretación geométrica 1



$$C = \{\mathbf{x} : \mathbf{w}^T \mathbf{x} + w_0 = 0\}$$

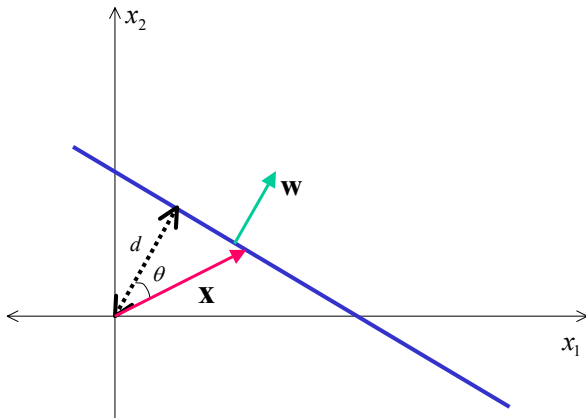
Interpretación geométrica 1



$$C = \{\mathbf{x} : \mathbf{w}^T \mathbf{x} + w_0 = 0\}$$

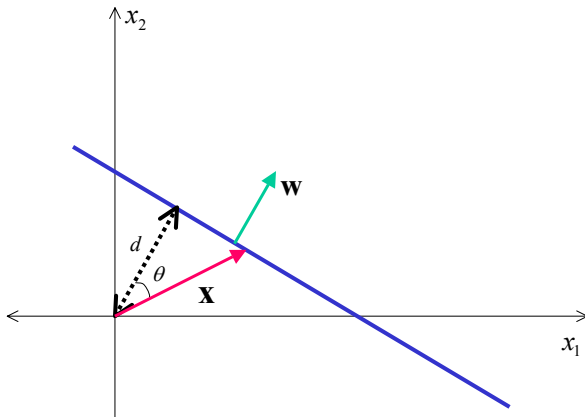
$$\mathbf{x}_1, \mathbf{x}_2 \in C \Rightarrow \mathbf{w}^T (\mathbf{x}_2 - \mathbf{x}_1) = 0$$

Interpretación geométrica 2



$$\mathbf{x} \in C \Rightarrow \mathbf{w}^T \mathbf{x} = \|\mathbf{w}\| \|\mathbf{x}\| \cos(\theta)$$

Interpretación geométrica 2



$$\begin{aligned}\mathbf{x} \in C \Rightarrow \mathbf{w}^T \mathbf{x} &= \|\mathbf{w}\| \|\mathbf{x}\| \cos(\theta) \\ &= \|\mathbf{w}\| d = -w_0 \Rightarrow d = \frac{-w_0}{\|\mathbf{w}\|}\end{aligned}$$

Cuándo puede ser este modelo óptimo?

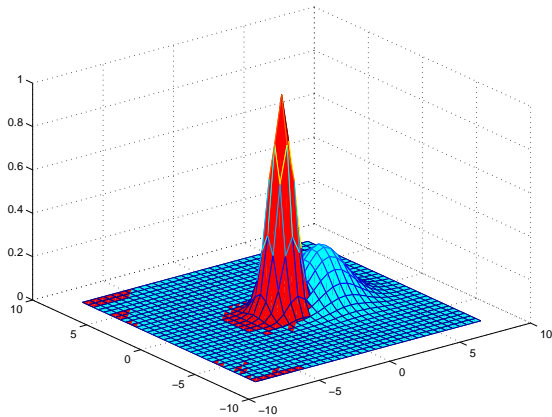
Cuándo puede ser este modelo óptimo?

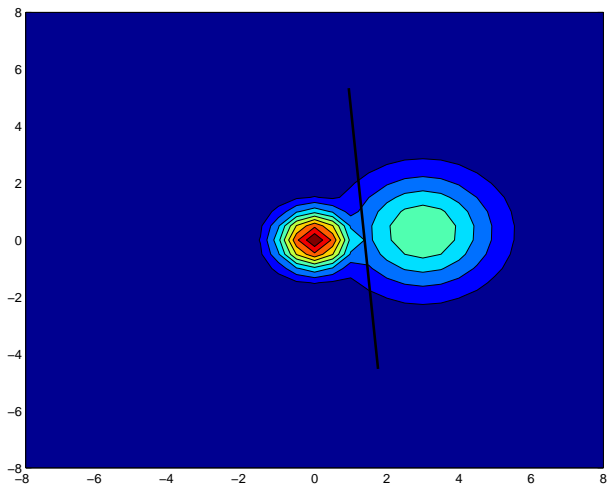
- Cuando $p_0(\mathbf{x}) \sim N(\mathbf{m}_0, \Sigma)$ y $p_1(\mathbf{x}) \sim N(\mathbf{m}_1, \Sigma)$ el clasificador óptimo de Bayes asigna \mathbf{x} a la clase 1 si:

Cuándo puede ser este modelo óptimo?

- Cuando $p_0(\mathbf{x}) \sim N(\mathbf{m}_0, \Sigma)$ y $p_1(\mathbf{x}) \sim N(\mathbf{m}_1, \Sigma)$ el clasificador óptimo de Bayes asigna \mathbf{x} a la clase 1 si:

$$\underbrace{(\mathbf{m}_1 - \mathbf{m}_0)^T \Sigma^{-1}}_{\mathbf{w}^T} \mathbf{x} > \underbrace{2 \ln \left(\frac{1 - \alpha}{\alpha} \right) + \frac{1}{2} (\mathbf{m}_1^T \Sigma^{-1} \mathbf{m}_1 - \mathbf{m}_0^T \Sigma^{-1} \mathbf{m}_0)}_{-w_0}$$





Problema de Aprendizaje

- Tenemos un conjunto de datos $\{\mathbf{x}_i, y_i\}, \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{0, 1\}$.

Problema de Aprendizaje

- Tenemos un conjunto de datos $\{\mathbf{x}_i, y_i\}$, $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{0, 1\}$.
- Queremos encontrar \mathbf{w} que nos de una buena regla de clasificación para datos futuros.

Problema de Aprendizaje

- Tenemos un conjunto de datos $\{\mathbf{x}_i, y_i\}$, $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{0, 1\}$.
- Queremos encontrar \mathbf{w} que nos de una buena regla de clasificación **para datos futuros**.
- Si $z_i = f_{LD}(g(\mathbf{w}, \mathbf{x}_i))$, el objetivo es minimizar:

$$\mathbb{P}\{y_i \neq z_i\}$$

Problema de Aprendizaje

- Tenemos un conjunto de datos $\{\mathbf{x}_i, y_i\}$, $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{0, 1\}$.
- Queremos encontrar \mathbf{w} que nos de una buena regla de clasificación **para datos futuros**.
- Si $z_i = f_{LD}(g(\mathbf{w}, \mathbf{x}_i))$, el objetivo es minimizar:

$$\mathbb{P}\{y_i \neq z_i\}$$

- En general, no podemos calcular $\mathbb{P}\{y_i \neq z_i\}$!

Problema de Aprendizaje

- Tenemos un conjunto de datos $\{\mathbf{x}_i, y_i\}, \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{0, 1\}$.
- Queremos encontrar \mathbf{w} que nos de una buena regla de clasificación **para datos futuros**.
- Si $z_i = f_{LD}(g(\mathbf{w}, \mathbf{x}_i))$, el objetivo es minimizar:

$$\mathbb{P}\{y_i \neq z_i\}$$

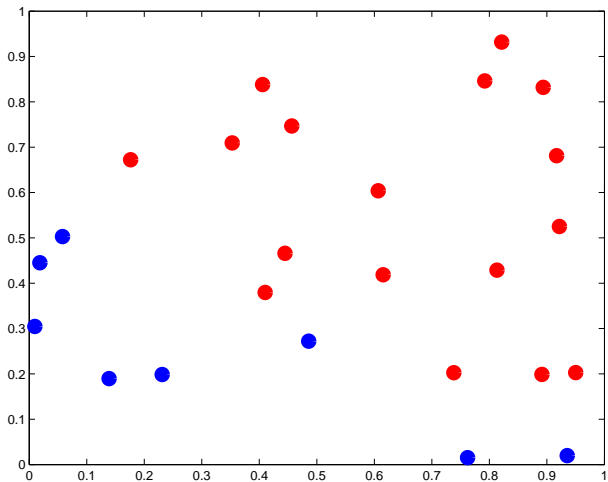
- En general, no podemos calcular $\mathbb{P}\{y_i \neq z_i\}$!
- Estrategia: minimizar función de error en los datos que sea calculable.

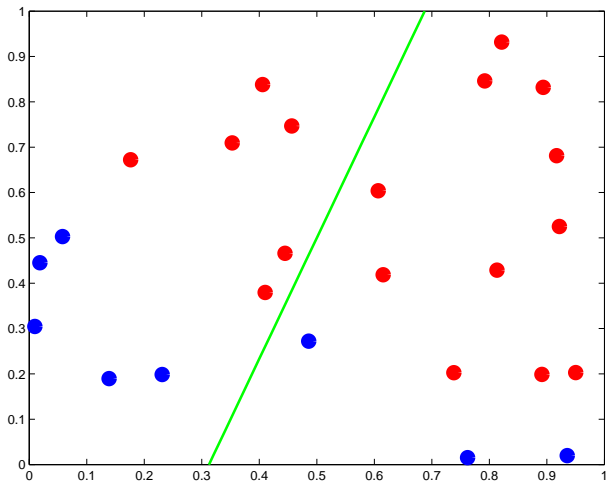
Problema de Aprendizaje

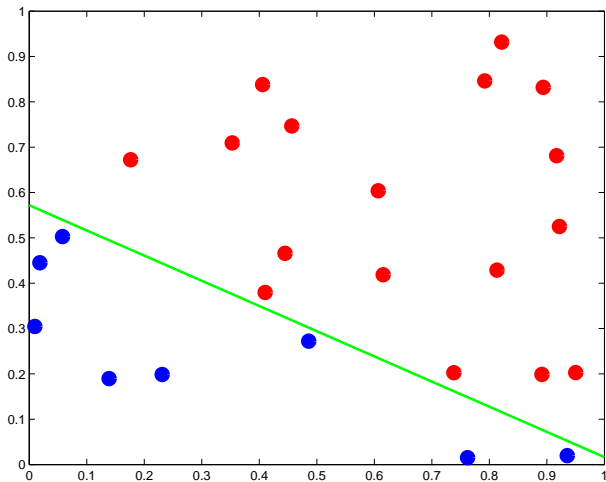
- Tenemos un conjunto de datos $\{\mathbf{x}_i, y_i\}, \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{0, 1\}$.
- Queremos encontrar \mathbf{w} que nos de una buena regla de clasificación **para datos futuros**.
- Si $z_i = f_{LD}(g(\mathbf{w}, \mathbf{x}_i))$, el objetivo es minimizar:

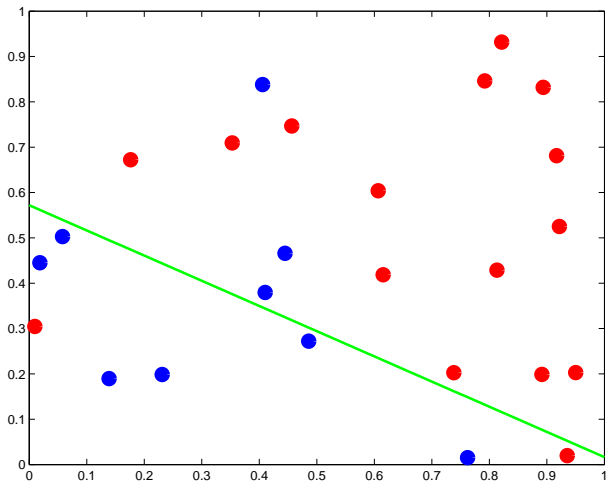
$$\mathbb{P}\{y_i \neq z_i\}$$

- En general, no podemos calcular $\mathbb{P}\{y_i \neq z_i\}$!
- Estrategia: minimizar función de error en los datos que sea calculable.









Algoritmos de entrenamiento

Algoritmos de entrenamiento

- Algoritmo LMS

Algoritmos de entrenamiento

- Algoritmo LMS
 - ▶ Widrow y Hoff (1960)

Algoritmos de entrenamiento

- Algoritmo LMS
 - ▶ Widrow y Hoff (1960)
 - ▶ Adaptive linear networks (ADALINE).

Algoritmos de entrenamiento

- Algoritmo LMS
 - ▶ Widrow y Hoff (1960)
 - ▶ Adaptive linear networks (ADALINE).
- Algoritmo del Perceptrón

Algoritmos de entrenamiento

- Algoritmo LMS
 - ▶ Widrow y Hoff (1960)
 - ▶ Adaptive linear networks (ADALINE).
- Algoritmo del Perceptrón
 - ▶ Rosemblatt (1962).

Algoritmos de entrenamiento

- Algoritmo LMS
 - ▶ Widrow y Hoff (1960)
 - ▶ Adaptive linear networks (ADALINE).
- Algoritmo del Perceptrón
 - ▶ Rosemblatt (1962).
 - ▶ Prueba de convergencia.

Algoritmos de entrenamiento

- Algoritmo LMS
 - ▶ Widrow y Hoff (1960)
 - ▶ Adaptive linear networks (ADALINE).
- Algoritmo del Perceptrón
 - ▶ Roseblatt (1962).
 - ▶ Prueba de convergencia.
- Perceptrón con bolsillo

Algoritmos de entrenamiento

- Algoritmo LMS
 - ▶ Widrow y Hoff (1960)
 - ▶ Adaptive linear networks (ADALINE).
- Algoritmo del Perceptrón
 - ▶ Roseblatt (1962).
 - ▶ Prueba de convergencia.
- Perceptrón con bolsillo
 - ▶ Gallant (1986)

Algoritmos de entrenamiento

- Algoritmo LMS
 - ▶ Widrow y Hoff (1960)
 - ▶ Adaptive linear networks (ADALINE).
- Algoritmo del Perceptrón
 - ▶ Roseblatt (1962).
 - ▶ Prueba de convergencia.
- Perceptrón con bolsillo
 - ▶ Gallant (1986)
 - ▶ Para datos no linealmente separables.

El algoritmo LMS

- Función de error:

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

El algoritmo LMS

- Función de error:

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

El algoritmo LMS

- Función de error:

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

- Problema de minimización:

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

El algoritmo LMS

- Función de error:

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

- Problema de minimización:

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

- Para el caso de una neurona, este problema admite una solución analítica.

Solución usando razonamiento geométrico

- Defina:

$$\mathbf{X} = [\mathbf{x}_1^T; \mathbf{x}_2^T; \cdots \mathbf{x}_n^T]$$

Solución usando razonamiento geométrico

- Defina:

$$\mathbf{X} = [\mathbf{x}_1^T; \mathbf{x}_2^T; \cdots \mathbf{x}_n^T]$$
$$\mathbf{y} = [y_1 \ y_2 \ \cdots \ y_n]^T$$

Solución usando razonamiento geométrico

- Defina:

$$\mathbf{X} = [\mathbf{x}_1^T; \mathbf{x}_2^T; \cdots \mathbf{x}_n^T]$$
$$\mathbf{y} = [y_1 \ y_2 \ \cdots \ y_n]^T$$

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

Solución usando razonamiento geométrico

- 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:

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

Solución usando razonamiento geométrico

- Defina:

$$\mathbf{X} = [\mathbf{x}_1^T; \mathbf{x}_2^T; \cdots \mathbf{x}_n^T]$$
$$\mathbf{y} = [y_1 \ y_2 \ \cdots \ y_n]^T$$

- 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} .

Solución usando razonamiento geométrico

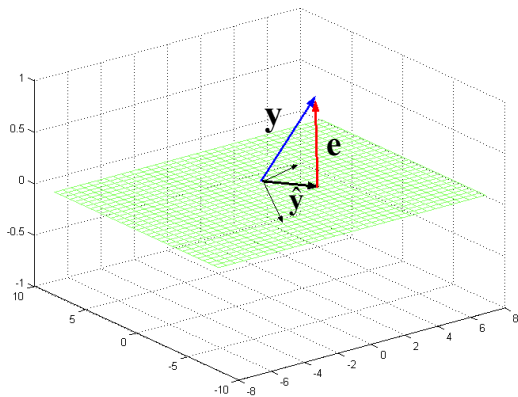
- Defina:

$$\mathbf{X} = [\mathbf{x}_1^T; \mathbf{x}_2^T; \cdots \mathbf{x}_n^T]$$
$$\mathbf{y} = [y_1 \ y_2 \ \cdots \ y_n]^T$$

- 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.



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

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

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

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

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

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

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

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\hat{\mathbf{w}} = \mathbf{X}^+\mathbf{y}$$

Optimización de la función de error

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

Optimización de la función de error

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

Optimización de la función de error

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

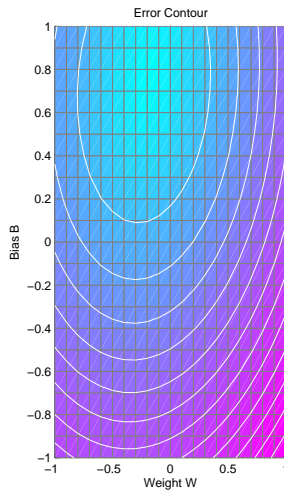
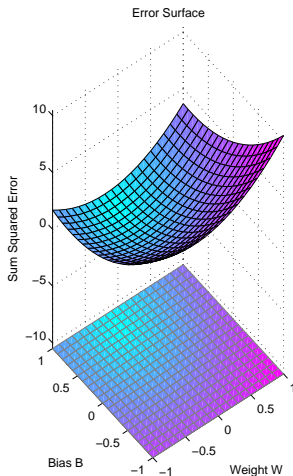
Optimización de la función de error

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

donde:

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

Función de error



Solución

Solución

- El mínimo de $E(\mathbf{w})$ cumple $\nabla_{\mathbf{w}} E = 0$:

Solución

- El mínimo de $E(\mathbf{w})$ cumple $\nabla_{\mathbf{w}}E = 0$:

$$\nabla_{\mathbf{w}}E = \mathbf{H}\mathbf{w} - \mathbf{b} = 0$$

Solución

- El mínimo de $E(\mathbf{w})$ cumple $\nabla_{\mathbf{w}}E = 0$:

$$\begin{aligned}\nabla_{\mathbf{w}}E &= \mathbf{H}\mathbf{w} - \mathbf{b} = 0 \\ \hat{\mathbf{w}} &= \mathbf{H}^{-1}\mathbf{b}\end{aligned}$$

Solución

- El mínimo de $E(\mathbf{w})$ cumple $\nabla_{\mathbf{w}}E = 0$:

$$\begin{aligned}\nabla_{\mathbf{w}}E &= \mathbf{H}\mathbf{w} - \mathbf{b} = 0 \\ \hat{\mathbf{w}} &= \mathbf{H}^{-1}\mathbf{b}\end{aligned}$$

- Se requiere invertir \mathbf{H} , que puede no ser invertible o ser mal condicionada.

Solución

- El mínimo de $E(\mathbf{w})$ cumple $\nabla_{\mathbf{w}}E = 0$:

$$\begin{aligned}\nabla_{\mathbf{w}}E &= \mathbf{H}\mathbf{w} - \mathbf{b} = 0 \\ \hat{\mathbf{w}} &= \mathbf{H}^{-1}\mathbf{b}\end{aligned}$$

- Se requiere invertir \mathbf{H} , que puede no ser invertible o ser mal condicionada.
- LMS: solución iterativa.

Algoritmo LMS

- Procedimiento iterativo:

Algoritmo LMS

- Procedimiento iterativo:
 - ➊ Comenzar en un punto (aleatorio):

$$\mathbf{w}_0 = \text{random}$$

Algoritmo LMS

- Procedimiento iterativo:

- 1 Comenzar en un punto (aleatorio):

$$\mathbf{w}_0 = \text{random}$$

- 2 Búsqueda de gradiente: Ir “hacia abajo de la colina”.

Algoritmo LMS

- Procedimiento iterativo:

- 1 Comenzar en un punto (aleatorio):

$$\mathbf{w}_0 = \text{random}$$

- 2 Búsqueda de gradiente: Ir “hacia abajo de la colina”.

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \mu \nabla_{\mathbf{w}} E|_{\mathbf{w}_k}$$

Algoritmo LMS

- Procedimiento iterativo:

- 1 Comenzar en un punto (aleatorio):

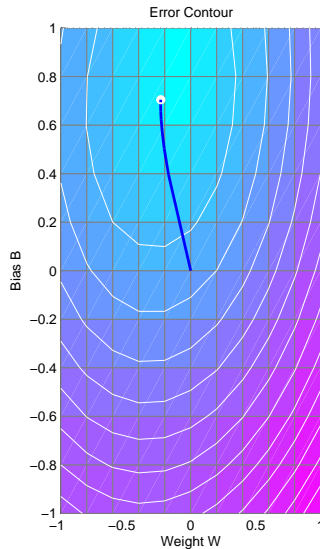
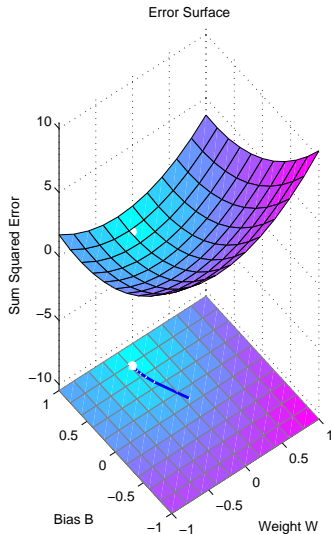
$$\mathbf{w}_0 = \text{random}$$

- 2 Búsqueda de gradiente: Ir “hacia abajo de la colina”.

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \mu \nabla_{\mathbf{w}} E|_{\mathbf{w}_k}$$

- $\nabla_{\mathbf{w}} E|_{\mathbf{w}_k}$ no se calcula exactamente.

Búsqueda iterativa



Estimación del gradiente

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

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

Estimación del gradiente

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

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

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

Estimación del gradiente

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

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

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

Estimación del gradiente

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

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

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

Estimación del gradiente

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

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

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

Inicialize \mathbf{w}_0 a valores pequeños.

repeat

 Escoja (\mathbf{x}_i, y_i)

Algoritmo LMS

Inicialize \mathbf{w}_0 a valores pequeños.

repeat

Escoja (\mathbf{x}_i, y_i)

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

Algoritmo LMS

Inicialize \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

Inicialize \mathbf{w}_0 a valores pequeños.

repeat

Escoja (\mathbf{x}_i, y_i)

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

$$e = g - y_i$$

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

Algoritmo LMS

Inicialize \mathbf{w}_0 a valores pequeños.

repeat

 Escoja (\mathbf{x}_i, y_i)

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

$$e = g - y_i$$

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

until Condición de terminación.

Condición de terminación

Condición de terminación

- Cuando el error sea suficientemente pequeño:

$$E \leq E_{min}$$

Condición de terminación

- Cuando el error sea suficientemente pequeño:

$$E \leq E_{min}$$

- Cuando el gradiente sea suficientemente cercano a cero:

$$\|\nabla_{\mathbf{w}} E\| \leq g_{\min}$$

Condición de terminación

- Cuando el error sea suficientemente pequeño:

$$E \leq E_{min}$$

- Cuando el gradiente sea suficientemente cercano a cero:

$$\|\nabla_{\mathbf{w}} E\| \leq g_{\text{mín}}$$

- Validación cruzada.

Condición de terminación

- Cuando el error sea suficientemente pequeño:

$$E \leq E_{min}$$

- Cuando el gradiente sea suficientemente cercano a cero:

$$\|\nabla_{\mathbf{w}} E\| \leq g_{\text{mín}}$$

- Validación cruzada.

El perceptrón

- Conjunto de datos:

$$\{\mathbf{x}_i, y_i\}, \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$$

El perceptrón

- Conjunto de datos:

$$\{\mathbf{x}_i, y_i\}, \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$$

- (\mathbf{x}_i, y_i) es clasificado correctamente si:

$$g(\mathbf{w}, \mathbf{x}_i)y_i > 0$$

El perceptrón

- Conjunto de datos:

$$\{\mathbf{x}_i, y_i\}, \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$$

- (\mathbf{x}_i, y_i) es clasificado correctamente si:

$$g(\mathbf{w}, \mathbf{x}_i)y_i > 0$$

$$(\mathbf{w}^T \mathbf{x}_i)y_i > 0$$

El perceptrón

- Conjunto de datos:

$$\{\mathbf{x}_i, y_i\}, \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$$

- (\mathbf{x}_i, y_i) es clasificado correctamente si:

$$g(\mathbf{w}, \mathbf{x}_i)y_i > 0$$

$$(\mathbf{w}^T \mathbf{x}_i)y_i > 0$$

- Criterio de error del perceptrón:

El perceptrón

- Conjunto de datos:

$$\{\mathbf{x}_i, y_i\}, \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$$

- (\mathbf{x}_i, y_i) es clasificado correctamente si:

$$g(\mathbf{w}, \mathbf{x}_i)y_i > 0$$

$$(\mathbf{w}^T \mathbf{x}_i)y_i > 0$$

- Criterio de error del perceptrón:

$$E(\mathbf{w}) = - \sum_{\mathbf{x}_i \in \mathcal{M}} (\mathbf{w}^T \mathbf{x}_i)y_i, \quad \mathcal{M} = \{\mathbf{x}_i : (\mathbf{w}^T \mathbf{x}_i)y_i < 0\}$$

Criterio de error del perceptrón

- $E(\mathbf{w})$ es una función lineal a trozos.

Criterio de error del perceptrón

- $E(\mathbf{w})$ es una función lineal a trozos.
- $E(\mathbf{w})$ es una función continua.

Criterio de error del perceptrón

- $E(\mathbf{w})$ es una función lineal a trozos.
- $E(\mathbf{w})$ es una función continua.
- $\nabla_{\mathbf{w}} E$ es una función discontinua.

Criterio de error del perceptrón

- $E(\mathbf{w})$ es una función lineal a trozos.
- $E(\mathbf{w})$ es una función continua.
- $\nabla_{\mathbf{w}}E$ es una función discontinua.
- Sin embargo, en los puntos donde es continua, $-\nabla_{\mathbf{w}}E$ es una dirección de descenso en la superficie de error.

Algoritmo del perceptrón

- Procedimiento iterativo:

Algoritmo del perceptrón

- Procedimiento iterativo:

- 1 Comenzar en:

$$\mathbf{w}_0 = 0$$

Algoritmo del perceptrón

- Procedimiento iterativo:

- ① Comenzar en:

$$\mathbf{w}_0 = 0$$

- ② Búsqueda de gradiente: Ir “hacia abajo de la colina”.

Algoritmo del perceptrón

- Procedimiento iterativo:

- ① Comenzar en:

$$\mathbf{w}_0 = 0$$

- ② Búsqueda de gradiente: Ir “hacia abajo de la colina”.

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \nabla_{\mathbf{w}} E|_{\mathbf{w}_k}$$

Algoritmo del perceptrón

- Procedimiento iterativo:

- 1 Comenzar en:

$$\mathbf{w}_0 = 0$$

- 2 Búsqueda de gradiente: Ir “hacia abajo de la colina”.

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \nabla_{\mathbf{w}} E|_{\mathbf{w}_k}$$

- Nuevamente, $\nabla_{\mathbf{w}} E|_{\mathbf{w}_k}$ no se calcula exactamente:

Algoritmo del perceptrón

- Procedimiento iterativo:

- 1 Comenzar en:

$$\mathbf{w}_0 = 0$$

- 2 Búsqueda de gradiente: Ir “hacia abajo de la colina”.

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \nabla_{\mathbf{w}} E|_{\mathbf{w}_k}$$

- Nuevamente, $\nabla_{\mathbf{w}} E|_{\mathbf{w}_k}$ no se calcula exactamente:

$$\nabla_{\mathbf{w}} E = - \sum_{\mathbf{x}_i \in \mathcal{M}} \mathbf{x}_i y_i$$

Algoritmo del perceptrón

- Procedimiento iterativo:

- 1 Comenzar en:

$$\mathbf{w}_0 = 0$$

- 2 Búsqueda de gradiente: Ir “hacia abajo de la colina”.

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \nabla_{\mathbf{w}} E|_{\mathbf{w}_k}$$

- Nuevamente, $\nabla_{\mathbf{w}} E|_{\mathbf{w}_k}$ no se calcula exactamente:

$$\begin{aligned}\nabla_{\mathbf{w}} E &= - \sum_{\mathbf{x}_i \in \mathcal{M}} \mathbf{x}_i y_i \\ &\approx -\mathbf{x}_i y_i\end{aligned}$$

Algoritmo del perceptrón

Incialize $\mathbf{w}_0 = 0$

Algoritmo del perceptrón

Incialize $\mathbf{w}_0 = 0$

repeat

 Escoja (\mathbf{x}_i, y_i) al azar

Algoritmo del perceptrón

Incialize $\mathbf{w}_0 = 0$

repeat

Escoja (\mathbf{x}_i, y_i) al azar

if $(\mathbf{w}^T \mathbf{x}_i)y_i < 0$ **then**

Algoritmo del perceptrón

Incialize $\mathbf{w}_0 = 0$

repeat

Escoja (\mathbf{x}_i, y_i) al azar

if $(\mathbf{w}^T \mathbf{x}_i)y_i < 0$ **then**

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \mathbf{x}_i y_i$$

Algoritmo del perceptrón

Incialize $\mathbf{w}_0 = 0$

repeat

Escoja (\mathbf{x}_i, y_i) al azar

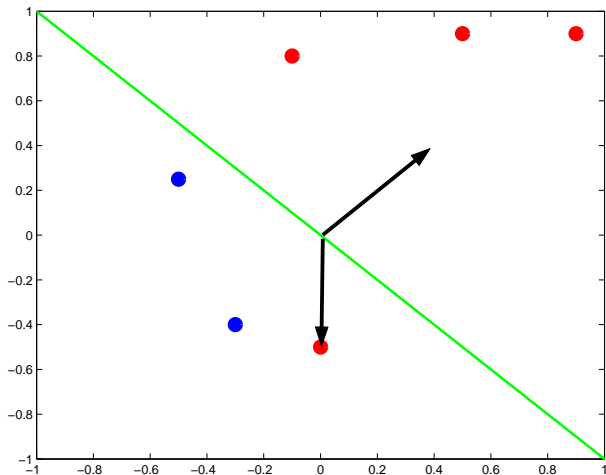
if $(\mathbf{w}^T \mathbf{x}_i)y_i < 0$ **then**

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \mathbf{x}_i y_i$$

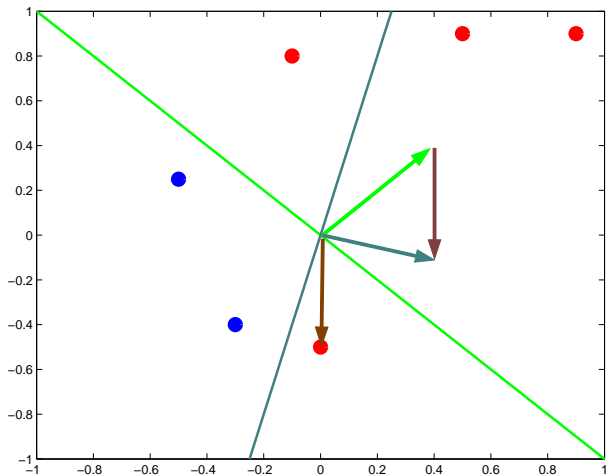
end if

until Convergencia.

Interpretación geométrica



Interpretación geométrica



Convergencia

Teorema

Sponga:

Convergencia

Teorema

Suponga:

- $\|\mathbf{x}_i\| \leq K \in \mathbb{R}, \quad i = 1 \dots, n.$

Convergencia

Teorema

Suponga:

- $\|\mathbf{x}_i\| \leq K \in \mathbb{R}, \quad i = 1 \dots, n.$
- $\exists \hat{\mathbf{w}} \in \mathbb{R}^{d+1}, \delta > 0 \quad \text{tal que} \quad \hat{\mathbf{w}}^T \mathbf{x}_i \geq \delta \quad i = 1, \dots, n.$

Convergencia

Teorema

Suponga:

- $\|\mathbf{x}_i\| \leq K \in \mathbb{R}, \quad i = 1 \dots, n.$
- $\exists \hat{\mathbf{w}} \in \mathbb{R}^{d+1}, \delta > 0 \quad \text{tal que} \quad \hat{\mathbf{w}}^T \mathbf{x}_i \geq \delta \quad i = 1, \dots, n.$

Entonces el algoritmo del perceptrón ejecuta el paso de actualización a lo sumo $\left(\frac{K\|\hat{\mathbf{w}}\|}{\delta}\right)^2$ veces.

Convergencia

Teorema

Suponga:

- $\|\mathbf{x}_i\| \leq K \in \mathbb{R}, \quad i = 1 \dots, n.$
- $\exists \hat{\mathbf{w}} \in \mathbb{R}^{d+1}, \delta > 0 \quad \text{tal que} \quad \hat{\mathbf{w}}^T \mathbf{x}_i \geq \delta \quad i = 1, \dots, n.$

Entonces el algoritmo del perceptrón ejecuta el paso de actualización a lo sumo $\left(\frac{K\|\hat{\mathbf{w}}\|}{\delta}\right)^2$ veces.

- Es decir, para datos linealmente separables, el algoritmo del perceptrón converge en un número finito de pasos.

Demostración

Demostración

- S.P.D.G Cambie los \mathbf{x}_i para los cuales $\hat{\mathbf{w}}^T \mathbf{x}_i < 0$ por $-\mathbf{x}_i$. Con este cambio $\mathbf{w}^T \mathbf{x}_i > 0$ indica clasificación correcta.

Demostración

- S.P.D.G Cambie los \mathbf{x}_i para los cuales $\hat{\mathbf{w}}^T \mathbf{x}_i < 0$ por $-\mathbf{x}_i$. Con este cambio $\mathbf{w}^T \mathbf{x}_i > 0$ indica clasificación correcta.
- Sea t el número de correcciones a \mathbf{w} . Con $\mathbf{w}_0 = 0$ tenemos:

Demostración

- S.P.D.G Cambie los \mathbf{x}_i para los cuales $\hat{\mathbf{w}}^T \mathbf{x}_i < 0$ por $-\mathbf{x}_i$. Con este cambio $\mathbf{w}^T \mathbf{x}_i > 0$ indica clasificación correcta.
- Sea t el número de correcciones a \mathbf{w} . Con $\mathbf{w}_0 = 0$ tenemos:

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \mathbf{x}_i$$

Demostración

- S.P.D.G Cambie los \mathbf{x}_i para los cuales $\hat{\mathbf{w}}^T \mathbf{x}_i < 0$ por $-\mathbf{x}_i$. Con este cambio $\mathbf{w}^T \mathbf{x}_i > 0$ indica clasificación correcta.
- Sea t el número de correcciones a \mathbf{w} . Con $\mathbf{w}_0 = 0$ tenemos:

$$\begin{aligned}\mathbf{w}_t &= \mathbf{w}_{t-1} + \mathbf{x}_i \\ \hat{\mathbf{w}}^T \mathbf{w}_t &= \hat{\mathbf{w}}^T \mathbf{w}_{t-1} + \hat{\mathbf{w}}^T \mathbf{x}_i \geq \hat{\mathbf{w}}^T \mathbf{w}_{t-1} + \delta\end{aligned}$$

Demostración

- S.P.D.G Cambie los \mathbf{x}_i para los cuales $\hat{\mathbf{w}}^T \mathbf{x}_i < 0$ por $-\mathbf{x}_i$. Con este cambio $\mathbf{w}^T \mathbf{x}_i > 0$ indica clasificación correcta.
- Sea t el número de correcciones a \mathbf{w} . Con $\mathbf{w}_0 = 0$ tenemos:

$$\begin{aligned}\mathbf{w}_t &= \mathbf{w}_{t-1} + \mathbf{x}_i \\ \hat{\mathbf{w}}^T \mathbf{w}_t &= \hat{\mathbf{w}}^T \mathbf{w}_{t-1} + \hat{\mathbf{w}}^T \mathbf{x}_i \geq \hat{\mathbf{w}}^T \mathbf{w}_{t-1} + \delta \\ &\vdots\end{aligned}$$

Demostración

- S.P.D.G Cambie los \mathbf{x}_i para los cuales $\hat{\mathbf{w}}^T \mathbf{x}_i < 0$ por $-\mathbf{x}_i$. Con este cambio $\mathbf{w}^T \mathbf{x}_i > 0$ indica clasificación correcta.
- Sea t el número de correcciones a \mathbf{w} . Con $\mathbf{w}_0 = 0$ tenemos:

$$\begin{aligned}\mathbf{w}_t &= \mathbf{w}_{t-1} + \mathbf{x}_i \\ \hat{\mathbf{w}}^T \mathbf{w}_t &= \hat{\mathbf{w}}^T \mathbf{w}_{t-1} + \hat{\mathbf{w}}^T \mathbf{x}_i \geq \hat{\mathbf{w}}^T \mathbf{w}_{t-1} + \delta \\ &\vdots \\ \hat{\mathbf{w}}^T \mathbf{w}_t &\geq (t\delta)\end{aligned}$$

Demostración

- S.P.D.G Cambie los \mathbf{x}_i para los cuales $\hat{\mathbf{w}}^T \mathbf{x}_i < 0$ por $-\mathbf{x}_i$. Con este cambio $\mathbf{w}^T \mathbf{x}_i > 0$ indica clasificación correcta.
- Sea t el número de correcciones a \mathbf{w} . Con $\mathbf{w}_0 = 0$ tenemos:

$$\begin{aligned}\mathbf{w}_t &= \mathbf{w}_{t-1} + \mathbf{x}_i \\ \hat{\mathbf{w}}^T \mathbf{w}_t &= \hat{\mathbf{w}}^T \mathbf{w}_{t-1} + \hat{\mathbf{w}}^T \mathbf{x}_i \geq \hat{\mathbf{w}}^T \mathbf{w}_{t-1} + \delta \\ &\vdots \\ \hat{\mathbf{w}}^T \mathbf{w}_t &\geq (t\delta)\end{aligned}$$

- Similarmente:

Demostración

- S.P.D.G Cambie los \mathbf{x}_i para los cuales $\hat{\mathbf{w}}^T \mathbf{x}_i < 0$ por $-\mathbf{x}_i$. Con este cambio $\mathbf{w}^T \mathbf{x}_i > 0$ indica clasificación correcta.
- Sea t el número de correcciones a \mathbf{w} . Con $\mathbf{w}_0 = 0$ tenemos:

$$\begin{aligned}\mathbf{w}_t &= \mathbf{w}_{t-1} + \mathbf{x}_i \\ \hat{\mathbf{w}}^T \mathbf{w}_t &= \hat{\mathbf{w}}^T \mathbf{w}_{t-1} + \hat{\mathbf{w}}^T \mathbf{x}_i \geq \hat{\mathbf{w}}^T \mathbf{w}_{t-1} + \delta \\ &\vdots \\ \hat{\mathbf{w}}^T \mathbf{w}_t &\geq (t\delta)\end{aligned}$$

- Similarmente:

$$\|\mathbf{w}_t\|^2 = \mathbf{w}_t^T \mathbf{w}_t = (\mathbf{w}_{t-1} + \mathbf{x}_i)^T (\mathbf{w}_{t-1} + \mathbf{x}_i)$$

Demostración

- S.P.D.G Cambie los \mathbf{x}_i para los cuales $\hat{\mathbf{w}}^T \mathbf{x}_i < 0$ por $-\mathbf{x}_i$. Con este cambio $\mathbf{w}^T \mathbf{x}_i > 0$ indica clasificación correcta.
- Sea t el número de correcciones a \mathbf{w} . Con $\mathbf{w}_0 = 0$ tenemos:

$$\begin{aligned}\mathbf{w}_t &= \mathbf{w}_{t-1} + \mathbf{x}_i \\ \hat{\mathbf{w}}^T \mathbf{w}_t &= \hat{\mathbf{w}}^T \mathbf{w}_{t-1} + \hat{\mathbf{w}}^T \mathbf{x}_i \geq \hat{\mathbf{w}}^T \mathbf{w}_{t-1} + \delta \\ &\vdots \\ \hat{\mathbf{w}}^T \mathbf{w}_t &\geq (t\delta)\end{aligned}$$

- Similarmente:

$$\begin{aligned}\|\mathbf{w}_t\|^2 &= \mathbf{w}_t^T \mathbf{w}_t = (\mathbf{w}_{t-1} + \mathbf{x}_i)^T (\mathbf{w}_{t-1} + \mathbf{x}_i) \\ &= \|\mathbf{w}_{t-1}\|^2 + 2\mathbf{w}_{t-1}^T \mathbf{x}_i + \|\mathbf{x}_i\|^2 \leq \|\mathbf{w}_{t-1}\|^2 + K^2\end{aligned}$$

Demostración

- S.P.D.G Cambie los \mathbf{x}_i para los cuales $\hat{\mathbf{w}}^T \mathbf{x}_i < 0$ por $-\mathbf{x}_i$. Con este cambio $\mathbf{w}^T \mathbf{x}_i > 0$ indica clasificación correcta.
- Sea t el número de correcciones a \mathbf{w} . Con $\mathbf{w}_0 = 0$ tenemos:

$$\begin{aligned}\mathbf{w}_t &= \mathbf{w}_{t-1} + \mathbf{x}_i \\ \hat{\mathbf{w}}^T \mathbf{w}_t &= \hat{\mathbf{w}}^T \mathbf{w}_{t-1} + \hat{\mathbf{w}}^T \mathbf{x}_i \geq \hat{\mathbf{w}}^T \mathbf{w}_{t-1} + \delta \\ &\vdots \\ \hat{\mathbf{w}}^T \mathbf{w}_t &\geq (t\delta)\end{aligned}$$

- Similarmente:

$$\begin{aligned}\|\mathbf{w}_t\|^2 &= \mathbf{w}_t^T \mathbf{w}_t = (\mathbf{w}_{t-1} + \mathbf{x}_i)^T (\mathbf{w}_{t-1} + \mathbf{x}_i) \\ &= \|\mathbf{w}_{t-1}\|^2 + 2\mathbf{w}_{t-1}^T \mathbf{x}_i + \|\mathbf{x}_i\|^2 \leq \|\mathbf{w}_{t-1}\|^2 + K^2 \\ &\vdots\end{aligned}$$

Demostración

- S.P.D.G Cambie los \mathbf{x}_i para los cuales $\hat{\mathbf{w}}^T \mathbf{x}_i < 0$ por $-\mathbf{x}_i$. Con este cambio $\mathbf{w}^T \mathbf{x}_i > 0$ indica clasificación correcta.
- Sea t el número de correcciones a \mathbf{w} . Con $\mathbf{w}_0 = 0$ tenemos:

$$\begin{aligned}\mathbf{w}_t &= \mathbf{w}_{t-1} + \mathbf{x}_i \\ \hat{\mathbf{w}}^T \mathbf{w}_t &= \hat{\mathbf{w}}^T \mathbf{w}_{t-1} + \hat{\mathbf{w}}^T \mathbf{x}_i \geq \hat{\mathbf{w}}^T \mathbf{w}_{t-1} + \delta \\ &\vdots \\ \hat{\mathbf{w}}^T \mathbf{w}_t &\geq (t\delta)\end{aligned}$$

- Similarmente:

$$\begin{aligned}\|\mathbf{w}_t\|^2 &= \mathbf{w}_t^T \mathbf{w}_t = (\mathbf{w}_{t-1} + \mathbf{x}_i)^T (\mathbf{w}_{t-1} + \mathbf{x}_i) \\ &= \|\mathbf{w}_{t-1}\|^2 + 2\mathbf{w}_{t-1}^T \mathbf{x}_i + \|\mathbf{x}_i\|^2 \leq \|\mathbf{w}_{t-1}\|^2 + K^2 \\ &\vdots \\ \|\mathbf{w}_t\|^2 &\leq tK^2\end{aligned}$$

- Combinando:

- Combinando:

$$t\delta \leq \hat{\mathbf{w}}^T \mathbf{w}_t$$

- Combinando:

$$\begin{aligned} t\delta &\leq \hat{\mathbf{w}}^T \mathbf{w}_t \\ &= \|\hat{\mathbf{w}}\| \|\mathbf{w}_t\| \cos\theta \end{aligned}$$

- Combinando:

$$\begin{aligned} t\delta &\leq \hat{\mathbf{w}}^T \mathbf{w}_t \\ &= \|\hat{\mathbf{w}}\| \|\mathbf{w}_t\| \cos\theta \\ &\leq \|\hat{\mathbf{w}}\| \|\mathbf{w}_t\| \end{aligned}$$

- Combinando:

$$\begin{aligned} t\delta &\leq \hat{\mathbf{w}}^T \mathbf{w}_t \\ &= \|\hat{\mathbf{w}}\| \|\mathbf{w}_t\| \cos\theta \\ &\leq \|\hat{\mathbf{w}}\| \|\mathbf{w}_t\| \\ &\leq \|\hat{\mathbf{w}}\| K\sqrt{t} \end{aligned}$$

- Combinando:

$$\begin{aligned} t\delta &\leq \hat{\mathbf{w}}^T \mathbf{w}_t \\ &= \|\hat{\mathbf{w}}\| \|\mathbf{w}_t\| \cos\theta \\ &\leq \|\hat{\mathbf{w}}\| \|\mathbf{w}_t\| \\ &\leq \|\hat{\mathbf{w}}\| K\sqrt{t} \\ t &\leq \left(\frac{K\|\hat{\mathbf{w}}\|}{\delta} \right)^2 \end{aligned}$$

- Combinando:

$$\begin{aligned} t\delta &\leq \hat{\mathbf{w}}^T \mathbf{w}_t \\ &= \|\hat{\mathbf{w}}\| \|\mathbf{w}_t\| \cos\theta \\ &\leq \|\hat{\mathbf{w}}\| \|\mathbf{w}_t\| \\ &\leq \|\hat{\mathbf{w}}\| K\sqrt{t} \\ t &\leq \left(\frac{K\|\hat{\mathbf{w}}\|}{\delta} \right)^2 \end{aligned}$$



Observaciones

Observaciones

- Convergencia (lo cual es bueno...)

Observaciones

- Convergencia (lo cual es bueno...).
- En la práctica no es posible determinar número de iteraciones antes de ejecutar el algoritmo.

Observaciones

- Convergencia (lo cual es bueno...).
- En la práctica no es posible determinar número de iteraciones antes de ejecutar el algoritmo.
- Para datos que son “menos” separables (δ pequeño) t es grande.

Observaciones

- Convergencia (lo cual es bueno...).
- En la práctica no es posible determinar número de iteraciones antes de ejecutar el algoritmo.
- Para datos que son “menos” separables (δ pequeño) t es grande.
- Condición de separabilidad lineal:

Observaciones

- Convergencia (lo cual es bueno...).
- En la práctica no es posible determinar número de iteraciones antes de ejecutar el algoritmo.
- Para datos que son “menos” separables (δ pequeño) t es grande.
- Condición de separabilidad lineal:
 - ▶ Ruido.

Observaciones

- Convergencia (lo cual es bueno...).
- En la práctica no es posible determinar número de iteraciones antes de ejecutar el algoritmo.
- Para datos que son “menos” separables (δ pequeño) t es grande.
- Condición de separabilidad lineal:
 - ▶ Ruido.
 - ▶ Generalización?

Observaciones

- Convergencia (lo cual es bueno...).
- En la práctica no es posible determinar número de iteraciones antes de ejecutar el algoritmo.
- Para datos que son “menos” separables (δ pequeño) t es grande.
- Condición de separabilidad lineal:
 - ▶ Ruido.
 - ▶ Generalización?
 - ▶ Cuando no hay separabilidad, determinar máximo conjunto separable es un problema difícil (NP completo).

Observaciones

- Convergencia (lo cual es bueno...).
- En la práctica no es posible determinar número de iteraciones antes de ejecutar el algoritmo.
- Para datos que son “menos” separables (δ pequeño) t es grande.
- Condición de separabilidad lineal:
 - ▶ Ruido.
 - ▶ Generalización?
 - ▶ Cuando no hay separabilidad, determinar máximo conjunto separable es un problema difícil (NP completo).
- Cuando no hay separabilidad, el algoritmo oscila y no termina.