Regresión logística

Fernando Lozano

Universidad de los Andes

29 de agosto de 2022

Problema de Clasificación binaria

- Datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$:

Problema de Clasificación binaria

- Datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$:
  - $\mathbf{x}_i$ objeto a clasificar.

Problema de Clasificación binaria

- Datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$:
  - $\mathbf{x}_i$ objeto a clasificar.
  - $y_i \in \{-1, 1\}$ etiqueta.

Problema de Clasificación binaria

- Datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$:
  - $\mathbf{x}_i$ objeto a clasificar.
  - $y_i \in \{-1,\ 1\}$ etiqueta.
- Queremos aprender regla de clasificación a partir de los datos:

Problema de Clasificación binaria

- Datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$:
  - $\mathbf{x}_i$ objeto a clasificar.
  - $y_i \in \{-1,\ 1\}$ etiqueta.
- Queremos aprender regla de clasificación a partir de los datos:

$$f\ :\ \mathcal{X} \longrightarrow \{-1,\ 1\}$$

Problema de Clasificación binaria

- Datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$:
  - $\mathbf{x}_i$ objeto a clasificar.
  - $y_i \in \{-1,\ 1\}$ etiqueta.
- Queremos aprender regla de clasificación a partir de los datos:

$$f\ :\ \mathcal{X} \longrightarrow \{-1,\ 1\}$$

- Separador lineal:

$$f(\mathbf{x}) = \operatorname{sign}(\mathbf{w}^T\mathbf{x})$$

Problema de Clasificación binaria

- Datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$:
  - $\mathbf{x}_i$ objeto a clasificar.
  - $y_i \in \{-1,\ 1\}$ etiqueta.
- Queremos aprender regla de clasificación a partir de los datos:

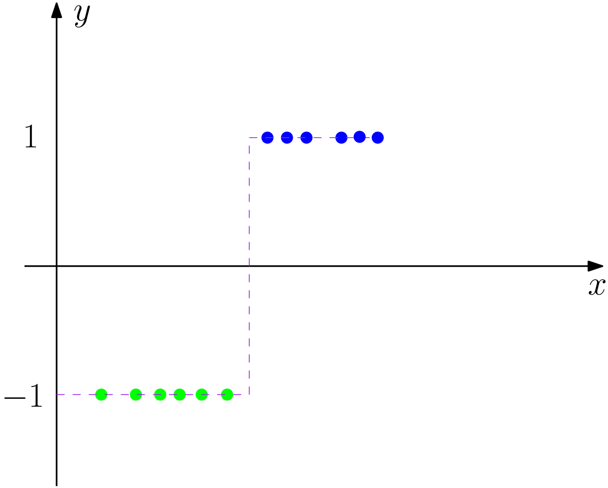
$$f\ :\ \mathcal{X} \longrightarrow \{-1,\ 1\}$$

- Separador lineal:

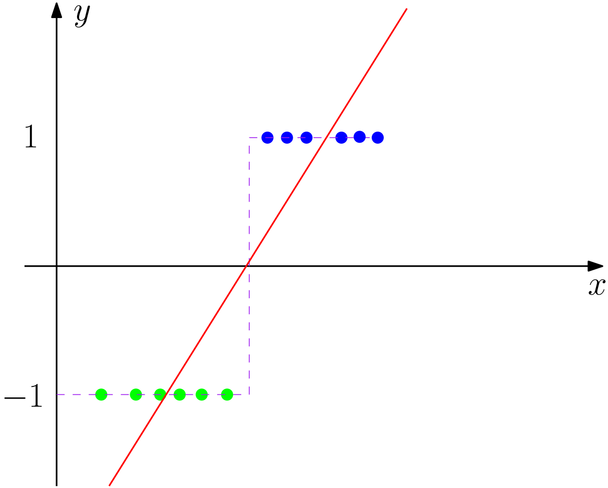
$$f(\mathbf{x}) = \operatorname{sign}(\mathbf{w}^T\mathbf{x})$$

- Cómo encontrar un buen clasificador?

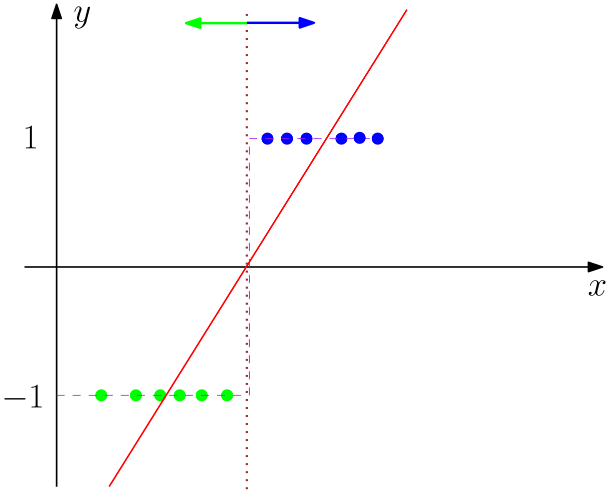
Regresión lineal + umbral



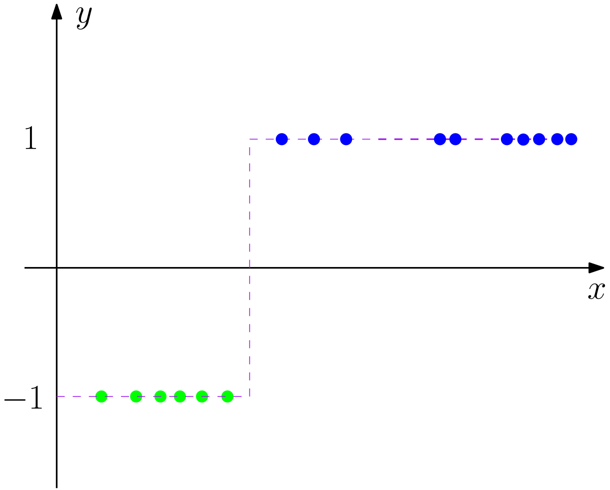
Regresión lineal + umbral



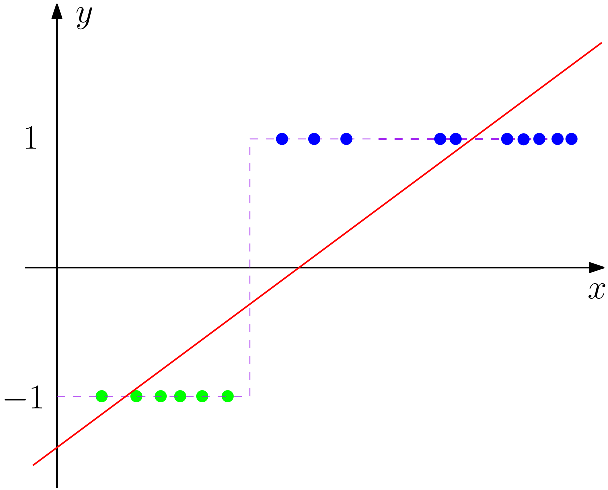
Regresión lineal + umbral



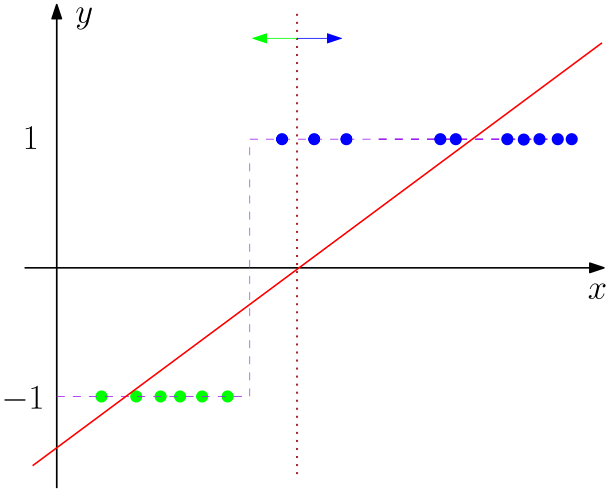
Regresión lineal + umbral



Regresión lineal + umbral



Regresión lineal + umbral



Modelo

Modelo

- Clasificación binaria: $y \in \{0, 1\}$.

Modelo

- Clasificación binaria: $y \in \{0, 1\}$.
- Queremos restringir $y \in [0\ ,\ 1]$

Modelo

- Clasificación binaria: $y \in \{0, 1\}$.
- Queremos restringir $y \in [0\ ,\ 1]$
- Usar función de activación:

Modelo

- Clasificación binaria: $y \in \{0, 1\}$.
- Queremos restringir $y \in [0\ ,\ 1]$
- Usar función de activación:

$$y = g(\mathbf{w}^T \mathbf{x})$$

Modelo

- Clasificación binaria: $y \in \{0, 1\}$.
- Queremos restringir $y \in [0\ ,\ 1]$
- Usar función de activación:

$$y = g(\mathbf{w}^T \mathbf{x})$$

donde:

Modelo

- Clasificación binaria: $y \in \{0, 1\}$.
- Queremos restringir $y \in [0\ ,\ 1]$
- Usar función de activación:

$$y = g(\mathbf{w}^T\mathbf{x})$$

donde:

  - $g\ :\ \mathbb{R} \longrightarrow [0\ ,\ 1]$

Modelo

- Clasificación binaria: $y \in \{0, 1\}$.
- Queremos restringir $y \in [0\ ,\ 1]$
- Usar función de activación:

$$y = g(\mathbf{w}^T \mathbf{x})$$

donde:

  - $g\ :\ \mathbb{R} \longrightarrow [0\ ,\ 1]$
  - g es monótona

Modelo

- Clasificación binaria: $y \in \{0, 1\}$.
- Queremos restringir $y \in [0\ ,\ 1]$
- Usar función de activación:

$$y = g(\mathbf{w}^T \mathbf{x})$$

  donde:
  - $g\ :\ \mathbb{R} \longrightarrow [0\ ,\ 1]$
  - g es monótona
- Clasificación:

Modelo

- Clasificación binaria: $y \in \{0, 1\}$.
- Queremos restringir $y \in [0\ ,\ 1]$
- Usar función de activación:

$$y = g(\mathbf{w}^T \mathbf{x})$$

  donde:
  - $g\ :\ \mathbb{R} \longrightarrow [0\ ,\ 1]$
  - g es monótona
- Clasificación: $g(z) \geq t$

Modelo

- Clasificación binaria: $y \in \{0, 1\}$.
- Queremos restringir $y \in [0\ ,\ 1]$
- Usar función de activación:

$$y = g(\mathbf{w}^T \mathbf{x})$$

  donde:

  - $g\ :\ \mathbb{R} \longrightarrow [0\ ,\ 1]$
  - g es monótona

- Clasificación: $g(z) \geq t \Rightarrow$ Superficie de separación lineal.

- Probabilidades a priori de cada clase:

- Probabilidades a priori de cada clase:

$$\mathbf{P}(y = 1) = \alpha, \qquad \mathbf{P}(y = 0) = 1 - \alpha$$

- Probabilidades a priori de cada clase:

$$\mathbf{P}(y=1)=\alpha, \qquad \mathbf{P}(y=0)=1-\alpha$$

- Probabilidades marginales

- Probabilidades a priori de cada clase:

$$\mathbf{P}(y=1)=\alpha, \qquad \mathbf{P}(y=0)=1-\alpha$$

- Probabilidades marginales

$$\mathbf{P}(\mathbf{x}|y=1)=p_1(\mathbf{x}), \qquad \mathbf{P}(\mathbf{x}|y=0)=p_0(\mathbf{x})$$

- Probabilidades a priori de cada clase:

$$\mathbf{P}(y = 1) = \alpha, \qquad \mathbf{P}(y = 0) = 1 - \alpha$$

- Probabilidades marginales

$$\mathbf{P}(\mathbf{x}|y = 1) = p_1(\mathbf{x}), \qquad \mathbf{P}(\mathbf{x}|y = 0) = p_0(\mathbf{x})$$

- Probabilidad (posterior) de que $\mathbf{x}$ pertenezca a la clase 1 es:

- Probabilidades a priori de cada clase:

$$\mathbf{P}(y=1)=\alpha, \qquad \mathbf{P}(y=0)=1-\alpha$$

- Probabilidades marginales

$$\mathbf{P}(\mathbf{x}|y=1)=p_1(\mathbf{x}), \qquad \mathbf{P}(\mathbf{x}|y=0)=p_0(\mathbf{x})$$

- Probabilidad (posterior) de que $\mathbf{x}$ pertenezca a la clase 1 es:

$$\mathbf{P}(y=1|\mathbf{x})=\frac{p_1(\mathbf{x})\alpha}{p_1(\mathbf{x})\alpha+p_0(\mathbf{x})(1-\alpha)}$$

- Probabilidades a priori de cada clase:

$$\mathbf{P}(y=1)=\alpha, \qquad \mathbf{P}(y=0)=1-\alpha$$

- Probabilidades marginales

$$\mathbf{P}(\mathbf{x}|y=1)=p_1(\mathbf{x}), \qquad \mathbf{P}(\mathbf{x}|y=0)=p_0(\mathbf{x})$$

- Probabilidad (posterior) de que $\mathbf{x}$ pertenezca a la clase 1 es:

$$\mathbf{P}(y=1|\mathbf{x})=\frac{p_1(\mathbf{x})\alpha}{p_1(\mathbf{x})\alpha+p_0(\mathbf{x})(1-\alpha)}=\frac{1}{1+\frac{p_0(\mathbf{x})(1-\alpha)}{p_1(\mathbf{x})\alpha}}$$

- Probabilidades a priori de cada clase:

$$\mathbf{P}(y=1)=\alpha, \qquad \mathbf{P}(y=0)=1-\alpha$$

- Probabilidades marginales

$$\mathbf{P}(\mathbf{x}|y=1)=p_1(\mathbf{x}), \qquad \mathbf{P}(\mathbf{x}|y=0)=p_0(\mathbf{x})$$

- Probabilidad (posterior) de que **x** pertenezca a la clase 1 es:

$$\mathbf{P}(y=1|\mathbf{x})=\frac{p_1(\mathbf{x})\alpha}{p_1(\mathbf{x})\alpha+p_0(\mathbf{x})(1-\alpha)}=\frac{1}{1+\frac{p_0(\mathbf{x})(1-\alpha)}{p_1(\mathbf{x})\alpha}}=\frac{1}{1+e^{-z}}$$

- Probabilidades a priori de cada clase:

$$\mathbf{P}(y=1)=\alpha, \qquad \mathbf{P}(y=0)=1-\alpha$$

- Probabilidades marginales

$$\mathbf{P}(\mathbf{x}|y=1)=p_1(\mathbf{x}), \qquad \mathbf{P}(\mathbf{x}|y=0)=p_0(\mathbf{x})$$

- Probabilidad (posterior) de que $\mathbf{x}$ pertenezca a la clase 1 es:

$$\mathbf{P}(y=1|\mathbf{x})=\frac{p_1(\mathbf{x})\alpha}{p_1(\mathbf{x})\alpha+p_0(\mathbf{x})(1-\alpha)}=\frac{1}{1+\frac{p_0(\mathbf{x})(1-\alpha)}{p_1(\mathbf{x})\alpha}}=\frac{1}{1+e^{-z}}$$

donde:

$$z=\ln\left(\frac{p_1(\mathbf{x})\alpha}{p_0(\mathbf{x})(1-\alpha)}\right)$$

Caso Especial

- Cuando $p_0(\mathbf{x})$ y $p_1(\mathbf{x})$ son Normales:

Caso Especial

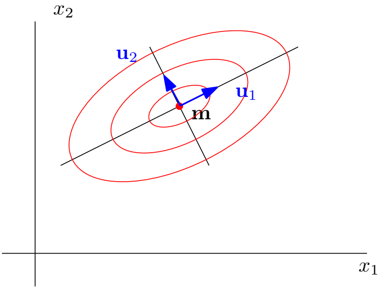
- Cuando $p_0(\mathbf{x})$ y $p_1(\mathbf{x})$ son Normales:

$$\frac{p_0(\mathbf{x})(1-\alpha)}{p_1(\mathbf{x})\alpha} = \frac{\frac{1-\alpha}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma_0|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_0)^T\Sigma_0^{-1}(\mathbf{x}-\mathbf{m}_0)\right\}}{\frac{\alpha}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma_1|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T\Sigma_1^{-1}(\mathbf{x}-\mathbf{m}_1)\right\}}$$

Caso Especial

- Cuando $p_0(\mathbf{x})$ y $p_1(\mathbf{x})$ son Normales:

$$\frac{p_0(\mathbf{x})(1-\alpha)}{p_1(\mathbf{x})\alpha} = \frac{\frac{1-\alpha}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma_0|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_0)^T\Sigma_0^{-1}(\mathbf{x}-\mathbf{m}_0)\right\}}{\frac{\alpha}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma_1|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T\Sigma_1^{-1}(\mathbf{x}-\mathbf{m}_1)\right\}}$$

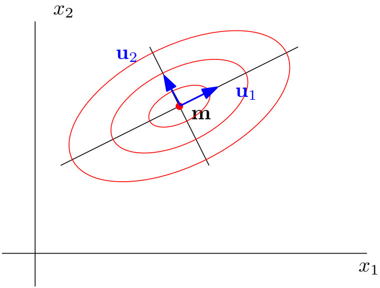


- Tomando logaritmos:

Caso Especial

- Cuando $p_0(\mathbf{x})$ y $p_1(\mathbf{x})$ son Normales:

$$\frac{p_0(\mathbf{x})(1-\alpha)}{p_1(\mathbf{x})\alpha} = \frac{\frac{1-\alpha}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma_0|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_0)^T\Sigma_0^{-1}(\mathbf{x}-\mathbf{m}_0)\right\}}{\frac{\alpha}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma_1|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T\Sigma_1^{-1}(\mathbf{x}-\mathbf{m}_1)\right\}}$$



- Tomando logaritmos:

$$z = -\frac{1}{2}(\mathbf{x}-\mathbf{m}_0)^T\Sigma_0^{-1}(\mathbf{x}-\mathbf{m}_0) + \frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T\Sigma_1^{-1}(\mathbf{x}-\mathbf{m}_1) + \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + \ln\left(\frac{1-\alpha}{\alpha}\right)$$

- Si además $\Sigma_0 = \Sigma_1 = \Sigma$:

$$z = -\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x} + \mathbf{m}_0^T\Sigma^{-1}\mathbf{x} - \frac{1}{2}\mathbf{m}_0^T\Sigma^{-1}\mathbf{m}_0 + \frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x} - \mathbf{m}_1^T\Sigma^{-1}\mathbf{x} + \frac{1}{2}\mathbf{m}_1^T\Sigma^{-1}\mathbf{m}_1 + \ln\left(\frac{1-\alpha}{\alpha}\right)$$

- Si además $\Sigma_0 = \Sigma_1 = \Sigma$:

$$z = -\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x} + \mathbf{m}_0^T\Sigma^{-1}\mathbf{x} - \frac{1}{2}\mathbf{m}_0^T\Sigma^{-1}\mathbf{m}_0 + \frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x} - \mathbf{m}_1^T\Sigma^{-1}\mathbf{x} + \frac{1}{2}\mathbf{m}_1^T\Sigma^{-1}\mathbf{m}_1 + \ln\left(\frac{1-\alpha}{\alpha}\right)$$

- entonces:

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

- Modelo:

$$h_{\mathbf{w}}(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x})$$

- Modelo:

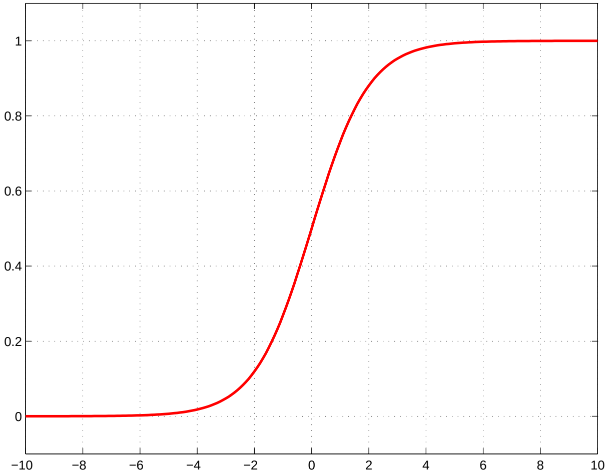
$$h_{\mathbf{w}}(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$$

- Modelo:

$$h_{\mathbf{w}}(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}}$$

$\sigma(.)$ es la función logística o sigmoide

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



Interpretación probabilística

- Interpretamos $\sigma(\mathbf{w}^T\mathbf{x})$ como el estimativo dado por el modelo con parámetros $\mathbf{w}$ de la probabilidad de que $\mathbf{x}$ pertenezca a la clase 1:

Interpretación probabilística

- Interpretamos $\sigma(\mathbf{w}^T\mathbf{x})$ como el estimativo dado por el modelo con parámetros $\mathbf{w}$ de la probabilidad de que $\mathbf{x}$ pertenezca a la clase 1:

$$\mathbf{P}(y = 1 \mid \mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x})$$

Interpretación probabilística

- Interpretamos $\sigma(\mathbf{w}^T\mathbf{x})$ como el estimativo dado por el modelo con parámetros $\mathbf{w}$ de la probabilidad de que $\mathbf{x}$ pertenezca a la clase 1:

$$\mathbf{P}(y = 1 \mid \mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x})$$
$$\mathbf{P}(y = 0 \mid \mathbf{x}; \mathbf{w}) = 1 - \sigma(\mathbf{w}^T\mathbf{x})$$

Interpretación probabilística

- Interpretamos $\sigma(\mathbf{w}^T\mathbf{x})$ como el estimativo dado por el modelo con parámetros **w** de la probabilidad de que $\mathbf{x}$ pertenezca a la clase 1:

$$\mathbf{P}(y = 1 \mid \mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x})$$
$$\mathbf{P}(y = 0 \mid \mathbf{x}; \mathbf{w}) = 1 - \sigma(\mathbf{w}^T\mathbf{x})$$

- Es decir, dado $\mathbf{x}$, y es una variable aleatoria de Bernoulli .

Interpretación probabilística

- Interpretamos $\sigma(\mathbf{w}^T\mathbf{x})$ como el estimativo dado por el modelo con parámetros $\mathbf{w}$ de la probabilidad de que $\mathbf{x}$ pertenezca a la clase 1:

$$\mathbf{P}(y = 1 \mid \mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x})$$
$$\mathbf{P}(y = 0 \mid \mathbf{x}; \mathbf{w}) = 1 - \sigma(\mathbf{w}^T\mathbf{x})$$

- Es decir, dado $\mathbf{x}$, y es una variable aleatoria de Bernoulli .
- Podemos escribir más compactamente

$$\mathbf{P}(y \mid \mathbf{x}; \mathbf{w}) = (\sigma(\mathbf{w}^T\mathbf{x}))^y(1 - \sigma(\mathbf{w}^T\mathbf{x}))^{1-y}$$

- Asumiendo datos i.i.d tenemos la función de verosimilitud:

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

- Asumiendo datos i.i.d tenemos la función de verosimilitud:

$$\begin{aligned} L(\mathbf{w}) &= \mathbf{P}(\mathbf{y} \mid \mathbf{X}; \mathbf{w}) \\ &= \prod_{i=1}^{n} \mathbf{P}(y_i \mid \mathbf{x}_i; \mathbf{w}) \end{aligned}$$

- Asumiendo datos i.i.d tenemos la función de verosimilitud:

$$\begin{aligned} L(\mathbf{w}) &= \mathbf{P}(\mathbf{y} \mid \mathbf{X}; \mathbf{w}) \\ &= \prod_{i=1}^{n} \mathbf{P}(y_i \mid \mathbf{x}_i; \mathbf{w}) \\ &= \prod_{i=1}^{n} (\sigma(\mathbf{w}^T \mathbf{x}_i))^{y_i} (1 - \sigma(\mathbf{w}^T \mathbf{x}_i))^{1-y_i} \end{aligned}$$

- Asumiendo datos i.i.d tenemos la función de verosimilitud:

$$\begin{aligned} L(\mathbf{w}) &= \mathbf{P}(\mathbf{y} \mid \mathbf{X}; \mathbf{w}) \\ &= \prod_{i=1}^{n} \mathbf{P}(y_i \mid \mathbf{x}_i; \mathbf{w}) \\ &= \prod_{i=1}^{n} (\sigma(\mathbf{w}^T \mathbf{x}_i))^{y_i} (1 - \sigma(\mathbf{w}^T \mathbf{x}_i))^{1-y_i} \end{aligned}$$

- Tomando logaritmo:

$$l(\mathbf{w}) = \log(L(\mathbf{w})) = \sum_{i=1}^{n} y_i \log(\sigma(\mathbf{w}^T \mathbf{x}_i)) + (1 - y_i) \log(1 - \sigma(\mathbf{w}^T \mathbf{x}_i))$$

- Asumiendo datos i.i.d tenemos la función de verosimilitud:

$$\begin{aligned} L(\mathbf{w}) &= \mathbf{P}(\mathbf{y} \mid \mathbf{X}; \mathbf{w}) \\ &= \prod_{i=1}^{n} \mathbf{P}(y_i \mid \mathbf{x}_i; \mathbf{w}) \\ &= \prod_{i=1}^{n} (\sigma(\mathbf{w}^T \mathbf{x}_i))^{y_i} (1 - \sigma(\mathbf{w}^T \mathbf{x}_i))^{1-y_i} \end{aligned}$$

- Tomando logaritmo:

$$l(\mathbf{w}) = \log(L(\mathbf{w})) = \sum_{i=1}^{n} y_i \log(\sigma(\mathbf{w}^T \mathbf{x}_i)) + (1 - y_i) \log(1 - \sigma(\mathbf{w}^T \mathbf{x}_i))$$

- Problema de optimización:

- Asumiendo datos i.i.d tenemos la función de verosimilitud:

$$
\begin{aligned}
L(\mathbf{w}) &= \mathbf{P}(\mathbf{y} \mid \mathbf{X}; \mathbf{w}) \\
&= \prod_{i=1}^{n} \mathbf{P}(y_i \mid \mathbf{x}_i; \mathbf{w}) \\
&= \prod_{i=1}^{n} (\sigma(\mathbf{w}^T \mathbf{x}_i))^{y_i} (1 - \sigma(\mathbf{w}^T \mathbf{x}_i))^{1-y_i}
\end{aligned}
$$

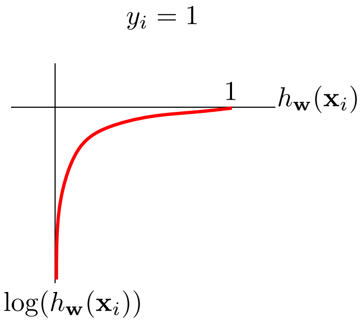
- Tomando logaritmo:

$$
l(\mathbf{w}) = \log(L(\mathbf{w})) = \sum_{i=1}^{n} y_i \log(\sigma(\mathbf{w}^T \mathbf{x}_i)) + (1 - y_i) \log(1 - \sigma(\mathbf{w}^T \mathbf{x}_i))
$$

- Problema de optimización:

$$
\mathbf{w}^* = \arg\max_{\mathbf{w}} l(\mathbf{w})
$$

Negativo de la Función de error (acierto!)



Ascenso de Gradiente

Inicialice $\mathbf{w}_0$

Ascenso de Gradiente

Inicialice $\mathbf{w}_0$
repeat

Ascenso de Gradiente

Inicialice $\mathbf{w}_0$

repeat

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

Ascenso de Gradiente

Inicialice $\mathbf{w}_0$

repeat

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

until Condición de terminación.

Gradiente

Gradiente

- Note que $\sigma'(z) = \sigma(z)(1 - \sigma(z))$

Gradiente

- Note que $\sigma'(z) = \sigma(z)(1 - \sigma(z))$. Denote $\sigma_i = \sigma(\mathbf{w}^T \mathbf{x}_i)$

Gradiente

- Note que $\sigma'(z) = \sigma(z)(1 - \sigma(z))$. Denote $\sigma_i = \sigma(\mathbf{w}^T \mathbf{x}_i)$
- Un término en la suma:

Gradiente

- Note que $\sigma'(z) = \sigma(z)(1 - \sigma(z))$. Denote $\sigma_i = \sigma(\mathbf{w}^T \mathbf{x}_i)$
- Un término en la suma:

$$[\nabla_{\mathbf{w}} l(\mathbf{w})]_i = y_i \frac{\mathbf{x}_i \sigma_i (1 - \sigma_i)}{\sigma_i} + (1 - y_i) \frac{-\mathbf{x}_i \sigma_i (1 - \sigma_i)}{1 - \sigma_i}$$

Gradiente

- Note que $\sigma'(z) = \sigma(z)(1 - \sigma(z))$. Denote $\sigma_i = \sigma(\mathbf{w}^T \mathbf{x}_i)$
- Un término en la suma:

$$\begin{aligned} [\nabla_{\mathbf{w}} l(\mathbf{w})]_i &= y_i \frac{\mathbf{x}_i \sigma_i (1 - \sigma_i)}{\sigma_i} + (1 - y_i) \frac{-\mathbf{x}_i \sigma_i (1 - \sigma_i)}{1 - \sigma_i} \\ &= y_i \mathbf{x}_i (1 - \sigma_i) + (y_i - 1) \mathbf{x}_i \sigma_i \end{aligned}$$

Gradiente

- Note que $\sigma'(z) = \sigma(z)(1 - \sigma(z))$. Denote $\sigma_i = \sigma(\mathbf{w}^T \mathbf{x}_i)$
- Un término en la suma:

$$\begin{aligned}
[\nabla_{\mathbf{w}} l(\mathbf{w})]_i &= y_i \frac{\mathbf{x}_i \sigma_i (1 - \sigma_i)}{\sigma_i} + (1 - y_i) \frac{-\mathbf{x}_i \sigma_i (1 - \sigma_i)}{1 - \sigma_i} \\
&= y_i \mathbf{x}_i (1 - \sigma_i) + (y_i - 1) \mathbf{x}_i \sigma_i \\
&= (y_i - \sigma(\mathbf{w}^T \mathbf{x}_i)) \mathbf{x}_i
\end{aligned}$$

Gradiente

- Note que $\sigma'(z) = \sigma(z)(1 - \sigma(z))$. Denote $\sigma_i = \sigma(\mathbf{w}^T\mathbf{x}_i)$
- Un término en la suma:

$$\begin{aligned}
[\nabla_{\mathbf{w}} l(\mathbf{w})]_i &= y_i \frac{\mathbf{x}_i \sigma_i (1 - \sigma_i)}{\sigma_i} + (1 - y_i) \frac{-\mathbf{x}_i \sigma_i (1 - \sigma_i)}{1 - \sigma_i} \\
&= y_i \mathbf{x}_i (1 - \sigma_i) + (y_i - 1)\mathbf{x}_i \sigma_i \\
&= (y_i - \sigma(\mathbf{w}^T\mathbf{x}_i))\mathbf{x}_i \\
&= e_i \mathbf{x}_i
\end{aligned}$$

Ascenso de Gradiente estocástico

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

Ascenso de Gradiente estocástico

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

Ascenso de Gradiente estocástico

Incialize $\mathbf{w}_0$ a valores pequeños.
repeat
    Escoja $(\mathbf{x}_i, y_i)$
    $g = \sigma(\mathbf{w}_k^T \mathbf{x}_i)$

Ascenso de Gradiente estocástico

Incialize $\mathbf{w}_0$ a valores pequeños.
repeat
  Escoja $(\mathbf{x}_i, y_i)$
  $g = \sigma(\mathbf{w}_k^T \mathbf{x}_i)$
  $e = y_i - g$

Ascenso de Gradiente estocástico

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

repeat

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

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

  $e = y_i - g$

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

Ascenso de Gradiente estocástico

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

- Hessiana de $l(\mathbf{w})$:✍

$$\nabla^2 l(\mathbf{w}) = -\sum_{i=1}^{n} \sigma_i(1-\sigma_i)\mathbf{x}_i\mathbf{x}_i^T$$

- $\nabla^2 l(\mathbf{w})$ es positiva definida

- Hessiana de $l(\mathbf{w})$:✍

$$\nabla^2 l(\mathbf{w}) = -\sum_{i=1}^{n} \sigma_i(1-\sigma_i)\mathbf{x}_i\mathbf{x}_i^T$$

- $\nabla^2 l(\mathbf{w})$ es positiva definida$\Rightarrow$ $l(\mathbf{w})$ es cóncava ($-l(\mathbf{w})$ es convexa).