Aprendizaje automatizado

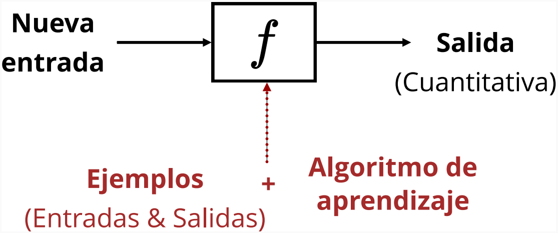
Mínimos cuadrados

Gibran Fuentes Pineda

Febrero 2023

Regresión

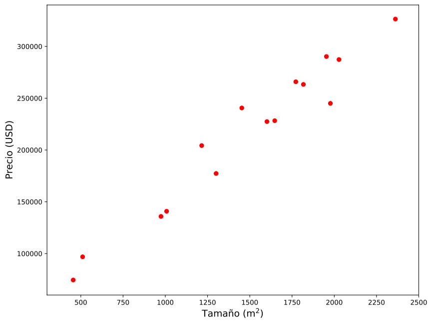
- Salida continua (cuantitativa)
- Ejemplos: predicción de temperatura de un cuarto, etc.



Prediciendo el precio de casas

- ¿Cómo podemos ajustar nuestra función f para modelar la relación entre el tamaño y el precio de casas?

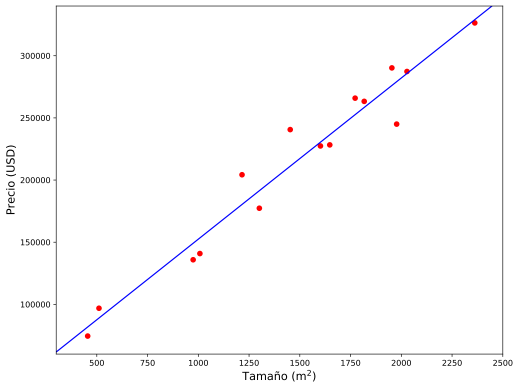
Tamaño (m^2)	Precio (USD)
489.59	489.59
556.08	556.08
570.35	570.35
772.84	772.84
970.95	970.95
1162.00	1162.00
1263.10	1263.10
⋮	⋮



Prediciendo el precio de casas

- Podemos hacer presuposiciones sobre f, por ejemplo que la relación es lineal:

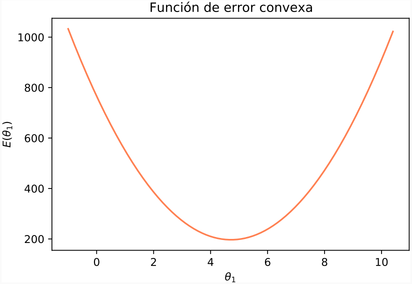
$$f_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 \cdot x$$



¿Cómo medimos la calidad del ajuste?

- Definimos una función de error, por ejemplo la suma de errores cuadráticos:

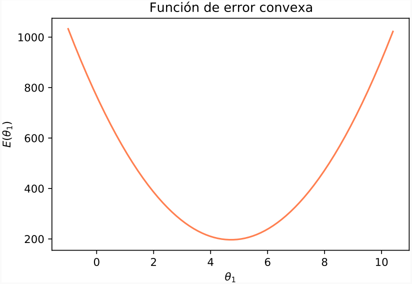
$$E(\boldsymbol{\theta}) = \frac{1}{2} \cdot \sum_{i=1}^{n} \left\{ f_{\boldsymbol{\theta}}(\mathrm{x}^{(i)}, \boldsymbol{\theta}) - y^{(i)} \right\}^2$$



¿Cómo medimos la calidad del ajuste?

- Definimos una función de error, por ejemplo la suma de errores cuadráticos:

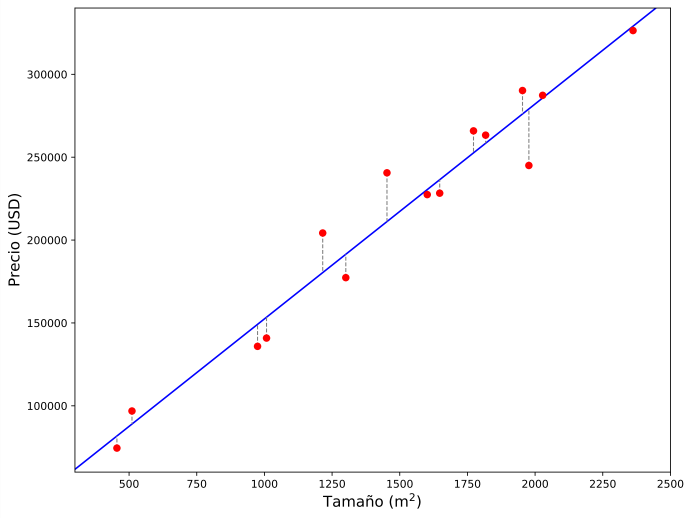
$$E(\boldsymbol{\theta}) = \frac{1}{2} \cdot \sum_{i=1}^{n} \left\{ f_{\boldsymbol{\theta}}(\mathrm{x}^{(i)}, \boldsymbol{\theta}) - y^{(i)} \right\}^2$$



- Objetivo: encontrar el valor de $\boldsymbol{\theta}$ que minimice $E(\boldsymbol{\theta})$

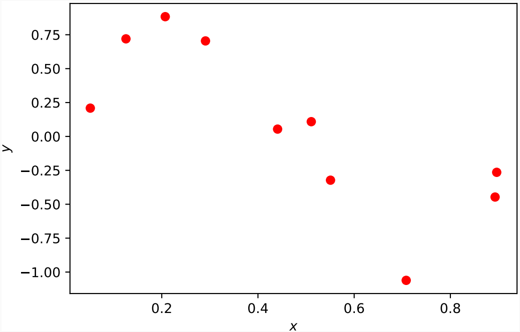
$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} E(\boldsymbol{\theta})$$

¿Cómo medimos la calidad del ajuste?



Modelando relaciones no lineales

- ¿Qué función se ajusta a estos datos?



Modelando relaciones no lineales

- Podemos ajustar un polinomio de la siguiente forma[1]

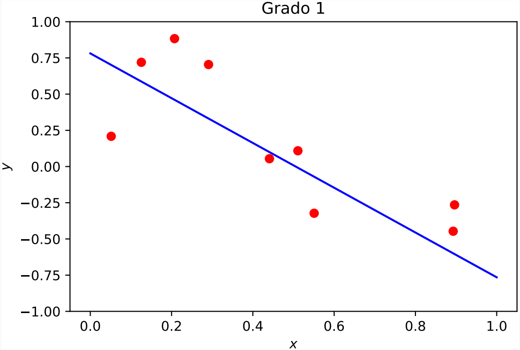
$$f_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 \cdot x + \theta_1 \cdot x^2 + \cdots + \theta_d \cdot x^d$$

[1]Nota que esta forma no está considerando interacciones

¿Qué grado del polinomio es adecuado?

- Podemos usar uno lineal nuevamente

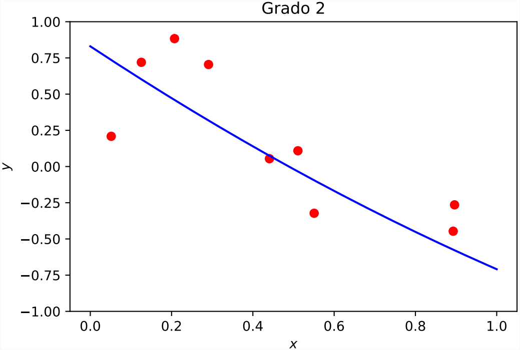
$$f_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 \cdot x$$



¿Qué grado del polinomio es adecuado?

- O uno cuadrático

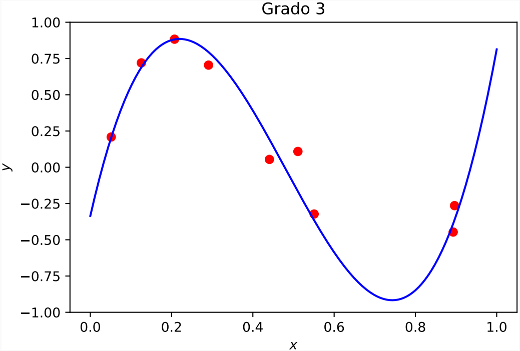
$$f_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 \cdot x + \theta_2 \cdot x^2$$



¿Qué grado del polinomio es adecuado?

- Grado 3

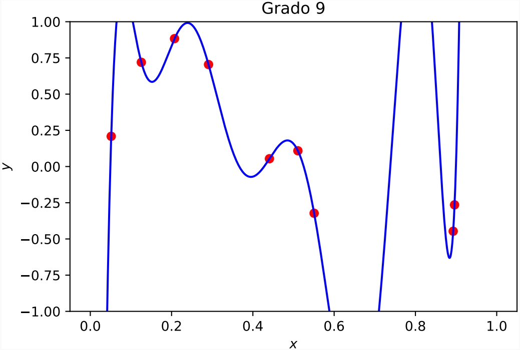
$$f_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 + \theta_2 \cdot x^2 + \theta_3 \cdot x^3$$



¿Qué grado del polinomio es adecuado?

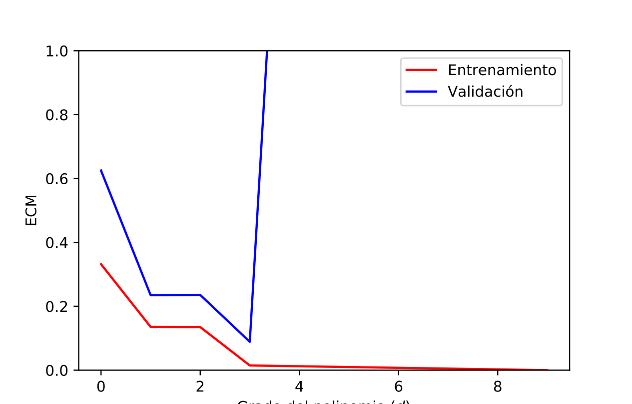
- O grado 9

$$f_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 + \theta_2 \cdot x^2 + \cdot x + \cdots + \theta_9 \cdot x^9$$



- Comparamos los desempeños con distintos grados de polinomio usando el error cuadrático medio (ECM)

$$E(\boldsymbol{\theta}) = \frac{1}{n} \cdot \sum_{i=1}^{n} \left\{ f_{\boldsymbol{\theta}}(\mathrm{x}^{(i)}, \boldsymbol{\theta}) - y^{(i)} \right\}^2$$



¿Por qué está sobreajustando?

	$d = 0$	$d = 1$	$d = 3$	$d = 9$
θ_0	0.05	0.78	−0.33	−17.62
θ_1		−1.54	12.32	762.18
θ_2			−36.32	12071.82
θ_3			25.14	98135.73
θ_4				−459092.41
θ_5				1301097.36
θ_6				−2263938.71
θ_7				2358449.27
θ_8				−1347197.15
θ_9				324015.43

¿Cómo evito el sobreajuste?

- Penalizando parámetros con valores grandes

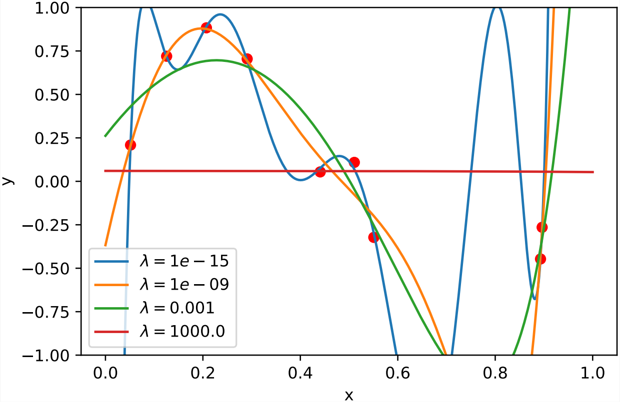
$$\tilde{E}(\boldsymbol{\theta}) = \frac{1}{2} \cdot \sum_{i=1}^{n} \left\{ f_{\boldsymbol{\theta}}(\mathsf{x}^{(i)}, \boldsymbol{\theta}) - y^{(i)} \right\}^2 + \frac{\lambda}{2} \cdot \|\boldsymbol{\theta}\|_2^2$$

- λ determina la ponderación que se le da al término de penalización

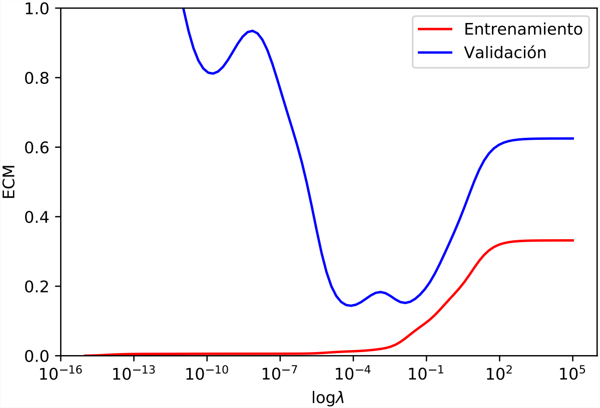
¿Cómo evito el sobreajuste?

	$\log \lambda = -\infty$	$\log \lambda = -18$	$\log \lambda = 0$
θ_0	0.35	0.35	−17.62
θ_1	232.37	4.74	−0.05
θ_2	−5321.83	−0.77	−0.06
θ_3	48568	−31.97	−0.05
θ_4	−231639.30	−3.89	−0.03
θ_5	640042.26	55.28	−0.02
θ_6	−1061800.52	41.32	−0.01
θ_7	1042400.18	−45.95	−0.00
θ_8	−557682.99	−91.53	0.00
θ_9	125201.43	72.68	0.01

Mínimos cuadrados penalizados



Mínimos cuadrados penalizados



Regresión lineal

- Modelo lineal

$$f_{\boldsymbol{\theta}}(\mathsf{x}, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathsf{x} = \sum_{i=1}^{d} \theta_i \cdot x_i$$

Regresión lineal

- Modelo lineal

$$f_{\boldsymbol{\theta}}(\mathsf{x}, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathsf{x} = \sum_{i=1}^{d} \theta_i \cdot x_i$$

- Con expansión de funciones base ϕ

$$f_{\boldsymbol{\theta}}(\mathsf{x}, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \phi(\mathsf{x}) = \sum_{i=1}^{d} \theta_i \cdot \phi(\mathsf{x})_i$$

Regresión lineal

- Modelo lineal

$$f_{\boldsymbol{\theta}}(\mathsf{x}, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathsf{x} = \sum_{i=1}^{d} \theta_i \cdot x_i$$

- Con expansión de funciones base ϕ

$$f_{\boldsymbol{\theta}}(\mathsf{x}, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \phi(\mathsf{x}) = \sum_{i=1}^{d} \theta_i \cdot \phi(\mathsf{x})_i$$

- Lineal en los parámetros $\boldsymbol{\theta}$

Interpretación probabilística

- Asumiendo ruido ϵ con distribución normal en el modelo

$$y = f_{\boldsymbol{\theta}}(\mathsf{x}, \boldsymbol{\theta}) + \epsilon$$

Interpretación probabilística

- Asumiendo ruido ϵ con distribución normal en el modelo

$$y = f_{\boldsymbol{\theta}}(\mathsf{x}, \boldsymbol{\theta}) + \epsilon$$

- Tratamos de modelar la probabilidad condicional de la salida dados los datos y parámetros

$$P(y|\mathsf{x}, \boldsymbol{\theta}, \sigma^2) = \mathcal{N}(y|f_{\boldsymbol{\theta}}(\mathsf{x}, \boldsymbol{\theta}), \sigma^2)$$

Obteniendo el estimador de máxima verosimilitud

- Se busca minimizar el negativo de la verosimilitud logarítmica

$$
\begin{aligned}
NVL(\boldsymbol{\theta}) &= -\sum_{i=1}^{n} \log P(y^{(i)}|\mathrm{x}^{(i)}, \boldsymbol{\theta}) \\
&= -\sum_{i=1}^{n} \log \mathcal{N}(y^{(i)}|f_{\boldsymbol{\theta}}(\mathrm{x}^{(i)}, \boldsymbol{\theta}), \sigma^2) \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} \left\{ f_{\boldsymbol{\theta}}(\mathrm{x}^{(i)}, \boldsymbol{\theta}) - y^{(i)} \right\}^2 - \frac{n}{2} log 2\pi\sigma^2
\end{aligned}
$$

Obteniendo el estimador de máxima verosimilitud

- Se busca minimizar el negativo de la verosimilitud logarítmica

$$
\begin{aligned}
NVL(\boldsymbol{\theta}) &= -\sum_{i=1}^{n} \log P(y^{(i)}|\mathrm{x}^{(i)}, \boldsymbol{\theta}) \\
&= -\sum_{i=1}^{n} \log \mathcal{N}(y^{(i)}|f_{\boldsymbol{\theta}}(\mathrm{x}^{(i)}, \boldsymbol{\theta}), \sigma^2) \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} \left\{f_{\boldsymbol{\theta}}(\mathrm{x}^{(i)}, \boldsymbol{\theta}) - y^{(i)}\right\}^2 - \frac{n}{2} log 2\pi\sigma^2
\end{aligned}
$$

- Equivalente a minimizar suma de errores cuadráticos

$$
E(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}\{f_{\boldsymbol{\theta}}(\mathrm{x}^{(i)}, \boldsymbol{\theta}) - y^{(i)}\}^2
$$

Obteniendo el estimador de máxima verosimilitud

- Reformulando NVL

$$
\begin{aligned}
NVL(\boldsymbol{\theta}) &= \frac{1}{2}(\mathrm{y} - \mathrm{X}\boldsymbol{\theta})^\top(\mathrm{y} - \mathrm{X}\boldsymbol{\theta}) \\
&= \frac{1}{2}\mathrm{y}^\top\mathrm{y} - \frac{1}{2}\mathrm{y}^\top\mathrm{X}\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^\top\mathrm{X}^\top\mathrm{y} + \frac{1}{2}\boldsymbol{\theta}^\top\mathrm{X}^\top\mathrm{X}\boldsymbol{\theta} \\
&= \frac{1}{2}\mathrm{y}^\top\mathrm{y} - \boldsymbol{\theta}^\top\mathrm{X}^\top\mathrm{y} + \frac{1}{2}\boldsymbol{\theta}^\top\mathrm{X}^\top\mathrm{X}\boldsymbol{\theta}
\end{aligned}
$$

Obteniendo el estimador de máxima verosimilitud

- Reformulando NVL

$$\begin{aligned} NVL(\boldsymbol{\theta}) &= \frac{1}{2}(\mathrm{y} - \mathrm{X}\boldsymbol{\theta})^\top(\mathrm{y} - \mathrm{X}\boldsymbol{\theta}) \\ &= \frac{1}{2}\mathrm{y}^\top\mathrm{y} - \frac{1}{2}\mathrm{y}^\top\mathrm{X}\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^\top\mathrm{X}^\top\mathrm{y} + \frac{1}{2}\boldsymbol{\theta}^\top\mathrm{X}^\top\mathrm{X}\boldsymbol{\theta} \\ &= \frac{1}{2}\mathrm{y}^\top\mathrm{y} - \boldsymbol{\theta}^\top\mathrm{X}^\top\mathrm{y} + \frac{1}{2}\boldsymbol{\theta}^\top\mathrm{X}^\top\mathrm{X}\boldsymbol{\theta} \end{aligned}$$

- Derivando con respecto a $\boldsymbol{\theta}$ e igualando a cero

$$\mathrm{X}^\top\mathrm{X}\boldsymbol{\theta} = \mathrm{X}^\top\mathrm{y}$$

Obteniendo el estimador de máxima verosimilitud

- Reformulando NVL

$$
\begin{aligned}
NVL(\boldsymbol{\theta}) &= \frac{1}{2}(\mathrm{y} - \mathrm{X}\boldsymbol{\theta})^\top(\mathrm{y} - \mathrm{X}\boldsymbol{\theta}) \\
&= \frac{1}{2}\mathrm{y}^\top\mathrm{y} - \frac{1}{2}\mathrm{y}^\top\mathrm{X}\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^\top\mathrm{X}^\top\mathrm{y} + \frac{1}{2}\boldsymbol{\theta}^\top\mathrm{X}^\top\mathrm{X}\boldsymbol{\theta} \\
&= \frac{1}{2}\mathrm{y}^\top\mathrm{y} - \boldsymbol{\theta}^\top\mathrm{X}^\top\mathrm{y} + \frac{1}{2}\boldsymbol{\theta}^\top\mathrm{X}^\top\mathrm{X}\boldsymbol{\theta}
\end{aligned}
$$

- Derivando con respecto a $\boldsymbol{\theta}$ e igualando a cero

$$\mathrm{X}^\top\mathrm{X}\boldsymbol{\theta} = \mathrm{X}^\top\mathrm{y}$$

- El estimador de máxima verosimilitud es

$$\hat{\boldsymbol{\theta}}_{EMV} = (\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}^\top\mathrm{y}$$

¿Y si tenemos múltiples variables de salida?

- Solución de mínimos cuadrados

$$\hat{\Theta}_{EMV} = (\mathrm{X}^\top \mathrm{X})^{-1}\mathrm{X}^\top \mathrm{Y}$$

- Equivalente a

$$\hat{\theta}_{k\,EMV} = (\mathrm{X}^\top \mathrm{X})^{-1}\mathrm{X}^\top \mathrm{y}_k$$

Obteniendo el estimador de máximo a posteriori

- Asumiendo distribución a priori normal sobre $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} log\mathcal{N}(y^{(i)}|\theta_0 + \boldsymbol{\theta}^\top \phi(\mathrm{x}^{(i)}), \sigma^2)$$
$$+ \sum_{j=0}^{d} log\mathcal{N}(\theta_j|0, \tau^2)$$

Obteniendo el estimador de máximo a posteriori

- Asumiendo distribución a priori normal sobre $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} log\mathcal{N}(y^{(i)}|\theta_0 + \boldsymbol{\theta}^\top \phi(\mathsf{x}^{(i)}), \sigma^2) + \sum_{j=0}^{d} log\mathcal{N}(\theta_j|0, \tau^2)$$

- Equivalente a minimizar suma de errores cuadráticos con los parámeros penalizados con la norma ℓ_2

$$\tilde{E}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}\{f_{\boldsymbol{\theta}}(\mathsf{x}^{(i)}, \boldsymbol{\theta}) - y^{(i)}\}^2 + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$

Obteniendo el estimador de máximo a posteriori

- Asumiendo distribución a priori normal sobre $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} log\mathcal{N}(y^{(i)}|\theta_0 + \boldsymbol{\theta}^\top \phi(\mathrm{x}^{(i)}), \sigma^2)$$
$$+ \sum_{j=0}^{d} log\mathcal{N}(\theta_j|0, \tau^2)$$

- Equivalente a minimizar suma de errores cuadráticos con los parámeros penalizados con la norma ℓ_2

$$\tilde{E}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}\{f_{\boldsymbol{\theta}}(\mathrm{x}^{(i)}, \boldsymbol{\theta}) - y^{(i)}\}^2 + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$

- Derivando $\tilde{E}(\boldsymbol{\theta})$ con respecto a $\boldsymbol{\theta}$ e igualando a cero

$$\hat{\boldsymbol{\theta}}_{ridge} = (\lambda \cdot \mathrm{I}_D + \mathrm{X}^\top \mathrm{X})^{-1}\mathrm{X}^\top \mathbf{y}$$

- Cuando la regularización es por norma ℓ_1 se conoce como LASSO

$$\hat{\boldsymbol{\theta}}_{LASSO} = \arg\min_{\boldsymbol{\theta}} \left[\frac{1}{2} \cdot \sum_{i=1}^{n} \left\{ f_{\boldsymbol{\theta}}(\mathrm{x}^{(i)}, \boldsymbol{\theta}) - y^{(i)} \right\}^2 + \frac{\lambda}{2} \cdot \|\boldsymbol{\theta}\|_1 \right]$$

Regularización con norma ℓ_1

- Cuando la regularización es por norma ℓ_1 se conoce como LASSO

$$\hat{\boldsymbol{\theta}}_{LASSO} = \arg\min_{\boldsymbol{\theta}} \left[\frac{1}{2} \cdot \sum_{i=1}^{n} \left\{ f_{\boldsymbol{\theta}}(\mathrm{x}^{(i)}, \boldsymbol{\theta}) - y^{(i)} \right\}^2 + \frac{\lambda}{2} \cdot \|\boldsymbol{\theta}\|_1 \right]$$

- Optimización cuadrática: no existe solución cerrada pero existen algoritmos eficientes

Regularización con diferentes normas

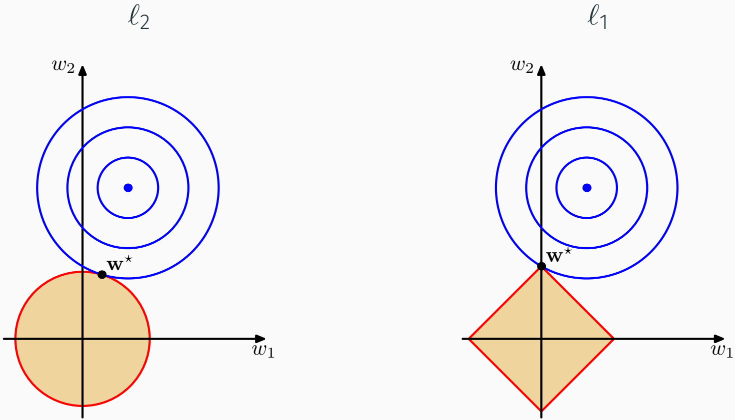


Imagen tomada de C. Bishop. PRML, 2009