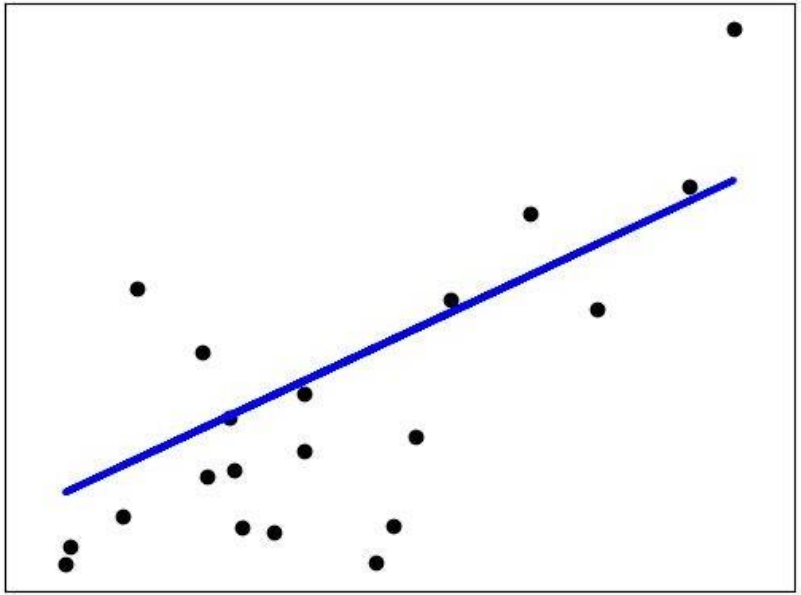
Machine Learning 101

Regularización

Índice

1. **Regresión lineal (revisited)**
2. El problema de overfitting
3. Regularización: ridge regression
4. Least Absolute Shrinkage and Selection Operator (LASSO)

Regresión lineal (una variable)



$$f_\omega(x) = \omega_0 + \omega_1 x$$

$$J(\omega_0, \omega_1) = \frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} - f_\omega(x^{(i)}) \right)^2$$

$$\min_{\omega_0, \omega_1} J(\omega_0, \omega_1)$$ Descenso por gradiente

Regresión lineal (varias variables)

	sqm_living	bedrooms	floors	years	price
0	109.625587	3	1.0	62	221900.0
1	238.760813	3	2.0	66	538000.0
2	71.535341	2	1.0	84	180000.0
3	182.089958	4	1.0	52	604000.0
4	156.077107	3	1.0	30	510000.0

$$\hat{y} = f_\omega(x) = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + \omega_4 x_4$$

$$\hat{y} = f_\omega(x) = 80 + 0.1 x_1 + 0.01 x_2 + 3 x_3 - 2 x_4$$

¡cuidado con la interpretación de estos coeficientes!

Regresión lineal (varias variables)

	sqm_living	bedrooms	floors	years	price
0	109.625587	3	1.0	62	221900.0
1	238.760813	3	2.0	66	538000.0
2	71.535341	2	1.0	84	180000.0
3	182.089958	4	1.0	52	604000.0
4	156.077107	3	1.0	30	510000.0

$$\hat{y}^{(0)} = \omega_0 + \omega_1 x_1^{(0)} + \omega_2 x_2^{(0)} + \omega_3 x_3^{(0)} + \omega_4 x_4^{(0)}$$

$$\hat{y}^{(1)} = \omega_0 + \omega_1 x_1^{(1)} + \omega_2 x_2^{(1)} + \omega_3 x_3^{(1)} + \omega_4 x_4^{(1)}$$

$$\hat{y}^{(4)} = \omega_0 + \omega_1 x_1^{(4)} + \omega_2 x_2^{(4)} + \omega_3 x_3^{(4)} + \omega_4 x_4^{(4)}$$

$$\begin{bmatrix} \hat{y}^{(0)} \\ \hat{y}^{(1)} \\ \vdots \\ \hat{y}^{(4)} \end{bmatrix} \begin{bmatrix} 1 & x_1^{(0)} & x_2^{(0)} & x_3^{(0)} & x_4^{(0)} \\ 1 & x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^{(4)} & x_2^{(4)} & x_3^{(4)} & x_4^{(4)} \end{bmatrix} \begin{bmatrix} \omega_0 \\ \omega_1 \\ \vdots \\ \omega_4 \end{bmatrix}$$

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\omega}$$

Regresión lineal (varias variables)

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\omega}$$

$$J(\boldsymbol{\omega}) = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 \equiv ||\mathbf{y} - \mathbf{X}\boldsymbol{\omega}||_2^2$$

1.1.1. Ordinary Least Squares

`LinearRegression` fits a linear model with coefficients $w = (w_1, \ldots, w_p)$ to minimize the residual sum of squares between the observed responses in the dataset, and the responses predicted by the linear approximation. Mathematically it solves a problem of the form:

$$\min_w ||Xw - y||_2^2$$

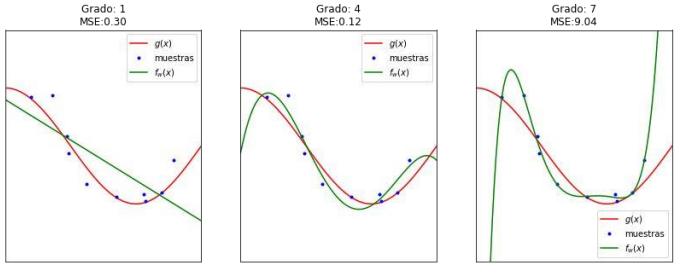
Regresión lineal (varias variables)

- La solución al problema de optimización anterior tiene una expresión analítica cerrada:

$$\boldsymbol{\omega} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

- ¿Qué pasa si $\mathbf{X}^T\mathbf{X}$ no es invertible? La solución no es única, debido a:
  - Variables redundantes (linealmente dependientes)
  - Más variables/características que muestras

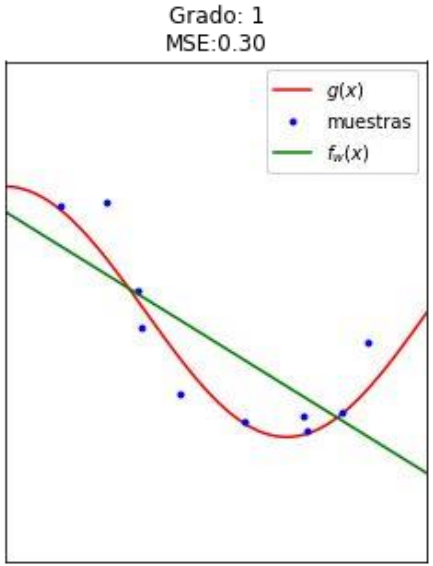
- Consecuencias
  - Inestabilidad de la solución (overfitting)
  - Coeficientes del modelo contraintuitivos

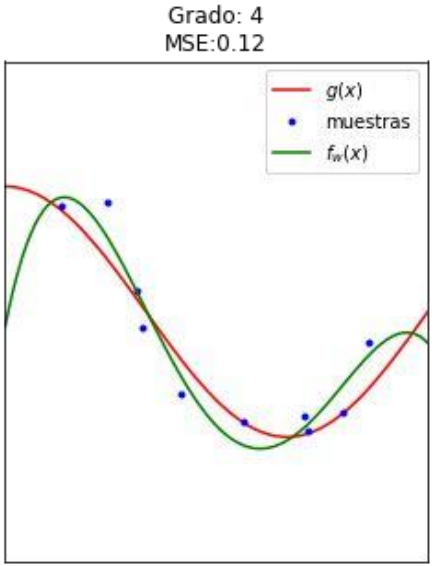


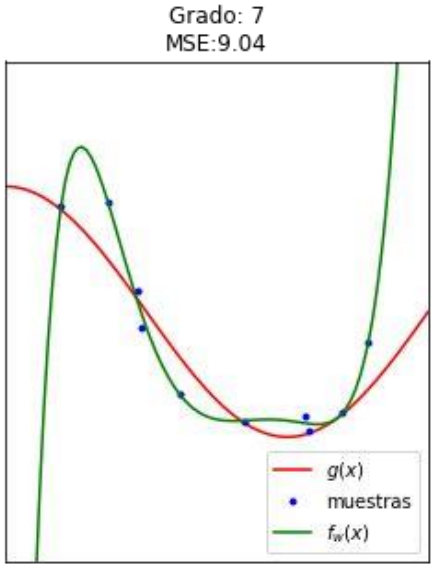
Índice

1. Regresión lineal (revisited)
2. **El problema de overfitting**
3. Regularización: ridge regression
4. Least Absolute Shrinkage and Selection Operator (LASSO)

El problema de overfitting







$$\boldsymbol{\omega} = [0, -2.1]$$

$$\boldsymbol{\omega} = [0, -16.8, -82.7, 118.2, -52.3]$$

$$\boldsymbol{\omega} = [0, 243.9, -1553.3, 4943.9, -9040.2, 9816.6, -5974.5, 1585.7]$$

El problema de *overfitting*

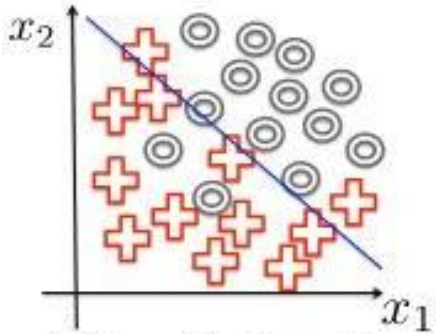
¿Cómo abordar el sobreajuste? Dos estrategias

- **Regularización**: penalizar coeficientes grandes
- **Selección de características**: reducir la dimensionalidad del problema

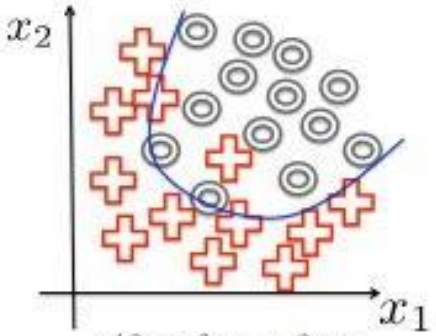
En ambos casos, se busca reducir la complejidad del modelo (a costa de aumentar el sesgo).

El problema de *overfitting*

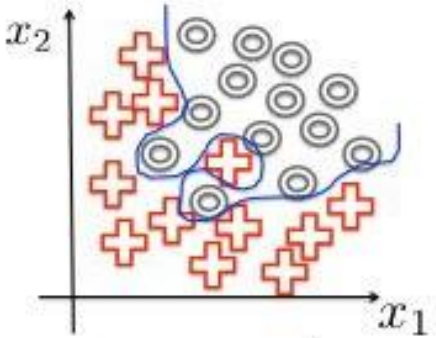
También aplica en clasificación: regresión logística



$$h_\beta(x) = g(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$



$$g(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2)$$



$$g(\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^2 x_2 + \beta_4 x_1^2 x_2^2 + \beta_5 x_1^2 x_2^3 + \beta_6 x_1^3 x_2 + \ldots)$$

Índice

1. Regresión lineal (revisited)
2. El problema de overfitting
3. **Regularización: *ridge regression***
4. Least Absolute Shrinkage and Selection Operator (LASSO)

Motivación

- La solución de mínimos cuadrados es inestable (coeficientes de alto valor)
- Solución: penalizar los coeficientes de alto valor
- ¿Cómo? Modificando la función de coste

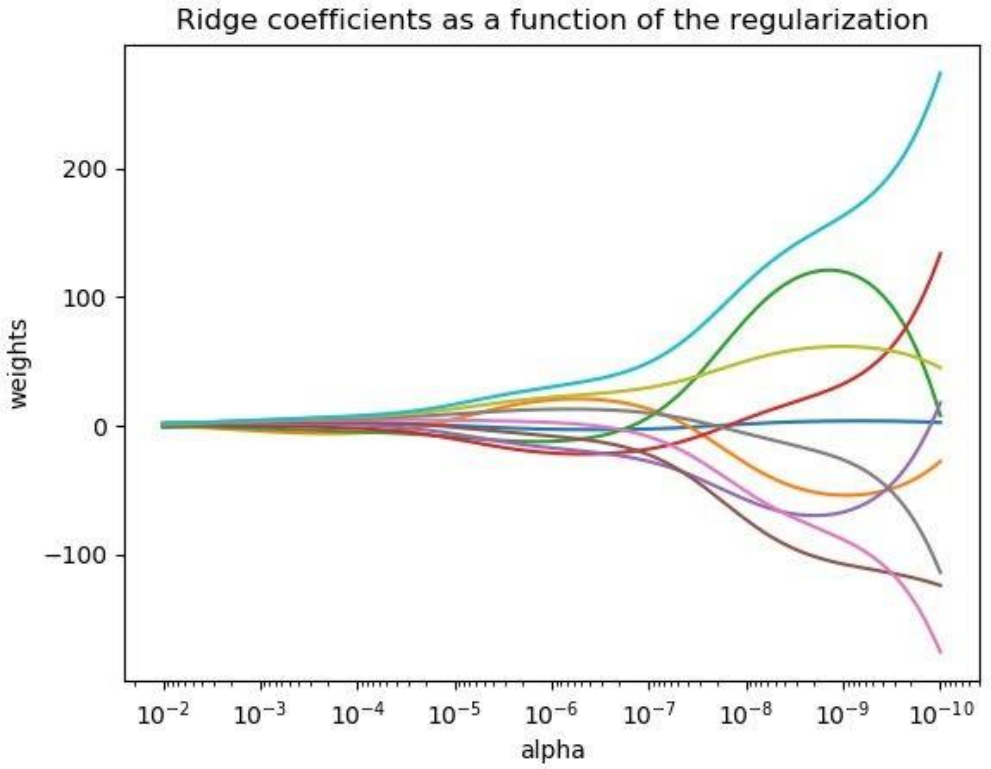
$$\min_{\omega} \frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} - f_{\omega}(\mathbf{x}^{(i)}) \right)^2 + \alpha \sum_{k=1}^{D} (\omega_k)^2$$

α: parámetro de regularización (hay que fijarlo a priori)

Parámetro de regularización

- Compromiso entre magnitud de los coeficientes y ajuste de la solución
  - Si α es muy grande $\rightarrow$ todos los coeficientes nulos (*underfitting*)
  - Si α es nulo $\rightarrow$ no hay regularización (posibilidad de sobreajuste)
- Ha de fijarse a priori (k-fold CV)
- Se cumple que $\alpha \geq 0$

Parámetro de regularización



Fuente: http://scikit-learn.org/stable/auto_examples/linear_model/plot_ridge_path.html#sphx-glr-auto-examples-linear-model-plot-ridge-path-py

Ridge regression

- En forma matricial

$$\min_{\omega} \frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} - f_{\omega}(\mathbf{x}^{(i)}) \right)^2 + \alpha \sum_{k=1}^{D} (\omega_k)^2$$

$$\min_{\boldsymbol{\omega}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\omega}||_2^2 + \alpha ||\boldsymbol{\omega}||_2^2$$

1.1.2. Ridge Regression

`Ridge` regression addresses some of the problems of Ordinary Least Squares by imposing a penalty on the size of coefficients. The ridge coefficients minimize a penalized residual sum of squares,

$$\min_{w} ||Xw - y||_2^2 + \alpha ||w||_2^2$$

Here, $\alpha \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of α, the greater the amount of shrinkage and thus the coefficients become more robust to collinearity.

Índice

1. Regresión lineal (revisited)
2. El problema de overfitting
3. Regularización: *ridge regression*
4. **Least Absolute Shrinkage and Selection Operator (LASSO)**

Lasso

$$\min_{\boldsymbol{\omega}} \underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - f_{\omega}(\mathbf{x}^{(i)})\right)^2}_{\text{Regresión lineal}} + \underbrace{\alpha\sum_{k=1}^{D}|\omega_k|}_{\text{Regularización}}$$

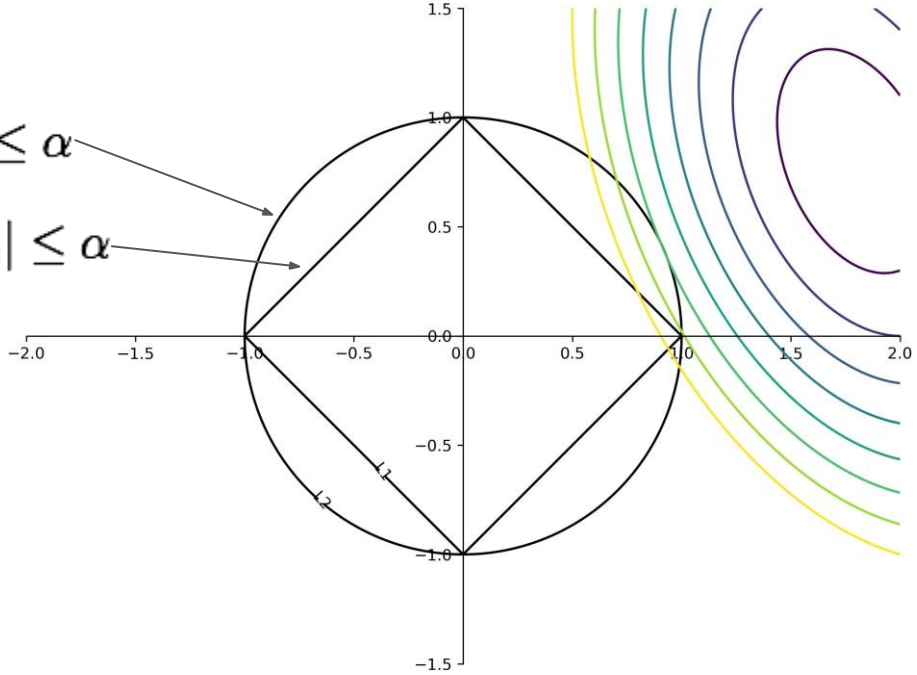
$$\min_{\boldsymbol{\omega}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\omega}||_2^2 + \alpha||\boldsymbol{\omega}||_1$$

Norma L1 vs Norma L2

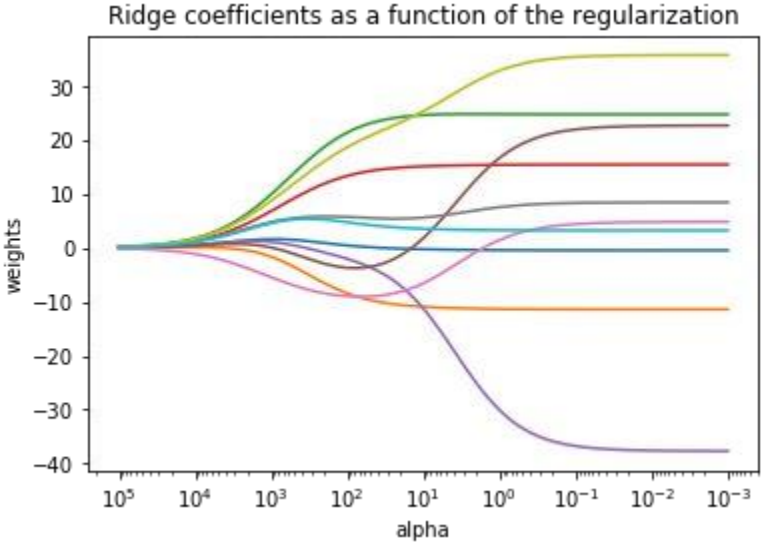
$$\min_{\boldsymbol{\omega}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\omega}||_2^2 + \alpha||\boldsymbol{\omega}||_2^2 \Rightarrow \min_{\boldsymbol{\omega}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\omega}||_2^2 \text{ sujeto a } ||\boldsymbol{\omega}||_2^2 \leq \alpha$$

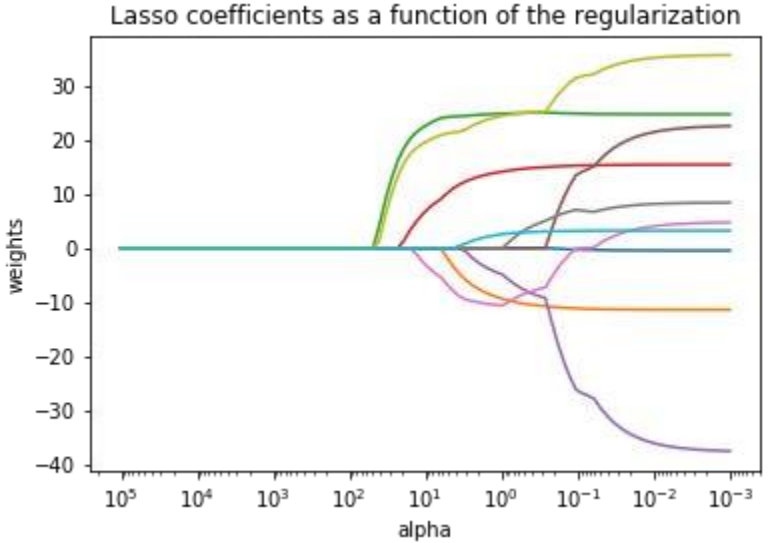
$$||\boldsymbol{\omega}||_2^2 = \omega_1^2 + \omega_2^2 \leq \alpha$$

$$||\boldsymbol{\omega}||_1 = |\omega_1| + |\omega_2| \leq \alpha$$



Lasso vs Ridge






```
from sklearn import datasets
diabetes = datasets.load_diabetes()
X = diabetes.data
y = diabetes.target
```

The Lasso path

Ridge

$$\min_{\omega} ||\mathbf{y} - \mathbf{X}\boldsymbol{\omega}||_2^2 + \alpha||\boldsymbol{\omega}||_2^2$$

- Afecta a todos los coeficientes (incluye todos o ninguno)
- Computacionalmente eficiente y previene *overfitting*
- Buen punto de partida para analizar un problema: por defecto, usarlo

Lasso

$$\min_{\omega} ||\mathbf{y} - \mathbf{X}\boldsymbol{\omega}||_2^2 + \alpha||\boldsymbol{\omega}||_1$$

- Capaz de anular algunos coeficientes (solución dispersa)
- Selección de características e interpretabilidad del modelo (también previene *overfitting*)

ElasticNet: https://en.wikipedia.org/wiki/Elastic_net_regularization

Lo que puedes hacer ahora ...

- Implementar algoritmos de regresión lineal (+regularización) y Lasso
- Entender cómo afecta el parámetro de regularización a los coeficientes del modelo
- Entender las diferencias entre Ridge y Lasso
- Entender cómo afecta el parámetro de regularización en regresión logística a la frontera de separación

Referencias

- An Introduction to Statistical Learning.
  - Capítulo 3.

- Hands On Machine Learning.
  - Capítulo 4

Let’s code!