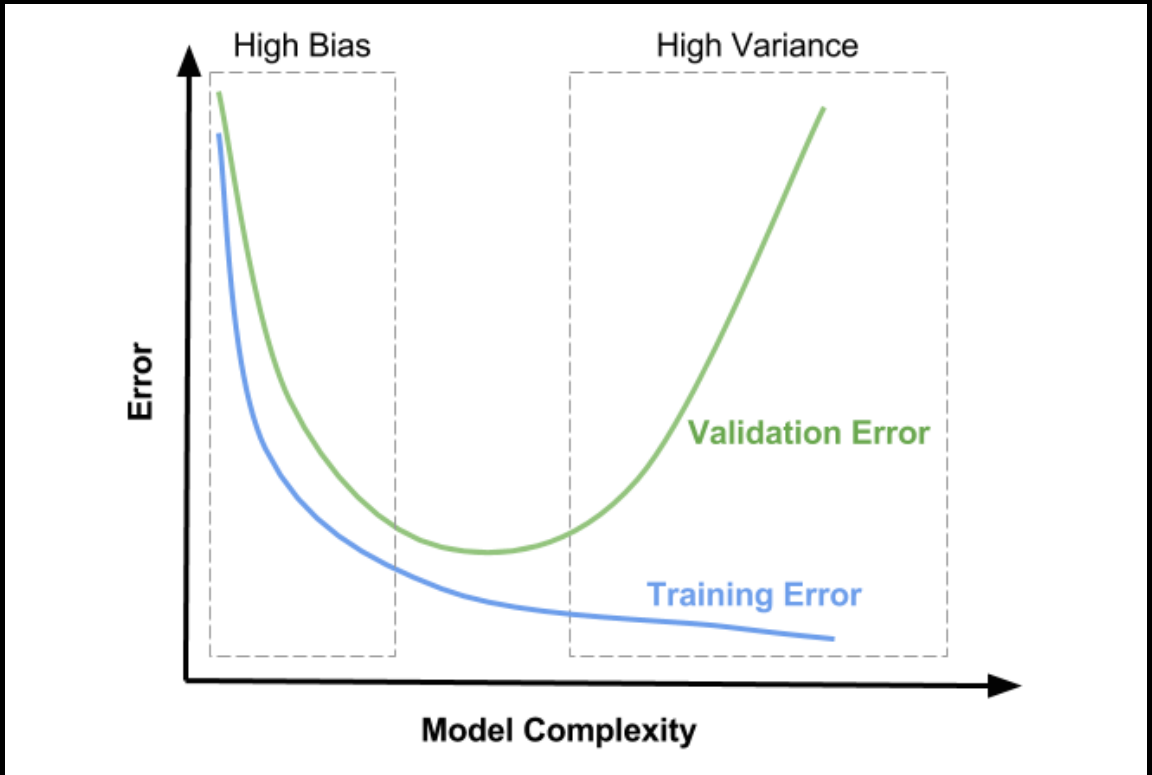
Machine Learning – Regularización

Underfitting Overfitting



Definición

Proceso que altera ligeramente la formulación matemática de un modelo, con la intención de **prevenir el overfitting**. Una manera de regularizar puede ser eliminando grados de una regresión polinómica o aplanando los pesos (w).

De esta manera simplificamos los modelos para que haya menos overfitting y generalicen mejor. A cambio sufriremos un pequeño aumento en los errores. De nuevo, Bias vs Variance.

Se suele aplicar en regresiones lineales y logísticas, pero también existen en otros modelos. Las penalizaciones más populares son:

1. **L2**: consigue que los parámetros estimados por el modelo (w) no tengan (en valor absoluto) un valor demasiado grande, de manera que aplana los pesos y evita los extremos.
2. **L1**: tiende a eliminar los pesos con menor importancia, es como si estuviese realizando un feature selection.

Regresiones que utilizan regularización:

1. **Ridge**: utiliza L2
2. **Lasso**: utiliza L1
3. **Elastic Net**: combinación lineal de L1 y L2

Ridge

Añade este nuevo término a la función de costes

$$\lambda \sum_{j=0}^{p} w_j^2$$

De tal manera que:

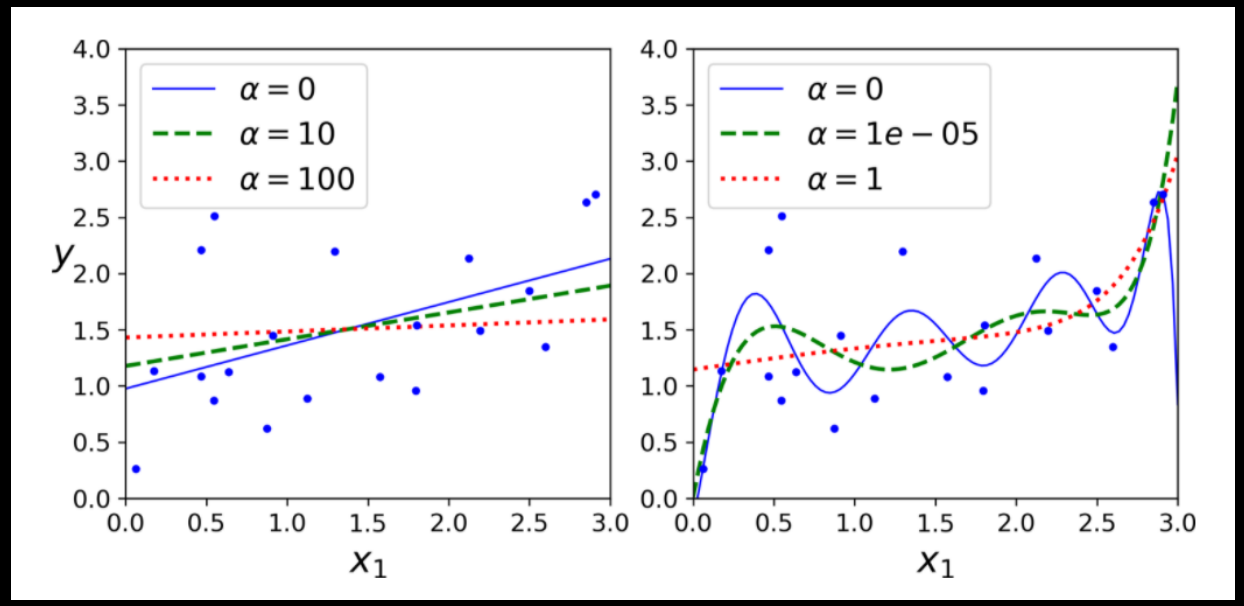
$$\sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j \times x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} w_j^2$$

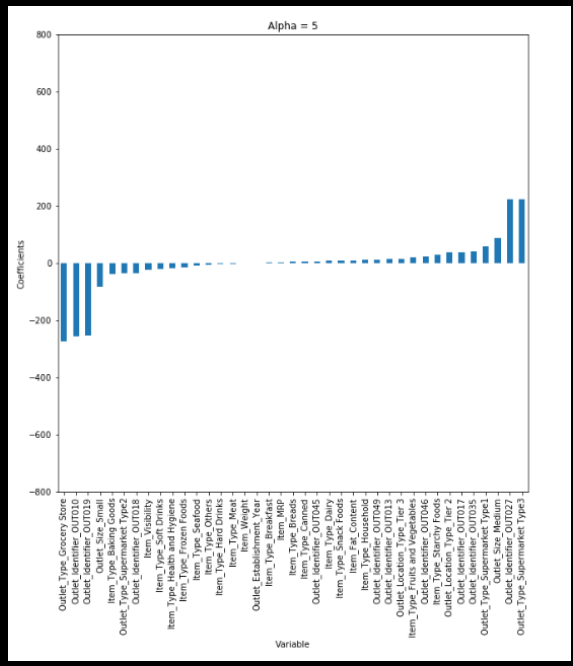
Cuanto mayor es Alpha, más estoy regularizando el modelo y la generalización debería ser mejor.

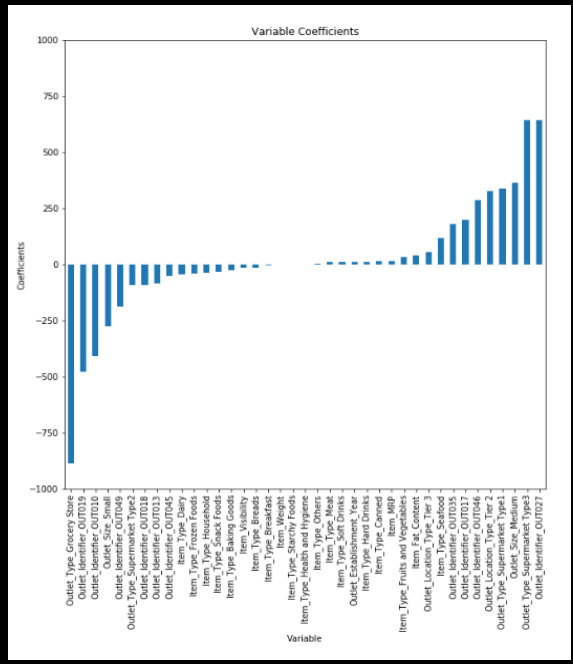
Por otro lado, el modelo final contará con todos los predictores

El hiperparámetro alpha controla cuánto regularizamos el modelo. Si alpha es 0, sería una regresión lineal normal. Sin embargo, si alpha es muy grande todos los pesos serían cercanos a 0 y el resultado de la regresión equivaldría a una línea plana.

Ridge


$\alpha = 0$
$\alpha = 10$
$\alpha = 100$
$\alpha = 0$
$\alpha = 1e-05$
$\alpha = 1$
y
x_1
x_1


Alpha = 5
Coefficients
Variable


Variable Coefficients
Coefficients
Variable

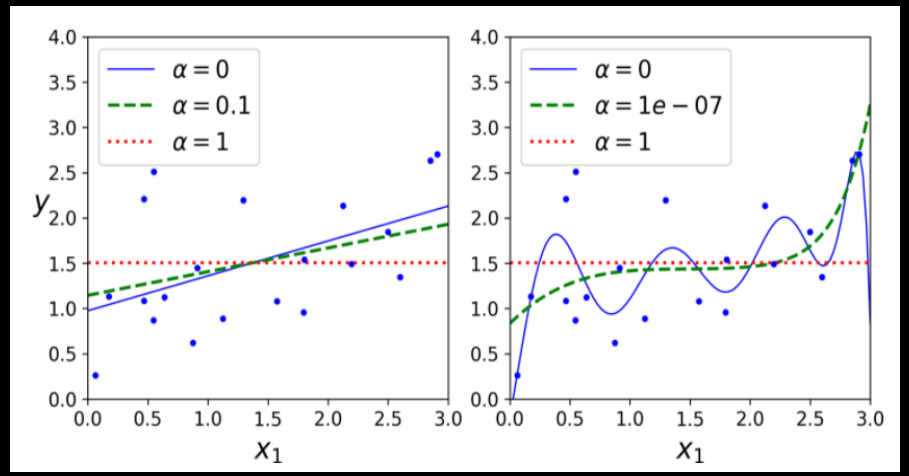
Lasso

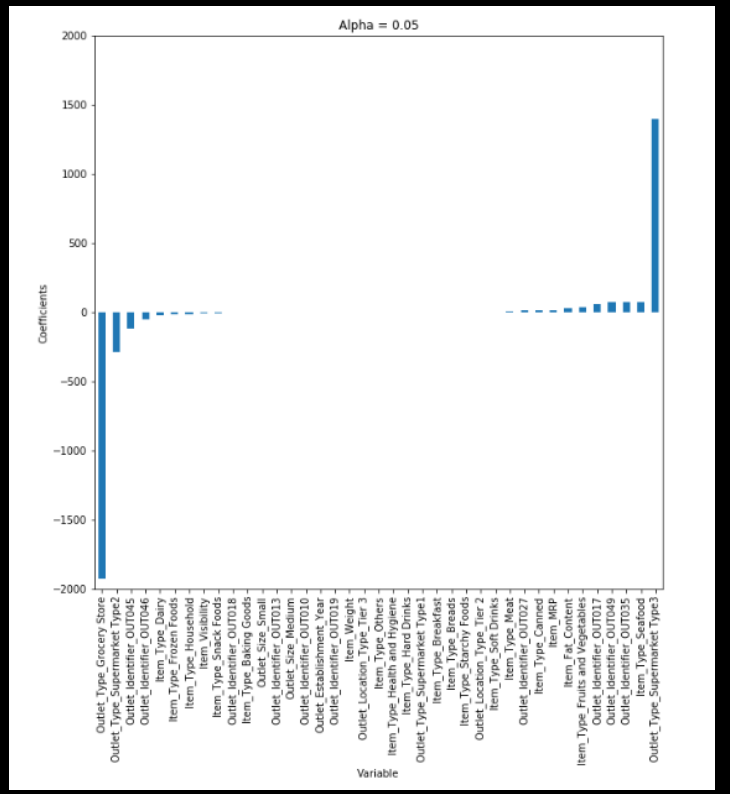
Least Absolute Shrinkage and Selection Operator Regression (Lasso). Añade un término de regularización a la función de costes, que en este caso es la norma l1 del vector de pesos

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p}|w_j|$$

La regresión de Lasso elimina los pesos de las variables menos importantes, como en la siguiente imagen donde pone a 0 los pesos de los grados altos de la regresión polinómica. **Es una forma de hacer feature selection.**

En presencia de correlaciones entre las variables Lasso puede tener comportamientos inestables





Elastic Net

Término medio entre la regresión de Ridge y la de Lasso. El término de regularización es una mezcla entre ambos:

$$J(\boldsymbol{\theta}) = \mathrm{MSE}(\boldsymbol{\theta}) + r\alpha \sum_{i=1}^{n} |\theta_i| + \frac{1-r}{2}\alpha \sum_{i=1}^{n} {\theta_i}^2$$

¿Cuándo usamos Ridge, Lasso o Elastic Net? En toda regresión siempre se recomienda algo de regularización para que no se produzca overfitting, y por tanto el modelo generalice mejor.

Por defecto Ridge funciona bastante bien aunque si sospechas que solo unas pocas features serán las buenas, quizá te encaje mejor Lasso o Elastic Net, ya que van a reducir o eliminar esas features.

Igualmente te encajarían estos dos últimos si tienes muchas features y quieres que la regularización realice una selección.
Elastic Net y Lasso funciona muy bien cuando tenemos muchas features.

Estandarización
Se recomienda utilizar el StandardScaler de sklearn, ya que los modelos que utilizan gradient descent son sensibles a las escalas de las variables.

Bibliografía

https://learning.oreilly.com/library/view/hands-on-machine-learning/9781492032632/

https://www.cienciadedatos.net/documentos/py14-ridge-lasso-elastic-net-python.html

https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a

https://towardsdatascience.com/ridge-lasso-and-elasticnet-regression-b1f9c00ea3a3