2.1.

1.

$$\hat{y} = \beta_0 + \beta_1 x + \cdots + \beta_p x^p =$$

$$= [1 \quad x \quad x^2 \quad \ldots \quad x^p] \begin{bmatrix} \beta_0 \\ \ldots \\ \beta_p \end{bmatrix}$$

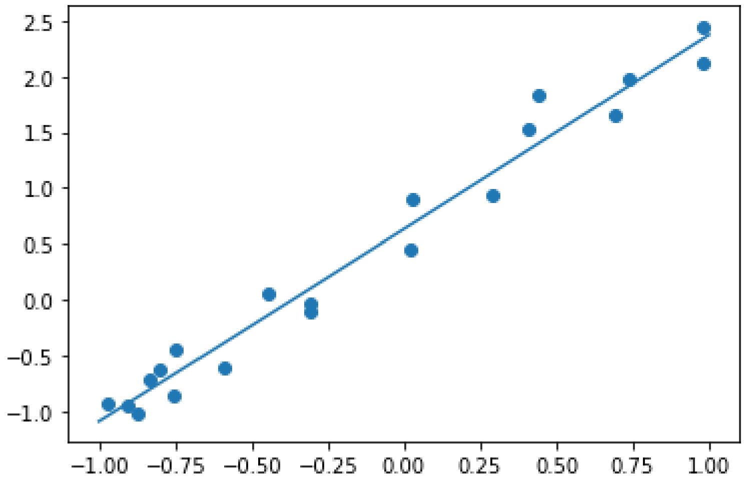
$$SSE = \sum_{i=1}^{N} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \left|\left|Y - X\beta\right|\right|^2 ,$$

$$com\ Y = \begin{bmatrix} y^{(1)} \\ \ldots \\ y^{(N)} \end{bmatrix},\ X = \begin{bmatrix} 1 & x_{(1)} & \ldots & x^p{}_{(1)} \\ 1 & \ldots & \ldots & \ldots \\ 1 & x_{(N)} & \ldots & x^p{}_{(N)} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_p \end{bmatrix}$$

$$\nabla_\beta SSE = 0 => \beta = (X^T X)^{-1} X^T Y$$

3.

a)



A reta obtida ajusta-se ao conjunto de pontos de treino

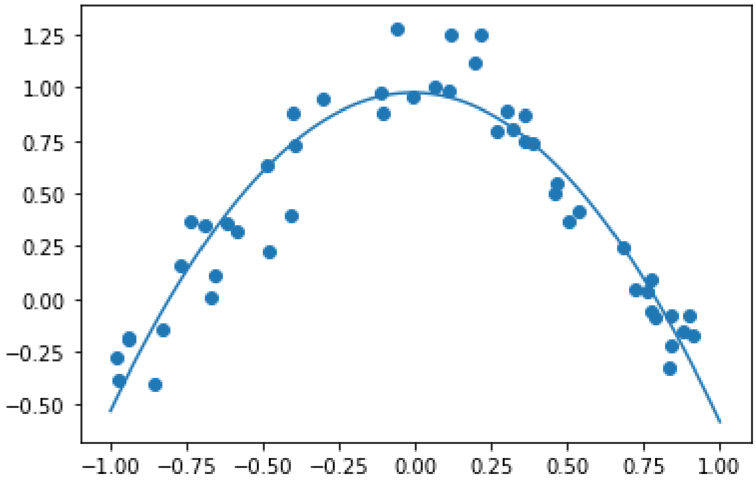
b) coeficientes: beta_0 = 0.63512224

beta_1 = 1.7332128

SSE = 0.74333541

4.

a)



A função parece ajustar-se razoavelmente ao conjunto de dados.

b) beta_0 = 0.97572

beta_1 = -0.0257195

beta_2 = -1.53224

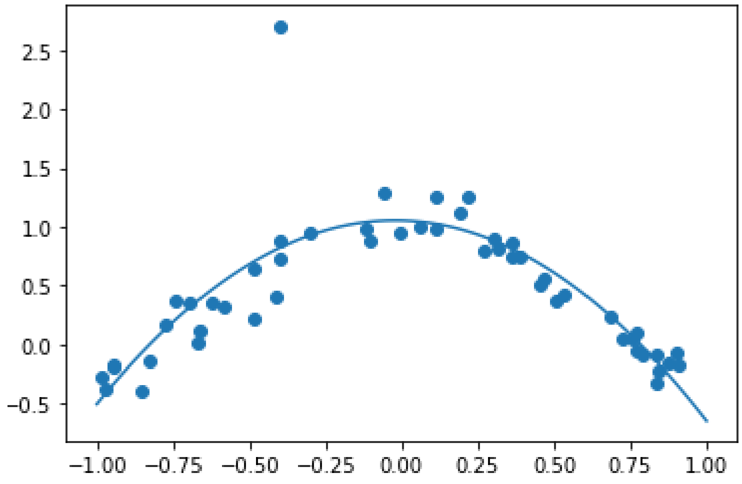
SSE = 1.34159372

A regressão polinomial de ordem 2 ajusta-se bastante bem ao conjunto de dados. Observamos que os dados se disponibilizam no espaço cartesiano de forma próxima a uma curva hiperbólica, pelo que deduzimos que seria melhor ajustar os dados a um polinómio de ordem 2.

Deste modo, foi melhor utilizar uma regressão para ordem 2 do que 1, já que verificámos também que não existe um ajuste demasiado aos dados e a curva generaliza suficientemente a sua distribuição espacial, tal como é demonstrado pelo valor bastante reduzido, mas também já bastante superior a 0 para o SSE.

5.

a)



A curva ajusta-se bastante bem à quase totalidade dos pontos, ao qual se exclui um ponto, denominado "outlier".

b) beta_0 = 1.05233523

beta_1 = -0.07159754

beta_2 = -1.63133339

SSE = 5.02487271

SSE_inliers = 1.49032572

Todos os outros pontos se encontram próximos da curva de ajuste e são denominados "inliers". Comparando a curva sem o ponto "outlier", com a outra com o ponto "inlier", verificamos que as curvas continuam a ser bastante próximas uma vez que existe apenas um ponto "outlier", que não tem muita influência na mudança dos coeficientes de cada uma das curvas.

2.2.

2.

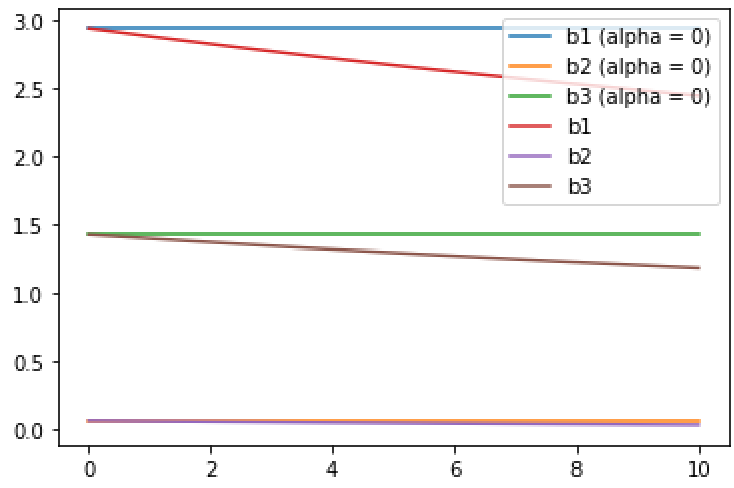
No método Ridge Regression, em vez de se tentar minimizar a expressão $||Y - X\beta||^2$, tenta-se minimizar a expressão $||Y - X\beta||^2 + \lambda||\beta||^2$. O termo $||\beta||^2$ penaliza a utilização de coeficientes de valor elevado. O objetivo é manter os coeficientes com o menor valor possível.

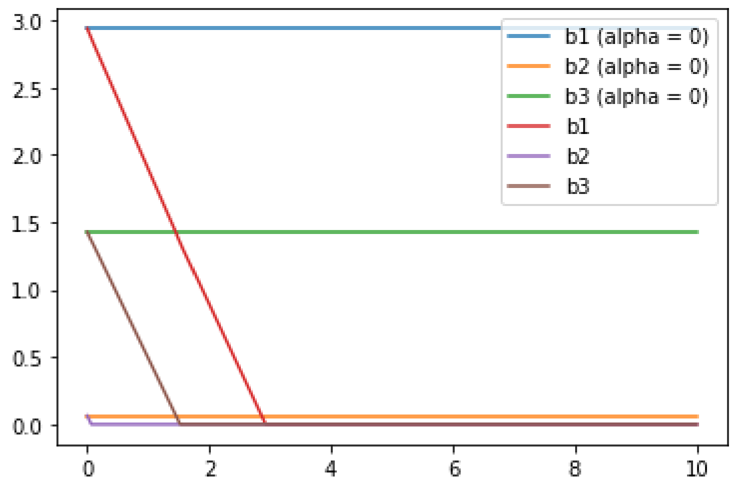
Já o método Lasso tenta minimizar a expressão $||Y - X\beta||_2^{\ 2} + \lambda||\beta||_1$

Ao contrário da Ridge Regression, em que os coeficientes tendem assintoticamente para 0 à medida que se aumenta lambda, no método Lasso os coeficientes podem assumir o valor 0.

Assim, este método pode ser utilizado para seleção de features: á medida que se aumenta lambda, os primeiros coeficientes a anularem-se serão os que correspondem a features irrelevantes.

6.



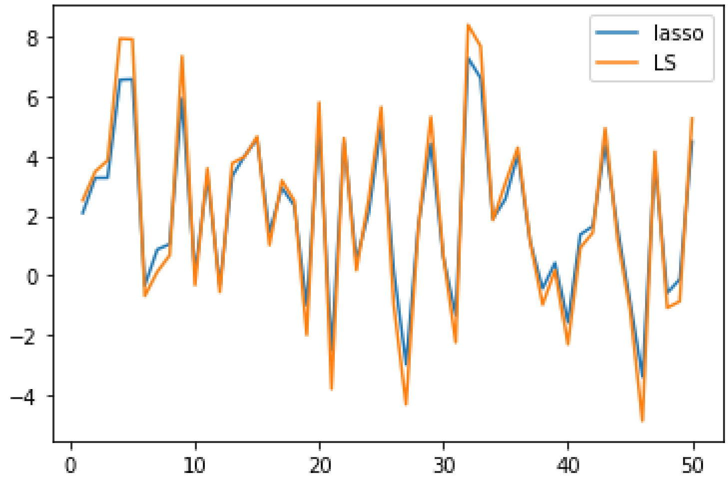


À medida que o valor de α aumenta, em ambos os casos os valores de todos os coeficientes diminuem, como seria de esperar.

Na regressão de Lasso, verificamos que para α suficientemente grande, os coeficientes anulam-se. O primeiro coeficiente a anular-se é o β_2, pelo que a feature irrelevante é x_2.

7.

Por exemplo α = 0.5, uma vez que $\beta_2 = 0$ e $\beta_1, \beta_3 \neq 0$



Comparando o plot dos valores de y obtidos pelo método dos mínimos quadrados e pelo método de Lasso, verificamos que os valores pelo método de Lasso têm menor variância que os obtidos pelo método dos mínimos quadrados. Isto resulta do facto de que os coeficientes obtidos pelo método de Lasso serem inferiores aos do método dos mínimos quadrados

SSE_LS = 14.9520

SSE_LASSO = 39.8905

O SSE é mais baixo na regressão linear de mínimos quadrados porque o objetivo dessa regressão é precisamente minimizar o SSE.