Aprendizagem Automática
Laboratório 1: **Regressão Linear**

N.º:90007 Nome: Alice Rosa
N.º:90026 Nome: Aprígio Malveiro

Turno: 3ª feira - 14h00

2.1 Least Squares Fitting

Questão 2.1.1
Expressão utilizada na estimativa dos coeficientes:

$$\beta = (X^T X)^{-1} X^T y \tag{1}$$

Com,

$$X = \begin{bmatrix} 1 & x_1^{(1)} & \dots & x_p^{(1)} \\ 1 & x_1^{(2)} & \dots & x_p^{(2)} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_1^{(n)} & \dots & x_p^{(n)} \end{bmatrix}, \quad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

Onde X é a matriz de delineamento, y é o vector com N valores de treino e β são os coeficientes. E SSE é dado por:

$$SSE(\beta) = ||y - X\beta||^2 \tag{2}$$

Questão 2.1.3

a) Os dados fornecidos nesta questão são aproximadamente lineares. O que leva à escolha do valor P igual a 1. Desta forma, obteve-se uma reta ajustada aos valores fornecidos.

b)

$$\beta_0 \approx 0.63512$$
$$\beta_1 \approx 1.73321$$
$$SSE \approx 0.74334$$

Questão 2.1.4

a) O modelo quadrático obtido nessa regressão adapta-se bem aos valores de treino dados, uma vez que estes pertencem apenas ao meio período positivo da função cosseno. No entanto, podemos verificar a partir do gráfico que alguns dos pontos se encontram ligeiramente afastados, tal pode ser justificado pelo ruído presente nos dados.

b)

$$\beta_0 \approx 0.97572$$
$$\beta_1 \approx -0.02572$$
$$\beta_2 \approx -1.53224$$
$$\text{SSE} \approx 1.34159$$

O coeficiente β_2 é negativo, que é coerente com a concavidade da parábola para baixo. O SSE não é muito elevado, sendo este erro devido principalmente ao ruído gaussiano presente nos dados.

Questão 2.1.5

a) O gráfico obtido nesta questão não é muito diferente do da alínea anterior. O modelo obtido continua a adaptar-se bem aos pontos "inliers", que significa que, neste caso, os pontos "outliers" não têm muita influência no modelo adquirido.

b)

$$\beta_0 \approx 1.08684$$
$$\beta_1 \approx 0.05616$$
$$\beta_2 \approx -1.60380$$
$$\text{SSE c/outlier} \approx 9.89011$$
$$\text{SSE s/outlier} \approx 1.81391$$

Pode-se verificar que o valor de SSE obtido nesta alínea é bastante superior ao calculado em 2.1.4. b), passando de 1.34 para 9.89. Assim, evidencia-se a grande sensibilidade do SSE relativamente aos pontos "outliers", como era de esperar. Uma vez que estes pontos estão bastante afastados do modelo obtido, a distância eucledian é elevada e, consequentemente, o SSE também será elevado.

Relativamente ao SSE obtido depois de se eliminar os "outliers", de aproximadamente 1.81, verifica-se que desceu bastante em relação ao valor anterior, no entanto, é ligeiramente superior ao erro "original". Visto que os "outliers" também influenciaram os coeficientes, β.

2.2 Regularization

Questão 2.2.2

O método de Ridge e Lasso são técnicas de regularização úteis quando a matriz $X^T X$ é singular e, consequentemente há infinitas soluções possíveis para o mesmo problema. Também ajudam a reduzir o over-fitting.

Os coeficientes de Ridge são determinados por:

$$\beta_{ridge} = \arg\min_{\beta} ||y - X\beta||^2 + \lambda||\beta||^2 \tag{3}$$

Onde o novo termo $||\beta||^2$ penaliza o uso de coeficientes com valores altos e é designado por termo de regularização. O parâmetro λ representa a "troca" entre os dois objectivos, que são a adaptação do modelo aos dados de treino (associado ao primeiro termo) e de teste (segundo termo). Quanto maior o valor de λ maior a adaptação aos dados de teste e menor aos de treino, e vice-versa para valores de λ baixos. No entanto, apesar dos coeficientes deste método tenderem para zero, nunca o atingem.

No modelo de Lasso os coeficientes são dados por:

$$\beta_{ridge} = \arg\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_1 \tag{4}$$

O último termo pode ser interpretado como o termo de regularização. Este método tem as seguintes características: não pode ser resolvido analiticamente e o vector β de coeficientes obtido é normalmente disperso ("sparse"), ou seja, quando uma característica não é importante o coeficiente correspondente é colocado a zero. Tal pode ser interpretado como selecção de características ("feature selection").

Questão 2.2.6

A partir dos gráficos adquiridos verifica-se que, para valores de α baixos, tanto o método de Ridge como o de Lasso apresentam coeficientes iguais ao método LS. Quando o valor de α começa a aumentar, os coeficientes de Ridge e de Lasso começam a diminuir.

Enquanto que no método de Ridge, os coeficientes diminuem lentamente e nunca atingem zero, os de Lasso divergem rapidamente atingindo todos valores nulos, pouco depois de $\alpha = 1$.

O coeficiente β_2, que já apresenta um valor baixo quando obtido pelo método LS, no método de Ridge e de Lasso é o mais rápido a chegar a zero, logo pode-se concluir que a característica irrelevante é x_2.

Questão 2.2.7

$$\alpha \approx 0.0410$$
$$\text{SSE LS} \approx 14.98201$$
$$\text{SSE Lasso} \approx 15.23260$$

A partir dos gráficos obtidos confirma-se que as curvas estão quase sobrepostas. Isto acontece porque, para o α escolhido, os parâmetros β são praticamente idênticos.

O SSE obtido pelo método de Lasso é ligeiramente superior em relação ao de LS, o que comprova que x_2 tinha uma certa influência no valor de y estimado, quando maior a relevância de x_2 maior será este valor de SSE.

A escolha do α foi feita de modo a reduzir o peso dos dados irrelevantes, sem comprometer o peso dos restantes.