1 Least squares Fitting

1.1

$\beta = (X^T X)^{-1} X^T y$, se $X^T X$ for invertível, ou seja, $det(X^T X) \neq 0$.

$$A = \begin{bmatrix} 1 & x^{(1)} & x^{(1)^2} & \dots & x^{(1)^p} \\ 1 & x^{(2)} & x^{(2)^2} & \dots & x^{(2)^p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x^{(N)} & x^{(N)^2} & \dots & x^{(N)^p} \end{bmatrix}$$

$$Y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix} \in R^N$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_P \end{bmatrix} \in R^{P+1}$$

1.3

1.3.1 Comentário

Os dados do training set variam de forma linear, por isso, foi escolhido grau 1 para a regressão de Least Square. Como era de se esperar, a regressão adapata-se bem ao training set, o que é visível no gráfico.

1.3.2 Coeficientes e SSE obtidos

$$\beta = \begin{bmatrix} 0.63512224 \\ 1.7332128 \end{bmatrix}$$
$$SSE = 0.74333541$$

1.4

1.4.1 Comentário

Existem alguns valores afastados do ajuste modelado devido principalmente ao ruído dos dados de entrada. Apesar disso, o modelo encontra-se bem aproximado em relação aos dados de entrada. Como descrito no enunciado, existe um ruído Gaussiano com soma nula. Ou seja verifica-se que para uma modelão polinomial de grau 1, ou seja regressão linear, este ruído não vai fazer muita diferença, pois a soma do ruído seré aproximadamente nula. No entanto, para um modelo polinomial de grau 2, a presença de ruído já vai ter impacto no modelo gerado, devido à parcela elevada ao quadrado.

1.4.2 Coeficientes e SSE obtidos

$$\beta = \begin{bmatrix} 0.97571963 \\ -0.02571951 \\ -1.53223529 \end{bmatrix}$$

$$SSE = 1.34159372$$

Verifica-se uma Soma do Erro Quadrático, SSE, relativamente elevada, devido principalmente ao ruído existente nos dados de treino.

1.5

1.5.1 Comentário

Verifica-se um modelo mais próximo da maioria dos dados, apesar da presença de alguns pontos, ditos outliers, bastante afastados do modelo.

1.5.2 Coeficientes e SSE obtidos

$$\beta = \begin{bmatrix} 0.61055486 \\ -0.02906402 \\ -1.60379935 \end{bmatrix}$$

$$SSE = 9.89010685$$

Verifica-se um SSE bastante mais elevado que anteriormente, pois com a remoção de alguns pontos com ruído o modelo foi capaz de se aproximar mais à maioria dos dados de teste, mas ao existirem 2 dados bastante atípicos, estes encontram-se muito afastados do modelo gerado o que leva a um valor elevado do SSE.

Retirando os outliers obtém-se um SSE = 1.331138, verificando-se assim um SSE menor que na questão 4, o que faz sentido, pois os dados de teste sem ruído e sem a presença de outliers, permitem gerar um modelo mais preciso. Isto demonstra a elevada sensiblidade do SSE à presença de outliers, mesmo que em baixo número.

2 Regularization

2.1 Métodos de Regularização de Ridge e Lasso

As regressões de Ridge e Lasso são técnicas de regularização, que pretendem reduzir a complexidade dos modelos, ajudando a prevenir *over-fitting*, que podem resultar de regressões lineares mais simples, como o modelo dos Least Squares. Há dois parâmetros que se pertendem diminuir quando se quer estudar cada modelo, sendo o bias e variância. Um bias baixo equivale a uma boa adaptação ao training set, já uma variância baixa equivale a uma boa adaptação aos testing sets. É necessário referir que quando um parâmetro diminui, o outro aumenta. As regressões como a de Least Squares têm um bias menor,

pois dá relevância a todas as features, contudo possui variância alta. É preferível encontrar um ponto onde se combinem bias e variância baixas para se ter uma boa adaptação tanto a training sets como a testing sets.

Na regressão de Ridge obtêm-se os coeficientes da seguinte forma

$$\beta^{ridge} = argmin_\beta \|y - X\beta\| + \alpha \|\beta\|_2{}^2. \tag{1}$$

Nesta regressão, os coeficientes beta são reduzidos, diminuindo a variância e aumentando o bias um pouco. Esta diminuição dos valores dos coeficiente faz com que haja uma redução na multicolinearidade entre features. É de notar que os coeficientes tendem para zero, nunca o atingem.

Na regressão de Lasso obtêm-se os coeficientes da seguinte forma

$$\beta^{lasso} = argmin_\beta \|y - X\beta\| + \alpha \|\beta\|_1. \tag{2}$$

Este modelo pode ter coeficientes nulos, então pode-se considerar irrelevantes as features cujo coeficiente é nulo. Assim, é possível usar o modelo de Lasso para seleção de features.

2.6 Irrelevant Feature

Inicialmente, para um valor de alfa = 0, os coeficientes dos métodos utilizados são idênticos aos da regressão de Least Squares. À medida que alfa aumenta,os coeficientes tanto do método Ridge como do Lasso, começam a diminuir.

Na regressão de Lasso, verifica-se uma descida abrupta dos coeficientes, atingindo todos zero a certa altura. Assim, consegue-se identificar a feature irrelevante do conjunto de dados verificando qual o coeficiente que se anula primeiro, com a crescimento de alfa. Neste caso, identifica-se a 2a feature como sendo irrelevante, pois é o coeficiente B_2 que se anula primeiro.

Tem-se assim o melhor alfa = 0.071, para quando o B_2 é igual a 0 pela primeira vez. Para este valor de alfa temos os seguintes coeficientes de Lasso:

$$\beta = \begin{bmatrix} 2.85952 \\ 0 \\ 1.36337 \end{bmatrix}$$

2.7

O valor adequado de alfa é o menor para o qual um dos coeficientes é 0. Assim o valor escolhido é alfa=0.071, sendo se fosse maior implicaria uma diminuição nos outros coeficientes.

No gráfico, é possível observar que para quase todos os casos, os valores previstos pela regressão Least Squares e a de Lasso são muito próximos, o que confirma que desprezando a feature x_2, o fit é na mesma muito próximo do que se obtém se não a desprezarmos. Utilizar a regressão de Lasso é vantajoso porque só se precisam de calcular dois coeficientes em vez de três, demorando menos tempo na sua execução.

Tem-se $LassoSSE = 15.69425$ e $LSSSSE = 14.98201$. Era de se esperar que o SSE obtido pela regressão de Lasso seja superior ao obtido pela regressão de LSS, pois, mesmo que seja pouco relevante, a feature desprezada tem alguma influência no valor de y.

Neste caso, era apenas necessário calcular três coeficientes, logo o ganho em tempo de execução é desprezável relativamente à perda de precisão do modelo. Contudo, se for necessário calcular muitos coeficientes a regressão de Lasso pode vir a ser útil.