Aula prática 4
Métodos dos Mínimos Quadrados

Will Sena*

Contents

*wllsena@protonmail.com

1)

Resolva o exercício 32, da seção 7.3, página 588, do livro Álgebra Linear, tradução da 4ª edição americana, David Poole. Relate a modelagem utilizada e use o SciLab para os cálculos

32. **Quando um objeto é arremessado para cima, a Segunda Lei de Newton para o movimento afirma que sua altura $s(t)$ no tempo t é dada por**

$$s(t) = s_0 + v_0 t + \frac{1}{2} g t^2$$

em que v_0 é sua velocidade inicial e g é constante de aceleração da gravidade. Considere as medidas mostradas na tabela a seguir:

Tempo (s)	0.5	1	1.5	2	3
Altura(m)	11	17	21	23	18

(a) **Encontre a aproximação quadrática por mínimos quadrados para esses dados.**

```
--> S = [0.5 11; 1 17; 1.5 21; 2 23; 3 18]
S =

0.5 11.
1. 17.
1.5 21.
2. 23.
3. 18.

--> A = [S(:,1)^0 S(:,1)^1 S(:,1)^2]
A =

1. 0.5 0.25
1. 1. 1.
1. 1.5 2.25
```

```
1. 2. 4.
1. 3. 9.

--> b = S(:,2)
b =

11.
17.
21.
23.
18.

--> x_ = Gaussian_Elimination(A'*A, A'*b)
x_ =

1.9175258
20.306333
-4.9720177
```

$$s(t) = 1.9175258 + 20.306333t - \frac{9.9440353}{2}t^2$$

(b) **Estime a altura na qual o objeto foi solto (em m), sua velocidade inicial (em m/s) e sua aceleração da gravidade (em m/s^2)**

- Altura inicial $= s(0) = s_0 = 1.9175258$ m
- Velocidade inicial $= s'(0) = v_0 = 20.306333$ m/s
- Aceleração da gravidade $= -s''(0) = -g = 9.9440353$ m/s^2

(c) **Quando, aproximadamente, o objeto atingirá o chão?**

$$0 = 1.9175258 + 20.306333t - \frac{9.9440353}{2}t^2 \to t = 4.1764653$$

Extra)

A seguir uma função genérica para encontrar uma determinada aproximação de mínimos quadrados e o respectivo erro quadrático:

```
function [w, e] = lms(S, combinations_for_products)
  m = size(S, 1);
  n = size(combinations_for_products);
  A = zeros(m, n);
  b = S(:, size(S,2));

  for i = 1:n
    comb = combinations_for_products(i);
    col = ones(m, 1);
    for j = 1:size(comb, 2)
      col = col .* S(:, comb(j));
    end
    A(:, i) = col;
  end

  w = Gaussian_Elimination(A'*A, A'*b);
  e = norm(A*w-b, 2);
endfunction
```

Função para prever o resultado de uma determinada aproximação de mínimos quadrados com um vetor de valores.

```
function y = prev(w, combinations_for_products, x)
  m = size(x, 1);
  n = size(combinations_for_products);
  A = zeros(m, n);

  for i = 1:n
    comb = combinations_for_products(i);
    col = ones(m, 1);
    for j = 1:size(comb, 2)
      col = col .* x(:, comb(j));
    end
    A(:, i) = col;
```

```
  end

  y = A * w;
endfunction
```

Exemplo:

32. (a) ***Questão anterior***

    Para a equação dada ***combinations_for_products*** é igual à ***list([], [1], [1 1])*** (uma constante multiplicada por nenhuma variável, uma constante multiplicada pela variável de posição 1, x no caso, e uma constante multiplicada pela variável de posição 1 duas vezes, x^2 no caso.

```
--> S = [0.5 11; 1 17; 1.5 21; 2 23; 3 18];

--> combinations_for_products = list([], [1], [1 1]);

--> [w, e] = lms(S, combinations_for_products)
w =

1.9175258
20.306333
-4.9720177
e =

0.5148745
```

$$s(t) = 1.9175258 + 20.306333t - \frac{9.9440353}{2}t^2$$

   (b) **Altura inicial**

```
--> y = prev(w, combinations_for_products, [0])
y =

1.9175258
```

36. **Encontre o plano $z = a + bx + cy$ que melhor ajusta os pontos $(0, -4, 0)$, $(5, 0, 0)$, $(4, -1, 1)$, $(1, -3, 1)$ e $(-1, -5, -2)$.**

Para a equação dada ***combinations_for_products*** é igual à ***list([], [1], [2])*** (uma constante multiplicada por nenhuma variável, uma constante multiplicada pela variável de posição 1, x no caso, e uma constante multiplicada pela variável de posição 2, y no caso).

```
--> S = [0 -4 0; 5 0 0; 4 -1 1; 1 -3 1; -1 -5 -2];

--> [w, e] = lms(S, list([], [1], [2]))
w =

14.333333
-2.6666667
3.6666667
e =

1.6329932
```

$$z = 14.333333 - 2.6666667x + 3.6666667y$$

2)

Resolva o exercício 33, da seção 7.3, página 588, do livro Álgebra Linear, tradução da 4ª edição americana, David Poole. No item b), pesquise a população real dos EEUU em 2010 e compare com a população que você estimou. Relate a modelagem utilizada e use o SciLab para os cálculos.

33. **A tabela a seguir mostra a população dos Estados Unidos em intervalos de dez anos, no período de 1950 a 2000:**

Ano	População (em milhões)
1950	150
1960	179
1970	203
1980	227
1990	250
2000	281

(a) **Supondo um modelo de crescimento exponencial da forma $p(t) = ce^{kt}$, em que $p(t)$ é a população em um tempo t, utilize mínimos quadrados para encontrar a equação para a taxa de crescimento da população. [Sugestão: considere $t = 0$ para 1950].**

$$ln(p(t)) = ln(c) + kt \tag{1}$$

$$y = c2 + kt \text{ para } y = ln(p(t)) \text{ e } c2 = ln(c) \tag{2}$$

```
--> S = [1950 log(150); 1960 log(179); 1970 log(203);
         1980 log(227); 1990 log(250); 2000 log(281)];

--> w = lms(S, list([], [1]))
w =

-18.647292
0.0121502
```

$$y = -18.647292 + 0.0121502t \tag{1}$$

$$ln(p(t)) = -18.647292 + 0.0121502t \tag{2}$$

$$p(t) = e^{-18.647292}e^{0.0121502t} \tag{3}$$

```
--> ts = [1950; 1960; 1970; 1980; 1990; 2000];

--> ps = [150; 179; 204; 227; 250; 281];

--> e = norm(exp(ts * 0.0121502) * exp(-18.647292) - ps, 2)
```

```
e =

10.570975
```

- Obs: considerando $t = 0$ para 1950 obtém-se a equação:

$$p(t) = 150e^{0.1310316t}$$

```
--> ts = [0; 1; 2; 3; 4; 5];

--> ps = [150; 179; 204; 227; 250; 281];

--> e = norm(exp(ts * 0.1310316) * 150 - ps, 2)
e =

15.526688
```

Um erro quadrático maior...

(b) **Use a equação obtida para estimar a população dos Estados Unidos em 2010.**

```
--> exp(2010 * 0.0121502) * exp(-18.647292)
ans =

322.01882
```

Valor real: 309.3 milhões

3)

Resolva o exercício 34, da seção 7.3, página 588, do livro Álgebra Linear, tradução da 4ª edição americana, David Poole. Relate a modelagem utilizada e use o SciLab para os cálculos.

34. **A tabela a seguir mostra a média salarial da liga adulta de beisebol para os anos de 1970 a 200:**

Ano	Média salarial (milhares de dólares)
1970	29.3
1975	44.7
1980	143.8
1985	371.6
1990	597.5
1995	1110.8
2000	1895.6
2005	2476.6

(a) **Encontre a aproximação quadrática por mínimos quadrados para esses dados.**

```
--> S = [1970 29.3; 1975 44.7; 1980 143.8; 1985 371.6;
         1990 597.5; 1995 1110.8; 2000 1895.6; 2005 2476.6];

--> [w, e] = lms(S, list([], [1], [1 1]))
w =

10086382.
-10219.589
2.5886432
e =

174.19173
```

$$s(t) = 10086382 - 10219.589t + 2.5886432t^2$$

(b) **Encontre a aproximação exponencial por mínimos quadrados para esses dados.**

$$s(t) = c + e^{kt} \quad (1)$$

$$ln(s(t)) = ln(c) + kt \quad (2)$$

$$y = c2 + kt \text{ para } y = ln(p(t)) \text{ e } c2 = ln(c) \quad (3)$$

```
--> S = [1970 log(29.3); 1975 log(44.7); 1980 log(143.8); 1985 log(371.6);
      1990 log(597.5); 1995 log(1110.8); 2000 log(1895.6); 2005 log(2476.6)];

--> w = lms(S, list([], [1]))
w =

-261.05853
0.1342956
```

$$y = -261.05853 + 0.1342956t \tag{1}$$

$$ln(s(t)) = -261.05853 + 0.1342956t \tag{2}$$

$$s(t) = e^{-261.05853}e^{0.1342956t} \tag{3}$$

```
--> ts = [1970; 1975; 1980; 1985; 1990; 1995; 2000; 2005];

--> ss = [29.3; 44.7; 143.8; 371.6; 597.5; 1110.8; 1895.6; 2476.6];

--> e = norm(exp(ts * 0.1342956) * exp(-261.05853) - ss, 2)
e =

1201.7103
```

(c) **Qual equação dá uma melhor aproximação? Por quê?**

A aproximação quadrática por mínimos quadrados, por ter um erro quadrático menor (174.19173 em relação ao erro de 1201.7103 da aproximação exponencial por mínimos quadrados).

(d) **Qual a sua estimativa para a média salarial da liga adulta de beisebol em 2010 e em 2015?**

- Para 2010:

  A aproximação quadrática por mínimos quadrados (estimativa principal):

```
--> 10086382 - 10219.589*2010 + 2.5886432*2010^2
ans =

3385.5023
```

A aproximação exponencial por mínimos quadrados:

```
--> exp(2010 * 0.1342956) * exp(-261.05853)
ans =

7155.4244
```

Valor real: 3014.572

- Para 2015:

  A aproximação quadrática por mínimos quadrados (estimativa principal):

```
--> 10086382 - 10219.589*2015 + 2.5886432*2015^2
ans =

4384.0017
```

A aproximação exponencial por mínimos quadrados:

```
--> exp(2015 * 0.1342956) * exp(-261.05853)
ans =

14004.080
```

Valor real: 3952.252

4)

Agora vamos usar o método dos mínimos quadrados para implementar um método rudimentar de "machine learning" para diagnosticar câncer de mama a partir de um conjunto de características fornecidas para cada paciente. São dados dois arquivos: um arquivo para "treinamento" (cancer_train.csv) do modelo e um arquivo para "teste" (cancer_test.csv).

O primeiro arquivo contém 300 registros e o segundo 260 registros, partes do "Wisconsin Diagnostic Breast Cancer dataset". Cada registro de cada arquivo contém 11 valores: os 10 primeiros correspondem a valores reais de 10 características dos núcleos celulares observados em imagens digitalizadas de uma fina camada de massa mamária coletada de cada paciente. O décimo primeiro valor é $+1$ se a paciente tem câncer de mama e -1, caso contrário.
Sendo $\boldsymbol{x}$ o vetor das 10 características de cada paciente (variáveis independentes) e $\boldsymbol{y}$ o valor ($+1$ ou -1) que indica o diagnóstico (variável dependente), a ideia é, usando o arquivo de treinamento, obter o hiperplano $\boldsymbol{y} = h(\boldsymbol{x})$ ($\boldsymbol{y} = \alpha_0 + \sum_{i=1}^{10} \alpha_i \boldsymbol{x_i}$) que "melhor se ajuste aos dados fornecidos" usando o método dos mínimos quadrados. Uma vez obtido o hiperplano, o mesmo será usado para classificar cada paciente da seguinte forma: se $h(x) \geq 0$, então o diagnóstico é $+1$ (tem câncer), caso contrário, o diagnóstico é -1 (não tem câncer). Use o seu classificador (hiperplano) e calcule a porcentagem de acertos sobre o arquivo de treinamento (de certa forma é uma medida do ajuste do seu modelo aos dados de treinamento) e sobre o arquivo de teste (de certa forma é uma medida da capacidade de generalização do seu modelo). Construa uma Matriz de Confusão (Confusion Matrix) (pesquise a respeito) com o conjunto de teste e calcule as diversas medidas daí decorrentes, tais como: acurácia, precisão, recall, probabilidade de falso alarme, probabilidade de falsa omissão de alarme. Interprete essas medidas e comente os resultados obtidos.

- Carregando os dados

```
--> cancer_train = csvRead("cancer_train.csv");

--> cancer_test = csvRead("cancer_test.csv");

--> size(cancer_train)
ans =

300. 11.
```

```
--> size(cancer_test)
ans =

260. 11.

--> cancer_train(1:5, :)
ans =

column 1 to 5

0.64 0.2643 0.6515 0.4006 0.8182
0.7318 0.4524 0.705 0.5306 0.5856
0.7005 0.541 0.6897 0.4814 0.7574
0.4063 0.5188 0.4116 0.1545 0.9848
0.7218 0.3651 0.7167 0.519 0.6932

column 6 to 11

0.8037 0.7031 0.7311 0.7957 0.8078 1.
0.2277 0.2036 0.3488 0.5961 0.5816 1.
0.4629 0.4625 0.6357 0.6806 0.6157 1.
0.8219 0.5656 0.5229 0.8543 1. 1.
0.3845 0.4639 0.5184 0.5951 0.6038 1.
```

- Treinando o modelo:

```
--> combinations_for_products = list([], [1], [2], [3], [4], [5],
[6], [7], [8], [9], [10]);

--> [w e] = lms(cancer_train, combinations_for_products)
w =

-6.7579731
29.311052
2.0765803
-18.730222
-7.3665161
1.2222756
0.2283419
```

```
0.0503253
2.2385058
0.0249405
0.7704282
e =

10.139367
```

- Prevendo os conjuntos de dados

```
--> cancer_prev_train = prev(w, combinations_for_products, cancer_train(:, 1:10));

--> cancer_prev_train = cancer_prev_train ./ abs(cancer_prev_train);

--> cancer_prev_test = prev(w, combinations_for_products, cancer_test(:, 1:10));

--> cancer_prev_test = cancer_prev_test ./ abs(cancer_prev_test);
```

- Calculado porcentagens de acertos

```
--> sum(cancer_prev_train == cancer_train(:, 11)) / 300
ans =

0.93

--> sum(cancer_prev_test == cancer_test(:, 11)) / 260
ans =

0.7115385
```

93% de acerto para o arquivo de treinamento e 71% para o arquivo de teste. O modelo se ajustou bem no conjunto de dados de treinamento, mas não teve resultados igualmente bons para o conjunto de dados de teste, provavelmente o modelo está em overfitting.

- Calculando a Matriz de Confusão

```
--> TP = sum(cancer_prev_test == 1 & cancer_test(:, 11) == 1)
TP =

60.

--> FN = sum(cancer_prev_test == -1 & cancer_test(:, 11) == 1)
FN =

0.

--> FP = sum(cancer_prev_test == 1 & cancer_test(:, 11) == -1)
FP =

75.

--> TN = sum(cancer_prev_test == -1 & cancer_test(:, 11) == -1)
TN =

125.
```

Matriz de confusão:

Total: 260.	PP: 135.	PN: 125.	BM: 0.625	BA: 0.8125
P: 60.	TP: 60.	FN: 0.	TPR: 1.	FNR: 0.
N: 200.	FP: 75.	TN: 125.	FPR: 0.375	TNR: 0.625
Prev: 0.2307692	PPV: 0.4444444	FOR: 0.	LR+: 2.6666667	LR-: 0.
PT: 0.3797959	FDR: 0.5555556	MPV: 1.	MK: 0.4444444	DOR: Inf
ACC: 0.7115385	F1: 0.6153846	FM: 0.6666667	TS: 0.4444444	MCC: 0.5270463

. Total = População total = $P + N$
. P = Condição positiva = $TP + FN$
. N = Condição negativa = $FP + TN$
. Prev = Prevalência = $\frac{P}{P+N}$
. PT = Prevalence threshold = $\frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
. ACC = Acurácia = $\frac{TP+TN}{P+N}$
. PP = Condição positiva prevista = $TP + FP$
. TP = Verdadeiro positivo

. FP = Falso positivo, Erro tipo 1
. PPV = Valor preditivo positivo, Precisão = $\frac{TP}{PP}$
. FDR = Taxa de falsa descoberta = $\frac{FP}{PP}$
. F1 = F1 score = $\frac{2TP}{2TP+FP+FN}$
. PN = Condição negativa prevista = $FN + TN$
. FN = Falso negativo, Erro do tipo II
. TN = Verdadeiro negativo
. FOR = Taxa de Falsa Omissão = $\frac{FN}{PN}$
. NPV = Valor preditivo negativo = $\frac{TN}{PN}$
. FM = Fowlkes–Mallows index = $\sqrt{PPV \times TPR}$
. BM = Bookmaker informedness, Informedness = $TPR + TNR - 1$
. TPR = Taxa de Verdadeiro Positivo, Revocação, Sensibilidade, probabilidade de detecção, Potência = $\frac{TP}{P}$
. FPR = Taxa de Falso Positivo, Fall-out, probabilidade de alarme falso = $\frac{FP}{N}$
. LR+ = Teste da razão de verossimilhança positiva = $\frac{TPR}{FPR}$
. MK = Markedness, deltaP = $PPV + NPV - 1$
. TS = Threat score, critical success index (CSI) = $\frac{TP}{TP+FN+FP}$
. BA = Balanced accuracy = $\frac{TPR+TNR}{2}$
. FNR = Taxa de Falso Negativo, Taxa de perda = $\frac{FN}{P}$
. TNR = True negative rate, specificity (SPC), selectivity = $\frac{TN}{N}$
. LR- = Teste da razão de verossimilhança negativa = $\frac{FNR}{TNR}$
. DOR = Razão de possibilidades de diagnóstico (DOR) = $\frac{LR+}{LR-}$
. MCC = Matthews correlation coefficient = $\sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$

A acurácia (ACC) é o valor obtido anteriormente (porcentagem de acertos).
A precisão (PPV) foi bem baixa (0.44%), ou seja, o modelo acertou uma menor parte das previsões positivas. Porém o recall (TPR) é de 100%, indiciando que o modelo previu corretamente todos os valores positivos. Entenda que é diferente acertar uma previsão positiva e prever corretamente um valor positivo.
A probabilidade de falso alarme (FPR) foi bem baixa (0.375%), mostra que o modelo errou a previsão de poucos dos valores negativos. Enquanto que a probabilidade de falsa omissão de alarme (FOR) foi nula, isto é, o modelo errou nenhuma das previsões negativas.