Relatório da Aula Prática 4

Sillas Rocha da Costa

21 de maio de 2024

Questão 01

Dados fornecidos:

Ano	P	L	K
1899	100	100	100
1900	101	105	107
1901	112	110	114
1902	122	117	122
1903	124	122	131
1904	122	121	138
1905	143	125	149
1906	152	134	163
1907	151	140	176
1908	126	123	185
1909	155	143	198
1910	159	147	208
1911	153	148	216
1912	177	155	226
1913	184	156	236
1914	169	152	244
1915	189	156	266
1916	225	183	298
1917	227	198	335
1918	223	201	366
1919	218	196	387
1920	231	194	407
1921	179	146	417
1922	240	161	431

Alternativa a

Temos:

$$P = bL^{\alpha}K^{1-\alpha}$$

E procuramos os parâmetros b e α que melhor representem o nosso modelo, deste modo, para achar os melhores parâmetros por mínimos quadrados, precisamos modelar o problema de forma a fazer os parâmetros virarem escalares e deixarem de serem expoentes, visando aplicar mínimos quadrados. Assim:

$$\ln(P) = \ln(bL^{\alpha}K^{1-\alpha}) = \ln(b) + \ln(L^{\alpha}) + \ln(K^{1-\alpha})$$

$$\ln(P) = \ln(b) + \alpha \cdot \ln(L) + (1-\alpha) \cdot \ln(K)$$

$$\ln(b) + \alpha \cdot \ln(L) + ln(K) - \alpha \cdot \ln(K) = \ln(P)$$

$$\alpha \cdot (\ln(L) - \ln(K)) + \ln(b) = \ln(P) - ln(K)$$

Assim, obtemos uma modelagem que nos permite achar, com um pouco de álgebra linear, por mínimos quadrados, os melhores parâmetros procurados. Para simplificar, seja $y = \ln(P) - \ln(K)$, $x = \ln(L) - \ln(K)$ e $c = \ln(b)$. Temos então a equação linear:

$$y = \alpha x + c$$

Ou, $Ax = b$, onde A é a matriz com a primeira coluna como $\ln(L) - \ln(K)$, e a segunda coluna de uns, x o vetor dos parâmetros $x = \begin{pmatrix} \alpha \\ \ln(b) \end{pmatrix}$, e, por fim, o vetor de previsões formado por $b = \ln(P) - \ln(K)$, deste modo, para obter a melhor aproximação dos parâmetros de x, resolveremos

$$A^T Ax = A^T b$$

No código, chegaremos em:

```
--> P = dados(:,2);

--> L = dados(:,3);

--> K = dados(:,4);


--> A = zeros(size(P, 1), 1);

--> A(:, 1) = log(L) - log(K);

--> A(:, 2) = ones(size(A, 1), 1);


--> b = log(P) - log(K);
```

Que nos levará à:

```
--> A'*A
 ans  =

   5.4619816  -9.2244845
  -9.2244845   24.

--> A'*b
 ans  =

   4.0020480
  -6.6995520
```

Finalmente obtendo os valores utilizando a Gaussian Elimination para resolver o sistema de:

```
--> [x, per, C] = Gaussian_Elimination_4(A'*A, A'*b);


--> alpha = x(1);


--> beta = exp(x(2));

--> [alpha; beta]
 ans  =

   0.7446062
   1.0070689
```

Resultando em parâmetros $\alpha = 0.7446062$ e $\beta = 1.0070689$

Alternativa b

Com os parâmetros encontrados anteriormente, ao calcular as produções em 1910 e 1920, obteremos os seguintes resultados:

```
--> L_1910 = 147;

--> K_1910 = 208;


--> P_1910 = beta * L_1910^(alpha) * K_1910^(1-alpha)
 P_1910  =

   161.76185


--> L_1920 = 194;

--> K_1920 = 407;


--> P_1920 = beta * L_1920^(alpha) * K_1920^(1-alpha)
 P_1920  =

   236.07215
```

Que, ao comparar com os resultados originais, onde a produção P de 1910 é, na verdade 159, e de 1920 é de 231, obtemos um erro aproximado de $2,7$ em 1910 e de 5 em 1920, o que nos leva a boas aproximações dos valores originais

Questão 02

Para fazer o classificador, leremos os dados de treino, fazendo a seguinte divisão, uma matriz A, com a primeira coluna de uns (os interceptadores) e as seguintes 10 colunas das 10 primeiras colunas dos dados, e o vetor y de previsões como a última coluna dos dados, obtendo os seguintes resultados abaixo:

```
--> data_train = csvRead("cancer_train_2024.csv", ";");


--> y_train = data_train(:, 11);


--> A_train = zeros(size(data_train, 1), 11);

--> A_train(:, 1) = ones(size(data_train, 1), 1);

--> A_train(:, 2:11) = data_train(:, 1:10);


--> [x, per, C] = Gaussian_Elimination_4(A_train'*A_train, A_train'*y_train);


--> [acertos_train, pct_train] = acertos(A_train*x, y_train)
 acertos_train  =

   258.
 pct_train  =

   0.9214286
```

Tendo como o vetor de pesos x:

```
--> x
 x  =

  -6.2101493
   15.902409
   1.5568757
  -5.0718598
  -7.1846562
   1.2702227
  -0.9298812
   0.5285964
   1.9535131
  -0.0470564
   0.7701829
```

Assim, ao aplicar o mesmo vetor x ao conjunto de teste para fazer as predições, fazendo as mesmas transformações da matriz A e do vetor de previsões y para este conjunto de testes, obteremos:

```
--> data_test = csvRead("cancer_test_2024.csv", ";");


--> y_test = data_test(:, 11);


--> A_test = zeros(size(data_test, 1), 11);

--> A_test(:, 1) = ones(size(data_test, 1), 1);

--> A_test(:, 2:11) = data_test(:, 1:10);


--> [acertos_test, pct_test] = acertos(A_test*x, y_test)
 acertos_test  =

   249.
 pct_test  =

   0.8892857
```

Obtendo assim, uma precisão no conjunto de teste de acertos de quase 89%.

Seguem os diversos tipos de medidas adquiridos da confusion matrix, onde dado TP predições corretas de positivos, FP predições incorretas de positivos, FN predições incorretas de negativos e TN predições corretas de negativos, com o total sendo a soma de todos, as medidas são dadas da seguinte forma:

$$Acurácia = \frac{TP + TN}{total}$$

$$Precisão = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$TaxaFalsosPositivos(FalsoAlarme) = \frac{FP}{FP + TN}$$

$$TaxaFalsosNegativos(FalsaEmissãodeAlarme) = \frac{FN}{FN + TP}$$

```
--> confusion_matrix_train = confusion_matrix(y_pred_train, y_train);


--> print_medidas(confusion_matrix_train);

 "Matriz de Confusão:"

  106.    8.
  14.     152.

 "Métricas:"

 "Acurácia: 0.9214286"

 "Precisão: 0.9298246"

 "Recall: 0.8833333"

 "Taxa de Falsos Positivos (Falso Alarme): 0.05"

 "Taxa de Falsos Negativos (Falsa Emissão de Alarme): 0.1166667"
```

E os mesmos resultados para o conjunto de testes:

```
--> confusion_matrix_test = confusion_matrix(y_pred_test, y_test);


--> print_medidas(confusion_matrix_test);

 "Matriz de Confusão:"

  80.    25.
  6.     169.

 "Métricas:"

 "Acurácia: 0.8892857"

 "Precisão: 0.7619048"

 "Recall: 0.9302326"

 "Taxa de Falsos Positivos (Falso Alarme): 0.128866"

 "Taxa de Falsos Negativos (Falsa Emissão de Alarme): 0.0697674"
```

No conjunto de testes, diferente do de treino, vemos um recall maior que a precisão, o que indica que, neste conjunto houve menos acertos em relação aos valores positivos, entretanto, menos valores que deveria ser positivos foram previstos incorretamente. Mesmo em relação a estes pontos, são semelhantes as taxas finais dos dois conjuntos de dados.

Questão 03

Para analisar a relevância das variáveis de entrada, um modo de fazer isto é observando o seu desvio padrão, em que, um desvio padrão baixo, pode significar que a variável, pelo menos no conjunto de treino, não possui grande variedade, logo, influênciando pouco ou muito os resultados finais, sem tanta consistência, devido a falta de diversividade de dados.

```
"Desvio padrão das colunas de treino:"

       column 1 to 7
 0.12796   0.116178   0.1317821   0.1437375   0.0966583   0.1656143   0.1980321
       column 8 to 10
 0.1973597   0.0987134   0.0776195

"Original:"

"Número de acertos: 249 e porcentagem: 0.8892857"
```

Também é possível analisar a covariância dos dados e as médias da covariância de cada variável para analisar sua correlação entre as outras:

```
"Covariância das colunas de treino:"

       column 1 to 7
 0.0149802   0.0026856   0.0151411   0.0163473   0.0021433   0.0105028   0.0166606
 0.0026856   0.0170951   0.0026808   0.0030863  -0.0018462   0.002092    0.0051048
 0.0151411   0.0026808   0.0153675   0.0165604   0.002642    0.0115493   0.0176905
 0.0163473   0.0030863   0.0165604   0.0183127   0.0025569   0.011658    0.0187834
 0.0021433  -0.0018462   0.002642    0.0025569   0.0087898   0.0099617   0.0085884
 0.0105028   0.002092    0.0115493   0.011658    0.0099617   0.0295652   0.0283793
 0.0166606   0.0051048   0.0176905   0.0187834   0.0085884   0.0283793   0.036069
 0.019298    0.0035002   0.0201447   0.0215482   0.0096562   0.0259317   0.0333909
 0.0008028   0.0010654   0.0010808   0.0011477   0.003516    0.0084012   0.0068325
-0.0029106  -0.0010391  -0.0025821  -0.0028987   0.003499    0.0063304   0.0029853
 0.0816305   0.0244532   0.0851454   0.0900361   0.0404492   0.0860561   0.1290988
       column 8 to 11
 0.019298    0.0008028  -0.0029106   0.0816305
 0.0035002   0.0010654  -0.0010391   0.0244532
 0.0201447   0.0010808  -0.0025821   0.0851454
 0.0215482   0.0011477  -0.0028987   0.0900361
 0.0096562   0.003516    0.003499    0.0404492
 0.0259317   0.0084012   0.0063304   0.0860561
 0.0333909   0.0068325   0.0029853   0.1290988
 0.0356336   0.0057289   0.0013518   0.139818
 0.0057289   0.0087096   0.0029366   0.0154824
 0.0013518   0.0029366   0.0046746  -0.002296
 0.139818    0.0154824  -0.002296    0.8542755

"Média da covariância das colunas de treino:"

       column 1 to 7
 0.0161165   0.0053526   0.0168564   0.0179217   0.0081778   0.020948   0.0275985
       column 8 to 11
 0.0287275   0.005064   0.0009138   0.1403772
```

Finalmente, analisar como são os resultados das predições dos conjuntos de teste com o treinamento removendo as colunas, realizado para cada coluna a seguir:

```
"Sem a coluna 1:"

"Número de acertos: 242 e porcentagem: 0.8642857"

"Sem a coluna 2:"

"Número de acertos: 260 e porcentagem: 0.9285714"

"Sem a coluna 3:"

"Número de acertos: 248 e porcentagem: 0.8857143"

"Sem a coluna 4:"

"Número de acertos: 251 e porcentagem: 0.8964286"

"Sem a coluna 5:"

"Número de acertos: 242 e porcentagem: 0.8642857"

"Sem a coluna 6:"

"Número de acertos: 235 e porcentagem: 0.8392857"

"Sem a coluna 7:"

"Número de acertos: 235 e porcentagem: 0.8392857"

"Sem a coluna 8:"

"Número de acertos: 247 e porcentagem: 0.8821429"

"Sem a coluna 9:"

"Número de acertos: 246 e porcentagem: 0.8785714"

"Sem a coluna 10:"

"Número de acertos: 249 e porcentagem: 0.8892857"

"Sem a coluna 11:"

"Número de acertos: 246 e porcentagem: 0.8785714"
```

Assim, conseguimos inferir, que, ao se comparar com os 249 acertos do conjunto completo com todas as colunas, a coluna que ao ser removida gerou melhores resultados finais aparenta ser a coluna 2, elevando o número de acertos para 260 (lembrando que a coluna 1 é formanda de uns representando o intercepto, ou viés), assim, a primeira coluna dos dados originais, ao ser removida, aparenta levar a uma predição mais eficiente. Além disso, também observamos que algumas variáveis não aparentam afetar tanto o resultado, como a coluna 3, 4, 8 e 10, correspondentes as colunas da base de dados 2, 3, 7 e 9, respectivamente, que aparentam manter a quantidade de acertos próxima de 249. Entretanto, para conclusões mais precisas e apuradas, seriam necessários mais dados, tanto para treino, quanto para teste.