Maximizing Correlation To Identify Features With Linear Predictive Power: A Study of Cancer With Mathematica

Brian Tenneson

Source Data

The website this data comes from is here: https://www.kaggle.com/competitions/icr-identify-age-related-conditions/data

This command loads the xls and stores it in the variable g1:

```
g1 = Import["G:\\Other computers\\My Laptop
      (1)\\bak\\datasets\\icr-identify-age-related-conditions\\train-unclean-8.xls",
    "Data"][[1]];
```

Cleaning

Even after row and column labels are removed, the file train.csv still has many ? characters; so we implement the following module to build (rather than select) a new table without any ? characters. In doing so, around 70 of approximately 600 rows get weeded out. The variable will assign the cleaned version to the variable g.

```
In[7]:= m1 = Dimensions[g1][[1]];
n1 = Dimensions[g1][[2]];
g = Module[{p = {}}, Do[If[! MemberQ[g1[[k]], "?"], p = Join[p, {g1[[k]]}]], {k, 1, m1}];
   p];
```

The Dimensions of the Cleaned Data

```
In[12]:= m = Dimensions[g][[1]];
n = Dimensions[g][[2]];
```

Analysis

```
y[i_] := g[[i, 1]] (* ith row of first column of g *)
s[i_] := Sum[(b[j] × g[[i, j]]), {j, 2, n}]; (*ith Synthetic score *)
r = Correlation[
    Table[s[i], {i, 1, m}],
    Table[y[i], {i, 1, m}]
   ]; (* r is a function of the b[j]*)
(*We seek to maximize the absolute value
 of correlation as a function of the bias vector *)
t2 = NMaximize[Abs[r], Table[b[j], {j, 2, n}], MaxIterations → 1000, Method → "NelderMead"]
 (*numerically finds the maximum absolute value of correlation r*)
```

Out[24]=

$\{0.721053, \{b[2] \to -1.90766 \times 10^{8}, b[3] \to -26976.6, b[4] \to 41019.8, b[5] \to 368559.,$
$b[6] \to -1.56398 \times 10^{6}, b[7] \to -4.14353 \times 10^{6}, b[8] \to -9.88111 \times 10^{7}, b[9] \to -537719.,$
$b[10] \to -2.23629 \times 10^{6}, b[11] \to 1208.16, b[12] \to 1.23303 \times 10^{6}, b[13] \to 4227.26,$
$b[14] \to -768383., b[15] \to -3100.1, b[16] \to 66019.7, b[17] \to 44923.5,$
$b[18] \to 1.95028 \times 10^{8}, b[19] \to -2.20183 \times 10^{6}, b[20] \to -3.07515 \times 10^{6},$
$b[21] \to 1.57276 \times 10^{7}, b[22] \to -2.19118 \times 10^{6}, b[23] \to 2.71239 \times 10^{8},$
$b[24] \to 2.52191 \times 10^{6}, b[25] \to 5.99691 \times 10^{7}, b[26] \to 1.55315 \times 10^{6}, b[27] \to 604632.,$
$b[28] \to 32792.9, b[29] \to -3.53327 \times 10^{7}, b[30] \to 3.27598 \times 10^{7}, b[31] \to -707480.,$
$b[32] \to 1.34101 \times 10^{6}, b[33] \to 1.10765 \times 10^{7}, b[34] \to -7.1856 \times 10^{6},$
$b[35] \to -5.58464 \times 10^{7}, b[36] \to -2.32345 \times 10^{6}, b[37] \to -6.23218 \times 10^{6},$
$b[38] \to 3.59548 \times 10^{7}, b[39] \to 13861.1, b[40] \to -1.43954 \times 10^{8}, b[41] \to -1.91567 \times 10^{7},$
$b[42] \to -482669., b[43] \to 1.45247 \times 10^{6}, b[44] \to -350.114, b[45] \to 65039.4,$
$b[46] \to 5.96563 \times 10^{6}, b[47] \to -3915.35, b[48] \to 2.46712 \times 10^{7}, b[49] \to -2.19308 \times 10^{7},$
$b[50] \to -842073., b[51] \to 4.79262 \times 10^{6}, b[52] \to -1.68585 \times 10^{6}, b[53] \to 237977.,$
$b[54] \to 1351.57, b[55] \to -2.8573 \times 10^{6}, b[56] \to 58906., b[57] \to -3.82275 \times 10^{6}\}\}$

Analysis Interpretation

In[25]:=
```
Table[B[j] = (Table[b[j], {j, 2, n}] //. t2〚2〛)〚j - 1〛, {j, 2, n}];
(*Sets B[j] to t2〚2〛〚j-1〛*)
```

In[26]:=
```
Table[{j, B[j]}, {j, 2, n}] // MatrixForm
```

Out[26]//MatrixForm=

$$\begin{pmatrix} 2 & -1.90766\times10^{8} \\ 3 & -26976.6 \\ 4 & 41019.8 \\ 5 & 368559. \\ 6 & -1.56398\times10^{6} \\ 7 & -4.14353\times10^{6} \\ 8 & -9.88111\times10^{7} \\ 9 & -537719. \\ 10 & -2.23629\times10^{6} \\ 11 & 1208.16 \\ 12 & 1.23303\times10^{6} \\ 13 & 4227.26 \\ 14 & -768383. \\ 15 & -3100.1 \\ 16 & 66019.7 \\ 17 & 44923.5 \\ 18 & 1.95028\times10^{8} \\ 19 & -2.20183\times10^{6} \\ 20 & -3.07515\times10^{6} \\ 21 & 1.57276\times10^{7} \\ 22 & -2.19118\times10^{6} \\ 23 & 2.71239\times10^{8} \\ 24 & 2.52191\times10^{6} \\ 25 & 5.99691\times10^{7} \\ 26 & 1.55315\times10^{6} \\ 27 & 604632. \\ 28 & 32792.9 \\ 29 & -3.53327\times10^{7} \\ 30 & 3.27598\times10^{7} \\ 31 & -707480. \\ 32 & 1.34101\times10^{6} \\ 33 & 1.10765\times10^{7} \\ 34 & -7.1856\times10^{6} \\ 35 & -5.58464\times10^{7} \\ 36 & -2.32345\times10^{6} \\ 37 & -6.23218\times10^{6} \\ 38 & 3.59548\times10^{7} \\ 39 & 13861.1 \\ 40 & -1.43954\times10^{8} \\ 41 & -1.91567\times10^{7} \\ 42 & -482669. \end{pmatrix}$$

43	1.45247×10^{6}
44	-350.114
45	65039.4
46	5.96563×10^{6}
47	-3915.35
48	2.46712×10^{7}
49	-2.19308×10^{7}
50	$-842073.$
51	4.79262×10^{6}
52	-1.68585×10^{6}
53	$237977.$
54	1351.57
55	-2.8573×10^{6}
56	$58906.$
57	-3.82275×10^{6}

In[31]:= **syntheticscore[i_] := $\sum_{j=2}^{n}$(B[j] × g⟦i, j⟧)**

```
temp2 = Table[syntheticscore[i], {i, 1, m}];
temp3 = Table[g⟦i, 1⟧, {i, 1, m}];
Correlation[temp2, temp3]
% // Abs
```

Out[34]=

−0.721053

Out[35]=

0.721053

In[36]:=
```
f[s_] := Evaluate[Fit[Table[{syntheticscore[i], g[[i, 1]]}, {i, 1, m}], {1, s}, s]]
f[s]
```

Out[37]=

$0.4916 - 6.7999 \times 10^{-10}\, s$

In[43]:=
```
δ = 0.500;
prediction[i_] := If[f[ syntheticscore[i]] > δ, 1., 0.]
temp1 = Tally[Table[prediction[i] == g[[i, 1]], {i, 1, m}]] // Sort
```

Out[45]=

```
{{False, 49}, {True, 499}}
```

In[71]:=
```
δ = 0.365;
prediction[i_] := If[f[ syntheticscore[i]] > δ, 1., 0.]
temp1 = Tally[Table[prediction[i] == g[[i, 1]], {i, 1, m}]] // Sort
```

Out[73]=

```
{{False, 43}, {True, 505}}
```