Chapter 24
Discriminant Analysis for Making a Diagnosis from Multiple Outcomes (45 Patients)

General Purpose

To assess whether discriminant analysis can be used to make a diagnosis from multiple outcomes both in groups and in individual patients.

Specific Scientific Question

Laboratory screenings were performed in patients with different types of sepsis (urosepsis, bile duct sepsis, and airway sepsis). Can discriminant analysis of laboratory screenings improve reliability of diagnostic processes.

Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10	Var 11
8,00	5,00	28,00	4,00	2,50	79,00	108,00	19,00	18,00	16,00	2,00
11,00	10,00	29,00	7,00	2,10	94,00	89,00	18,00	15,00	15,00	2,00
7,00	8,00	30,00	7,00	2,20	79,00	96,00	20,00	16,00	14,00	2,00
4,00	6,00	16,00	6,00	2,60	80,00	120,00	17,00	17,00	19,00	2,00
1,00	6,00	15,00	6,00	2,20	84,00	108,00	21,00	18,00	20,00	2,00
23,00	5,00	14,00	6,00	2,10	78,00	120,00	18,00	17,00	21,00	3,00
12,00	10,00	17,00	5,00	3,20	85,00	100,00	17,00	20,00	18,00	3,00
31,00	8,00	27,00	5,00	,20	68,00	113,00	19,00	15,00	18,00	3,00
22,00	7,00	26,00	5,00	1,20	74,00	98,00	16,00	16,00	17,00	3,00
30,00	6,00	25,00	4,00	2,40	69,00	90,00	20,00	18,00	16,00	3,00

(continued)

This chapter was previously published in “Machine learning in medicine-cookbook 1” as Chap. 9, 2013.

T.J. Cleophas, A.H. Zwinderman, *Machine Learning in Medicine - a Complete Overview*, DOI 10.1007/978-3-319-15195-3_24

Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10	Var 11
2,00	12,00	21,00	4,00	2,80	75,00	112,00	11,00	14,00	19,00	1,00
10,00	21,00	20,00	4,00	2,90	70,00	100,00	12,00	15,00	20,00	1,00

Var 1 gammagt
Var 2 asat
Var 3 alat
Var 4 bilirubine
Var 5 ureum
Var 6 creatinine
Var 7 creatinine clearance
Var 8 erythrocyte sedimentation rate
Var 9 c-reactive protein
Var 10 leucocyte count
Var 11 type of sepsis (1–3 as described above)

The first 12 patients are shown only, the entire data file is entitled "optscalingfactorplscanonical" and is in extras.springer.com.

The Computer Teaches Itself to Make Predictions

SPSS 19.0 is used for training and outcome prediction. It uses XML (eXtended Markup Language) files to store data. Start by opening the data file.

Command:

Click Transform….click Random Number Generators….click Set Starting Point…. click Fixed Value (2000000)….click OK….click Analyze….Classify…. Discriminant Analysis….Grouping Variable: enter diagnosisgroup….Define Range: Minimum enter 1…Maximum enter 3….click Continue….Independents: enter all of the 10 laboratory variables….click Statistics….mark Unstandardized ….mark Separate-groups covariance….click Continue….click Classify….mark All groups equal….mark Summary table….mark Within-groups….mark Combined groups…. click Continue….click Save….mark Predicted group memberships….in Export model information to XML file enter: exportdiscriminant….click Browse and save the XML file in your computer….click Continue….click OK.

The scientific question "is the diagnosis group a significant predictor of the outcome estimated with 10 lab values" is hard to assess with traditional multivariate methods due to interaction between the outcome variables. It is, therefore, assessed with the question "is the clinical outcome a significant predictor of the odds of having had a particular prior diagnosis. This reasoning may seem incorrect, using an

outcome for making predictions, but, mathematically, it is no problem. It is just a matter of linear cause-effect relationships, but just the other way around, and it works very conveniently with "messy" outcome variables like in the example given. However, first, the numbers of outcome variables have to be reduced. SPSS accomplishes this by orthogonal modeling of the outcome variables, which produces novel composite outcome variables. They are the y-values of linear equations. The x-values of these linear equations are the original outcome variables, and their regression coefficients are given in the underneath table.

Structure matrix		
	Function	
	1	2
As at	,574*	,184
Gammagt	,460*	,203
C-reactive protein	−.034	,761*
Leucos	,193	,537*
Ureum	,461	,533*
Creatinine	,462	,520*
Alat	,411	,487*
Bili	,356	,487*
Esr	,360	,487*
Creatinine clearance	−,083	−.374*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.
*Largest absolute correlation between each variable and any discriminant function

Wilks' Lambda				
Test of function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,420	32,500	20	,038
2	,859	5,681	9	,771

The two novel outcome variables significantly predict the odds of having had a prior diagnosis with $p = 0.038$ as shown above. When minimizing the output sheets we will return to the data file and observe that the novel outcome variables have been added (the variables entitled Dis1_1 and Dis1_2), as well as the predicted diagnosis group predicted from the discriminant model (the variable entitled Dis_1). For convenience the XML file entitled "exportdiscriminant" is stored in extras.springer. com.

The saved XML file can now be used to predict the odds of having been in a particular diagnosis group in five novel patients whose lab values are known but whose diagnoses are not yet obvious.

Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10
1049,00	466,00	301,00	268,00	59,80	213,00	−2,00	109,00	121,00	42,00
383,00	230,00	154,00	120,00	31,80	261,00	13,00	80,00	58,00	30,00
9,00	9,00	31,00	204,00	34,80	222,00	10,00	60,00	57,00	34,00
438,00	391,00	479,00	127,00	31,80	372,00	9,00	69,00	56,00	33,00
481,00	348,00	478,00	139,00	21,80	329,00	15,00	49,00	47,00	32,00

Var 1 gammagt
Var 2 asat
Var 3 alat
Var 4 bilirubine
Var 5 ureum
Var 6 creatinine
Var 7 creatinine clearance
Var 8 erythrocyte sedimentation rate
Var 9 c-reactive protein
Var 10 leucocyte count

Enter the above data in a new SPSS data file.

Command:

Utilities….click Scoring Wizard….click Browse….click Select….Folder: enter the exportdiscriminant.xml file….click Select….in Scoring Wizard click Next….click Use value substitution….click Next….click Finish.

The above data file now gives predicted odds of having been in a particular diagnosis group computed by the discriminant analysis module with the help of the xml file.

Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10	Var 11
1049,00	466,00	301,00	268,00	59,80	213,00	−2,00	109,00	121,00	42,00	2,00
383,00	230,00	154,00	120,00	31,80	261,00	13,00	80,00	58,00	30,00	2,00
9,00	9,00	31,00	204,00	34,80	222,00	10,00	60,00	57,00	34,00	1,00
438,00	391,00	479,00	127,00	31,80	372,00	9,00	69,00	56,00	33,00	1,00
481,00	348,00	478,00	139,00	21,80	329,00	15,00	49,00	47,00	32,00	2,00

Var 1 gammagt
Var 2 asat
Var 3 alat
Var 4 bilirubine
Var 5 ureum
Var 6 creatinine
Var 7 creatinine clearance
Var 8 erythrocyte sedimentation rate
Var 9 c-reactive protein
Var 10 leucocyte count
Var 11 predicted odds of having been in a particular diagnosis group

Conclusion

The discriminant analysis module can be readily trained to provide from the laboratory values of individual patients the best fit odds of having been in a particular diagnosis group. In this way discriminant analysis can support the hard work of physicians trying to make a diagnosis.

Note

More background, theoretical and mathematical information of discriminant analysis is available in Machine learning part one, Chap. 17, entitled "Discriminant analysis for supervised data", pp 215–224, Springer Heidelberg Germany, 2013, from the same authors.