Session 2

Exercise Answers

Exercise A. (Optional)

To confirm that the result obtained from COREXpress (and XLSTAT-CCR) for the 10-component CCR-LDA (saturated) model corresponds to the traditional LDA solution, use SPSS (or XLSTAT) to estimate the 10-gene LDA model. Table 2 below shows how to compute the logit model coefficients from the associated coefficients obtained from the SPSS discriminant classification output. To get the logit coefficient for a given predictor, simply subtract the coefficient associated with Z=0 from the corresponding coefficient for Z=1. Note: To get the correct intercept, in the SPSS Classification section menu option, select 'Compute from Group Sizes' which utilizes the sample proportions in the training data as the prior for group membership.

Table 2.

Classification Function Coefficients			
	Z		
	0	1	Coefficient
X436	.929	-3.191	-4.12
X456	1.005	-3.211	-4.22
X956	-1.722	.555	2.28
X979	-4.101	6.593	10.69
X1182	1.547	-.947	-2.49
X1693	.745	1.679	0.93
X2323	2.247	-4.084	-6.33
X2481	-1.162	3.176	4.34
X2911	.226	-1.067	-1.29
X3126	-3.274	.813	4.09
(Constant)	-3.884	-10.703	-6.82

COREXpress Exercise A Answer:

SPSS Syntax:

```
DISCRIMINANT
  /GROUPS=Z(0 1)
  /VARIABLES=X436 X456 X956   X979   X1182 X1693 X2323 X2481 X2911 X3126
  /SELECT=Validation(0)
  /ANALYSIS ALL
  /PRIORS SIZE
  /STATISTICS=MEAN STDDEV COEFF RAW
  /CLASSIFY=NONMISSING POOLED.
```

XLSTAT Exercise A Answer:

In XLSTAT, go to Analyzing Data→ Discriminant Analysis (DA) and open the following saved model definition: ExerciseAAnswers.txt

COREXpress Exercise E1:

1. Repeat the analysis for 5 components. Is the CV-ACC higher than that for the 4-component model?
2. Repeat the analysis using CCR.logistic instead of CCR.lda. Unlike LDA, which utilizes the assumption that the predictors follow a multivariate normal distribution (with common covariances within each dependent variable group), logistic regression does not make this assumption. Since the data were in fact generated according to the LDA assumptions, the LDA models would be expected to be somewhat better. Is the CV-ACC higher for LDA?

   Notes:

   The logistic regression utilizes an iterative algorithm and is therefore much slower than LDA. Therefore, you may wish to reduce the number of rounds to 2.

   It is possible to obtain perfect separation between the dependent variable groups with logistic regression. In an attempt to avoid such, instead of maximizing the likelihood,

a penalized likelihood is maximized, where the 'Ridge Parameter' controls the size of the penalty. By default, the penalty is set to 0.001.

CORExpress Exercise E1 Answer:

1. Yes. 4-component CV-ACC = .8690; 5-component CV-ACC = .8750
2. Yes.

XLSTAT-CCR Exercise E1:

1. Repeat the analysis for 6 components. Is the CV-ACC higher than that for the 5-component model?
2. Repeat the analysis using CCR.logistic instead of CCR.lda. Unlike LDA, which utilizes the assumption that the predictors follow a multivariate normal distribution (with common covariances within each dependent variable group), logistic regression does not make this assumption. Since the data were in fact generated according to the LDA assumptions, the LDA models would be expected to be somewhat better. Is the CV-ACC higher for LDA?

Note:

It is possible to obtain perfect separation between the dependent variable groups with logistic regression. In an attempt to avoid such, instead of maximizing the likelihood, a penalized likelihood is maximized, where the 'Ridge Parameter' controls the size of the penalty. By default, the penalty is set to 0.001.

XLSTAT-CCR Exercise E1 Answer:

1. NO. 5-component CV-ACC = .820; 6-component CV-ACC = .800
2. Yes.