Classification & Classification Statistics

Jeroen K. Vermunt
Department of Methodology and Statistics, Tilburg University
www.jeroenvermunt.nl

Introduction

- Based of the estimated parameters of a LC model, it is possible to determine to which class each individual belongs.
- When the LC model used for clustering or diagnosing. In that case, the model building stage is just the first step.
- When you want to see how classes are related to other variables using a 3-step approach.

- In this video, I will show how this classification works.
- I will also discuss classification statistics, which quantify how good the classification is; that is, how certain we are about the class assignments.

Bayes rule for reversing the prediction

- The LC model yields information on $P(\mathbf{y}|X=\text{c})$, the probability of a response pattern given class membership.

- For classification, we need the reversed probability $P(X=c|\mathbf{y})$, the probability of belonging to a latent class given the response pattern.

- Prediction can be reversed using the Bayes rule:

$$P(B|A) = P(B)\ P(A|B)\ /\ P(A)$$

Posterior class membership probabilities

- Applying the Bayes rule yields:

$$P(X=c \mid y_1,\ldots,y_J) = \frac{P(X=c)P(y_1,\ldots,y_J \mid X=c)}{P(y_1,\ldots,y_J)} = \frac{P(X=c)\prod_{j=1}^{J} P(y_j \mid X=c)}{\sum_{c'=1}^{C} P(X=c')\prod_{j=1}^{J} P(y_j \mid X=c')}$$

- Modal classification rule: assign every individual to the class for which $P(X=c \mid y_1,\ldots,y_J)$ is largest.

Classification for GSS82.sav (3-class model)

accuracy	cooperat	understa	purpose	ObsFreq	Modal	Cluster1	Cluster2	Cluster3
mostly true	interested	Good	GOOD PURPOSE	535.0000	1	0.8848	0.1147	0.0005
mostly true	interested	Good	DEPENDS	29.0000	1	0.8631	0.1255	0.0113
mostly true	interested	Good	WASTE OF TIME AND $	32.0000	1	0.8974	0.0687	0.0340
mostly true	interested	Fair/Poor	GOOD PURPOSE	105.0000	2	0.0467	0.9523	0.0009
mostly true	interested	Fair/Poor	DEPENDS	9.0000	2	0.0411	0.9405	0.0184
mostly true	interested	Fair/Poor	WASTE OF TIME AND $	4.0000	2	0.0698	0.8404	0.0898
mostly true	cooperative	Good	GOOD PURPOSE	49.0000	2	0.4101	0.5877	0.0022
mostly true	cooperative	Good	DEPENDS	5.0000	2	0.3662	0.5892	0.0446
mostly true	cooperative	Good	WASTE OF TIME AND $	3.0000	1	0.4550	0.3853	0.1597
mostly true	cooperative	Fair/Poor	GOOD PURPOSE	44.0000	2	0.0044	0.9948	0.0008
mostly true	cooperative	Fair/Poor	DEPENDS	3.0000	2	0.0039	0.9801	0.0160
mostly true	cooperative	Fair/Poor	WASTE OF TIME AND $	3.0000	2	0.0068	0.9115	0.0816
mostly true	Impatient,Hostile	Good	GOOD PURPOSE	5.0000	2	0.0156	0.9770	0.0075
mostly true	Impatient,Hostile	Good	DEPENDS	1.0000	2	0.0121	0.8549	0.1330

Computation of posterior class membership probabilities using Profile output

- For the first data pattern:

Cluster Size	0.5677	0.2612	0.1712		
mostly true	0.5959	0.6453	0.0135		
interested	0.9595	0.6413	0.6439		
Good	0.9897	0.3788	0.7383		
GOOD PURPOSE	0.8863	0.9013	0.1488		
P(y|X=c)	0.5015	0.1413	0.0010		
P(X=c) * P(y|X=c)	0.2847	0.0369	0.0002	**P(y)**	0.3218
P(X=c|y)	0.8848	0.1147	0.0005		

Classification statistics

- How well are we doing when assigning individuals to latent classes?
- How well can we predict the persons' class memberships based on the observed indicators?

- Three types of statistics:
  - Estimated proportion of classification errors
  - Classification table
  - Pseudo R-squared measures based on classification errors, entropy, or qualitative variance

Classification errors and table

- Estimated proportion of classification errors:
  - Compute 1 - max[$P(X = c \mid y_1, \ldots, y_J)$] for each individual
  - Average these over individuals
  - 0.1533
- Classification table:
  - The entry corresponding to true class *X*=*c* and assigned modal class *W*=*t* is the sum of $P(X = c \mid y_1, \ldots, y_J)$ for those assigned to class *t*.

Classification Table	**Modal**			
Latent	**Cluster1**	**Cluster2**	**Cluster3**	**Total**
Cluster1	841.3035	43.6934	48.5065	933.5034
Cluster2	101.1836	310.6107	17.4839	429.2782
Cluster3	21.5129	19.6959	240.0096	281.2184
Total	964	374	306	1644

Pseudo R-squared measures

- Compare of prediction based on $P(X = c)$ with prediction based on $P\left(X = c|y_1, \dots, y_J\right)$.

- Pseudo $R^2 = \frac{Loss(X) - Loss(X|y_1,\dots,y_J)}{Loss(X)} = 1 - \frac{Loss(X|y_1,\dots,y_J)}{Loss(X)}$

- *Loss*(…) can be proportion of classification errors (Lambda), average entropy, or average qualitative variance.

- Entropy: sum over classes of -$P \ln(P)$
- Qualitative variance: 1 - sum of over classes of P^2

Pseudo R-squared for GSS82.sav (3-class)

- Lambda = (0.4323-0.1533)/0.4323=0.6453
- Entropy R^2 =(0.9742-0.4348)/0.9742= 0.5537
- “Standard” R^2 = (0.5802-0.2420)/0.5802=0.5830
- Entropy R^2 is the most popular measure

Writing classifications to a new data file