`lcda`: Local Classification of Discrete Data by Latent Class Models

Michael Bücker

`buecker@statistik.tu-dortmund.de`

July 9, 2009

technische universität
dortmund

Introduction

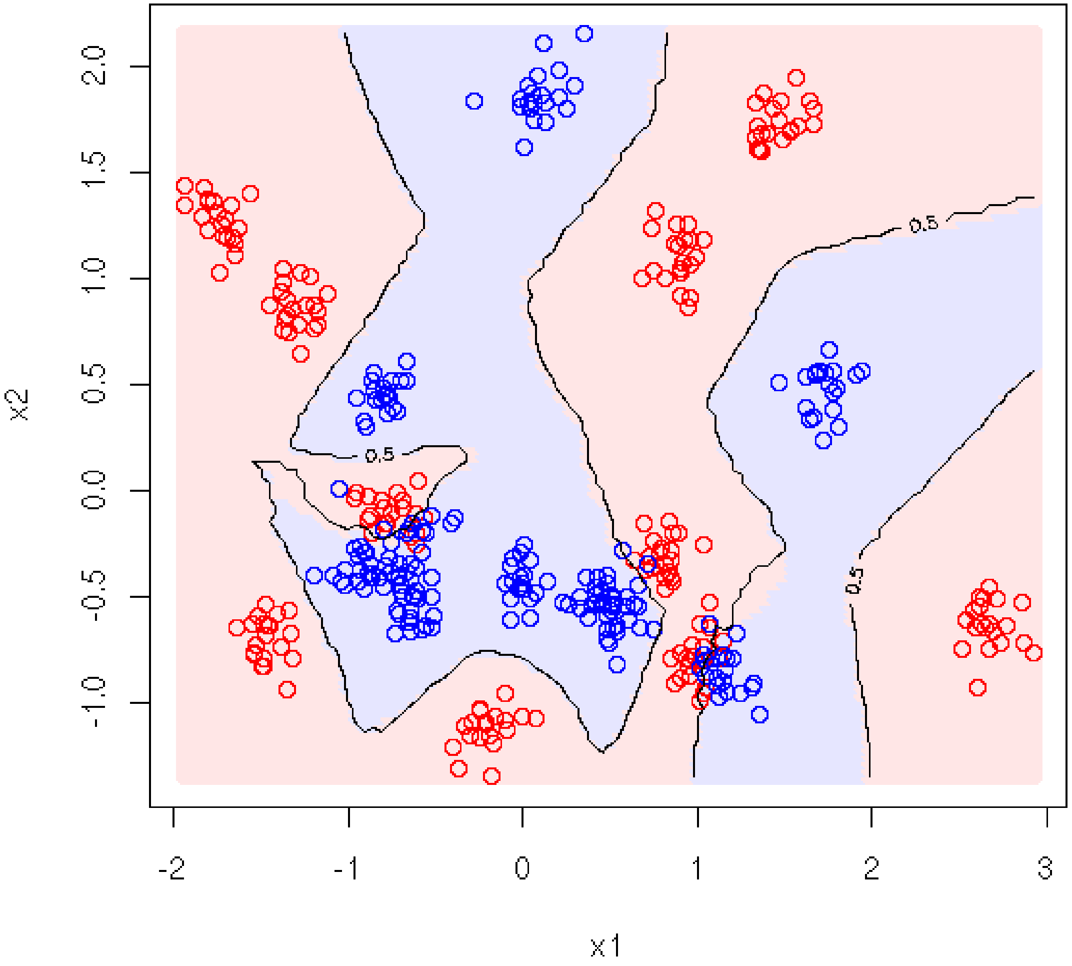
- common global classification methods may be inefficient when groups are heterogeneous
  ⇒ need for more flexible, local models

- continuous models that allow for subclasses:
  - Mixture Discriminant Analysis (MDA): assumption of class conditional mixtures of (multivariate) normals
  - Common Components (Titsias and Likas 2001) imply a mixture of normals with common components

- in this talk: discrete counterparts based on Latent Class Models (see Lazarsfeld and Henry 1968) implemented in R-package `lcda`

- application to SNP data

Local structures



Mixture Discriminant Analysis and Common Components

- class conditional density (MDA)

$$f(x|Z = k) = f_k(x) = \sum_{m=1}^{M_k} w_{mk}\phi(x; \mu_{mk}, \Sigma)$$

Mixture Discriminant Analysis and Common Components

- class conditional density (MDA)

$$f(x|Z=k) = f_k(x) = \sum_{m=1}^{M_k} w_{mk}\phi(x;\mu_{mk},\Sigma)$$

- class conditional density of the Common Components Model (Titsias and Likas 2001)

$$P(X=x|Z=k) = f_k(x) = \sum_{m=1}^{M} w_{mk}\phi(x;\mu_m,\Sigma)$$

Mixture Discriminant Analysis and Common Components

- class conditional density (MDA)

$$f(x|Z=k) = f_k(x) = \sum_{m=1}^{M_k} w_{mk}\phi(x;\mu_{mk},\Sigma)$$

- class conditional density of the Common Components Model (Titsias and Likas 2001)

$$P(X=x|Z=k) = f_k(x) = \sum_{m=1}^{M} w_{mk}\phi(x;\mu_m,\Sigma)$$

- posterior based on Bayes' rule

$$P(Z=k|X=x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K}\pi_l f_l(x)}$$

Latent Class Model

- latent (unobservable) variable Y with categorical outcomes in $\{1, \ldots, M\}$ with probability $P(Y = m) = w_m$

Latent Class Model

- latent (unobservable) variable Y with categorical outcomes in $\{1, \ldots, M\}$ with probability $P(Y = m) = w_m$
- manifest (observable) variables $X_1, \ldots, X_D$, X_d with outcomes in $\{1, \ldots, R_d\}$ with probability $P(X_d = r|Y = m) = \theta_{mdr}$

Latent Class Model

- latent (unobservable) variable Y with categorical outcomes in $\{1, \dots, M\}$ with probability $P(Y = m) = w_m$
- manifest (observable) variables $X_1, \dots, X_D$, X_d with outcomes in $\{1, \dots, R_d\}$ with probability $P(X_d = r|Y = m) = \theta_{mdr}$
- define $X_{dr} = 1$ if $X_d = r$ and $X_{dr} = 0$ else and **assume stochastic independence of manifest variables conditional on** Y, then the conditional probability mass function is given by

$$f(x|m) = \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}}$$

Latent Class Model

- latent (unobservable) variable Y with categorical outcomes in $\{1, \ldots, M\}$ with probability $P(Y = m) = w_m$
- manifest (observable) variables $X_1, \ldots, X_D$, X_d with outcomes in $\{1, \ldots, R_d\}$ with probability $P(X_d = r | Y = m) = \theta_{mdr}$
- define $X_{dr} = 1$ if $X_d = r$ and $X_{dr} = 0$ else and **assume stochastic independence of manifest variables conditional on** Y, then the conditional probability mass function is given by

$$f(x|m) = \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}}$$

- unconditional probability mass function of manifest variables is

$$f(x) = \sum_{m=1}^{M} w_m \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}}$$

Identifiability

Proposition 1. *The LCM* $f(x) = \sum_{m=1}^{M} w_m \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}}$ *is not identifiable.*

Identifiability

Proposition 1. *The LCM* $f(x) = \sum_{m=1}^{M} w_m \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}}$ *is not identifiable.*

Proof.

- the LCM is a finite mixture of products of multinomial distributions
- each mixture component $f(x|m)$ is the product of $\mathbb{M}(1, \theta_{md_1}, \ldots, \theta_{mdR_d})$-distributed random variables
- mixtures of M multinomials $\mathbb{M}(N, \theta_1, \ldots, \theta_p)$ are identifiable iff $N \geq 2M - 1$ (Elmore and Wang 2003)
- mixtures of the product of marginal distributions are identifiable if mixtures of the marginal distributions are identifiable (Teicher 1967)

$\Rightarrow$ the LCM is not identifiable.

□

Estimation of the LCM

- estimation by EM-algorithm:

Estimation of the LCM

- estimation by EM-algorithm:
- **E step:** Determination of conditional expectation of Y given $X = x$

$$\tau_{mn} = \frac{w_m f(x_n|m)}{f(x_n)}$$

Estimation of the LCM

- estimation by EM-algorithm:
- **E step:** Determination of conditional expectation of Y given $X = x$

$$\tau_{mn} = \frac{w_m f(x_n|m)}{f(x_n)}$$

- **M step:** Maximization of the log-Likelihood and estimation of

$$w_m = \frac{1}{N} \sum_{n=1}^{N} \tau_{mn}$$

and

$$\theta_{mdr} = \frac{1}{N w_m} \sum_{n=1}^{N} \tau_{mn} x_{ndr}$$

Model selection criteria

- information criteria
  - AIC

$$-2\log\mathcal{L}(w,\theta|x)+2\eta$$

  - BIC

$$-2\log\mathcal{L}(w,\theta|x)+\eta\log N$$

where $\eta = M\left(\sum_{d=1}^{D} R_d - D + 1\right) - 1$ (=number of parameters)

- goodness-of-fit test statistics (predicted vs. observed frequencies)
  - Pearson's χ^2
  - likelihood ratio χ^2

Local Classification of Discrete Data

- two ways to use LCM for local classification:
  - class conditional mixtures (like in MDA)
  - common components

Local Classification of Discrete Data

- two ways to use LCM for local classification:
  - class conditional mixtures (like in MDA)
  - common components
- class conditional mixtures

$$P(X = x|Z = k) = f_k(x) = \sum_{m=1}^{M_k} w_{mk} \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mkdr}^{x_{kdr}},$$

Local Classification of Discrete Data

- two ways to use LCM for local classification:
  - class conditional mixtures (like in MDA)
  - common components
- class conditional mixtures

$$P(X = x|Z = k) = f_k(x) = \sum_{m=1}^{M_k} w_{mk} \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mkdr}^{x_{kdr}},$$

- common components

$$P(X = x|Z = k) = f_k(x) = \sum_{m=1}^{M} w_{mk} \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}},$$

Estimation of a common components model (option 1)

- let π_k be the class prior, then

$$P(X = x) = \sum_{k=1}^{K} \pi_k \sum_{m=1}^{M} w_{mk} \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}}$$

$$= \sum_{m=1}^{M} w_m \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}}$$

since

$$w_m := P(m) = \sum_{k=1}^{K} P(k)P(m|k) = \sum_{k=1}^{K} \pi_k w_{mk}$$

Estimation of a common components model (option 1)

- let π_k be the class prior, then

$$P(X = x) = \sum_{k=1}^{K} \pi_k \sum_{m=1}^{M} w_{mk} \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}}$$

$$= \sum_{m=1}^{M} w_m \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}}$$

since

$$w_m := P(m) = \sum_{k=1}^{K} P(k)P(m|k) = \sum_{k=1}^{K} \pi_k w_{mk}$$

- this is a common Latent Class Model
- hence, estimate a global Latent Class model and determine parameter w_{mk} of the common components model by

$$\hat{w}_{mk} = \frac{1}{N_k} \sum_{i=1}^{N_k} \hat{P}(Y = m | Z = k, X = x_i)$$

Estimation of a common components model (option 2)

- **E step:** Determination of conditional expectation

$$\tau_{mkn} = \frac{w_{mk} f(x_n|m)}{f(x_n)}$$

Estimation of a common components model (option 2)

- **E step:** Determination of conditional expectation

$$\tau_{mkn} = \frac{w_{mk} f(x_n|m)}{f(x_n)}$$

- **M step:** Maximization of the log-Likelihood and estimation of

$$w_{mk} = \frac{1}{N_k} \sum_{n=1}^{N_k} \tau_{mkn}$$

and

$$\theta_{mdr} = \sum_{k=1}^{K} \frac{1}{N_k w_{mk}} \sum_{n=1}^{N_k} \tau_{mkn} x_{ndr}$$

Classification capability in Common Components Models

- measure for the ability to separate classes adequately
- impurity measures handling the subgroups like nodes in decision trees

Classification capability in Common Components Models

- measure for the ability to separate classes adequately
- impurity measures handling the subgroups like nodes in decision trees
- standardized mean entropy

$$H = -\sum_{m=1}^{M} w_m \sum_{k=1}^{K} P(k|m) \cdot \log_K \left(P(k|m)\right)$$

Classification capability in Common Components Models

- measure for the ability to separate classes adequately
- impurity measures handling the subgroups like nodes in decision trees
- standardized mean entropy

$$H = -\sum_{m=1}^{M} w_m \sum_{k=1}^{K} P(k|m) \cdot \log_K \left(P(k|m)\right)$$

- mean Gini impurity

$$G = \sum_{m=1}^{M} w_m \left[1 - \sum_{k=1}^{K} \left(P(k|m)\right)^2\right]$$

Implementation in R

- Package: `lcda` (requires `poLCA`, `scatterplot3d` and `MASS`)
- main functions: `lcda`, `cclcda`, `cclcda2`
- syntax like `lda(MASS)` (including `predict` method)
- example:

```
lcda(x, ...)

## Default S3 method:
lcda(x, grouping=NULL, prior=NULL,
                              probs.start=NULL, nrep=1, m=3,
                              maxiter = 1000, tol = 1e-10,
                              subset, na.rm = FALSE, ...)
```

Application: simulation study

- intention: discrete MDA can be seen as localized Naive Bayes, it assumes local independence instead of "global" independence
- simulation of data by the discrete MDA model with and without existing subgroups
- probabilities θ_{mkdr} are defined in a way so that the subgroups are not existent
- in the case of existing subgroups discrete MDA classifies more adequately than Naive Bayes
- otherwise discrete MDA and Naive Bayes lead to the same decision

Application: SNP data

- GENICA study: aims at identifying genetic and gene-environment associated breast cancer risks
- 1166 observations, 605 controls and 561 cases, of 68 SNP variables and 6 categorical epidemiological variables
- application of the presented local classification methods
- comparison to the classification results of Schiffner et al. (2009) on the same data set with
  - localized logistic regression
  - CART
  - random forests
  - logic regression
  - logistic regression

Results: SNP-data

Table 1: Tenfold cross-validated error rates of the presented methods (with number of subclasses in parentheses)

method	10 cv error (sd)
lcda (10/10)	0.220 (0.030)
cclcda (4)	0.345 (0.056)
cclcda2 (10)	0.471 (0.049)

Table 2: Tenfold cross-validated error rates as noted in Schiffner et al. (2009)

method	10 cv error
localized logistic regression	0.367
CART	0.379
random forests	0.382
logic regression	0.385
logistic regression	0.366

Conclusion

- three models based on Latent Class Analysis that provide a flexible approach to local classification
- the models can handle missing values without imputation
- discrete MDA can be seen as a localized version of the Naive Bayes method
- further research: extend the methods to data of mixed type by assuming normality of the continuous variables

References

- R. Elmore and S. Wang. *Identifiability and estimation in finite mixture models with multinomial components*. Technical Report 03–04, Department of Statistics, Pennsylvania State University, 2003.

- P.F. Lazarsfeld and N.W. Henry. *Latent structure analysis*. Houghton Miflin, Boston, 1968.

- J. Schiffner, G. Szepannek, Th. Monthé, and C. Weihs. Localized Logistic Regression for Categorical Influential Factors. To appear in A. Fink, B. Lausen, W. Seidel and A. Ultsch, editors, *Advances in Data Analysis, Data Handling and Business Intelligence*. Springer-Verlag, Heidelberg-Berlin, 2009.

- H. Teicher. Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics*, 38:1300–1302, 1967.

- M.K. Titsias and A.C. Likas. Shared kernel models for class conditional density estimation. *IEEE Transactions on Neural Networks*, 12:987–997, 2001.