Classification Using Mixtures

Prof. Sharon McNicholas

STATS 780/CSE 780

Introduction

- We have looked at clustering using Gaussian, and other, mixture models.
- This includes the GPCM family, the MFA model and PGMM family, and an approach for clustering longitudinal data in a mixture framework.
- In this "lecture", we will look at model-based classification and mixture discriminant analysis.
- Again, some of the material is taken from McNicholas (2016), and references are given at the end.

Classification Framework

- Suppose n p-dimensional data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are observed.
- Suppose that k of the n observations are labelled as belonging to one of G classes and order them WLOG so that the first k are labelled.
- As with the labelled points, each of the unlabelled $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_n$ is usually taken to come from one of G classes.
- The goal is to predict the labels for $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_n$.

Model-Based Classification

- Use $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to predict the labels for $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_n$.
- This is a **semi-supervised** approach.
- The (Gaussian) model-based classification likelihood is

$$\mathcal{L}(\boldsymbol{\vartheta}) = \prod_{i=1}^{k} \prod_{g=1}^{G} [\pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)]^{z_{ig}} \times \prod_{j=k+1}^{n} \sum_{h=1}^{H} [\pi_h \phi(\mathbf{x}_j \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)], \tag{1}$$

  where $H \geq G$ but it is typically assumed that $H = G$.
- The EM algorithm is used for parameter analysis.

Mixture Discriminant Analysis

- Use $\mathbf{x}_1, \ldots, \mathbf{x}_k$ to predict the labels for $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_n$.
- This is a **supervised** approach.
- Also called "model-based discriminant analysis".
- Very briefly, model-based clustering is performed on the labelled observations in each known class and then the resulting rule is used to predict the classes for the unlabelled observations $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_n$.

Mixture Discriminant Analysis contd.

Summary of Gaussian Model-Based Discriminant Analysis

for g in 1 to G
  carry out model-based clustering for $\mathbf{x}_1, \ldots, \mathbf{x}_k$ in class g
  choose a $\mathcal{G}_g$-component model, e.g., using the BIC
  record corresponding labels LAB_g
end for
compute $\mathcal{G} = \mathcal{G}_1 + \cdots + \mathcal{G}_G$
fit a $\mathcal{G}$-component mixture to $\mathbf{x}_1, \ldots, \mathbf{x}_k$ using labels $\text{LAB}_1, \ldots, \text{LAB}_G$
record resulting parameter estimates $\hat{\boldsymbol{\pi}}$, $\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_{\mathcal{G}}$, $\hat{\boldsymbol{\Sigma}}_1, \ldots, \hat{\boldsymbol{\Sigma}}_{\mathcal{G}}$
for $j = k + 1$ to n
  for g in 1 to $\mathcal{G}$
    compute $\hat{z}_{jg} = \hat{\pi}_g \phi(\mathbf{x}_j \mid \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g) / \sum_{h=1}^{\mathcal{G}} \hat{\pi}_h \phi(\mathbf{x}_j \mid \hat{\boldsymbol{\mu}}_h, \hat{\boldsymbol{\Sigma}}_h)$
  end for
  assign $\mathbf{x}_j$ to the class corresponding to component $h = \arg\max_g \hat{z}_{jg}$
end for

Italian Olive Oil Data

- Forina and Tiscornia (1982) report the percentage composition of eight fatty acids found by lipid fraction of 572 Italian olive oils.

Palmitic acid	Palmitoleic acid	Stearic acid
Oleic acid	Linoleic acid	Linolenic acid
Arachidic acid	Eicosenoic acid	

- Broadly, the data come from three regions: Southern Italy, Sardinia, and Northern Italy.
- Within Southern Italy are North Apulia, Calabria, South Apulia, and Sicily; Sardinia is broken into Inland Sardinia and Costal Sardinia; and Northern Italy comprises Umbria, East Liguria, and West Liguria.
- These data are available within the `pgmm` package.

MixtureDA, Labelled Data

	A	B	C	D	E	F	G	H	I
North Apulia	19	0	0	0	0	0	0	0	0
Calabria	0	41	1	0	0	0	0	0	0
South Apulia	0	1	154	0	0	0	0	0	0
Sicily	0	0	0	27	0	0	0	0	0
Inland Sardinia	0	0	0	0	49	0	0	0	0
Costal Sardinia	0	0	0	0	0	24	0	0	0
East Liguria	0	0	0	0	0	0	38	0	0
West Liguria	0	0	0	0	0	0	0	37	0
Umbria	0	0	0	0	0	0	0	0	38

MixtureDA, Predicted Classifications

	A	B	C	D	E	F	G	H	I
North Apulia	5	0	1	0	0	0	0	0	0
Calabria	0	14	0	0	0	0	0	0	0
South Apulia	0	0	51	0	0	0	0	0	0
Sicily	2	2	3	2	0	0	0	0	0
Inland Sardinia	0	0	0	0	16	0	0	0	0
Costal Sardinia	0	0	9	0	0	0	0	0	0
East Liguria	0	0	0	0	0	0	12	0	0
West Liguria	0	0	0	0	0	0	7	6	0
Umbria	0	1	0	0	0	0	9	0	3

MixtureDA, Discussion

Classes	n_g (labelled)	Model	G	Misclassified
North Apulia	19	XXX	1	1
Calabria	42	XXX	1	0
South Apulia	155	XXX	1	0
Sicily	27	VEV	4	7
Inland Sardinia	49	EEI	2	0
Costal Sardinia	24	VEV	4	9
East Liguria	38	XXX	1	0
West Liguria	37	EEV	5	7
Umbria	38	EEV	5	10

Model-Based Class., Pred. Class

	A	B	C	D	E	F	G	H	I
North Apulia	6	0	0	0	0	0	0	0	0
Calabria	0	14	0	0	0	0	0	0	0
South Apulia	0	0	50	1	0	0	0	0	0
Sicily	0	1	0	8	0	0	0	0	0
Inland Sardinia	0	0	0	0	16	0	0	0	0
Coastal Sardinia	0	0	0	0	0	9	0	0	0
East Liguria	1	0	0	0	0	0	11	0	0
West Liguria	0	0	0	0	0	0	1	12	0
Umbria	0	0	0	0	0	0	0	0	13

Comments

- The olive oil examples are taken from McNicholas (2016).
- They illustrate a potential danger in mixture discriminant analysis.
- This danger is related to what we have discussed about merging Gaussian components in model-based clustering.
- Perhaps unsurprisingly, using straightforward discriminant analysis with a more flexible distribution (e.g., parameterizing concentration and skewness) for each class leads to better results on the olive oil data than mixture DA (McNicholas, 2016).
- Let's look at some examples in R.

References

- Forina, M. and E. Tiscornia (1982). 'Pattern recognition methods in the prediction of Italian olive oil origin by their fatty acid content'. *Annali di Chimica* **72**, 143–155.

- McNicholas, P.D. (2016). *Mixture Model-Based Classification*. Boca Raton: Chapman & Hall/CRC Press.