Model-based clustering and classification

Julien JACQUES

http://eric.univ-lyon2.fr/~jjacques
Plan

Introduction

The mixture model
   Definition and notation
   Clustering and classification rule
   Continuous features: the Gaussian mixture
   Categorical features: the multinomial mixture

Model selection

Classification
   Continuous features
   Categorical features
   Continuous and categorical (mixed) features

Clustering
   Continuous features
   Categorical and mixed features

Assignment
Plan

Introduction
The mixture model
  Definition and notation
  Clustering and classification rule
  Continuous features: the Gaussian mixture
  Categorical features: the multinomial mixture

Model selection

Classification
  Continuous features
  Categorical features
  Continuous and categorical (mixed) features

Clustering
  Continuous features
  Categorical and mixed features

Assignment
Introduction

Contents:

- **clustering** (unsupervised):
  
  *task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters)*

- **classification** (discrimination, scoring / supervised):
  
  *to predict the group of a new observation from a labeled sample*
Introduction

Contents:

- **clustering** (unsupervised):
  *task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters)*

- **classification** (discrimination, scoring / supervised):
  *to predict the group of a new observation from a labeled sample*

Notations:

- observations are described by $p$ features $\mathbf{X} = (X_1, \ldots, X_p) \in E$
  
  ($E = \mathbb{R}^p, \ldots$)

- $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$ is the features for observation $i$ ($1 \leq i \leq n$)

- $Z_i \in \{1, \ldots, K\}$ is the group of observation $i$
Clustering vs classification

Clustering

- $Z_i$ unknown
- goal: to predict $Z_1, \ldots, Z_n$ from $X_1, \ldots, X_n$
- $Z_1, \ldots, Z_n$ are a posteriori interpreted in order to give significance to the clusters

Classification

- $Z_i$ observed
- goal: to build a classification rule $r$ from $(X_1, Z_1), \ldots, (X_n, Z_n)$:

$$r : X \rightarrow r(X) = Z$$

- to use this rule in order to classify new observation for which the group is unknown
Some applications

Clustering
- exploratory analysis: to give a simplified representation of data in order to understand them
- example: customer typology in marketing (Customer Relationship Management)

Classification
- predictive analysis: to predict $Z$ (categorical) from covariates $X$ (categorical, continuous...)
- example: to predict the probability (score) ...
  - marketing: ...for a new customer to buy a product
  - medicine: ...for a patient to be suffering from a disease
  - finance: ...for a firm to enter bankruptcy
Different methods

Clustering

- geometric methods
  - kmeans, hierarchical clustering
- probabilistic methods
  - mixture models

Classification

- generative methods: estimation of $p(X, Z)$
  - mixture models (linear/quadratic discriminant analysis, ...)
- predictive methods: estimation of $p(Z|X)$
  - logistic regression, K-nearest neighbors, classification tree, SVM, neural networks...
Different methods

Clustering

- geometric methods
  - kmeans, hierarchical clustering
- probabilistic methods
  - mixture models

Classification

- generative methods: estimation of \( p(\mathbf{X}, Z) \)
  - mixture models (linear/quadratic discriminant analysis, ...)
- predictive methods: estimation of \( p(Z|\mathbf{X}) \)
  - logistic regression, K-nearest neighbors, classification tree, SVM, neural networks...
Plan

Introduction

The mixture model
- Definition and notation
- Clustering and classification rule
- Continuous features: the Gaussian mixture
- Categorical features: the multinomial mixture

Model selection

Classification
- Continuous features
- Categorical features
- Continuous and categorical (mixed) features

Clustering
- Continuous features
- Categorical and mixed features

Assignment
Plan

Introduction

The mixture model
  Definition and notation
  Clustering and classification rule
  Continuous features: the Gaussian mixture
  Categorical features: the multinomial mixture

Model selection

Classification
  Continuous features
  Categorical features
  Continuous and categorical (mixed) features

Clustering
  Continuous features
  Categorical and mixed features

Assignment
The mixture model

- Idea: each group is described by its own probability distribution

\[
X|Z = k \sim f(x, \theta_k) = f_k(x)
\]

for instance, \( f(\cdot, \theta_k) \) can be
- Continuous features: the Gaussian distrib. \((\theta_k = (\mu_k, \Sigma_k))\), the Student distribution for more heavy tails...
- Binary features: multivariate Bernoulli distrib., \(\theta_k = (\alpha_{kj})_{1 \leq j \leq p}\)
- Categorical features: multinomial distribution...

- Mixing proportion

\[
Z = k \iff \tilde{Z} = (0, \ldots, 0, \underbrace{1}_{k-\text{th position}}, 0, \ldots, 0)
\]

\[
\tilde{Z} \sim \mathcal{M}(1, p_1, \ldots, p_K)
\]

where \(p_k = P(Z = k) = P(\tilde{Z}_k = 1)\) is the mixing proportion of group \(k\)
The mixture model

- marginal distribution of $\mathbf{X}$ (mixture density)

$$\mathbf{X} \sim \sum_{k=1}^{K} p_k f_k(\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}).$$

proof: $P(\mathbf{X} \in I) = P(\mathbf{X} \in I \cap Z \in \{1, \ldots, K\}) = \sum_{k=1}^{K} P(\mathbf{X} \in I \cap Z = k) = \sum_{k=1}^{K} P(\mathbf{X} \in I|Z = k)P(Z = k)$

- conditional probability that $\mathbf{x}$ belongs to group $k$ (via Bayes theorem):

$$t_k(\mathbf{x}) = \frac{p_k f_k(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}.$$

proof: $t_k(\mathbf{x}) = P(Z = k|\mathbf{X} = \mathbf{x}) = \frac{P(\mathbf{x}|Z=k)P(Z=k)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|Z=k)P(Z=k)}{\sum_{\ell=1}^{K} p(\mathbf{x}|Z=\ell)P(Z=\ell)}$
Plan

Introduction

The mixture model
  Definition and notation
  Clustering and classification rule
    Continuous features: the Gaussian mixture
    Categorical features: the multinomial mixture

Model selection

Classification
  Continuous features
  Categorical features
  Continuous and categorical (mixed) features

Clustering
  Continuous features
  Categorical and mixed features

Assignment
Clustering and classification rule

Let assume that all the mixture model parameters ($p_k$ and the parameters of $f_k$) are known (they will be estimated in practice from data).

Clustering:

each observation $x$ is classified into the group $k$ maximizing the conditional probability $t_k(x) = P(Z = k|X = x)$:

$$Z = \arg\max_k t_k(x)$$
Classification rule and classification error

Classification rule: \( r : \mathbb{X} \rightarrow r(\mathbf{x}) \in \{1, \ldots, K\} \).

To define \( r \) \( \iff \) to divide the \( \mathbb{X} \) into \( K \) subsets \( \Omega_k \) s.t.
\[
\Omega_1 \cup \ldots \cup \Omega_K = \mathbb{R}^p, \quad \Omega_k \cap \Omega_\ell = \emptyset \quad \text{and} \quad \mathbf{x} \in \Omega_k \iff r(\mathbf{x}) = k.
\]

Probability to classify an observation of group \( G_k \) into \( G_\ell \) (\( \ell \neq k \)) with \( r \):
\[
e_k \ell (r) = P(r(\mathbf{X}) = \ell | Z = k) = \int_{\Omega_\ell} f_k(\mathbf{x}) d\mathbf{x}.
\]

Probability of misclassification of an observation of \( G_k \) with \( r \):
\[
e_k(r) = P(r(\mathbf{X}) \neq k | Z = k) = \sum_{\ell \neq k} e_k \ell (r) = \int_{\mathbb{E} \setminus \Omega_k} f_k(\mathbf{x}) d\mathbf{x}.
\]

Global probability of misclassification (global misclassification error):
\[
e(r) = \sum_{k=1}^{K} p_k e_k(r).
\]
Misclassification cost

Cost of misclassifying an observation of $G_\ell$ in $G_k$:

$$C : (k, \ell) \in \{1, \ldots, K\} \times \{1, \ldots, K\} \to C(k, \ell) \in \mathbb{R}^+,$$

with $C(k, k) = 0$.

Misclassification cost

- generally not symmetric
- to be defined with practician (or fixed to 1 if you have no information).

Example:

- a consumer finance company predicts if the customer will have ($G_1$) problem in repaying loan or not ($G_2$)
- $C(1, 2)$: cost to classify a *good* customer ($G_2$) as a *bad* one ($G_1$)
- $C(2, 1)$: cost to classify a *bad* customer as a *good* one
- the company probably has to choose $C(2, 1) \gg C(1, 2)$
Bayes optimal classification rule

**Conditional risk** associated to $\mathbf{x}$: average cost of misclassification of $\mathbf{x}$

$$R(r(\mathbf{X})|\mathbf{X} = \mathbf{x}) = E[C(r(\mathbf{X}), Z)|\mathbf{X} = \mathbf{x}] = \sum_{k=1}^{K} C(r(\mathbf{x}), k) t_k(\mathbf{x}),$$

**Average risk**

$$R(r) = E_{\mathbf{x}}[R(r(\mathbf{X})|\mathbf{X} = \mathbf{x})] = \sum_{k=1}^{K} p_k \sum_{\ell=1}^{K} C(\ell, k) \int_{\Omega_{\ell}} f_k(\mathbf{x}) d\mathbf{x}.$$ 

Proofs: exercice.
Bayes optimal classification rule

We look for the optimal rule $r^*$ which minimize the average risk, which is equivalent to minimize the conditional risk since:

$$R(r^*) = \min_r E_x[R(r(X)|X=x)] \geq E_x[\min_r R(r(X)|X=x)].$$

The optimal rule classify $x$ into $G_k$ if

$$R(r(X) = k|X=x) < R(r(X) = \ell|X=x) \quad \forall \ell \neq k.$$

Since

$$R(r(X) = k|X=x) = E[C(k, Z)|X=x] = \sum_{\ell=1}^{K} C(k, \ell) t_\ell(x) = \sum_{\ell \neq k}^{K} C(k, \ell) t_\ell(x),$$

the optimal Bayes classification rule is:

$$r^*(x) = k \quad \text{if} \quad \sum_{\ell \neq k}^{K} C(k, \ell) t_\ell(x) < \sum_{\ell \neq k'}^{K} C(k', \ell) t_\ell(x) \quad \forall k' \neq k.$$
Bayes optimal rule for equal costs

If $C(k,\ell) = c \forall k \neq \ell$, the conditional risk is

$$R(r(X) = k|X = x) = c \sum_{\ell \neq k} t_\ell(x) = c(1 - t_k(x)),$$

and thus $r^*(x) = k$ if $c(1 - t_k(x)) < c(1 - t_{k'}(x)) \ \forall k' \neq k$ or equivalently

$$r^*(x) = k \quad \text{if} \quad t_k(x) > t_{k'}(x) \quad \forall k' \neq k.$$

$\Rightarrow x$ is classify into the group which has the greater posterior probability (maximum a posteriori).

If moreover $c = 1$, the average risk is

$$R(r) = \sum_{k=1}^{K} p_k \sum_{\ell \neq k} \int_{\Omega_\ell} f_k(x)\,dx = \sum_{k=1}^{K} p_k \int_{\Omega_\ell} f_k(x)\,dx = \sum_{k=1}^{K} p_k e_k(r) = e(r)$$
Bayes optimal rule for 2 groups

For 2 groups, we have

\[ r^*(x) = 1 \quad \text{if} \quad C(1, 2)t_2(x) < C(2, 1)t_1(x), \]
and \[ r^*(x) = 2 \quad \text{if} \quad C(2, 1)t_1(x) < C(1, 2)t_2(x), \]

and by noting \( g(x) = \frac{C(2, 1)t_1(x)}{C(1, 2)t_2(x)} \), the Bayes optimal rule is:

\[ r^*(x) = 1 \quad \text{if} \quad g(x) > 1, \]
and \[ r^*(x) = 2 \quad \text{if} \quad g(x) < 1. \]

\( g(x) = 1 \) is the equation of the *separating surface*. 
Plan

Introduction

The mixture model
  Definition and notation
  Clustering and classification rule
  Continuous features: the Gaussian mixture
  Categorical features: the multinomial mixture

Model selection

Classification
  Continuous features
  Categorical features
  Continuous and categorical (mixed) features

Clustering
  Continuous features
  Categorical and mixed features

Assignment
The Gaussian Mixture Model

The density of group $k$ is

$$f_k(x) = \frac{1}{(2\pi)^{p/2}\lvert\Sigma_k\rvert^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_k)^t\Sigma_k^{-1}(x - \mu_k)\right\}$$

where $\mu_k$ is the mean vector and $\Sigma_k$ the variance matrix of group $k$ ($|\Sigma_k|$ denotes the determinant of $\Sigma_k$):

$$\Sigma_k = \begin{pmatrix}
\text{var}(X_1|Z = k) & \text{cov}(X_1, X_2|Z = k) & \ldots & \text{cov}(X_1, X_p|Z = k) \\
\text{cov}(X_2, X_1|Z = k) & \text{var}(X_2|Z = k) & \ldots & \text{cov}(X_2, X_p|Z = k) \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}(X_p, X_1|Z = k) & \text{cov}(X_p, X_2|Z = k) & \ldots & \text{var}(X_p|Z = k)
\end{pmatrix}$$
The Gaussian Mixture Model
Model complexity

Number of parameters of the model:

- $\Sigma_k$: $K \times p(p + 1)/2$
- $\mu_k$: $K \times p$
- $p_k$: $K - 1$ (since $\sum_{k=1}^{K} p_k = 1$
- total: $K(p(p + 1)/2 + p + 1) - 1$

example: $K = 3, p = 10 \Rightarrow 197$ parameters to estimate
  $K = 6, p = 100 \Rightarrow 30905$ parameters to estimate

There is a need to reduce the number of parameters

Two ways to do that:

⇒ feature selection
⇒ parsimonious models
Parsimonious models

Most of the parameters are dedicated to the variance matrices $\Sigma_k$.

**Parsimonious Gaussian models** (Banfield & Raftery 1993, Celeux & Govaert 1995)

- spectral decomposition $\Sigma_k = \lambda_k D_k A_k D_k^t$ where
  - $\lambda_k$: largest eigenvalue
  - $D_k$: orthogonal matrix of eigenvectors
  - $A_k$: diagonal matrix of normalized eigenvalues, such that $A_k = diag(a_{1k}, \ldots, a_{pk})$ with $1 = a_{1k} \geq \ldots \geq a_{pk}$

**Interpretation:**
  - $\lambda_k$: volume of (space occupied by) group $k$
  - $D_k$: orientation of group $k$
  - $A_k$: shape of group $k$

Restrictions on $\lambda_k$, $D_k$, $A_k \Rightarrow$ parsimonious models
Parsimonious Gaussian models

ellipsoidal, equal volume, shape, and orientation ($\lambda$, $D$, $A$)

diagonal, equal volume and shape ($\lambda$, $A$, $D + DAD^t$ diagonal)

diagonal, equal volume, varying shape ($\lambda$, $A_k$, $D + DA_k D^t$ diagonal)

ellipsoidal, equal shape ($\lambda_k$, $D_k$, $A$)
Plan

Introduction

The mixture model
  Definition and notation
  Clustering and classification rule
  Continuous features: the Gaussian mixture
  Categorical features: the multinomial mixture

Model selection

Classification
  Continuous features
  Categorical features
  Continuous and categorical (mixed) features

Clustering
  Continuous features
  Categorical and mixed features

Assignment
The Multinomial Model

- each categorical feature $X_j$ is coded as follows: $X_j = (X_j^1, \ldots, X_j^{m_j})$ with $X_j^h = 1$ if the $j$th categorical feature takes the $h$th category, 0 otherwise.
- the full multinomial model (for group $k$) is defined by probabilities:

$$f_k(x) = p(x_1^{h_1} = 1, \ldots, x_p^{h_p} = 1 | Z = k) = \alpha_k^{h_1 \ldots h_p}$$

- number of parameters for each conditional distribution: $\prod_{j=1}^{p} m_j - 1$
where $m_j$ is the number of categories of feature $X_j$
(10 features with 5 categories $\Rightarrow 10^5 - 1$)
- this model is never used due to its too large number of parameters
The Latent Class Model

- The Latent Class Model assumes that the categorical feature are independent conditionally to $Z$

$$f_k(\mathbf{x}) = p(x_1^{h_1} = 1, \ldots, x_p^{h_p} = 1 | Z = k)$$

$$= \prod_{j=1}^{p} p(x_j^{h_j} = 1 | Z = k) = \alpha_k^{h_1} \times \ldots \times \alpha_k^{h_p} = \prod_{j=1}^{p} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_j^h}$$

- number of parameters for each conditional distribution: $\sum_{j=1}^{p} (m_j - 1)$
  (10 features with 5 categories $\Rightarrow$ 40 parameters)

- the marginal distribution is:

$$f(\mathbf{x}) = \sum_{k=1}^{K} p_k \prod_{j=1}^{p} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_j^h}$$

- more parsimonious models can be considered that for each $X_j$, only the probability of the majority category is free (all the others categories are assumed to be equally distributed)
Plan

Introduction

The mixture model
  Definition and notation
  Clustering and classification rule
  Continuous features: the Gaussian mixture
  Categorical features: the multinomial mixture

Model selection

Classification
  Continuous features
  Categorical features
  Continuous and categorical (mixed) features

Clustering
  Continuous features
  Categorical and mixed features

Assignment
Feature and model selection

We have to select:

- which features to include in the model
- which model is the a **good model** (among the parsimonious ones)
  - model too simple: low variance estimation but high biais
  - model **too complex**: low biais by high variance (**overfitting**)  

- additionally for clustering: which number of clusters
Feature and model selection

**Model selection** for classification:
- estimation of the misclassification error $e(r)$
  - training / test samples
  - cross-validation
- penalized likelihood criteria
  (only for comparing models with the same features)

**Feature selection** for classification:
- stepwise algorithms

**Model selection** for clustering:
- penalized likelihood criteria
Model selection for classification

Goal: to estimate the misclassification error \( e(r) \) and to select the model with the smallest one.

- estimation by reclassifying the training set underestimate \( e(r) \), and lead to select the most complex model
- partition method: estimation using a test set:
  - if the available sample is large enough: divide it into \textit{training} (70\%) and \textit{test} (30\%) sets
  - estimate the model parameters using the \textit{training} set
  - evaluate the error on the \textit{test} set

\[
\hat{e}_{\text{training}}(r) = 1 - \frac{1}{n_t} \sum_{i \in \text{test}} \tilde{z}_{ik} \hat{r}_{\text{training}}(x_i)
\]

where

- \( n_t = \#\text{test} \),
- \( \hat{r}_{\text{training}}(x_i) \) is the classification rule for \( x_i \) estimated using the training set,
- \( \tilde{z}_{ik} = 1 \) if \( x_i \) belongs to group \( k \).
Model selection for classification

- **Leave-One-Out cross-validation method:**
  - if the available sample is not large enough, the idea is to maximize the size of the training set
  - \( test = \{x_i\} \) and \( training = \{x_1, \ldots, x_n\} \setminus \{x_i\} \Rightarrow \hat{e}_{\{x_i\}}(r) \)
  - this individual error is now averaged over all the possible test sets (⇔ all the observations):

\[
\hat{e}_{CV}(r) = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_{\{x_i\}}(r)
\]

- **V-fold cross-validation method:**
  - to reduce the computing time, the whole sample is partitioned into \( V \) folds: \( S_1, \ldots, S_V \)
  - \( test = S_i \) and \( training = \{x_1, \ldots, x_n\} \setminus S_i \Rightarrow \hat{e}_{S_i}(r) \)
  - this individual error is now averaged over the \( V \) folds:

\[
\hat{e}_{CV}(r) = \frac{1}{V} \sum_{i=1}^{V} \hat{e}_{S_i}(r)
\]
Model selection for classification

- **Penalized-likelihood criteria**
  - probabilistic framework of mixture model allows to use likelihood
  - model log-likelihood

  \[
  \ell(\mathbf{x}_1, \ldots, \mathbf{x}_n, \theta) = \ln \prod_{i=1}^{n} f_X(\mathbf{x}_i, \theta) = \sum_{i=1}^{n} \ln f_X(\mathbf{x}_i, \theta)
  \]

  can not be used for model selection since it would select the most complex one

  - BIC and AIC criteria penalize \( \ell(\theta) \) by the model complexity

    \[
    BIC = -2\ell(\mathbf{x}, \theta) + \nu \ln n \quad \text{AIC} = -2\ell(\mathbf{x}, \theta) + 2\nu
    \]

    with \( \nu \) the number of model parameters.

    Model with smallest BIC or AIC should be selected

    - minimizing BIC \( \Leftrightarrow \) maximizing the probability a posteriori of the model (Bayesian paradigm)
    - minimizing AIC \( \Leftrightarrow \) minimizing the information lost (information theory)
    - BIC generally selects most simple models than AIC (\( \ln n > 2 \))
Model selection for classification in practice

In practice:
■ large sample: partition method
■ small sample: Leave-One-Out cross-validation
■ intermediate sample: $V$-fold cross-validation ($V = 3, 10...$)
■ or, for any sample size: BIC *generally preferred to AIC in a classification or clustering context*
Feature selection for classification

Feature selection = model selection

- each subset of features defines a new model
- models can be compared using cross-validation / partition methods
  but not AIC / BIC (if the features change, the likelihood are not computed
  on the same data and so are not comparable)

Stepwise algorithm

- number of subsets of \{X_1, \ldots, X_p\} is combinatorial \Rightarrow exhaustive
  enumeration intractable
- forward stepwise algorithm defines a list of \(p\) models

\begin{align*}
1: & \textbf{while} \ \text{number of features in the model} < p \ \textbf{do} \\
2: & \text{find the best feature (using } \hat{e}^{CV}(r) \text{ or } \hat{e}^{P}(r)\text{)} \\
3: & \text{include it in the model} \\
4: & \textbf{end while}
\end{align*}

- the best model among this short list is retained (\(\hat{e}^{CV}(r)\) or \(\hat{e}^{P}(r)\))
Model selection for clustering

- unsupervised setting $\Rightarrow$ none labeled sample is available: cross-validation / partition methods not usable
- penalized likelihood criteria as BIC can be used to:
  - select the **best model** among several models using the same features
  - select the **number $K$ of clusters**

- no feature selection task for clustering: new features $\Rightarrow$ new data $\Rightarrow$ new clustering study
Exercices

Theoretical

- proofs on slide 11, 16.
- explain the expression of $e(r)$ on slide 14.
Plan

Introduction

The mixture model
- Definition and notation
- Clustering and classification rule
- Continuous features: the Gaussian mixture
- Categorical features: the multinomial mixture

Model selection

Classification
- Continuous features
- Categorical features
- Continuous and categorical (mixed) features

Clustering
- Continuous features
- Categorical and mixed features

Assignment
Plan

Introduction

The mixture model
  Definition and notation
  Clustering and classification rule
  Continuous features: the Gaussian mixture
  Categorical features: the multinomial mixture

Model selection

Classification
  Continuous features
  Categorical features
  Continuous and categorical (mixed) features

Clustering
  Continuous features
  Categorical and mixed features

Assignment
Parameter estimation

Full mixture of Gaussian:

\[ f_X(x, \theta) = \sum_{k=1}^{K} p_k \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_k)^t \Sigma_k^{-1} (x - \mu_k)\right\} \]

where the model parameters are \( \theta = (p_k, \mu_k, \Sigma_k)_{1 \leq k \leq K} \)

- Estimation of the classification rule \( r^* \) is obtained by estimating \( \theta \) by maximum likelihood
- log-likelihood in the classification context (\( x, z \) available):

\[ \ell(x, z, \theta) = \sum_{k=1}^{K} \sum_{i=1}^{n} \tilde{z}_{ik} \left( \ln p_k - \frac{p}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^t \Sigma_k^{-1} (x_i - \mu_k) \right) \]

- maximization leads to the usual empirical estimates:
  - \( \hat{p}_k = \frac{n_k}{n} \) with \( n_k = \sum_{i=1}^{n} \tilde{z}_{ik} \) the number of observations of group \( k \)
  - \( \hat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n} \tilde{z}_{ik} x_i \)
  - \( \hat{\Sigma}_k = \frac{1}{n_k} \sum_{i=1}^{n} \tilde{z}_{ik} (x_i - \mu_k)^t (x_i - \mu_k) \) or dividing by \( n_k - 1 \) for unbiased estimator

Proof: use Lagrange multipliers for the constraint \( \sum_{k=1}^{K} p_k = 1 \) and matrix derivatives
Linear and Quadratic Discriminant Analysis

The first parsimonious model has been introduced by Fisher (1936) assuming

- $\Sigma_k = \Sigma$
- the estimation of $\Sigma$ is:

$$
\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} \tilde{z}_{ik}(x_i - \mu_k)^t(x_i - \mu_k)
= \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_k)^t(x_i - \mu_k)
$$

or dividing by $n - K$ to obtain an unbiased estimator
Linear and Quadratic Discriminant Analysis

Impact of the assumption $\Sigma_k = \Sigma$ on the classification rule ($K = 2$):

- if $\Sigma_1 \neq \Sigma_2$, the equation of the **separating surface** is $g(x) = 1$ with

$$
\ln g(x) = \ln \frac{C(2, 1)p_1 f_1(x)}{C(1, 2)p_2 f_2(x)}
= \ln \frac{f_1(x)}{f_2(x)} + \ln \frac{C(2, 1)p_1}{C(1, 2)p_2}
= \frac{1}{2} \left( \ln \frac{\Sigma_2}{\Sigma_1} + (x - \mu_2)^t \Sigma_2^{-1} (x - \mu_2) - (x - \mu_1)^t \Sigma_1^{-1} (x - \mu_1) \right) + s.
$$

which is **quadratic** in $x \Rightarrow$ **Quadratic Discriminant Analysis (QDA)**

- if now $\Sigma_1 = \Sigma_2 = \Sigma$:

$$
\ln g(x) = (\mu_1 - \mu_2)^t \Sigma^{-1} (x - \frac{\mu_1 + \mu_2}{2}) + s,
$$

which is **linear** in $x \Rightarrow$ **Linear Discriminant Analysis (LDA)**
Linear and Quadratic Discriminant Analysis
LDA and QDA only

- `lda` or `qda` function of package `{MASS}`

Most parsimonious models in

- `MclustDA` function of package `{mclust}`
  

- `mixmodLearn` function of package `{Rmixmod}`
Application: Fisher’s Iris

> plot(iris[,1:4], col=iris$Species)
Application: Fisher’s Iris

Partition of the dataset into train and test sets

```r
tr <- sample(1:150, 100, replace=F)
X.train <- iris[tr,-5]
Class.train <- iris[tr,5]
X.test <- iris[-tr,-5]
Class.test <- iris[-tr,5]
```

Estimation of the LDA model (with `mclust` parametrization)

```r
irisMclustDA <- MclustDA(X.train, Class.train, modelType = "EDDA",
modelNames = "EEE")
summary(irisMclustDA, parameters = TRUE)
summary(irisMclustDA, newdata = X.test, newclass = Class.test)
```

Training error = 0.02
Test error = 0.02
Application: Fisher’s Iris

> plot(irisMclustDA)
Application: Fisher’s Iris

```r
> plot(irisMclustDA, dimens = 3:4)
```
Application: Fisher’s Iris

> plot(irisMclustDA, dimens = 4)
Application: Fisher’s Iris

Selection of the best Gaussian mixture model using BIC

> irisMclustDA <- MclustDA(X.train, Class.train,G=1)
> summary(irisMclustDA, parameters = TRUE)

> summary(irisMclustDA, newdata = X.test, newclass = Class.test)
Training error = 0.02
Test error = 0.04
Application: Fisher’s Iris

> plot(irisMclustDA)
Application: Fisher’s Iris

```r
> plot(irisMclustDA, dimens = 3:4)
```
Application: Fisher’s Iris

> plot(irisMclustDA, dimens = 4)
Application: Fisher’s Iris

> plot(irisMclustDA, what = "classification", newdata = X.test)
Application: Fisher’s Iris

> plot(irisMclustDA, what = "error", newdata = X.test)
Plan

Introduction
The mixture model
  Definition and notation
  Clustering and classification rule
  Continuous features: the Gaussian mixture
  Categorical features: the multinomial mixture

Model selection

Classification
  Continuous features
  Categorical features
  Continuous and categorical (mixed) features

Clustering
  Continuous features
  Categorical and mixed features

Assignment
Parameter estimation

- Latent Class Model assumes that the categorical features are independent conditionally to $Z$

$$f(x) = \sum_{k=1}^{K} p_k \prod_{j=1}^{p} \prod_{h=1}^{m_j} (\alpha_k^{jh} x_j^h)$$

- maximum likelihood estimation

$$\hat{\alpha}_k^{jh} = \frac{1}{n} \sum_{i=1}^{n} \tilde{z}_{ik} x_{ij}^h$$
mixmodLearn function of package \{Rmixmod\}
Application

> library(Rmixmod)
> data(birds)

> summary(birds)

<table>
<thead>
<tr>
<th>gender</th>
<th>eyebrow</th>
<th>collar</th>
<th>sub-caudal</th>
<th>border</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>none</td>
<td>none</td>
<td>white</td>
<td>none</td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>6</td>
<td>40</td>
<td>65</td>
</tr>
<tr>
<td>female</td>
<td>poor pronounced</td>
<td>dotted</td>
<td>black</td>
<td>few</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>29</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>pronounced</td>
<td>dashed</td>
<td>black &amp; white</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>38</td>
<td>0</td>
<td>3</td>
<td>many</td>
</tr>
<tr>
<td></td>
<td>very pronounced</td>
<td>longdashed</td>
<td>black &amp; WHITE</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>continuous</td>
<td></td>
<td>BLACK &amp; white</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Application

> plot(birds)
Application

```r
> learn.birds<-mixmodLearn(data=birds, knownLabels=birds.partition)
> summary(learn.birds)
```
Plan

Introduction

The mixture model
  Definition and notation
  Clustering and classification rule
  Continuous features: the Gaussian mixture
  Categorical features: the multinomial mixture

Model selection

Classification
  Continuous features
  Categorical features
  Continuous and categorical (mixed) features

Clustering
  Continuous features
  Categorical and mixed features

Assignment
A latent class mixture model

Let assume that continuous and categorical features are available

- $X_1, \ldots, X_c$: categorical
- $X_{c+1}, \ldots, X_p$: continuous

bad idea

- to discretize continuous feature into categorical ones
  $\Rightarrow$ information loss

simple but good idea

- assume that continuous and categorical features are independent conditionally to $Z = k$:

$$f_k(x) = \prod_{j=1}^{c} \prod_{h=1}^{m_j} (\alpha_k^{jh}) x_j^h \times \frac{1}{2 \pi (p-c)/2 |\Sigma_k|^{1/2}} \exp \left\{ -\frac{1}{2} (\tilde{x} - \mu_k)^t \Sigma_k^{-1} (\tilde{x} - \mu_k) \right\}$$

with $\tilde{x} = (x_{c+1}, \ldots, x_p)$
due to the independence assumption, parameter estimation can be done independently for categorical and continuous features:

\[
\ell(x, z, \theta) = \sum_{k=1}^{K} \sum_{i=1}^{n} \tilde{z}_{ik} \left( \ln f_{k}^{\text{categ.}}(x_{i1}, \ldots, x_{ic}) + \ln f_{k}^{\text{contin.}}(x_{ic+1}, \ldots, x_{ip}) \right)
\]
mixmodLearn function of package {Rmixmod}
Application

> library(Rmixmod)
> data(heterodatatrain)
> plot(heterodatatrain)
Application

```r
> learn.hetero <- mixmodLearn(heterodatatrain[-1],
knownLabels = heterodatatrain$V1)
> summary(learn.hetero)
```
Plan

Introduction

The mixture model
  Definition and notation
  Clustering and classification rule
  Continuous features: the Gaussian mixture
  Categorical features: the multinomial mixture

Model selection

Classification
  Continuous features
  Categorical features
  Continuous and categorical (mixed) features

Clustering
  Continuous features
  Categorical and mixed features

Assignment
Plan

Introduction

The mixture model
  Definition and notation
  Clustering and classification rule
  Continuous features: the Gaussian mixture
  Categorical features: the multinomial mixture

Model selection

Classification
  Continuous features
  Categorical features
  Continuous and categorical (mixed) features

Clustering
  Continuous features
  Categorical and mixed features

Assignment
Parameter estimation

Full mixture of Gaussian:

\[ f_{\mathbf{x}}(\mathbf{x}, \theta) = \sum_{k=1}^{K} p_k \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \mu_k)^t \Sigma_k^{-1} (\mathbf{x} - \mu_k) \right\} \]

where the model parameters are \( \theta = (p_k, \mu_k, \Sigma_k)_{1 \leq k \leq K} \)

- Estimation of the groups of individual \( z_1, \ldots, z_n \) is obtained by estimating \( \theta \) by maximum likelihood
- Log-likelihood in the clustering context (only \( \mathbf{x} \) is available)

\[ \ell(\mathbf{x}, \theta) = \sum_{i=1}^{n} \ln \left( \sum_{k=1}^{K} p_k \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \mu_k)^t \Sigma_k^{-1} (\mathbf{x} - \mu_k) \right\} \right) \]

- Maximization is not so easy than in the classification context
Parameter estimation: EM algorithm

The idea of EM algorithm:

- to maximize the completed-likelihood (by the unobserved data $\mathbf{z}$) is easier than the observed-data likelihood:

$$
\ell_c(\mathbf{x}, \mathbf{z}, \theta) = \sum_{k=1}^{K} \sum_{i=1}^{n} \tilde{z}_{ik} \left( \ln p_k - \frac{p}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_k| - \frac{1}{2} (\mathbf{x} - \mu_k)^t \Sigma_k^{-1} (\mathbf{x} - \mu_k) \right)
$$

- since $\mathbf{z}$ is unobserved, the $q$-th iteration of the EM algorithm consists in alternating:
  - E step: computation of
    $$
    Q(\theta, \theta^{(q)}) = E[\ell_c(\mathbf{x}, \mathbf{z}, \theta)|\mathbf{x}, \theta^{(q)}]
    $$
  - M step: maximisation of $Q(\theta, \theta^{(q)})$ according to $\theta$:
    $$
    \theta^{(q+1)} = \arg\max_{\theta} Q(\theta, \theta^{(q)})
    $$

until convergence of the log-likelihood: $|\ell(\mathbf{x}, \theta^{(q+1)}) - \ell(\mathbf{x}, \theta^{(q)})| < \epsilon$
EM algorithm - E step

Computation of $Q(\theta, \theta^{(q)}) = E[\ell_c(x, z, \theta)|x, \theta^{(q)}]$: 

$$Q(\theta, \theta^{(q)}) = \sum_{k=1}^{K} \sum_{i=1}^{n} E[\tilde{z}_{ik} | x, \theta^{(q)}] \left( \ln p_k - \frac{p}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_k| - \frac{1}{2} (x - \mu_k)^t \Sigma_k^{-1} (x - \mu_k) \right)$$

with

$$E[\tilde{z}_{ik} | x, \theta^{(q)}] = 1 \times P(\tilde{z}_{ik} = 1|x_i, \theta^{(q)}) + 0 \times P(\tilde{z}_{ik} = 0|x_i, \theta^{(q)})$$

$$= \frac{f_{|\tilde{z}_{ik}=1}(x_i, \theta^{(q)}) P(\tilde{z}_{ik} = 1|\theta^{(q)})}{f(x_i, \theta^{(q)})}$$

$$= \frac{f_k(x_i, \theta^{(q)}) p_k^{(q)}}{f_{x}(x_i, \theta^{(q)})}$$

$$= t_k^{(q)}(x_i)$$
EM algorithm - M step

Maximisation of $Q(\theta, \theta^{(q)})$ according to $\theta$:

$$Q(\theta, \theta^{(q)}) = \sum_{k=1}^{K} \sum_{i=1}^{n} t_k^{(q)}(x_i) \left( \ln p_k - \frac{p}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_k| - \frac{1}{2} (x - \mu_k)^t \Sigma_k^{-1} (x - \mu_k) \right)$$

is equivalent to the log-likelihood maximization in the classification context, but by ponderating each observation with $t_k^{(q)}(x_i)$

- $\hat{p}_k = \frac{n_k^{(q)}}{n}$ with $n_k = \sum_{i=1}^{n} t_k^{(q)}(x_i)$
- $\hat{\mu}_k = \frac{1}{n_k^{(q)}} \sum_{i=1}^{n} t_k^{(q)}(x_i)x_i$
- $\hat{\Sigma}_k = \frac{1}{n_k^{(q)}} \sum_{i=1}^{n} t_k^{(q)}(x_i)(x_i - \mu_k)^t(x_i - \mu_k)$
EM algorithm - properties

Properties

- the EM algorithm converges to a local maximum of the likelihood
- convergence to the global maximum is expected to be achieved with multiple initializations of the algorithm
- in practice, the most efficient initialization strategy is:
  - run several small EM (with 10 iterations)
  - run a long EM starting from the small EM solution with highest log-likelihood
The Classification EM algorithm

The CEM algorithm

- is a variant of the EM algorithm, obtained by rounding the $t_k^{(q)}(x_i)$:
  - $t_k^{(q)}(x_i) = 1$ for the group $k$ s.t. $t_k^{(q)}(x_i)$ is maximum
  - $t_k^{(q)}(x_i) = 0$ for the other groups
- CEM performs hard classification whereas EM performs soft classification
- the convergence of CEM is faster than EM, but leads to a biased estimator
- nevertheless, for large samples and well separated groups, the CEM is very efficient
Gaussian mixture model - link with k-means

It can be shown that the k-means algorithm is equivalent to

- a Gaussian mixture model with $\Sigma_k = \alpha I_p$
- estimated by the CEM algorithm

Mixture models sometimes generalized well known clustering algorithm
Clustering of the iris of Fisher with R

- dataset: iris of Fisher without using the species feature
- Mclust function of mclust package
- computation of BIC for all parsimonious models with 1 to 9 groups

```r
> mod1 = Mclust(iris[,1:4])
> plot(mod1)
```
Application: Fisher’s Iris

⇒ the best model is VEV with 2 groups
Application: Fisher’s Iris

VEV model (ellipsoidal, equal shape, varying volume and orientation) with 2 groups
Application: Fisher’s Iris

Density contour of the VEV model with 2 groups
Application: Fisher’s Iris

Comparison of the clustering with the species partition
Plan

Introduction
The mixture model
  Definition and notation
  Clustering and classification rule
  Continuous features: the Gaussian mixture
  Categorical features: the multinomial mixture

Model selection
Classification
  Continuous features
  Categorical features
  Continuous and categorical (mixed) features

Clustering
  Continuous features
  Categorical and mixed features

Assignment
Categorical and mixed features

- the same mixture models than in classification are used
- maximum likelihood estimation can be performed thanks to the EM (or CEM) algorithm
- \texttt{R} package: \texttt{Rmixmod}
Plan

Introduction

The mixture model
  Definition and notation
  Clustering and classification rule
  Continuous features: the Gaussian mixture
  Categorical features: the multinomial mixture

Model selection

Classification
  Continuous features
  Categorical features
  Continuous and categorical (mixed) features

Clustering
  Continuous features
  Categorical and mixed features

Assignment
Exercice

1. slide 41: prove the expression of the log-likelihood and of the estimators
2. predict the group of $X_5 = 4$ with LDA and QDA, using this training set:
   - group 1: $X_1 = 0, X_2 = 2$
   - group 2: $X_3 = 6, X_4 = 11$
Exercice

1. slide 41: prove the expression of the log-likelihood and of the estimators
2. predict the group of $X_5 = 4$ with LDA and QDA, using this training set:
   - group 1: $X_1 = 0$, $X_2 = 2$
   - group 2: $X_3 = 6$, $X_4 = 11$

Corrections of question 2:

**QDA**
- $\hat{p}_1 = \hat{p}_2 = 0.5$
- group 1: $\hat{\mu}_1 = 1$, $\hat{\sigma}_1^2 = 1$
- group 2: $\hat{\mu}_2 = 8.5$, $\hat{\sigma}_2^2 = \frac{25}{4}$
- $t_1(4) = \frac{0.5 \times \text{dnorm}(4, 1, 1)}{0.5 \times \text{dnorm}(4, 1, 1) + 0.5 \times \text{dnorm}(4, 8.5, \frac{5}{2})} = 0.08254814$

**LDA**
- $\hat{\sigma}^2 = 17.6875$
- $t_1(4) = \frac{0.5 \times \text{dnorm}(4, 1, \sqrt{17.6875})}{0.5 \times \text{dnorm}(4, 1, \sqrt{17.6875}) + 0.5 \times \text{dnorm}(4, 8.5, \sqrt{17.6875})} = 0.5044949$
Assignment: application in classification

Marketing application

- the VisaPremier.txt dataset provides information about bank customers and if they possess or not the VisaPremier credit card
  
  http://eric.univ-lyon2.fr/~jjacques/Download/DataSet/VisaPremier.txt

  Meaning of the attributes are available (in french) here:
  
  http://eric.univ-lyon2.fr/~jjacques/Download/Cours/FDD-TPclassification.pdf

- goal: to predict the probability for a new customer to buy this card

- for that, use
  
  1. only the continuous features
  2. only the categorical features
  3. both continuous and categorical features
Assignment: application in classification

Pre-processing with R:
- start by identifying which feature is categorical, and which is continuous (I can help for this)
- replace missing values by the mean (cont. feat.) or the mode (categ. feat.)
- perform a factorial analysis in order to have a first idea of data and to detect potential outliers. For this, you can either:
  - perform PCA on cont. feat. and MCA for categ. feat.
  - perform PCAmix on both kinds of data simultaneously (package PCAmixdata)
- some variables can be constant (or almost constant), you may have to omit them
Assignment: application in classification

Model estimation:
- you can either use the function for R:
  - `MculstDA` function of package `{mclust}` (for cont. features only)
  - `mixmodLearn` function of package `{Rmixmod}` (for cont. and/or categ. features)
- you can also use the following SaaS:
  https://modal-research.lille.inria.fr/BigStat/public/project/mixtcomp

Model evaluation:
- evaluate your model(s) on a test dataset (randomly chosen among the whole sample) (or with cross-validation).
References

Books


Journal papers