Model-Based Clustering, Discriminant Analysis, and Density Estimation

Presentation by C. Simon

Paper by Chris Fraley and Adrian E. Raftery
Published in J. of American Statistical Association
June 2002, vol 97, No.458, Review Paper

Applied Math and Statistics, UCSC

March 9, 2010

Outline

- Mixture Model
- EM algorithm
- Model Selection
- Cluster Analysis
- Discriminant Analysis
- Density Estimation
- future Work

Mixture Model

$$l(\theta_1, ..., \theta_G; \tau_1, ..., \tau_G | \mathbf{y}) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(y_i | \theta_k)$$

usually

$$f_k(\mathbf{y_i} | \theta) = \phi(\mathbf{y_i} | \mu_k, \Sigma_k)$$

such that

$$\mathbf{y_i} | \mu_k, \Sigma_k \sim N(\mu_k, \Sigma_k)$$

Mixture Model

For flexibility Σ_k is allowed different geometrical features.

- $\Sigma_k = \lambda I$ spherical clusters (for all k)
- $\Sigma_k = \lambda_k I$ spherical clusters (possibly different)
- $\Sigma_k = \lambda_k A_k$ covariance diagonal (shape, size, orientation allowed to vary.)
- Σ_k no constraints

The Expectation Maximization Algorithm for Mixtures

Rate of convergence can be slow. Typically gives good results if data conforms well to model.
Complete likelihood

$$l_c(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(\mathbf{x}_i|\theta)$$

Observed likelihood

$$l_o(y|\theta) = \int l_c(\mathbf{x}|\theta)d\mathbf{z} \qquad \mathbf{x} = (\mathbf{y}, \mathbf{z})$$

For clustering,

$$z_{i,k} = \begin{cases} 1 & \text{if } x_i \text{ belongs to group } k \\ 0 & \text{otherwise} \end{cases}$$

$\mathbf{z}_i$ is from a multinomial distribution.

EM algorithm (normals are great)

Typically,

$$\sum_{k=1}^{G} \tau_k f_k(y_i|\theta_k) \Rightarrow \prod_{k=1}^{G} f_k(\mathbf{y}_i|\theta_k)^{z_{i,k}}$$

$$\Rightarrow l(\theta, \tau, \mathbf{z}|\mathbf{y}) = \sum_{i=1}^{n}\sum_{k=1}^{G} z_{i,k} \log\left(\tau_k f_k(\mathbf{y}_i|\theta_k)\right)$$

E-step, conditional expectation of the complete data log-likelihood with current parameter estimates

$$\hat{z}_{i,k} \leftarrow \frac{\hat{\tau}_k f_k(\mathbf{y}_i|\hat{\theta_k})}{\sum_{j=1}^{G} \hat{\tau}_j f_j(\mathbf{y}_i|\hat{\theta_k})}$$

M-step, parameters are maximized with respect to the expected log-likelihood from E-step

$$\hat{\tau}_k \leftarrow \frac{n_k}{n} \qquad \hat{\mu}_k \leftarrow \frac{\sum_i \hat{z_{i,k}} \mathbf{y}_i}{n_k} \qquad n_k = \sum_{i=1}^{n} \hat{z}_{i,k}$$

Model Selection

$$P(M_k|D) \propto P(D|M_k)P(M_k)$$

D is the data, and M_k refers to the kth model.

$$P(D|M_k) = \int p(D|\theta_k, M_k)p(\theta_k, M_k)d\theta_k$$

$$B_{12} = \frac{p(D|M_1)}{p(D|M_2)}$$

M_1 if $B_{12} > 1$ often $2\log(B_{12})$ is reported.

Model Seclection

The integrated likelihood can be approximated by BIC.

$$2log(p(D|M_k)) \approx 2\log(p(D|\hat{\theta}_k, M_k)) - \nu_k \log(n)$$

ν_k = number of independent parameters estimated in M_k

Approximation is particularly good when unit information prior is used.

No proof of regularity for mixture, but several examples suggest give good performance in model based clustering context.

Cluster Strategy

- Determine a max # of cluster M. (also, a set of mixture models to consider)
- Apply EM.
- Compute BIC for optimal parameters found in EM step for 1,...,M clusters.

Discriminant Analysis

Also known as *supervised classification*. Known classifications of some observations are used to classify others. The number of classes, G is assumed known

$$\Pr[\mathbf{y} \in \text{ class } j] = \frac{\tau_j f_j(\mathbf{y}_i)}{\sum_{k=1}^{G} \tau_k f_k(\mathbf{y}_i)}$$

τ_j is the proportion of members of the population that are in class k.

Assign y to class which has highest posterior probability of belonging minimizes the expected misclassification rate; this is called the Bayes classifier

Discriminant Analysis

$$\text{again, } \mathbf{y}|\mu_k, \Sigma_k \sim N(\mu_k, \Sigma_k)$$

- If $\Sigma_k = \Sigma$ for $k = 1, ..., G$ and μ_k, Σ are MLE $\rightarrow$ the conditional Bayes classifier is equivalent to Fisher's linear discriminant analysis.
- If Σ_k's are not constrained then it is quadratic discrimination analysis

(Single EM step can be used to classify)

Mixture Discriminant Analysis

Each class is a mixture,

$$f_j(\mathbf{y}|\theta) = \sum_{k=1}^{G_j} \tau_{jk} N(\mathbf{y}|\mu_{jk}, \Sigma_{jk})$$

- Allows much more flexibility than is possible with traditional LDA and QDA.
- Model Based Clustering can be used for the training data!!!

Density Estimation

The model-based clustering method can be viewed as leading to a multivariate extension of their method, because the parameter estimates define a multivariate mixture density for the data,

$$l(\theta_1, ..., \theta_G; \tau_1, ..., \tau_G | \mathbf{y}) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(\mathbf{y}_i | \theta_k)$$

Futurework

- Bayesianify it!!! Prior, prior, prior!!! Prior this, Prior that. Lots of room for prior. I would be very interested in seeing priors for mixture discriminant analysis.
- Reversible Jump MCMC for clustering, as well as, model selection. Needs comparison for simple models. Does the computational time of RJMCMC lead to significant improvements from BIC analysis?