Model based Clustering and Chinese Restaurant process

Jessica Tressou (tressou@ust.hk)

Hong Kong University of Science and Technology, Department of Information and Systems Management

ABARI, INA-PG, Paris, June 29, 2006

Outline

- Bayesian "'‘Updating"'’
- Model based clustering
- Chinese Restaurant Process and Sequential Seating Procedure
- Gibbs Weighted Chinese Restaurant process
- Extension: Mixture Methods, Estimation of the Mixing Distribution
- Ideas of Applications to Food Risk Analysis

Bayesian "'Updating"': a basic parametric case

- A data set $\mathbf{x} = (x_1, \ldots, x_n)$
- A parametric model assumption (A "'Must Have"')

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

- Choose a prior $\pi(\theta)$ for the parameter θ (conjugate)
- Obtain the posterior distribution of θ given $\mathbf{x}$

$$\pi(\theta|\mathbf{x}) \propto \prod_{i=1}^{n} f(x_i|\theta)\pi(\theta) = \frac{\prod_{i=1}^{n} f(x_i|\theta)\pi(\theta)}{\int \prod_{i=1}^{n} f(x_i|\theta)\pi(\theta)d\theta}$$

- Compute Posterior Maximum, Mean, Median,... from $\pi(\theta|\mathbf{x})$
- Do prediction for a new t given the data using

$$k(t|\mathbf{x}) = \int f(t|\theta)\pi(\theta|\mathbf{x})d\theta$$

- MCMC techniques

Bayesian Nonparametrics & Dirichlet Process Prior

- $X_i \sim_{\text{iid}} F$
- "*The Dirichlet family of distributions is a conjugate family for Multinomial models*"

$$\text{Model} \propto \prod_j p_j^{n_j}, \text{ Prior} \propto \prod_j p_j^{\alpha_j - 1}$$

$$\text{Posterior} \propto \prod_j p_j^{n_j + \alpha_j - 1}$$

- Natural choice for the prior $\pi(F)$: the Dirichlet process $F \sim \mathcal{D}(dF|\alpha)$

Model based clustering

THE DATA

- n objects with measurements $\mathbf{x} = (x_1, \ldots, x_n)$

THE MODEL

- If $\mathbf{p} = \{C_1, \ldots, C_{n(\mathbf{p})}\}$ is a partition of $\{1, \ldots, n\}$, $\mathbf{p} \in \mathcal{P}_n$
- A 'classification likelihood'

$$f(\mathbf{x} \mid \mathbf{p}) = \prod_{j=1}^{n(\mathbf{p})} k(x_i, i \in C_j),$$

where $k(x_i, i \in \{1, \ldots, n\}) = k(\mathbf{x})$ is the joint density of the data $\mathbf{x}$

Example

$k(\mathbf{x}) = \int f(\mathbf{x} \mid \theta)\, \pi(\theta)(d\theta)$, *where* θ *is an unknown parameter and* $\pi(d\theta)$ *is the prior distribution for this parameter.*

PRIOR CHOICE

The conjugate idea: model density viewed as a density in $\mathbf{p}$

$$f(\mathbf{x} \mid \mathbf{p}) = \prod_{j=1}^{n(\mathbf{p})} k(x_i, i \in C_j)$$

$$\pi(\mathbf{p}) = \prod_{j=1}^{n(\mathbf{p})} g(C_j)$$

POSTERIOR

$$\pi(\mathbf{p}|\mathbf{x}) \propto \prod_{j=1}^{n(\mathbf{p})} g^*(C_j) \text{ with } g^*(C_j) = g(C_j) \times k(x_i, i \in C_j)$$

Model based clustering

OPTIMAL CLUSTERING

Posterior Mode: $\mathbf{p}^* = \max_{\mathbf{p} \in \mathcal{P}_n} \pi(\mathbf{p}|\mathbf{x})$

Number of clusters: $n(\mathbf{p}^*)$

- PROBLEM: the set $\mathcal{P}_n$ is finite but large
- A SOLUTION:
  1. Simulate $\mathbf{p}_1, ..., \mathbf{p}_M$ from $\pi(\mathbf{p}|\mathbf{x})$
  2. Choose the $\mathbf{p}_k$ with larger $\pi(\mathbf{p}_k|\mathbf{x})$ for $\mathbf{p}^*$

$\Rightarrow$ Need a procedure to simulate from a distribution proportional to $\prod_{j=1}^{n(\mathbf{p})} g(C_j)$, i.e. which has "'the product form"'

Chinese Restaurant Process

- A natural candidate for prior distribution of partition
- n customers seat in a Chinese Restaurant one after the other
  - Cust. 1 seats at an empty table
  - Cust. i seat at occupied table j with probability $\propto e_j$, number of cust. already seated at table j, or at an empty table with probability $\propto e_0$

Definition (Chinese Restaurant (CR) Process with parameter $e_0 > 0$)

A CR(e_0) Process has a probability mass function of the "'product form"'

$$\pi(\mathbf{p}) \propto \prod_{j=1}^{n(\mathbf{p})} [e_0(e_j - 1)!] = \frac{\Gamma(e_0)}{\Gamma(e_0 + n)} \prod_{j=1}^{n(\mathbf{p})} [e_0(e_j - 1)!]$$

Sequential Seating Procedure

$$\text{Simulation from } \pi(\mathbf{p}) \propto \prod_{j=1}^{n(\mathbf{p})} g(C_j)$$

- Cust. 1 seats at an empty table C_0 with probability 1 (conv: $g(C_0) = 1$)
- Suppose i are seated, $\Rightarrow \mathbf{p}_{[i]} = \left\{ C_1^{\mathbf{p}_{[i]}}, \ldots, C_{n\left(\mathbf{p}_{[i]}\right)}^{\mathbf{p}_{[i]}} \right\}$, then for $j = 0, 1, \ldots, n\left(\mathbf{p}_{[i]}\right)$, Cust. $i+1$ seat table $C_j^{\mathbf{p}_{[i]}}$ with probability proportional to

$$g\left(\{i+1\} | C_j^{\mathbf{p}_{[i]}}\right) = \frac{g\left(\{i+1\} \cup C_j^{\mathbf{p}_{[i]}}\right)}{g\left(C_j^{\mathbf{p}_{[i]}}\right)} \; (= e_j \text{ in the } CR(e_0) \text{ case})$$

- The process goes on until the n cust. are seated

A good alternative to the Sequential Seating Procedure

The Sequential Seating Procedure is exact sampling (no need for "‘warm up"’)from

$$q(\mathbf{p}\,|g) = \frac{\prod_{j=1}^{n(\mathbf{p})} g(C_j)}{\prod_{i=1}^{n} \lambda\left(\mathbf{p}_{[i-1]}\right)}, \tag{1}$$

where $\lambda\left(\mathbf{p}_{[i-1]}\right) = \sum_{j=0}^{n\left(\mathbf{p}_{[i-1]}\right)} g\left(\{i\} \mid C_j^{\mathbf{p}_{[i-1]}}\right)$

This is an approximation of

$$\pi(\mathbf{p}|g) = \frac{\prod_{j=1}^{n(\mathbf{p})} g(C_j)}{\sum_{\mathbf{p}} \prod_{j=1}^{n(\mathbf{p})} g(C_j)}$$

Difference between the two: the denominators!
$\rightarrow$ The order counts in the Sequential Seating Procedure.

$\Rightarrow$ Propose a Gibbs approximation (warm up needed but seating order does not count)

Gibbs Sampler for partitions

Choose an initial partition $\mathbf{p}_0$ (the one with n clusters $\mathbf{p}_0 = \{\{1\}, ..., \{n\}\}$ is better)
Then repeat M times the following Gibbs cycle
For $i = 1, ...n$, do

- Remove $\{i\}$ from the current partition of $\{1, ..., n\}$ to get a partition of $\{1, ..., i-1, i+1, ..., n\}$ ($n-1$ elements)
- Reseat $\{i\}$ according to the seating probabilities

$$g\left(\{i\} \,|\, C_j^{\mathbf{p}_{[n-1]}}\right)$$

to get a new partition of $\{1, ..., n\}$.

Gibbs Weighted Chinese Restaurant Process

Using the Chinese Restaurant process as a prior, we obtain seating probabilities prop. to

$$g\left(\{i+1\}|C_j^{\mathbf{p}[i]}\right) = \frac{g^*\left(\{i+1\}\cup C_j^{\mathbf{p}[i]}\right)}{g^*\left(C_j^{\mathbf{p}[i]}\right)} = e_j \times k(x_{i+1}|x_q, q \in C_j^{\mathbf{p}[i]})$$

where $k(x_{i+1}|x_q, q \in C_j^{\mathbf{p}[i]})$ is the conditional (predictive) density of x_t given the $x_q, q \in C_j^{\mathbf{p}[i]}$ and "‘weights"’ the Chinese Restaurant Process.

Gibbs Weighted Chinese Restaurant process

Example (The Gaussian Case)

The classication likelihood $k(\mathbf{x})$ *is obtained as a mixture of Gaussians with mixing distribution being the Gamma-Normal* $(\alpha, 1/\beta, m, 1/v)$ *prior*

$$k(\mathbf{x}) = \int \phi(\mathbf{x}|\mu,\tau)\,\pi(\mu,\tau)(d\mu,d\tau)$$

The seating probabilities in case of a $CR(e_0)$ *prior is*

$$g\left(\{i+1\}\,|C_j^{\mathbf{p}_{[i]}}\right) = e_j \times k(x_{i+1}|x_q, q \in C_j^{\mathbf{p}_{[i]}})$$

where $k(x_{i+1}|x_q, q \in C_j^{\mathbf{p}_{[i]}})$ *is a* t*-distribution with df* $2\alpha^*$*, location* m^* *and precision* $\alpha^*/\beta^* \times \frac{v^*}{v^*+1}$*, with*

$$\alpha^* = \alpha + n/2, v^* = v + n, m^* = \frac{vm + n\bar{x}}{v+n}, \beta^* = \beta + \frac{ns^2}{2} + (m-\bar{x})^2\frac{nv}{v+n}$$

Extension: Mixture Methods, Estimation of the Mixing Distribution

$$f(x \mid G) = \int k(x \mid u) G(du),$$

- (Feller-Shepp) Any decreasing density on $[0, \infty)$ can be represented as a scale mixture of $U(0, u)$ densities
- Any symmetric unimodal density on $(-\infty, \infty)$ can be represented as a mixture of $U(-u, u)$ densities
- (Bernstein) Any "completely monotone" density on $[0, \infty)$ can be represented as a scale mixture of exponentials
- (Bickel, Diaconis, Freedman) Scale mixture of Normal $N(0, 1/\tau)$
- Any density can be approximated by a location-scale mixture of Normal $N(\mu, 1/\tau)$

Extension: Mixture Methods, Estimation of the Mixing Distribution

- Model: $f(\mathbf{x} \mid G) = \prod_{i=1}^{n} \int k(x_i \mid u_i) G(du_i)$
- Prior: $G \sim \mathcal{D}(dG \mid \alpha)$
- Goal: evaluate quantities

$$\mathbb{E}\left(h(G) \mid \mathbf{x}\right) = \frac{\int h(G) f(\mathbf{x} \mid G) \mathcal{D}(dG \mid \alpha)}{\int f(\mathbf{x} \mid G) \mathcal{D}(dG \mid \alpha)}$$

- Examples: $h(G) = G(t)$, $h(G) = \int k(t \mid u) G(du)$

infinite dimensional

Two key results

The Fubini theorem

For any nonnegative function $h(x, F)$

$$\int \int h(x,F)F(dx)\mathcal{D}(dF \mid \alpha) = \int \int h(x,F)\mathcal{D}(dF \mid \alpha + \delta_x)\frac{\alpha(dx)}{\alpha(\infty)}$$

A combinatorics result

For any nonnegative functions $g_1, \ldots, g_n$

$$\int \cdots \int \prod_{i=1}^{n} g_i(u_i) \left(\alpha + \sum_{j=1}^{i-1} \delta_{u_j} \right) (du_i) = \sum_{\mathbf{p}} \Phi\left(\mathbf{p} \mid n, \mathbf{g}\right),$$

where $\mathbf{g} = (g_1, .., g_n)$ and

$$\Phi\left(\mathbf{p} \mid n, \mathbf{g}\right) = \prod_{j=1}^{n(\mathbf{p})} (e_j - 1)! \int \prod_{i \in C_j} g_i(u)\alpha(du)$$

Computation of $\mathbb{E}\left(h(G) \mid \mathbf{x}\right)$

- Apply Fubini n times

$$\int \cdots \mathcal{D}(dG \mid \alpha) \rightarrow \int_{\mathbb{R}^n} \cdots d\mathbf{u}$$

- Define $h(\mathbf{u}) := \int h(G)\mathcal{D}\left(dG \mid \alpha + \sum_{i=1}^{n} \delta_{u_i}\right)$
- Apply the combinatorics result

$$\int_{\mathbb{R}^n} \cdots d\mathbf{u} \rightarrow \sum_{\mathbf{p}} \cdots$$

$$\mathbb{E}\left(h(G) \mid \mathbf{x}\right) = \sum_{\mathbf{p}} w\left(\mathbf{p}\right) \mathbb{E}\left[h(\mathbf{u}) \mid \mathbf{p}\right]$$

$$w(\mathbf{p}) = \prod_{j=1}^{n(\mathbf{p})} (e_j - 1)! \int \prod_{i \in C_j} k(x_i \mid u)\alpha(du),$$

$\mathbf{u} \mid \mathbf{p}$ is the product of $\mathbf{u} \mid \mathbf{u}^*, \mathbf{p}$ and $\mathbf{u}^* \mid \mathbf{p}$

Computation of $\mathbb{E}\left(h(G) \mid \mathbf{x}\right)$

- For $j = 1, ..., n(\mathbf{p})$, u_j^* are iid $\pi(du \mid C_j)$, with

$$\pi(du \mid C_j) \propto \prod_{i \in C_j} k(x_i \mid u)\alpha(du)$$

- For $j = 1, ..., n(\mathbf{p})$, $u_i = u_j^*$ if $i \in C_j$

$$\mathbb{E}\left(h(G) \mid \mathbf{x}\right) = \sum_{\mathbf{p}} w\left(\mathbf{p}\right) \mathbb{E}\left[h(\mathbf{u}) \mid \mathbf{p}\right]$$

Ideas of Applications/Extensions related to Food Risk Analysis

Obs: one exposure per individual

- A 'classification likelihood' chosen as a mixture of Pareto distribution with mixing distribution being the classical conjugate prior (Gamma-Pareto)
- Characterization of a general mixture of Pareto (All densities in Fréchet domain?)

$$f(x \mid G) = \int k(x \mid u) G(du) \text{ with } k(. \mid u) = \text{ Pareto Density}$$

- A covariate dependent prior for partition

Obs: a distribution of exposure (m obs.) per individual

- Pb: write the model!

Reference

- HKUST-ISMT Lecture notes of "Bayesian Statistical Inference"
- Lo (1984) Ann. Stat. Vol. 12 for Density Estimates
- For extensions:
  - Lo & Weng (1989) Annals of the Institute of Statistical Mathematics, Vol 41 for Hazard rate Estimates
  - James (2002) Poisson Process Partition Calculus with applications to Exchangeable models and Bayesian Nonparametrics (arXiv) or James (2005) Ann. Stat.

(Additional tools: Fubini results for Gamma processes or Poisson process + Laplace functionals)