Advanced Bayesian Learning

Lecture 4 - Dirichlet Process Mixtures

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

Finite mixture models

- **Mixture of normals**

$$p(y) = \sum_{j=1}^{k} \pi_j \cdot \phi(y; \mu_j, \sigma_j^2)$$

- **Allocation variables**: $I_i = j$ if y_i comes from $\phi(y; \mu_j, \sigma_j^2)$.
- Let $I = (I_1, ..., I_n)$ and $n_j = \sum_{i=1}^{n} (I_i = j)$.
- **Gibbs sampling**:
  - $\pi_1, ..., \pi_k \mid I, y \sim \text{Dirichlet}(a_1 + n_1, a_2 + n_2, ..., a_k + n_k)$
  - $\sigma_j^2 \mid I, y \sim \text{Inv-}\chi^2$ and $\mu_j | I, \sigma_j^2, y \sim \text{Normal}$ for $j = 1, ..., k$.
  - $I_i \mid \pi, \mu, \sigma^2, y \sim \text{Categorical}(\omega_{i1}, ..., \omega_{ik})$, $i = 1, ..., n$,

$$\omega_{ij} = \frac{\pi_j \cdot \phi(y_i; \mu_j, \sigma_j^2)}{\sum_{q=1}^{k} \pi_q \cdot \phi(y_i; \mu_q, \sigma_q^2)}.$$

Infinite mixture models - DP mixtures

- **General mixture**

$$f(y|P) = \int \mathcal{K}(y|\theta)dP(\theta)$$

where $\mathcal{K}(y|\theta)$ is a **kernel** and $P(\theta)$ is a **mixing measure**.

- **Student-t**: $y \sim t_\nu(\mu, \sigma^2)$.
  - $\mathcal{K}(y|\theta) = \phi(y|\mu, \lambda)$ where μ is fixed, $\theta = \lambda$ and
  - $P(\theta)$ is the $Inv - \chi^2(\nu, \sigma^2)$ distribution.
- **Finite mixture of normals**: $p(y) = \sum_{j=1}^{k} \pi_j \cdot \phi(y; \mu_j, \sigma_j^2)$.
  - $\mathcal{K}(y|\theta) = \phi(y|\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$.
  - $P(\theta)$ is discrete with $\Pr\{\theta = (\mu_j, \sigma_j^2)\} = \pi_j$, for $j = 1, ..., k$.
- Dirichlet Process Mixture: $P \sim DP(\alpha P_0)$, **infinite mixture**

$$f(y) = \sum_{h=1}^{\infty} \pi_h \mathcal{K}(y|\theta_h^*), \qquad \pi \sim \mathrm{Stick}(\alpha) \text{ and } \theta_h \overset{iid}{\sim} P_0$$

DP mixture is like a finite mixture with large k

- Infinite mixture: **every observation has its own parameter**

$$y_i \sim \mathcal{K}(\theta_i)$$

- DP is a.s. discrete $\Rightarrow$ **ties**: some θ_i exactly the same value.
- DP implies **clustering** of the θ_i
- Each observation has **potentially** its own parameter θ_i, but that **parameter may be shared by other observations**.
- **Finite mixtures**: observations share k parameter values.

$$y_i | I_i \sim \mathcal{K}(\theta_{I_i})$$

$$I_i | \pi \sim \text{Categorical}(\pi_1, ..., \pi_k)$$

$$\theta_1, \ldots, \theta_k \overset{iid}{\sim} P_0$$

$$\pi \sim \text{Dirichlet}(\alpha/k, ..., \alpha/k)$$

- Finite mixture approaches DP mixture as $k \to \infty$ [Neal2000].

Marginalizing out P from a DP - Polya scheme

- Hierarchical representation of DP mixtures

$$y_i \sim \mathcal{K}(\theta_i), \qquad \theta_i \sim P \qquad P \sim DP(\alpha P_0)$$

- We can marginalize out P to obtain the Polya scheme

$$p(\theta_i|\theta_1, ..., \theta_{i-1}) \sim \left(\frac{\alpha}{\alpha + i - 1}\right) P_0(\theta_i) + \left(\frac{1}{\alpha + i - 1}\right) \sum_{j=1}^{i-1} \delta_{\theta_j}$$

- So $p(\theta_i|\theta_1, ..., \theta_{i-1})$ is a mixture of the base measure P_0 and point masses at the previously "drawn" θ-values.

- 'Marginalizing out P': integrate out π in the finite mixture model and let $k \to \infty$ [Neal2000].

DPs and the Chinese restaurant process

- **Polya scheme**:

$$p(\theta_i|\theta_1, \ldots, \theta_{i-1}) \sim \left(\frac{\alpha}{\alpha + i - 1}\right) P_0(\theta_i) + \left(\frac{1}{\alpha + i - 1}\right) \sum_{j=1}^{i-1} \delta_{\theta_j}$$

- **Chinese restaurant process**:
  - first customer sits at empty table and gets the dish θ_1^* from P_0.
  - second customer has options:
    - sit at first customer's table with probability $\frac{1}{1+\alpha}$ and eat θ_1^*
    - sit at a new table with probability $\frac{\alpha}{1+\alpha}$ and eat $\theta_2^* \sim P_0$.

    $\vdots$
  - the ith customer has options:
    - sit at table with dish θ_j^* with a probability proportional to n_j, the number of customers sitting at table j
    - sit at a new table with probability proportional to α.

Gibbs sampling DP mixtures - marginalizing P

- Similar to Gibbs sampling for finite mixtures. Data augmentation with mixture component indicators I_i.

1. **Update component allocation** for ith observation y_i by sampling from multinomial

$$\Pr(I_i = j|\cdot) \propto \begin{cases} n_j^{(-i)} \mathcal{K}(y_i|\theta_j^*) & \text{for } j = 1, \ldots, k^{(-i)} \\ \alpha \int \mathcal{K}(y_i|\theta) dP_0(\theta) & \text{for } j = k^{(-i)} + 1 \end{cases}.$$

2. **Update** the unique **parameter values** θ^* by sampling from

$$p(\theta_j^*|\cdot) \propto P_0(\theta_c^*) \prod_{i: I_i = j} \mathcal{K}(y_i|\theta_j^*)$$

- Note that, unlike finite mixtures, the I_i **are not independent** conditional on θ^*. This because we have marginalized out P. They have to be sampled **sequentially**.

Gibbs sampling for truncated DP mixtures

- Set upper bound N for the number of components. Approximate DP mixture with $\pi_h = 0$ for $h = N+1, \ldots$
- Posterior samping for infinite mixtures is now very similar to finite mixture. The I_i can be sampled independently.

1. Update component allocation for ith observation y_i by sampling from multinomial

$$\Pr(I_i = j|\cdot) \propto \pi_j \mathcal{K}(y_i|\theta_j^*) \quad \text{for } j = 1, 2, \ldots, N.$$

2. Update the stick-breaking weights $[\pi_h = V_h \prod_{\ell<h}(1 - V_\ell)]$

$$V_j|\cdot \sim Beta\left(1 + n_j, \alpha + \sum_{q=j+1}^{N} n_q\right) \qquad \text{for } j = 1, \ldots, N-1.$$

3. Update $\theta_1^*, \ldots \theta_N^*$ by sampling like in the finite mixture model. Sample θ^* from prior $P_0(\theta)$ for empty clusters.

MCMC for DP mixtures

- Let's look at the updating step:

$$\Pr(I_i = j|\cdot) \propto \begin{cases} n_j^{(-i)} \mathcal{K}(y_i|\theta_c^*) & \text{for } j = 1, ..., k^{(-i)} \\ \alpha \int \mathcal{K}(y_i|\theta) dP_0(\theta) & \text{for } j = k^{(-i)} + 1 \end{cases}.$$

- **A customer chooses table based on**:
  - the number of existing customers at the tables (with imaginary α customers at a new table)
  - how compatible the taste of the customer (y_i) is to the different dishes served at occupied tables (θ_c^*)
  - how compatible the taste of the customer (y_i) is to the different dishes that *may* be served at a new table.
  - A $P_0(\theta)$ with large variance is equivalent to an very experimental cook. You never know what you get ...
- α matters for the number of clusters (tables), but so does P_0.
- α can be learned from data. Just add updating step.
- P_0 may have hyperparameters (e.g. $P_0 = N(\mu, \sigma^2)$). Just add updating steps for those.

Mixture of multivariate regressions - Model

- The response vector $\mathbf{y}$ is p-dim. Covariates $\mathbf{x}$ is q-dim.
- The model is of the form

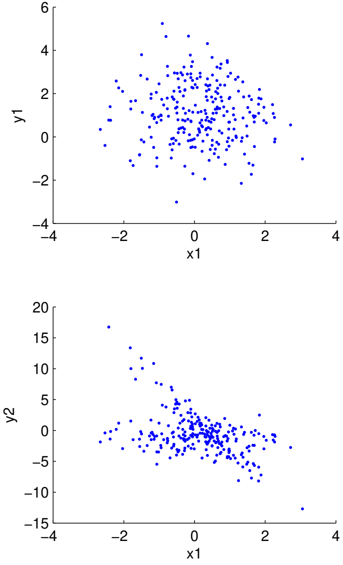
$$p(\mathbf{y}|\mathbf{x}) = \sum_{j=1}^{\infty} \pi_j \cdot N(\mathbf{y}_i|\mathbf{B}_j\mathbf{x}_i, \Sigma_j)$$

- Each mixture component is a Gaussian multivariate regression with its own regression coefficient and covariance matrix:

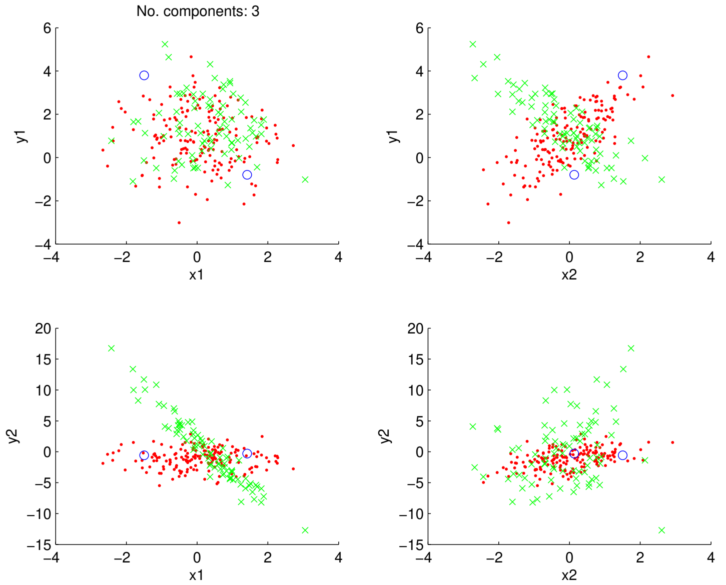
$$\underset{p\times 1}{\mathbf{y}_i} = \underset{p\times q}{\mathbf{B}_j} \underset{q\times 1}{\mathbf{x}_i} + \underset{p\times 1}{\varepsilon_i}, \; \varepsilon_i \overset{iid}{\sim} N\left(0, \Sigma_j\right)$$

- The mixture weights follow a DP stick prior $\pi \sim Stick(\alpha)$.

Mixture of multivariate regressions - Data



Mixture of multivariate regressions - DPM



Mixture of multivariate regressions - DPM