Markov Chain Sampling Methods for Dirichlet Process Mixture Models

K component mixture model

$$\pi \sim \text{Dirichlet}(\alpha/K, ..., \alpha/K) \quad (1)$$
$$z_i|\pi \sim \text{Categorical}(\pi) \quad (2)$$
$$\theta_k|H \sim H \quad (3)$$
$$y_i|z_i, \theta \sim F(\theta_{z_i}), \quad (4)$$

$\alpha > 0$ and H is the prior over the parameters $\theta_k \in \Theta$.

K component mixture model

When K is known, we have seen that EM-algorithm can be applied to estimate the parameters.

But in many settings, K is unknown and we need to experiment with different values of K.

Applications

- Topic modeling: organize/label a corpus of documents into K topics. How do you choose K?

Applications

- Topic modeling: organize/label a corpus of documents into K topics. How do you choose K?
- Cancer clonal detection: given an admixture of cells, detect the number of cancer subpopulations.

Applications

- Topic modeling: organize/label a corpus of documents into K topics. How do you choose K?
- Cancer clonal detection: given an admixture of cells, detect the number of cancer subpopulations.
- Density estimation: modeling multi-modal distribution with unknown components.

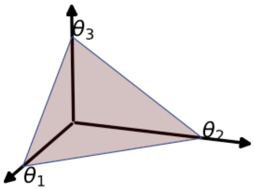
Dirichlet distribution

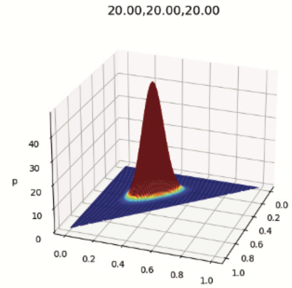
- "Distribution" of (discrete) distributions over K categories.

$$\pi \sim \text{Dirichlet}(\alpha)$$

- $\pi_k \in [0, 1]$.
- $\sum_{k=1}^{K} \pi_k = 1$.

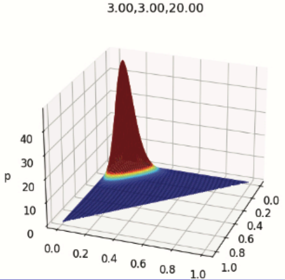
Dirichlet distribution

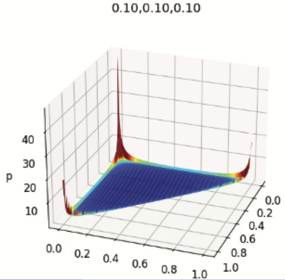




(a)

(b)





Dirichlet Process

When K is known, we use Dirichlet distribution.

When K is not known, we use a Dirichlet Process to place a prior over distributions (or, an unbounded mixture). Let's see how that works and what is a distribution over distributions?

Dirichlet Process

Distribution over infinite-dimensional discrete probability measures:

$$G \sim \mathsf{DP}(H, \alpha),$$

where H is the base measure defined on Θ and $\alpha > 0$ is the concentration parameters.

- G is a random probability measure defined on Θ.

Dirichlet Process

G is Dirichlet process distributed with base distribution H and concentration parameter α, if and only if

$$(G(A_1), ..., G(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), ..., \alpha H(A_K))$$

for every finite measurable partition $A_1, ..., A_K$ of Θ.

Note: $G(A_k)$ is a random variable because G is a random measure.

Dirichlet Process

- $\mathbb{E}[G(A)] = H(A)$.
- $\mathrm{var}(G(A)) = H(A)(1 - H(A))/(\alpha + 1)$.

Larger the value of α, the smaller the variance (concentrated around the mean $H(A)$).

A measure G sampled from DP is discrete with probability 1.

Posterior distribution of G

Since G is a distribution, we can draw samples from G.

Let $\theta_i \sim G$.

Given $\theta_1, ..., \theta_N$, what is the posterior distribution $p(G|\theta_{1:N})$?

Posterior distribution of G

Let $n_k = |\{i : \theta_i \in A_k\}|$, the number of points that fall in A_k.

- $(G(A_1), ..., G(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), ..., \alpha H(A_K))$
- $(n_1, ..., n_K) \sim \text{Multinomial}(G(A_1), ..., G(A_K))$
- Dirichlet and Multinomial are conjugate distributions:

$$(G(A_1), ..., G(A_K))|\theta_1, ..., \theta_N \sim \text{Dirichlet}(\alpha'_k),$$

where

$$\alpha'_k = \alpha H(A_k) + n_k.$$

Posterior over G

$G \sim DP(\alpha, H)$ if and only if for disjoint partition $A_1, ..., A_K$ of Θ such that,

$$(G(A_1), ..., G(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), ..., \alpha H(A_K)).$$

Since the result from previous slide holds for arbitrary partition $(A_1, ..., A_K)$, the posterior distribution $G|\theta_{1:N}$ is also a Dirichlet Process.

Therefore, DP provides a conjugate family of priors over (discrete) probability distributions.

Posterior over G

How do we update the hyperparameters?

$$G|\theta_1, ..., \theta_N \sim DP(\alpha', H')$$

$\alpha' = \alpha + N$ and

$$H' = \frac{\alpha H + \sum_{i=1}^{N} \delta_{\theta_i}}{\alpha + N},$$

weighted measure between the base measure H and empirical measure $\delta = \sum \delta_{\theta_i}$.

Why?

Posterior over G

We know that,

$$(G(A_1), ..., G(A_K))|\theta_{1:N} \sim \text{Dirichlet}(\alpha H(A_k) + n_k),$$

which has density

$$\prod_{k=1}^{K} G(A_k)^{\alpha H(A_k)+n_k-1}.$$

This implies that $\alpha' = \sum_{k=1}^{K}(\alpha H(A_k) + n_k) = \alpha \cdot 1 + N$.

$$\alpha' H'(A_k) = \alpha H(A_k) + n_k \Rightarrow H'(A_k) = \frac{\alpha H(A_k) + n_k}{\alpha + N}.$$

Note: $n_k = \sum_{i=1}^{N} \delta_{\theta_i}(A_k)$.

Stick breaking process

Does such stochastic process exist? Yes, Sethuraman's stick breaking construction.

Let u be a unit stick (length 1). We will break this stick infinite number of times.

For $i = 1, ..., \infty$,

- Sample $\beta_i \sim \text{Beta}(1, \alpha)$
- Set $\pi_i = \beta_i \prod_{n=1}^{i-1} (1 - \beta_i)$; $\pi_1 = \beta_1$.
- Sample $\theta_i \sim H$.

$$G = \sum_i \pi_i \delta_{\theta_i}$$

is a realization from $DP(\alpha, H)$

[Simulate this process N times for different values of α]

Dirichlet process mixture model

Generative model.

Dirichlet process mixture model

Generative model.

We first sample a random measure: $G \sim \mathsf{DP}(\alpha, H)$.

Dirichlet process mixture model

Generative model.

We first sample a random measure: $G \sim \mathsf{DP}(\alpha, H)$.

We then sample parameters for each datum i: $\theta_i | G \sim G$.

Dirichlet process mixture model

Generative model.

We first sample a random measure: $G \sim \mathsf{DP}(\alpha, H)$.

We then sample parameters for each datum i: $\theta_i | G \sim G$.

Finally, sample the datum: $y_i | \theta_i \sim F(\theta_i)$.

Dirichlet process mixture model

Generative model.

We first sample a random measure: $G \sim \text{DP}(\alpha, H)$.

We then sample parameters for each datum i: $\theta_i | G \sim G$.

Finally, sample the datum: $y_i | \theta_i \sim F(\theta_i)$.

How does this model solve the clustering problem with unknown K?

Dirichlet process mixture model

To simplify discussion,

- Let H be Normal distribution defined on Θ where $\Theta = (\mathbb{R}, \mathbb{R}^+)$ (parameter space of location and scale).

Dirichlet process mixture model

To simplify discussion,

- Let H be Normal distribution defined on Θ where $\Theta = (\mathbb{R}, \mathbb{R}^+)$ (parameter space of location and scale).
- Let F also be Normal distribution with $\theta_i = (\mu_i, \sigma_i^2) \in \Theta$.

Dirichlet process mixture model

To simplify discussion,

- Let H be Normal distribution defined on Θ where $\Theta = (\mathbb{R}, \mathbb{R}^+)$ (parameter space of location and scale).
- Let F also be Normal distribution with $\theta_i = (\mu_i, \sigma_i^2) \in \Theta$.

G is a random distribution of infinite dimension $(K = 1, 2, 3, ...)$.

Dirichlet process mixture model

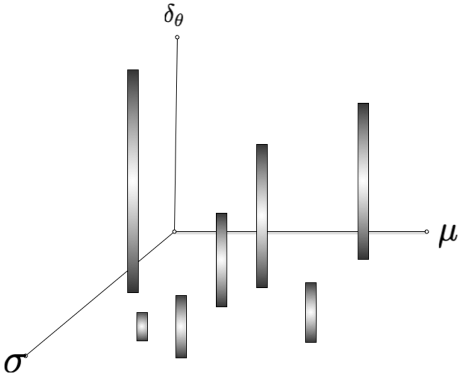
To simplify discussion,

- Let H be Normal distribution defined on Θ where $\Theta = (\mathbb{R}, \mathbb{R}^+)$ (parameter space of location and scale).
- Let F also be Normal distribution with $\theta_i = (\mu_i, \sigma_i^2) \in \Theta$.

G is a random distribution of infinite dimension $(K = 1, 2, 3, ...)$.

G is discrete with probability $1 \Rightarrow$ some θ_i will be repeated $\Rightarrow$ clustering but with undetermined dimension K.

Dirichlet process mixture model



Each bar represents a unique θ_k and the length indicates $\sum_i 1[\theta_i = \theta_k]$.

Chinese Restaurant Process

How do we do posterior inference? We need a prior over partitions.

- Consider a Chinese restaurant with infinitely many tables

Chinese Restaurant Process

How do we do posterior inference? We need a prior over partitions.

- Consider a Chinese restaurant with infinitely many tables
- First customer enters the restaurant and chooses a table to seat and selects a dish $\theta \sim H$.

Chinese Restaurant Process

How do we do posterior inference? We need a prior over partitions.

- Consider a Chinese restaurant with infinitely many tables
- First customer enters the restaurant and chooses a table to seat and selects a dish $\theta \sim H$.
- Each subsequent customer enters the restaurant, selects one of K tables based on probability:

Chinese Restaurant Process

How do we do posterior inference? We need a prior over partitions.

- Consider a Chinese restaurant with infinitely many tables
- First customer enters the restaurant and chooses a table to seat and selects a dish $\theta \sim H$.
- Each subsequent customer enters the restaurant, selects one of K tables based on probability:

$$p(z_i = k | z_{1:i-1}) = \frac{n_k}{i - 1 + \alpha}$$

and shares the dish θ_k or seat on a new table

$$p(z_i = k' | z_{1:i-1}) = \frac{\alpha}{i - 1 + \alpha}$$

and sample a new dish $\theta \sim H$.

Chinese Restaurant Process

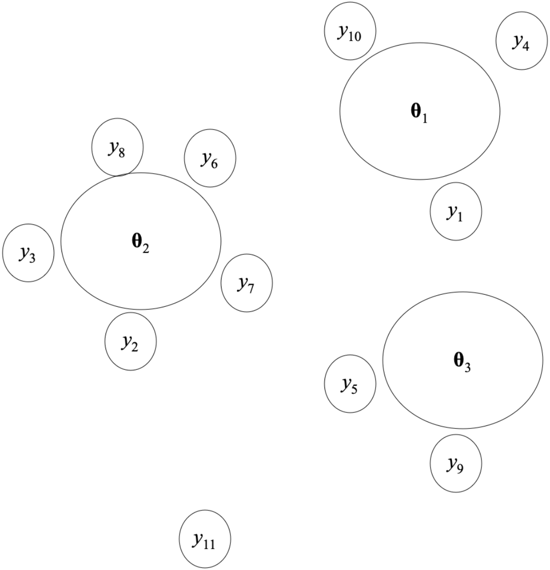


Figure 2: Compute $P(z_{11} = k | z_{1:10})$

Clustering

CRP is a prior over partition. To cluster data, we need:

$$p(z_i = k|y_i, y_{-i}, z_{-i}) \propto p(y_i|z_i = k, y_{-i}, z_{-i})p(z_i = k|z_{-i}) \qquad (5)$$

Clustering

CRP is a prior over partition. To cluster data, we need:

$$p(z_i = k|y_i, y_{-i}, z_{-i}) \propto p(y_i|z_i = k, y_{-i}, z_{-i})p(z_i = k|z_{-i}) \qquad (5)$$

- $p(y_i|z_i, y_{-i}, z_{-i})$: likelihood.

Clustering

CRP is a prior over partition. To cluster data, we need:

$$p(z_i = k|y_i, y_{-i}, z_{-i}) \propto p(y_i|z_i = k, y_{-i}, z_{-i})p(z_i = k|z_{-i}) \tag{5}$$

- $p(y_i|z_i, y_{-i}, z_{-i})$: likelihood.
- $p(z_i = k|z_{-i})$: CRP prior over assignment probability.

Clustering

CRP is a prior over partition. To cluster data, we need:

$$p(z_i = k|y_i, y_{-i}, z_{-i}) \propto p(y_i|z_i = k, y_{-i}, z_{-i})p(z_i = k|z_{-i}) \quad (5)$$

- $p(y_i|z_i, y_{-i}, z_{-i})$: likelihood.
- $p(z_i = k|z_{-i})$: CRP prior over assignment probability.

Perform the above step sequentially for datum $i = 1, ..., N$.

Clustering

CRP is a prior over partition. To cluster data, we need:

$$p(z_i = k | y_i, y_{-i}, z_{-i}) \propto p(y_i | z_i = k, y_{-i}, z_{-i}) p(z_i = k | z_{-i}) \tag{5}$$

- $p(y_i | z_i, y_{-i}, z_{-i})$: likelihood.
- $p(z_i = k | z_{-i})$: CRP prior over assignment probability.

Perform the above step sequentially for datum $i = 1, ..., N$.

Note: the order in which we assign datum does not matter. Why?

Clustering

CRP is a prior over partition. To cluster data, we need:

$$p(z_i = k|y_i, y_{-i}, z_{-i}) \propto p(y_i|z_i = k, y_{-i}, z_{-i})p(z_i = k|z_{-i}) \tag{5}$$

- $p(y_i|z_i, y_{-i}, z_{-i})$: likelihood.
- $p(z_i = k|z_{-i})$: CRP prior over assignment probability.

Perform the above step sequentially for datum $i = 1, ..., N$.

Note: the order in which we assign datum does not matter. Why?

Exchangeability (De Finetti's theorem).

Gibbs algorithm for DPMM (Algorithm 2)

For $t = 1, ..., T$ (MCMC chain length):

Gibbs algorithm for DPMM (Algorithm 2)

For $t = 1, ..., T$ (MCMC chain length):

1. Assign datum $i = 1, ..., N$ according to Eq~(5).

Gibbs algorithm for DPMM (Algorithm 2)

For $t = 1, ..., T$ (MCMC chain length):

1. Assign datum $i = 1, ..., N$ according to Eq~(5).
2. Sample $\theta_k | \{y_j : z_j = k\}$ for each table k.

Gibbs algorithm for DPMM (Algorithm 2)

For $t = 1, ..., T$ (MCMC chain length):

1. Assign datum $i = 1, ..., N$ according to Eq~(5).
2. Sample $\theta_k | \{y_j : z_j = k\}$ for each table k.

This is known as Metropolis-within-Gibbs. The overall procedure of assigning datum is Gibbs; sampling parameters to explain the data for each table is done using MH.

Note: MH-w-Gibbs preserves maintains detailed balance condition.

Collapsed Gibbs sampling

If H and F are conjugate, then we do not need to explicitly represent θ_k, we can marginalize it out to obtain the predictive distribution:

$$p(y_i|z_i = k, z_{-i}, y_{-i}) = \int F(y_i|\theta')p(\theta'|\{y_j : z_j = k\})d\theta' \qquad (6)$$

where $p(\theta'|\{y_j : z_j = k\})$ represents the posterior distribution of θ_k given the data points assigned to k: $\{y_j : z_j = k\}$.

Example: F and H are Normally distributed, then the posterior is also Normally distributed.

Collapsed Gibbs algorithm for DPMM (Algorithm 3)

For $t = 1, ..., T$ (MCMC chain length):

Collapsed Gibbs algorithm for DPMM (Algorithm 3)

For $t = 1, ..., T$ (MCMC chain length):

- Assign datum $i = 1, ..., N$ according to Eq~(5). Update the posterior distribution $p(\theta_k | \{y_j : z_j = k\})$.

Implementation notes

How do we implement this?

Additional slides

- CRP as predictive distribution.
- Exchangeability.
- De Finetti's theorem.

Predictive distribution (CRP)

The predictive distribution, with G marginalized:

$$p(\theta_{N+1} \in A|\theta_{1:N}) = \mathbb{E}[1[\theta_{N+1} \in A]|\theta_{1:N}].$$

$$\mathbb{E}[1[\theta_{N+1} \in A]|\theta_{1:N}] = \int 1[\theta_{N+1} \in A] p(G|\theta_{1:N}) dG.$$

Since $\theta_{N+1} \sim G$, $\mathbb{E}[1[\theta_{N+1} \in A]|G] = G(A)$. Hence, $p(\theta_{N+1} \in A|\theta_{1:N}) = \mathbb{E}[G(A)|\theta_{1:N}]$.

$$\mathbb{E}[G(A)|\theta_{1:N}] = H'(A) = \frac{\alpha}{\alpha + N} H(A) + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i}(A).$$

Exchangeability

A sequence of random variables $Y_1, ..., Y_N$ is exchangeable if for some permutation σ:

$$p(y_1, ..., y_N) =_d p(y_{\sigma(1)}, ..., y_{\sigma(N)})$$

De Finetti's theorem

Any infinite exchangeable sequence of random variables can be viewed as i.i.d. draws from a latent distribution G.

$$p(y_1, ..., y_N|G) = \prod_i p(y_i|G)$$

De Finetti's theorem

Any infinite exchangeable sequence of random variables can be viewed as i.i.d. draws from a latent distribution G.

$$p(y_1, ..., y_N|G) = \prod_i p(y_i|G)$$

- i.i.d $\Rightarrow$ Exchangeability but reverse is not true.