Bayesian Non-Parametrics

Advanced Machine Learning for NLP

Jordan Boyd-Graber

OVERVIEW

Clustering as Probabilistic Inference

- GMM is a probabilistic model (unlike K-means)
- There are several latent variables:
  - Means
  - Assignments
  - (Variances)

Clustering as Probabilistic Inference

- GMM is a probabilistic model (unlike K-means)
- There are several latent variables:
  - Means
  - Assignments
  - (Variances)
- Corresponds to representation in unbounded space

Nonparametric Clustering

- What if the number of clusters is not fixed?
- Nonparametric: can grow if data need it
- Probabilistic distribution over number of clusters

Dirichlet Process

- Distribution over distributions
- Parameterized by: α, G

Dirichlet Process

- Distribution over distributions
- Parameterized by: α, G
- Concentration parameter

Dirichlet Process

- Distribution over distributions
- Parameterized by: α, G
- Concentration parameter
- Base distribution

Dirichlet Process

- Distribution over distributions
- Parameterized by: α, G
- Concentration parameter
- Base distribution
- You can then draw observations from $x \sim \mathrm{DP}(\alpha, G)$.

Defining a DP

- Break off sticks

$$V_1, V_2, \cdots \sim_{\text{iid}} \text{Beta}(1, \alpha) \tag{1}$$

$$C_k \equiv V_k \prod_{j=1}^{k-1} (1 - V_k) \tag{2}$$

Defining a DP

- Break off sticks

$$V_1, V_2, \cdots \sim_{\text{iid}} \text{Beta}(1, \alpha) \tag{1}$$

$$C_k \equiv V_k \prod_{j=1}^{k-1} (1 - V_k) \tag{2}$$

- Draw atoms

$$\Phi_1, \Phi_2, \cdots \sim_{\text{iid}} G \tag{3}$$

Defining a DP

- Break off sticks

$$V_1, V_2, \dots \sim_{\text{iid}} \text{Beta}(1, \alpha) \tag{1}$$

$$C_k \equiv V_k \prod_{j=1}^{k-1} (1 - V_k) \tag{2}$$

- Draw atoms

$$\Phi_1, \Phi_2, \dots \sim_{\text{iid}} G \tag{3}$$

- Merge into complete distribution

$$\Theta = \sum_k C_k \delta_{\Phi_k} \tag{4}$$

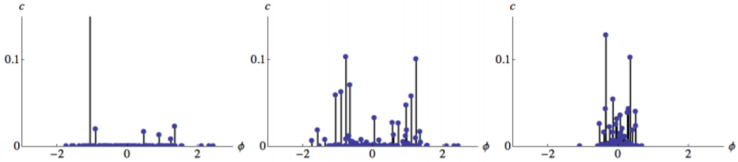
Properties of a DPMM

- Expected value is the same as base distribution

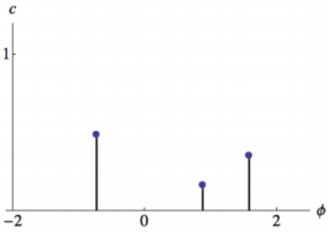
$$\mathbb{E}_{\mathrm{DP}(\alpha,G)}[x] = \mathbb{E}_G[x] \quad (5)$$

- As $\alpha \rightarrow \infty$, $\mathrm{DP}(\alpha, G) = G$
- Number of components unbounded
- Impossible to represent fully on computer (truncation)
- You can nest DPs

Effect of scaling parameter α

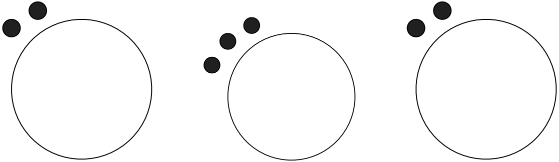


DP as mixture Model



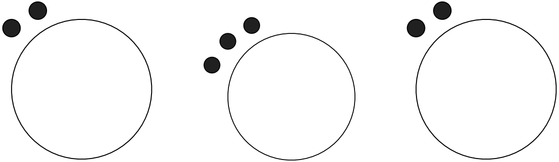
The Chinese Restaurant as a Distribution

To generate an observation, you first sit down at a table. You sit down at a table proportional to the number of people sitting at the table.



The Chinese Restaurant as a Distribution

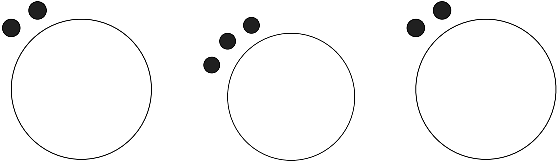
To generate an observation, you first sit down at a table. You sit down at a table proportional to the number of people sitting at the table.



$\frac{2}{7}$ $\frac{3}{7}$ $\frac{2}{7}$

The Chinese Restaurant as a Distribution

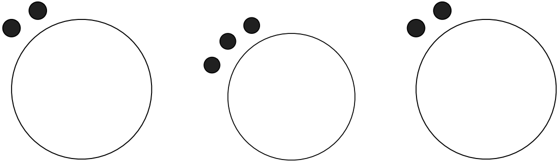
To generate an observation, you first sit down at a table. You sit down at a table proportional to the number of people sitting at the table.



$\frac{2}{7}$ $\frac{3}{7}$ $\frac{2}{7}$

The Chinese Restaurant as a Distribution

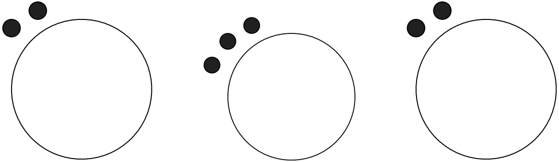
To generate an observation, you first sit down at a table. You sit down at a table proportional to the number of people sitting at the table.



$\frac{2}{7}$	$\frac{3}{7}$	$\frac{2}{7}$
$x \sim \mu_1$	$x \sim \mu_2$	$x \sim \mu_3$

The Chinese Restaurant as a Distribution

To generate an observation, you first sit down at a table. You sit down at a table proportional to the number of people sitting at the table.



$\frac{2}{7}$ $\qquad$ $\frac{3}{7}$ $\qquad$ $\color{red}{\frac{2}{7}}$

$x \sim \mu_1$ $\qquad$ $x \sim \mu_2$ $\qquad$ $x \sim \mu_3$

But this is just Maximum Likelihood

Why are we talking about Chinese Restaurants?

Always can squeeze in one more table . . .

- The *posterior* of a DP is CRP
- A new observation has a new table / cluster with probability proportional to α
- But this must be balanced against the probability of an observation *given a cluster*

$$\Theta = \sum_k C_k \delta_{\Phi_k} \tag{6}$$

Gibbs Sampling

- We want to know the cluster assignment of each observation
- Take a random guess initially

Gibbs Sampling

- We want to know the cluster assignment of each observation
- Take a random guess initially
- This provides a mean for each cluster

Gibbs Sampling

- We want to know the cluster assignment of each observation
- Take a random guess initially
- This provides a mean for each cluster
- Let the number of clusters grow

Gibbs Sampling

- We want to know the cluster assignment of each observation (tables)
- Take a random guess initially
- This provides a mean for each cluster
- Let the number of clusters grow

Gibbs Sampling

- We want to know $\vec{z}$
- Compute $p(z_i \,|\, z_1 \dots z_{i-1}, z_{i+1}, \dots z_m, x, \alpha, G)$
- Update z_i by sampling from that distribution
- Keep going ...

Gibbs Sampling

- We want to know $\vec{z}$
- Compute $p(z_i \,|\, z_1 \dots z_{i-1}, z_{i+1}, \dots z_m, x, \alpha, G)$
- Update z_i by sampling from that distribution
- Keep going ...

Notation

$$p(z_i = k \,|\, z_{-i}) \equiv p(z_i \,|\, z_1 \dots z_{i-1}, z_{i+1}, \dots z_m) \tag{7}$$

Gibbs Sampling for DPMM

$$p(z_i = k \,|\, \vec{z}_{-i}, \vec{x}, \{\theta_k\}, \alpha) \tag{8}$$

$$\tag{9}$$

Gibbs Sampling for DPMM

$$p(z_i = k \,|\, \vec{z}_{-i}, \vec{x}, \{\theta_k\}, \alpha) \tag{8}$$

$$= p(z_i = k \,|\, \vec{z}_{-i}, x_i, \vec{x}, \theta_k, \alpha) \tag{9}$$

$$\tag{10}$$

Dropping irrelevant terms

Gibbs Sampling for DPMM

$$p(z_i = k \,|\, \vec{z}_{-i}, \vec{x}, \{\theta_k\}, \alpha) \tag{8}$$

$$= p(z_i = k \,|\, \vec{z}_{-i}, x_i, \vec{x}, \theta_k, \alpha) \tag{9}$$

$$= p(z_i = k \,|\, \vec{z}_{-i}, \alpha) p(x_i \,|\, \theta_k, \vec{x}) \tag{10}$$

$$\tag{11}$$

Chain rule

Gibbs Sampling for DPMM

$$
\begin{aligned}
& p(z_i = k \,|\, \vec{z}_{-i}, \vec{x}, \{\theta_k\}, \alpha) && (8)\\
=& p(z_i = k \,|\, \vec{z}_{-i}, x_i, \vec{x}, \theta_k, \alpha) && (9)\\
=& p(z_i = k \,|\, \vec{z}_{-i}, \alpha) p(x_i \,|\, \theta_k, \vec{x}) && (10)\\
=& \begin{cases} \left(\frac{n_k}{n_\cdot + \alpha}\right) \int_\theta p(x_i \,|\, \theta) p(\theta \,|\, G, \vec{x}) & \text{existing} \\ \frac{\alpha}{n_\cdot + \alpha} \int_\theta p(x_i \,|\, \theta) p(\theta \,|\, G) & \text{new} \end{cases} && (11)\\
& && (12)
\end{aligned}
$$

Applying CRP

Gibbs Sampling for DPMM

$$
\begin{aligned}
& p(z_i = k \,|\, \vec{z}_{-i}, \vec{x}, \{\theta_k\}, \alpha) && (8) \\
=& p(z_i = k \,|\, \vec{z}_{-i}, x_i, \vec{x}, \theta_k, \alpha) && (9) \\
=& p(z_i = k \,|\, \vec{z}_{-i}, \alpha) p(x_i \,|\, \theta_k, \vec{x}) && (10) \\
=& \begin{cases} \left(\frac{n_k}{n_\cdot + \alpha}\right) \int_\theta p(x_i \,|\, \theta) p(\theta \,|\, G, \vec{x}) & \text{existing} \\ \frac{\alpha}{n_\cdot + \alpha} \int_\theta p(x_i \,|\, \theta) p(\theta \,|\, G) & \text{new} \end{cases} && (11) \\
=& \begin{cases} \left(\frac{n_k}{n_\cdot + \alpha}\right) \mathcal{N}\left(x, \frac{n\bar{x}}{n+1}, \mathbb{1}\right) & \text{existing} \\ \frac{\alpha}{n_\cdot + \alpha} \mathcal{N}(x, 0, \mathbb{1}) & \text{new} \end{cases} && (12)
\end{aligned}
$$

Scary integrals assuming G is normal distribution with mean zero and unit variance. (Derived in optional reading.)

Algorithm for Gibbs Sampling

1. Random initial assignment to clusters
2. For iteration i:
   1. "Unassign" observation n
   2. Choose new cluster for that observation