Dirichlet Processes

Machine Learning: Jordan Boyd-Graber
University of Maryland
INTRODUCTION

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)
- Before, we were doing EM

Clustering as Probabilistic Inference

- GMM is a probabilistic model (unlike K-means)
- There are several latent variables:
  - Means
  - Assignments
  - (Variances)
- Before, we were doing EM
- Today, new models and new methods

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

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

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

Defining a DP

- Break off sticks

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

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

- Draw atoms

$$\Phi_1, \Phi_2, \ldots \sim_{\text{iid}} G$$

Defining a DP

- Break off sticks

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

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

- Draw atoms

$$\Phi_1, \Phi_2, \ldots \sim_{\text{iid}} G$$

- Merge into complete distribution

$$\Theta = \sum_{k \in \mathbb{N}} C_k \delta_{\Phi_k}$$

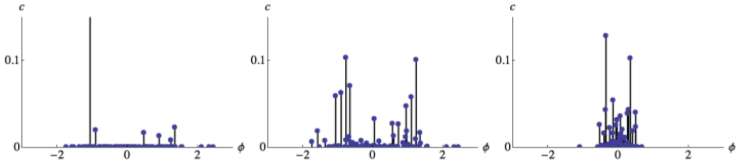
Properties of a DPMM

- Expected value is the same as base distribution

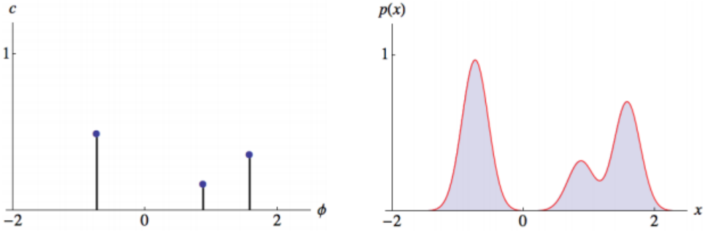
$$\mathbb{E}_{\text{DP}(\alpha,G)}[x] = \mathbb{E}_G[x] \tag{3}$$

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

Effect of scaling parameter α

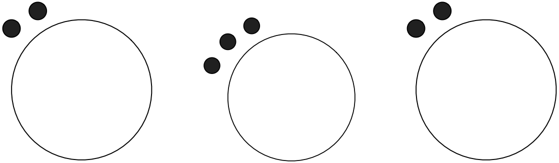


DP as mixture Model



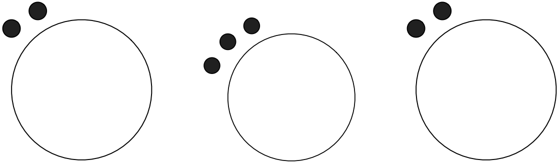
The Chinese Restaurant as a Distribution

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

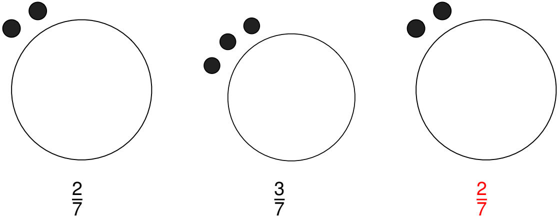
To generate an observation, you first sit down at a table. You sit down at a table proportional to the number of people sitting at the table.



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

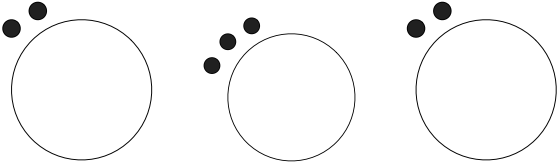
The Chinese Restaurant as a Distribution

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.



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.

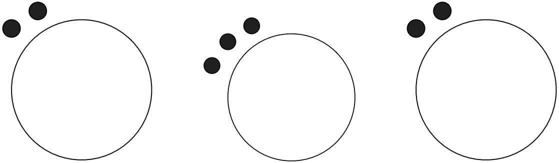


$\frac{2}{7}$ $\frac{3}{7}$ $\frac{2}{7}$

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

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.



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

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

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*

$$\Theta = \sum_{k \in \mathbb{N}} C_k \delta_{\Phi_k}$$

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

$$p(z_i = k \,|\, z_{-i}) \equiv p(z_i \,|\, z_1 \dots z_{i-1}, z_{i+1}, \dots z_m) \quad (4)$$

Gibbs Sampling for DPMM

$$p(z_i = k \,|\, \vec{z}_{-i}, \vec{x}, \{\theta_k\}, \alpha) \tag{5}$$

(6)

Gibbs Sampling for DPMM

$$p(z_i = k \,|\, \vec{z}_{-i}, \vec{x}, \{\theta_k\}, \alpha) \tag{5}$$

$$= p(z_i = k \,|\, \vec{z}_{-i}, x_i, \vec{x}, \theta_k, \alpha) \tag{6}$$

$$\tag{7}$$

Dropping irrelevant terms

Gibbs Sampling for DPMM

$$
\begin{aligned}
& p(z_i = k \,|\, \vec{z}_{-i}, \vec{x}, \{\theta_k\}, \alpha) && (5) \\
= & p(z_i = k \,|\, \vec{z}_{-i}, x_i, \vec{x}, \theta_k, \alpha) && (6) \\
= & p(z_i = k \,|\, \vec{z}_{-i}, \alpha) p(x_i \,|\, \theta_k, \vec{x}) && (7) \\
& && (8)
\end{aligned}
$$

Chain rule

Gibbs Sampling for DPMM

$$
\begin{aligned}
& p(z_i = k \,|\, \vec{z}_{-i}, \vec{x}, \{\theta_k\}, \alpha) && (5)\\
=& p(z_i = k \,|\, \vec{z}_{-i}, x_i, \vec{x}, \theta_k, \alpha) && (6)\\
=& p(z_i = k \,|\, \vec{z}_{-i}, \alpha) p(x_i \,|\, \theta_k, \vec{x}) && (7)\\
=& \begin{cases} \left(\frac{n_k}{n_\cdot + \alpha}\right) \int_\theta p(x_i \,|\, \theta) p(\theta \,|\, G, \vec{x}) & \text{existing} \\ \frac{\alpha}{n_\cdot + \alpha} \int_\theta p(x_i \,|\, \theta) p(\theta \,|\, G) & \text{new} \end{cases} && (8)\\
& && (9)
\end{aligned}
$$

Applying CRP

Gibbs Sampling for DPMM

$$p(z_i = k \,|\, \vec{z}_{-i}, \vec{x}, \{\theta_k\}, \alpha) \tag{5}$$

$$= p(z_i = k \,|\, \vec{z}_{-i}, x_i, \vec{x}, \theta_k, \alpha) \tag{6}$$

$$= p(z_i = k \,|\, \vec{z}_{-i}, \alpha) p(x_i \,|\, \theta_k, \vec{x}) \tag{7}$$

$$= \begin{cases} \left(\frac{n_k}{n_\cdot + \alpha}\right) \int_\theta p(x_i \,|\, \theta) p(\theta \,|\, G, \vec{x}) & \text{existing} \\ \frac{\alpha}{n_\cdot + \alpha} \int_\theta p(x_i \,|\, \theta) p(\theta \,|\, G) & \text{new} \end{cases} \tag{8}$$

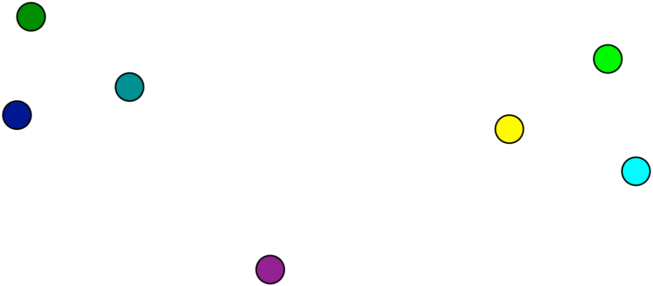
$$= \begin{cases} \left(\frac{n_k}{n_\cdot + \alpha}\right) \mathcal{N}\left(x, \frac{n\bar{x}}{n+1}, \mathbb{1}\right) & \text{existing} \\ \frac{\alpha}{n_\cdot + \alpha} \mathcal{N}(x, 0, \mathbb{1}) & \text{new} \end{cases} \tag{9}$$

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

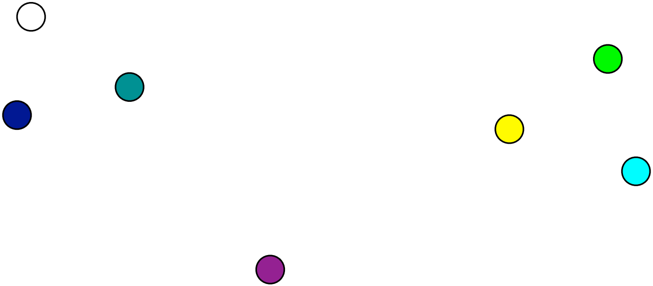
Algorithm for Gibbs Sampling

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

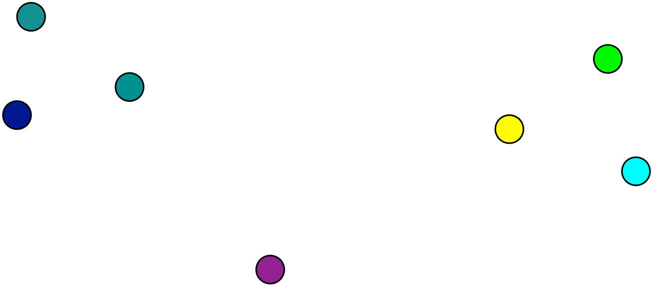
Toy Example



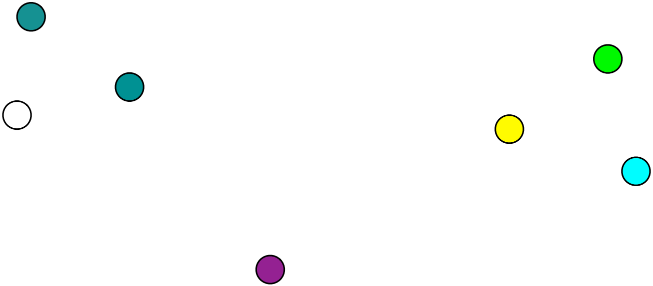
Toy Example



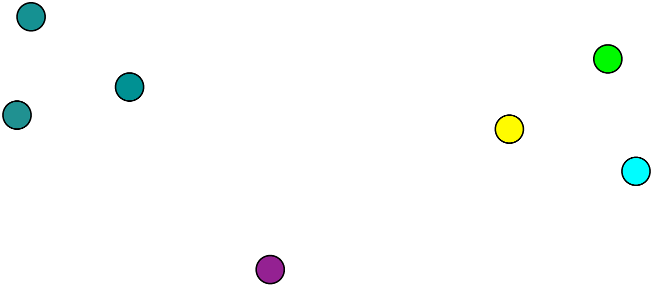
Toy Example



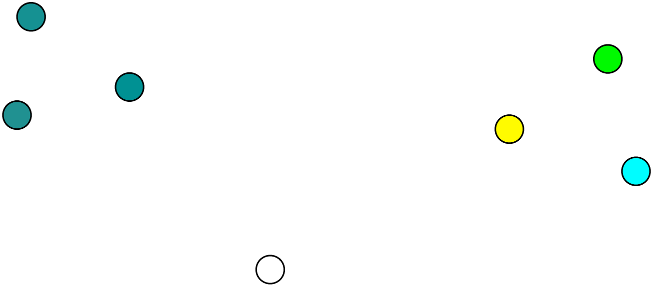
Toy Example



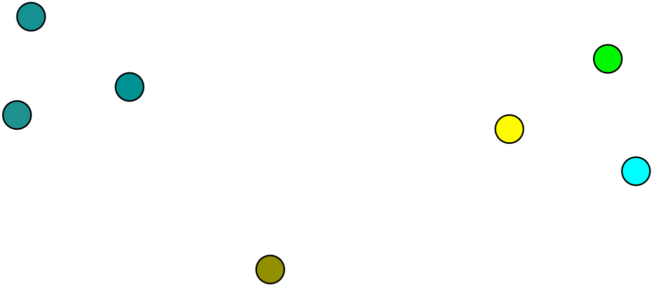
Toy Example



Toy Example

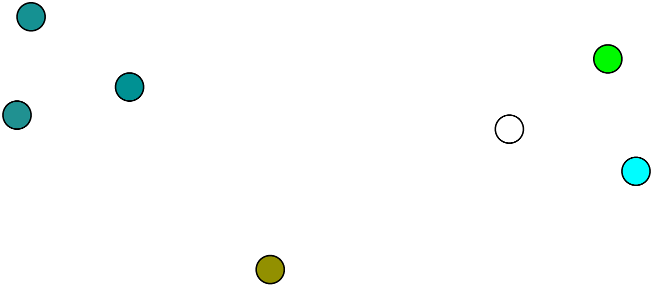


Toy Example

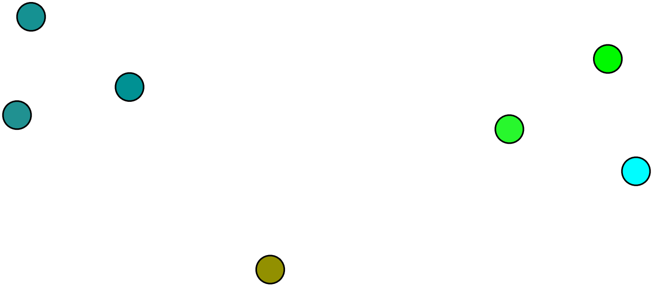


New cluster created!

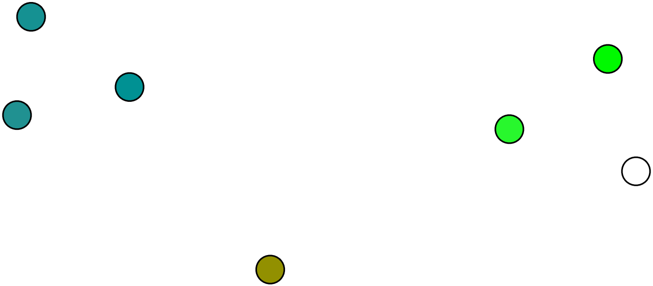
Toy Example



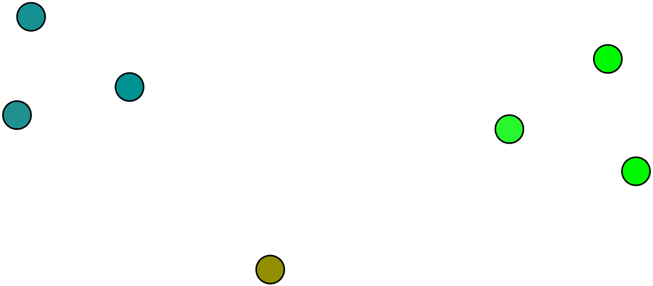
Toy Example



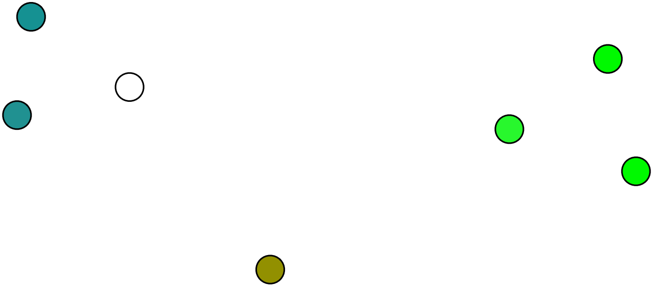
Toy Example



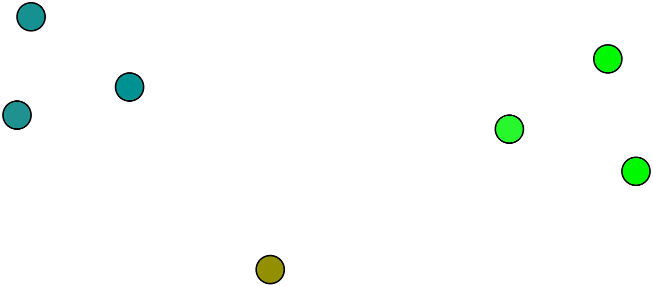
Toy Example



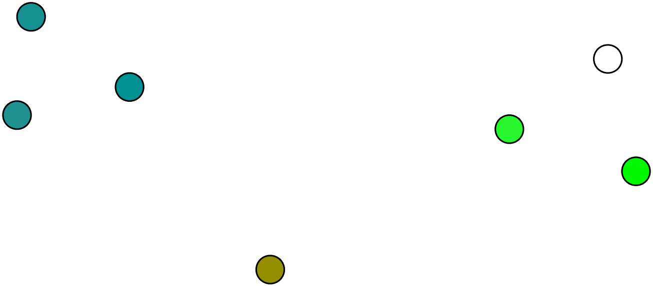
Toy Example



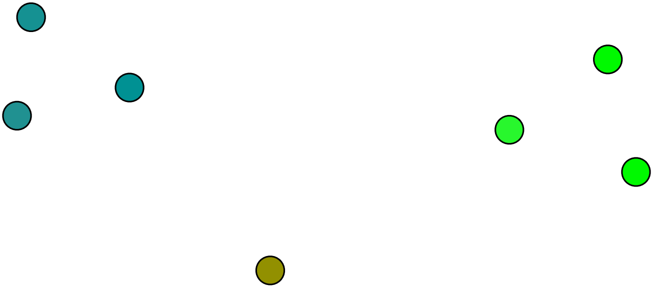
Toy Example



Toy Example



Toy Example



And repeat . . .

Differences between EM and Gibbs

- Gibbs often faster to implement
- EM easier to diagnose convergence
- EM can be parallelized
- Gibbs is more widely applicable

In class and next week

- Walking through DPMM clustering
- Clustering discrete data with more than one cluster per observation