Non-parametric Clustering with Dirichlet Processes

Timothy Burns

SUNY at Buffalo

Mar. 31 2009

Introduction

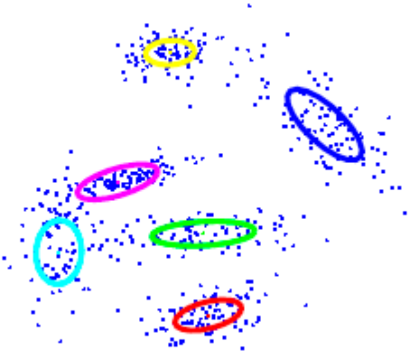
- Question: What should we do if we want to model data using a mixture, but we don't know the number of mixing elements k beforehand?

Introduction

- Question: What should we do if we want to model data using a mixture, but we don't know the number of mixing elements k beforehand?
- Good candidate for a **non-parametric** method!

Rational

- e.g. We want to select the number m of Gaussians in a mixture of Gaussians (right) or
- The number of means k in the k-means algorithm
- How can we do this?



Background

- First, we need to address a few things:

Background

- First, we need to address a few things:
    1. What is the Dirichlet distribution?

Background

- First, we need to address a few things:
    1. What is the Dirichlet distribution?
    2. What is a Dirichlet Process?

Background

- First, we need to address a few things:
    1. What is the Dirichlet distribution?
    2. What is a Dirichlet Process?
    3. How can it be represented?

Background

- First, we need to address a few things:
  1. What is the Dirichlet distribution?
  2. What is a Dirichlet Process?
  3. How can it be represented?
- Once we've done covered the basics we'll talk about a few examples!

The Dirichlet Distribution

- The **Dirichlet Distribution** is a distribution over the K-1 probability simplex.
- Let $\mathbf{p}$ be a K-dimensional vector s.t. $\forall j : p_j \geq 0$ and $\sum_{j=1}^{K} p_j = 1$, then

$$P(\mathbf{p}|\alpha) = \mathsf{Dir}(\alpha_1, ..., \alpha_K) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{K} p_j^{\alpha_j - 1} \qquad (1)$$

- The first term in the above equation is just a normalization constant.
- The Dirichlet Distribution is conjugate to the multinomial distribution. i.e if

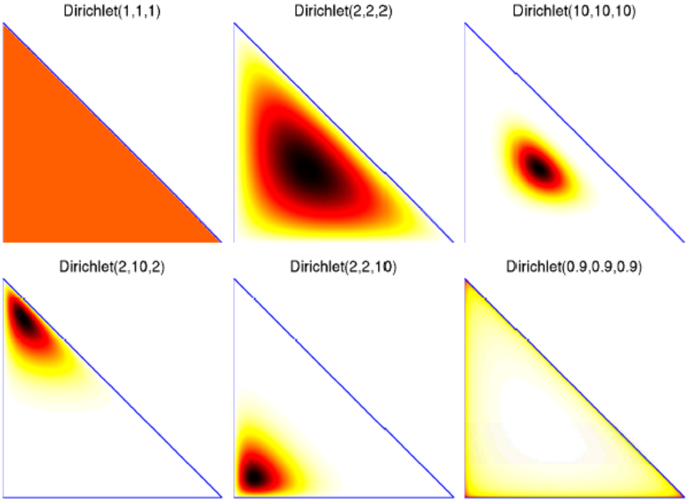
$$c|\mathbf{p} \sim \mathsf{Multinomial}(\cdot|\mathbf{p})$$

then the posterior is dirichlet

$$P(\mathbf{p}|c = j, \alpha) = \frac{P(c = j|\mathbf{p})P(\mathbf{p}|\alpha)}{P(c = j|\alpha)} = \mathsf{Dir}(alpha')$$

where $\alpha'_j = \alpha_j + 1$, and $\forall l \neq j : \alpha'_l = \alpha_l$

The Dirichlet Distribution



The Dirichlet Process

- Dirichlet Processes define a **distribution over distributions (or a measure on measures)**

$$G \sim \mathsf{DP}(\cdot|G_0, \alpha)$$

  where $\alpha > 0$ is a scaling parameter, and G_0 is the base distribution. Think of DP's as "infinite dimensional" Dirichlet distributions.
- It is important to note that G is an infinite dimensional object.

The Dirichlet Process

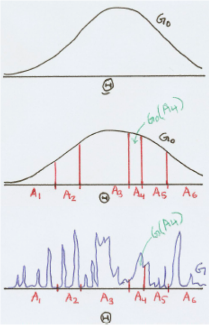
- Let Θ be a measurable space, G_0 be a probability measure on Θ, and α_0 be a real number.

- For all $(A_1, ..., A_K)$ finite partitions of Θ,

$$G \sim \mathrm{DP}(\cdot | G_0, \alpha)$$

means that

$$(G(A_1), ..., G(A_K)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), ..., \alpha_0 G_0(A_K))$$



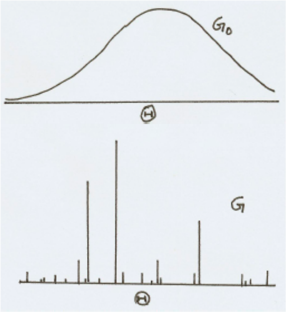
The Dirichlet Process

- Samples from a DP are discrete with probability one:

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta)$$

- Posterior $P(G|\theta)$ is also a DP! i.e. DP's are conjugate to themselves!

$$P(G|\theta) = \text{DP}\left(\frac{\alpha}{\alpha+1} G_0 + \frac{1}{\alpha+1}\delta_\theta, \alpha+1\right)$$



Urn Representation

- Given

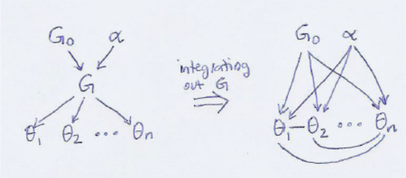
$$G \sim \mathsf{DP}(\cdot|G_0, \alpha) \quad \text{and} \quad \theta|G \sim G(\cdot)$$

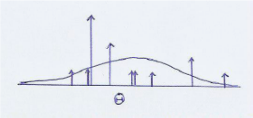
- Then

$$\theta_n|\theta_1, ..., \theta_{n-1}, G_0, \alpha \;\sim\; \frac{\alpha}{n-1+\alpha}G_0(\cdot) + \frac{1}{n-1+\alpha}\sum_{j=1}^{n-1}\delta_{\theta_j}(\cdot)$$

$$P(\theta_n|\theta_1, ..., \theta_{n-1}, G_0, \alpha) \;\propto\; \int \prod_{j=1}^{n} P(\theta_j|G)P(G|G_0, \alpha)dG$$

- This model exhibits a “clustering” kind of effect



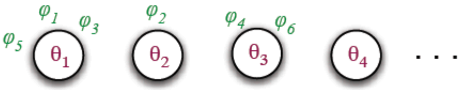


Chinese Restaurant Process (CRP)

The Chinese Restaurant Process is another representation of the DP. It can help us see this clustering effect more explicitly:

- Restaurant has potentially infinitely many tables $k = 1, \ldots$
- Customers are indexed by $i = 1, \ldots$, with values ϕ_i as they arrive
- Tables have values θ_k drawn from G_0
- $K =$ total number of occupied tables so far
- $n =$ total number of customers arrived thus far
- $n_k =$ number of customers seated at table k.

Chinese Restaurant Process (CRP)



Generating from a CRP:

customer 1 enters the restaurant and sits at table 1.

$\phi_1 = \theta_1$ where $\theta_1 \sim G_0$, $K = 1$, $n = 1$, $n_1 = 1$

for $n = 2, \ldots,$

customer n sits at table $\begin{cases} k & \text{with prob } \frac{n_k}{n-1+\alpha} \quad \text{for } k = 1 \ldots K \\ K+1 & \text{with prob } \frac{\alpha}{n-1+\alpha} \quad \text{(new table)} \end{cases}$

if new table was chosen **then** $K \leftarrow K+1$, $\theta_{K+1} \sim G_0$ **endif**

set ϕ_n to θ_k of the table k that customer n sat at; set $n_k \leftarrow n_k + 1$

endfor

Relationship between CRPs and DPs

- DP is a **distribution over distributions**.
- A DP results in discrete distributions, so drawing n points will likely result in repeat values.
- A DP induces a **partitioning** of the n points
- CRP is the corresponding **distribution over partitions**.

Stick Breaking Construction

- Samples $G \sim \text{DP}(\cdot|G_0, \alpha)$ can be represented as follows:

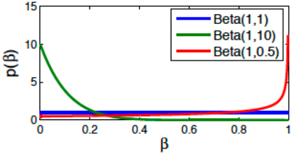
$$G(\cdot) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\cdot)$$

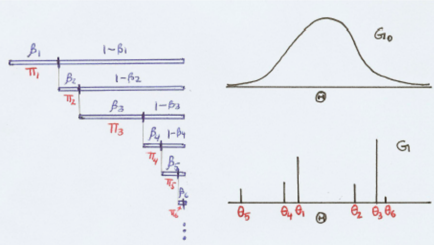
where $\theta_k \sim G_0(\cdot), \sum_{k=1}^{\infty} \pi_k = 1,$

$$\pi_k = \beta_k \prod_{j=1} k - 1(1 - \beta_j)$$

and

$$\beta_k \sim \text{Beta}(\cdot|1, \alpha)$$



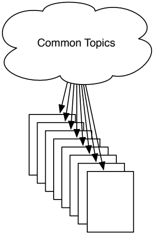


Extensions

- Hierarchical Dirichlet Processes (HDP) - For sharing statistical power between many different groups of clustering data.

Topic Modeling

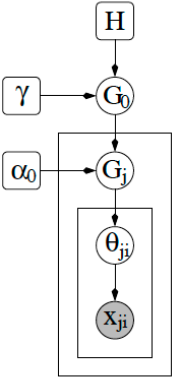
- Goal: To model topics (distributions of words) across an entire corpus.



- LDA (Latent Dirichlet Allocation) is an example of parametric solution to problem, but has shortcomings
  1. Documents each draw their own mixing proportions (mixture component is a topic) separately
  2. Words are then drawn independently from the mixture model.

Topic Modeling

- Not easy to extend like a simple mixture model since each document has essentially it's own mixture (and thus mixing proportions)
- In order to capture this we use a separate DP mixture to model each document where each DP is a draw from another DP. This is an application of HDPs



Dirichlet Process Mixtures

DPs are discrete with prob one, so they are not useful for use as a prior on continuous densities.

- In a DP **Mixture**, we draw the parameters of a mixture model from a draw from a DP:

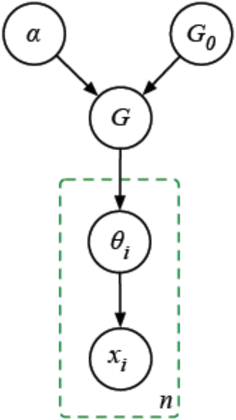
$$G \sim \mathrm{DP}(\cdot|G_0, \alpha)$$

$$\theta_i \sim G(\cdot)$$

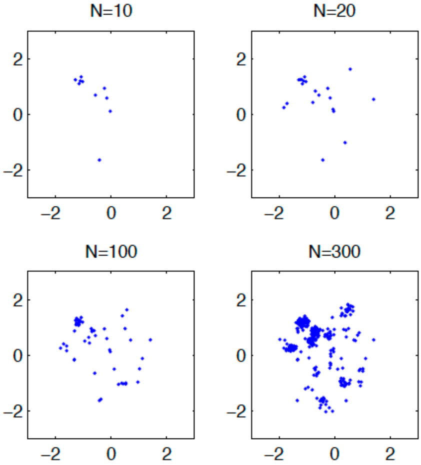
$$x_i \sim p(\cdot|\theta_i)$$

- For example, can be a DP-MoG if $p(\cdot|\theta) =$ Gaussian. However, $p(\cdot|\theta)$ could be any density.

Dirichlet Process Mixtures



Dirichlet Process Mixtures



Notice that more structure (clusters) appear as you draw more points.
(figure inspired by Neal)

Dirichlet Process Mixtures

We can think of Infinite DP Mixtures in terms of finite mixtures:

- Consider using a finite mixture of K components to model a data set $D = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}\}$

$$\begin{aligned} p(x^{(i)}|\theta) &= \sum_{j=1}^{K} \pi_j p_j(x^{(i)}|\theta_j) \\ &= \sum_{j=1}^{K} P(s^{(i)} = j|\pi) p_j(x^{(i)}|\theta_j, s^{(i)} = j) \end{aligned}$$

Dirichlet Process Mixtures

- The distribution of indicator variables (assignments of data to mixture) $s = (s^{(1)}, ..., s^{(n)})$ given π is **multinomial**

$$P(s^{(1)}, ..., s^{(n)}|\pi) = \prod_{j=1}^{K} \pi_j^{n_j}, \quad n_j = \sum_{i=1}^{n} \delta(s^{(i)}, j)$$

- Since we know the Dirichlet distribution is conjugate to the multinomial we can assume the mixing proportions π have a **Dirichlet prior**:

$$p(\pi|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{j=1}^{K} \pi_j^{\alpha/K-1}$$

- And integrating out the mixing proportions gives us:

$$P(s^{(1)}, ..., s^{(n)}|\alpha) = \int P(s|\pi)P(\pi|\alpha)d\pi = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{j=1}^{K} \frac{\Gamma(n_j + \alpha/K)}{\Gamma(\alpha/K)}$$

Dirichlet Process Mixtures

- **Conditionals: Finite K**

$$P(s^{(i)} = j | s_{-i}, \alpha) = \frac{n_{-i,j} + \alpha/K}{n - 1 + \alpha}$$

  where s_{-i} denotes all indeces except i, and $n_{-i,j} = \sum_{l \neq i} \delta(s^{(l)}, i)$

- **Conditionals: Infinite K**
  - Limit as $K \rightarrow \infty$

$$P(s^{(i)} = j | s_{-i}, \alpha) = \begin{cases} \frac{n_{-i,j}}{n-1+\alpha} & j \text{ represented} \\ \frac{\alpha}{n-1+\alpha} & \text{all } j \text{ not represented} \end{cases}$$

Summary

- DPs are essentially infinite dimensional Dirichlet distributions
- DPs can be constructed in a number of different ways (CRP,SB,Urn)
- DPs can be extended via HDPs (didn't talk a lot about this, but you can read yourself)
- Can use DPs to "Cluster" discrete data
- Can apply DPs as non-parametric priors over mixing proportions in non-discrete mixtures as a way to avoid "sticking" with a particular model.
- And that's all folks!

Sources

These slides have made extensive use of the following sources.

- Many slides were adapted directly from Zhoubin Ghahramani's *Non-parametric Bayesian Methods Tutorial*, 2005
- Yee Whye Teh, *Dirichlet Processes*
- *Hierarchical Dirichlet Processes*, Blei, Jordan, Y. W. Teh, Beal, JASA