Mixture Models, Chinese Restaurant Process and Dirichlet Process

Clint P. George

Computer and Information Science and Engineering
University of Florida

Acknowledgments: Tutorials by Michael I. Jordan and Yee Whye Teh

June 24, 2011

Outline

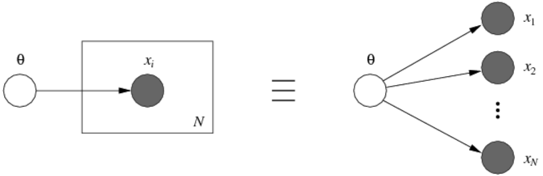
1. **Introduction**
   - Background
2. **Chinese Restaurant Process**
3. **Dirichlet Process**
   - Dirichlet Distribution
   - Dirichlet Process
   - Dirichlet Process Mixture Models
4. **Representing the Dirichlet Process**
   - Chinese Restaurant Process
   - Stick Breaking Construction

Parametric vs Nonparametric Models

- Parametric models
  - have finite dimensional parameter vectors
  - e.g. k-means, Gaussian mixtures, normal distribution $\mathcal{N}(\mu, \sigma^{\in})$, Latent Dirichlet Allocation
- Nonparametric models
  - nonparametric doesn't mean that no parameters in a model
  - roughly, it means that the number of parameters in a model increases with data points

Graphical Models – Review

- Given a graph $G = (V, E)$, where each node $v \in V$ is associated with a random variable
- A plate, a macro, represents replicated subgraphs



- The shaded nodes represent observed variables
- The above graph represents the following probability for observations $x_1, x_2, ..., x_n$:

$$P(x_1, x_2, ..., x_n) = \int \prod_{i=1}^{n} P(x_i|\theta) dP(\theta)$$

Model based Clustering

- A generative approach to clustering
  - choose a cluster from a distribution $\pi = (\pi_1, ..., \pi_K)$
  - draw a data point from the cluster-specific probability distribution
- This yields a mixture model:

$$p(x|\phi, \pi) = \sum_{k=1}^{K} \pi_k p(x|\phi_k)$$

  where π and $(\phi_1, ..., \phi_K)$ are model parameters
- This model assumes that each data point is generated from a single mixture component
  - i.e. k^{th} cluster is the set of data points drawn from the k^{th} mixture component

Finite Mixture Models

- Another way to express this model: define an underlying measure

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

  where δ_{ϕ_k} is a delta function (an atom) located at ϕ_k
- And, data generation is as follows:

$$\theta_i \sim G, i = 1, ..., n$$

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

- Note that each θ_i is equal to one of the underlying ϕ_k.
  - the k^{th} cluster is a subset of $(\theta_1, ..., \theta_n)$ that maps to ϕ_k

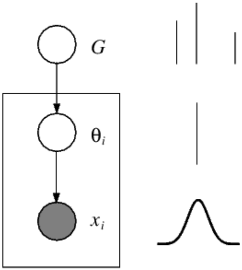
Finite Mixture Models – Graphical Model

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

$$\theta_i \sim G, i = 1, ..., n$$

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

Model selection is over ϕ_k, π, and K



Clustering – Choosing K

How do we chose K – the number of clusters in the data set ?

- Clustering based on objective functions
  - e.g. K-means, spectral clustering
  - hard to convert these into data-driven choices of K
- Clustering based on parametric log-likelihood
  - e.g. pLSI, Latent Dirichlet Allocation
  - underlying model assumptions are based on a fixed K
- what is next ?
  - Bayesian nonparametric methods, e.g., Dirichlet Process

Polya-urn Model

- Urn model:
  - an urn that contains x white balls and y black balls
  - one ball is drawn randomly from the urn, its colors is observed; it is then placed back in the urn
  - repeat the process
- Polya urn model:
  - differs only in – when a ball of a particular color is observed, that ball is put back along with a new ball of the same color.
  - the contents of the urn change over time, i.e., the rich get richer

Reference: Wikipedia

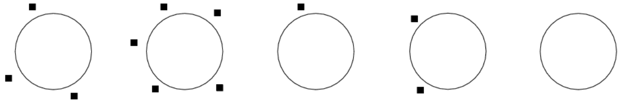
Chinese Restaurant Process (CRP)

- A modified Polya urn model
- A random process in which n customers sit down in a Chinese restaurant with an infinite number of tables
  - first customer sits at any table
  - m^{th} customer sits at a table with the probability:

$$P(a\, previously\, occupied\, table\, i | S_{m-1}) \propto n_i$$

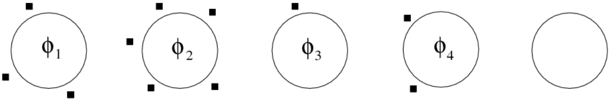
$$P(an\, unoccupied\, table\, j | S_{m-1}) \propto \alpha_0$$

  - n_i – the number of customers currently allocated to table i
  - S_{m-1} – the current state of the restaurant, after $(m-1)$ customers have been seated.



The CRP in Clustering

- Data points are like customers and table are like clusters
  - the CRP defines a prior on the data partitions and table counts
- We complete this prior with
  - a likelihood – associate a parameterized probability distribution with each table
  - a prior for the parameters – the first customer who sits at table i choses a parameter vector for that table (ϕ_i) from a prior distribution



- Now, we have a distribution which can be used in the clustering setting

The CRP – Properties

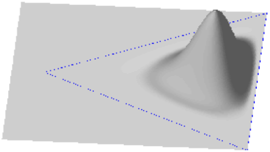
- The CRP can be used as an exchangeable prior on the data partitions and parameter vectors associated with tables
- As a prior on table counts, the CRP is nonparametric
  - i.e., the number of occupied tables grows with m, the number of customers
- Similarities to the Polya urn model – assuming θ_i as the parameter vector for i^{th} data point, we get:

$$\theta_i|\theta_1, ..., \theta_{i-1} \sim \alpha_0 G_0 + \sum_{j=1}^{i-1} \delta_{\theta_j}$$

- *How can we relate this to standard model based clustering?*

Dirichlet Distribution

- Let $\pi = (\pi_1, \pi_2, ..., \pi_m)$ be a point in the $m-1$ simplex
  - $0 < \pi_i < 1$
  - $\sum_{i=1}^{m} \pi_i = 1$



- Let $\alpha = (\alpha_1, \alpha_2, ..., \alpha_m)$ represents a set of hyper-parameters
  - where $\alpha_i > 0$
- Then, we can define the Dirichlet density as

$$p(\pi|\alpha) \propto \prod_{i=1}^{m-1} {\pi_i}^{\alpha_i - 1}$$

Dirichlet Distribution – Properties

- Agglomerative: combining the entries of probability vectors preserves Dirichlet property

$$(\pi_1, ..., \pi_K) \sim Dirichlet(\alpha_1, ..., \alpha_K)$$

$$\Rightarrow (\pi_1 + \pi_2, ..., \pi_K) \sim Dirichlet(\alpha_1 + \alpha_2, ..., \alpha_K)$$

  - prove this by using a representation of the Dirichlet as a normalized set of independent gamma random variables

- The converse is also true, i.e., decimative property

$$(\pi_1, ..., \pi_K) \sim Dirichlet(\alpha_1, ..., \alpha_K)$$

$$(\beta_1, \beta_2) \sim Beta(a, b)$$

$$(\gamma_1, \gamma_2) \sim Dirichlet(\alpha_1\beta_1, \alpha_1\beta_2)$$

$$\Rightarrow (\pi_1\gamma_1, \pi_1\gamma_2, \pi_2, ..., \pi_K) \sim Dirichlet(\alpha_1\beta_1, \alpha_1\beta_2, \alpha_2, ..., \alpha_K)$$

Dirichlet Process (DP)

A Dirichlet Process can be viewed as an infinitely decimated Dirichlet distribution

$$1 \sim Dirichlet(\alpha)$$

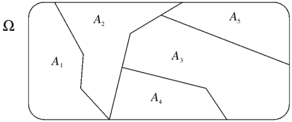
$$\pi_1, \pi_2 \sim Dirichlet(\alpha/2, \alpha/2)$$

$$\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22} \sim Dirichlet(\alpha/4, \alpha/4, \alpha/4, \alpha/4)$$

Dirichlet Process (DP) – Definition

- A measure is function from subsets to the nonnegative reals
- The DP is a distribution over probability measures:
  - let (X, Σ) be a measurable space, G_0 a probability measure on the space, and α_0 a positive real
  - a DP is the distribution of a random probability measure G over (X, Σ) such that, for any finite partition $A_1, ..., A_K$ of X, the random vector $(G(A_1), ..., G(A_K))$ is distributed as a finite dimensional Dirichlet distribution:

$$(G(A_1), ..., G(A_K)) \sim Dirichlet(\alpha_0 G_0(A_1), ..., \alpha_0 G_0(A_K))$$



  - we write $G \sim DP(\alpha_0 G_0)$, if G is DP distributed

Posterior Dirichlet Process (DP)

- Suppose, $G \sim DP(\alpha_0 G_0)$ and $\theta_i \sim G$.
- Then, what is the posterior DP ?
  - we get a Dirichlet-Multinomial update for a fixed partition, i.e., for the partition that contains θ_i the exponent increases by one.
  - for the sample θ_1, we have:

$$G|\theta_1 \sim DP(\alpha_0 G_0 + \delta_{\theta_1})$$

  - iterating through all θ_i, the posterior update yields:

$$G|\theta_1, ..., \theta_n \sim DP(\alpha_0 G_0 + \sum_{i=1}^{n} \delta_{\theta_i})$$

Posterior Dirichlet Process

- Based on the expectation formula of Dirichlet random variable, for a given set $A \subseteq \Omega$:

$$E[G(A)|\theta_1, ..., \theta_n] = \frac{\alpha_0 G_0(A) + \sum_{i=1}^{n} \delta_{\theta_i}(A)}{\alpha_0 + n} \to \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

  - ϕ_k are unique values of $(\theta_1, ..., \theta_n)$
  - $\pi_k = \lim_{n \to \infty} \frac{n_k}{n}$
  - n_k is the number of repeats of ϕ_k in $(\theta_1, ..., \theta_n)$
- This suggests that the DP random measures are discrete
  - this was proved using Stick Breaking construction by Sethuraman, 1994
  - there is a positive probability that θ_i's can have same value, ϕ_k, for some k, i.e., $(\theta_1, ..., \theta_n)$ cluster together into K partitions

DP Mixture Models

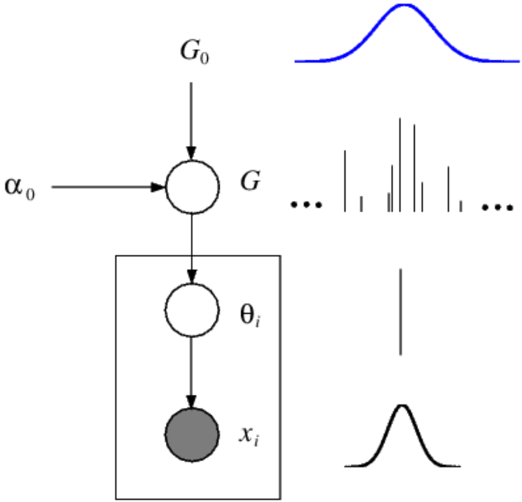
- In the mixture model setting, θ_i is the hidden parameter associated with x_i
- We use DP as prior on θ and complete model by introducing likelihood, as in finite mixture models
- This yields a model known as a DP mixture model

$$G \sim DP(\alpha_0, H)$$

$$\theta_i | G \sim G, i = 1, ..., n$$

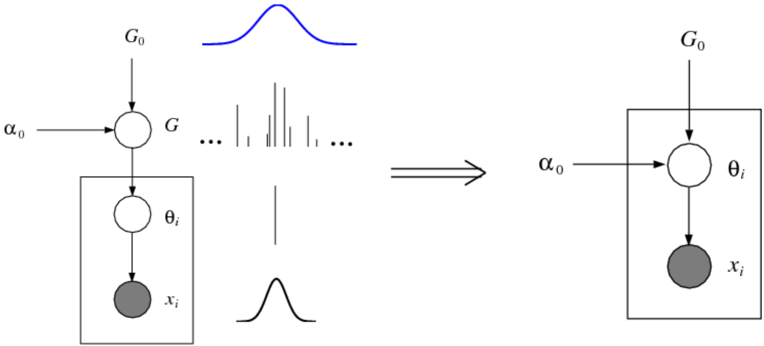
$$x_i | \theta_i \sim F(x_i | \theta_i), i = 1, ..., n$$

DP Mixture Models – Graphical Model



DP Mixture Models - Marginals

To obtain marginals on $\theta_1, ..., \theta_n$, we need to integrate out G



DP Mixture Models - Marginals

- Recall the expectation formula:

$$E[G(A)|\theta_1, ..., \theta_n] = \frac{\alpha_0 G_0(A) + \sum_{k=1}^{K} n_k \delta_{\phi_k}(A)}{\alpha_0 + n}$$

  - where A is the singleton set equal to one of ϕ_k
  - this says the marginal probability of observing $\phi_k \propto n_k$
  - also, the marginal probability of observing a new $\phi_{new} \propto \alpha_0$
- Thus, it is similar to the Polya urn model

Dirichlet Process – The CRP view

- Shows that draws from the DP are both discrete and exhibit a clustering property
- This do not refer to G directly; it refers to draws from G
  - suppose, $\theta_1, ..., \theta_n \sim G$
  - the conditional $\theta_i | \theta_1, ..., \theta_{i-1}$ is obtained as (after integrating out G, Blackwell and MacQueen 1973)

$$\theta_i | \theta_1, ..., \theta_{i-1}, \alpha_0, G_0 \sim \sum_{l=1}^{i-1} \frac{1}{i-1+\alpha_0} \delta_{\theta_l} + \frac{\alpha_0}{i-1+\alpha_0} G_0$$

- This conditional shows that θ_i has a positive probability of being equal to one of the previous draws

Dirichlet Process – The CRP view

- The CRP metaphor – generative process
  - first customer sits at any table
  - m^{th} customer sits at:
    - k^{th} table with probability $\frac{n_k}{\alpha+m-1}$, where n_k is the number of customers at table k
    - otherwise, a new table $K+1$ with probability $\frac{\alpha}{\alpha+m-1}$
- customers $\Leftrightarrow$ data points and tables $\Leftrightarrow$ cluster or topics

Stick Breaking Construction (SB)

- Based on the agglomerative and decimative property of Dirichlet distributions
- Suppose

$$G \sim DP(\alpha G_0), \theta_i \sim G$$

$$G = \beta_1 \delta_{\phi_1} + (1 - \beta_1) G_1$$

  - this means G has a point mass located at ϕ_1
  - G_1 is the (renormalized) DP probability measure with the point mass removed; by the properties of Dirichlet

$$G = \beta_1 \delta_{\phi_1} + (1 - \beta_1)(\beta_2 \delta_{\phi_2} + (1 - \beta_2) G_2)$$

  finally,

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

- We shall see that the coefficients π_k can be generated from a stick breaking construction.

Stick Breaking Construction (SB)

- Define infinite sequence of Beta random variables

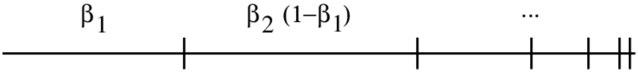
$$\beta_j \sim Beta(1, \alpha_0), j = 1, 2, ...$$

- Define infinite sequence of mixing proportions:

$$\pi_1 = \beta_1$$

$$\pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_j), k = 2, 3, ...$$

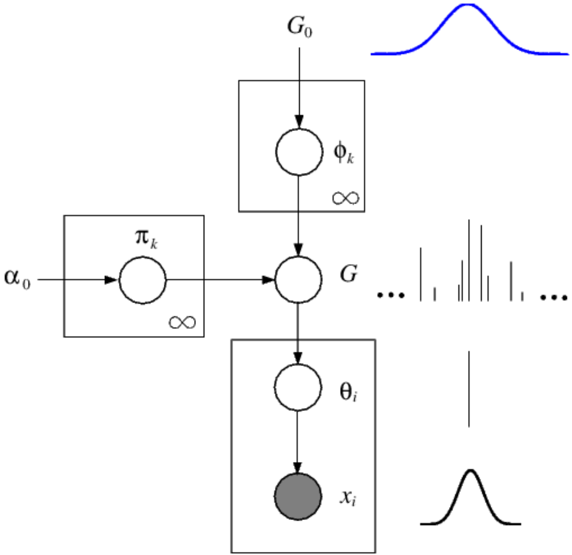
- This π_k's can be viewed as breaking off portions of a stick.



Stick Breaking Construction (SB)

- we can prove that $\sum_{k=1}^{\infty} \pi_k = 1$
- So now $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ has a clean definition as a random measure
- Sethuraman (1994) proved that SB is DP distributed, by taking the expected value of posterior DP
- The DP looks like a sum of point masses, where masses are drawn from a SB construction

SB – Graphical Model



DP - Density Estimation

- We assume

$$G \sim DP(\alpha, H)$$

$$x_i \sim G$$

- Since G is discrete there is no density, so we convolve the DP with a smooth distribution, i.e.,

$$G \sim DP(\alpha, H) \Rightarrow G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

$$F_x(.) = \int F(.|\theta) dG(\theta) \Rightarrow F_x(.) = \sum_{k=1}^{\infty} \pi_k F(.|\theta_k^*)$$

$$x_i \sim F_x \Rightarrow x_i \sim F_x$$

Note: This is similar to infinite mixture model

References and Useful Links

- Y.W. Teh - Dirichlet Processes: Tutorial and Practical Course
  http://videolectures.net/mlss07_teh_dp/
- Micheal I Jordan - Dirichlet Processes, Chinese Restaurant Processes, and all that
  http://videolectures.net/icml05_jordan_dpcrp/
- Y.W. Teh, Michael I Jordan, David Blei, and Matthew Beal - Hierarchical Dirichlet Process