Nonparametric Bayesian Models and Dirichlet Process

Huy Viet Nguyen

Department of Computer Science

University of Pittsburgh

Advanced Topics in Machine Learning

November 2011

What is the Dirichlet Process?

Outline

Bayes Rule

$$P(\theta|D,m) = \frac{P(D|\theta,m)P(\theta|m)}{P(D|m)}$$

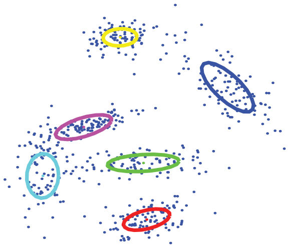
- Model Comparison

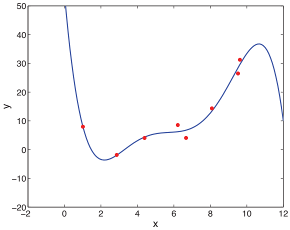
$$\begin{aligned} P(m|D) &= \frac{P(D|m)P(m)}{P(D)} \\ P(D|m) &= \int P(D|\theta,m)P(\theta|m)d\theta \end{aligned}$$

- Prediction

$$\begin{aligned} P(x|m,D) &= \int P(x|\theta,D,m)P(\theta|D,m)d\theta \\ &= \int P(x|\theta,m)P(\theta|D,m)d\theta \qquad \text{(if } x \text{ is i.i.d given } \theta) \end{aligned}$$

Model Selection





- Selecting *m*, the number of Gaussians in a mixture model
- Selection *m*, the order of a polynomial in a nonlinear regression model

Parametric Modeling and Model Selection

- Two criteria
  - How well the model fits the data
  - How complex the model is
- However, real data is complicated
  - Any small finite number seems unreasonable
  - Any order polynomial seems unreasonable

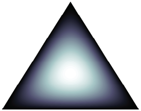
Bayesian Nonparametric Models

- Bayesian methods are the most powerful when prior distribution adequately captures the belief
- Inflexible model yields unreasonable inference: complex model often causes overfitting.
- Nonparametric models are a way of getting very flexible model
- Bayesian nonparametric models is to fit a single model that can adapt its complexity to the data
  - Complexity grows as more data are observed

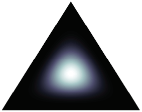
Dirichlet Distribution

- The Dirichlet distribution is a distribution over the K-dimensional probability simplex:
$$\Delta_K = \{(\pi_1, ..., \pi_K) : \pi_k \geq 0, \sum_k \pi_k = 1\}$$
- Then $\pi = (\pi_1, ..., \pi_K)$ is Dirichlet distributed
$$(\pi_1, ..., \pi_K) \sim \text{Dirichlet}(\alpha_1, ..., \alpha_K)$$
with parameters $\alpha = (\alpha_1, ..., \alpha_K), \alpha_k > 0$ if:
$$p(\pi_1, ..., \pi_K|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \pi_k^{\alpha_k - 1}$$
$$\begin{aligned} E[\pi_k] &= \frac{\alpha_k}{\alpha_0} \\ \text{Var}[\pi_k] &= \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \end{aligned}$$
where $\alpha_0 = \sum_k \alpha_k$

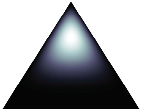
Dirichlet Distribution



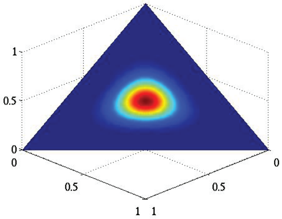
$\alpha = (2, 2, 2)$



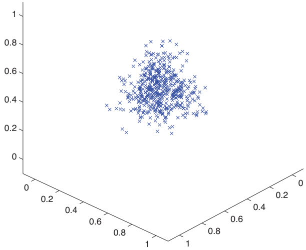
$\alpha = (5, 5, 5)$



$\alpha = (2, 2, 25)$



$\alpha = [10, 10, 10]$



Conjugate prior

- Dirichlet distribution is conjugate to multinomial distribution
- Let

$$\begin{aligned} \pi &\sim \text{Dirichlet}(\alpha) \\ c|\pi &\sim \text{Multinomial}(.|\pi) \\ p(c=k|\pi) &= \pi_k \end{aligned}$$

- Then we have

$$p(\pi|c=k,\alpha) = \text{Dirichlet}(\alpha')$$

where $\alpha'_k = \alpha_k + 1, \alpha'_i = \alpha_i \forall i \neq k$

Agglomerative property of Dirichlet distributions

- Combining entries by their sum

$$\begin{aligned}(\pi_1, \ldots, \pi_K) &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \\ \Rightarrow (\pi_1, \ldots, \pi_i + \pi_j, \ldots, \pi_K) &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_i + \alpha_j, \ldots, \alpha_K)\end{aligned}$$

- The converse of the agglomerative property is also true

$$\begin{aligned}(\pi_1, \ldots, \pi_K) &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \\ (\tau_1, \tau_2) &\sim \text{Dirichlet}(\alpha_i\beta_1, \alpha_i\beta_2) \\ \Rightarrow (\pi_1, \ldots, \pi_i\tau_1, \pi_i\tau_2, \ldots, \pi_K) &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_i\beta_1, \alpha_i\beta_2, \ldots, \alpha_K)\end{aligned}$$

where $\beta_1 + \beta_2 = 1$

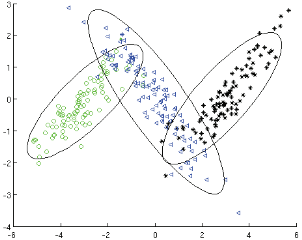
Finite Mixture Models

- Select one of K cluster from distribution $\pi = (\pi_1, ..., \pi_K)$
- Generate a data point from a cluster-specific probability distribution

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

where $\Phi = (\Phi_1, ..., \Phi_K)$ and Φ_k are parameters for cluster k.

- Frequentist approach: use maximize likelihood to learn parameters (π, Φ)



Finite Mixture Models (cont.)

- Define an underlying measure

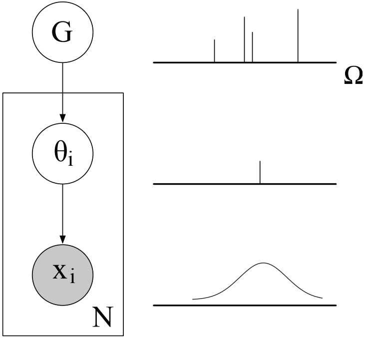
$$G = \sum_{k=1}^{K} \pi_k \delta_{\Phi_k}$$

where δ_{Φ_k} is an atom at Φ_k

- Process of drawing a sample from finite mixture model is as follow: $i = 1..N$

$$\begin{aligned} \theta_i &\sim G \\ x_i &\sim p(\cdot|\theta_i) \end{aligned}$$

where θ_i is one of underlying Φ_k.

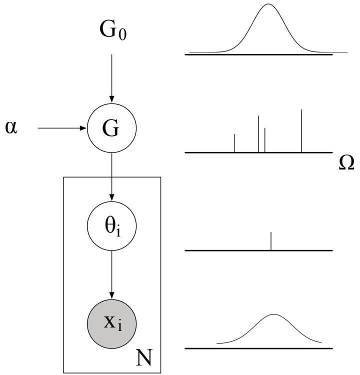


Bayesian Finite Mixture Models

- Need priors on parameters π and Φ
- Priors for Φ is model-specific
- Place Dirichlet prior on the mixing portions π

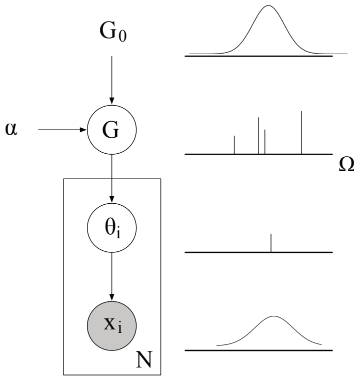
$$\pi \sim \text{Dirichlet}(\alpha_0/K, \ldots, \alpha_0/K)$$

  - The prior mean of π_k is equal to $1/K$
  - The prior variance of π_k is proportional to $1/\alpha_0$
  - α_0 is called concentration parameter



Bayesian Finite Mixture Models (cont.)

$$
\begin{aligned}
\Phi_k &\sim G_0 \\
\pi &\sim \text{Dirichlet}(\alpha_0/K, ..., \alpha_0/K) \\
G &= \sum_{k=1}^{K} \pi_k \delta_{\Phi_k} \\
\theta_i &\sim G \\
x_i &\sim p(\cdot|\theta_i)
\end{aligned}
$$



G_0 is a random measure.

Nonparametric or Infinite Mixture Models

- How to choose number of mixture components?
- Dirichlet Process provide a nonparametric Bayesian mixture models
- Define a countably infinite mixture model by taking K to infinity Dirichlet process is a flexible, nonparametric prior over an infinite number of clusters/classes as well as the parameters for those classes.

Gaussian Process (recall)

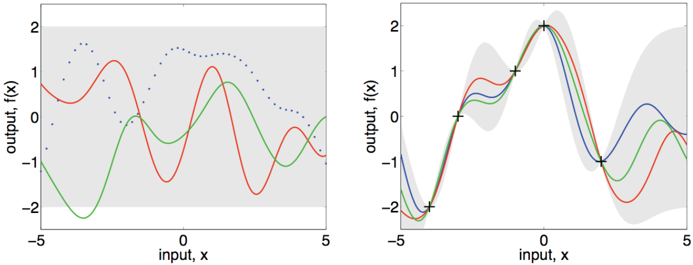
- GP defines a distribution over functions $f : X \to \mathbb{R}$
- For any input points $x_1, x_2, ...x_n$, we require:

$$(f(x_1), f(x_2), ...f(x_n)) \sim \mathcal{N}(\mu, \Sigma)$$

- Gaussian process for nonlinear regression

$$\begin{aligned} D &= (x, y) \\ y_i &= f(x_i) + \epsilon_i \\ f &\sim \mathrm{GP}(\cdot|0, c) \\ \epsilon_i &\sim \mathcal{N}(\cdot|0, \sigma^2) \end{aligned}$$

Gaussian Process



Dirichlet Process

- Dirichlet Process is a distribution over probability measures on a measurable space
- A draw from DP is a random distribution over that space
- **Definition:** Let G_0 be a probability measure on measurable space Ω and $\alpha \in \mathbb{R}^+$. The Dirichlet process is a distribution over probability measure G on Ω such that for any finite partition $(A_1, ..., A_K)$ of Ω, we have:

$$(G(A_1), ..., G(A_K)) \sim \text{Dirichlet}(\alpha G_0(A_1), ..., \alpha G_0(A_K))$$

Dirichlet Process

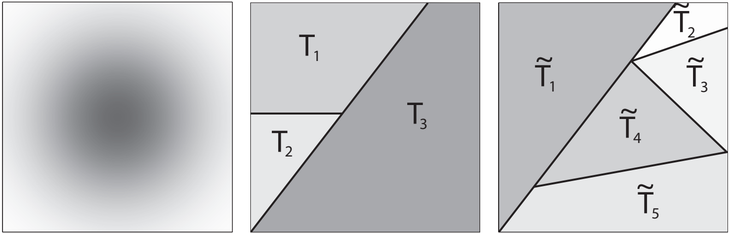


Figure: Left: base measure G_0, Middle: partition with $K = 3$, Right: partition with $K = 5$

Dirichlet Process

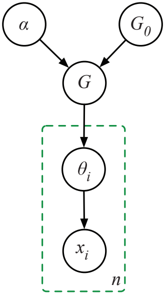
$$(G(A_1), ..., G(A_K)) \sim \text{Dirichlet}(\alpha G_0(A_1), ..., \alpha G_0(A_K))$$

- G_0 is called based distribution, likes the mean of DP

$$\forall A \subset \Omega, \text{E}[G(A)] = G_0(A)$$

- α is called concentration parameter

$$\text{Var}[G(A)] = \frac{G_0(A)(1 - G_0(A))}{\alpha + 1}$$



Posterior Dirichlet Process

Given

- $G \sim \text{DP}(\alpha, G_0)$
- $\theta \sim G$
- Fix a partition $(A_1, ..., A_K)$

$$
\begin{aligned}
(G(A_1), ..., G(A_K)) &\sim \text{Dirichlet}(\alpha G_0(A_1), ..., \alpha G_0(A_K)) \\
p(\theta \in A_k | G) &= G(A_k) \\
p(\theta \in A_k) &= G_0(A_k)
\end{aligned}
$$

Then the posterior is also DP

$$
\begin{aligned}
(G(A_1), ..., G(A_K))|\theta &\sim \text{Dirichlet}(\alpha G_0(A_1) + \delta_\theta(A_1), ..., \alpha G_0(A_K) + \delta_\theta(A_K)) \\
G|\theta &\sim \text{DP}\left(\alpha + 1, \frac{\alpha G_0 + \delta_\theta}{\alpha + 1}\right)
\end{aligned}
$$

Posterior Dirichlet Process

The posterior is also DP

$$
\begin{aligned}
(G(A_1), ..., G(A_K))|\theta &\sim \text{Dirichlet}(\alpha G_0(A_1) + \delta_\theta(A_1), ..., \alpha G_0(A_K) + \delta_\theta(A_K)) \\
G|\theta &\sim \text{DP}\left(\alpha + 1, \frac{\alpha G_0 + \delta_\theta}{\alpha + 1}\right)
\end{aligned}
$$

- For a fixed partition, we get a standard Dirichlet update
- For the "cell" $\mathcal{A}$ that contains θ, $\delta_\theta(\mathcal{A}) = 1$
- This is true no matter how small the cell is
- Generalize with n observation

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

The Dirichlet Process and Clustering

$$G|\theta_1, \ldots, \theta_n \sim \mathrm{DP}\left(\alpha + n, \frac{\alpha G_0 + \sum_{i=1}^n \delta_{\theta_i}}{\alpha + n}\right)$$

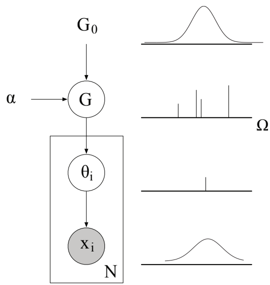
$$\mathrm{E}[G(A)|\theta_1, \ldots, \theta_n] = \frac{\alpha G_0 + \sum_{i=1}^n \delta_{\theta_i}}{\alpha + n}$$

- When $n \to \infty$,

$$\mathrm{E}[G(A)|\theta_1, \ldots, \theta_n] = \sum_{k=1}^{\infty} \pi_k \delta_{\Phi_k}(A)$$

- Φ_k is the "kth cluster" of unique values of θ_i
- $\pi_k = \lim_{n\to\infty} n_k/n$, n_k is number of "data point" in Φ_k

This suggests that random measure $G \sim \mathrm{DP}(\alpha, G_0)$ are discrete with probability 1



Blackwell-MacQueen Urn Scheme

- Blackwell-MacQueen urn scheme produces a sequence $\theta_1, \theta_2 ...$ with the following conditionals
- 1st step

$$
\begin{array}{rrcl rcl}
 & \theta_1 | G & \sim & G & G & \sim & \mathrm{DP}(\alpha, G_0) \\
\Rightarrow & \theta_1 & \sim & G_0 & G|\theta_1 & \sim & \mathrm{DP}\left(\alpha+1, \frac{\alpha G_0 + \delta_{\theta_1}}{\alpha+1}\right)
\end{array}
$$

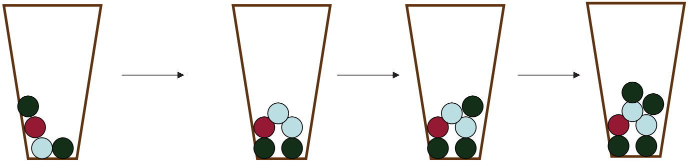
- 2nd step

$$
\begin{array}{rrcl rcl}
 & \theta_2 | \theta_1, G & \sim & G & & & \\
\Rightarrow & \theta_2 | \theta_1 & \sim & \frac{\alpha G_0 + \delta_\theta}{\alpha+1} & G|\theta_1, \theta_2 & \sim & \mathrm{DP}\left(\alpha+2, \frac{\alpha G_0 + \delta_{\theta_1} + \delta_{\theta_2}}{\alpha+1}\right)
\end{array}
$$

- nth step

$$
\begin{array}{rrcl rcl}
 & \theta_n | \theta_{1:n}, G & \sim & G & & & \\
\Rightarrow & \theta_n | \theta_{1:n} & \sim & \frac{\alpha G_0 + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha+n-1} & G|\theta_{1:n} & \sim & \mathrm{DP}\left(\alpha+n, \frac{\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha+n}\right)
\end{array}
$$

Blackwell-MacQueen Urn Scheme



Picking balls of different colors from an urn

- Start with no balls in the urn
- With prob. $\propto \alpha$ draw color $\theta_n \sim G_0$ and add a ball of that color into the urn
- With prob. $\propto n - 1$ pick a ball at random from the urn, record θ_n to be its color then place the ball with another ball of same color into urn

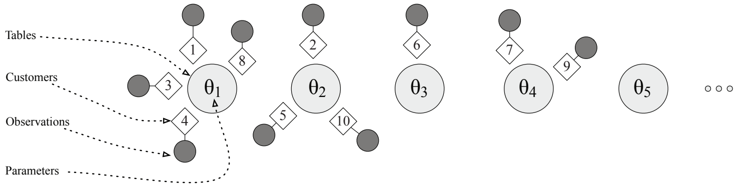
Blackwell-MacQueen Urn Scheme

- Starting with a DP, we constructed Blackwell-MacQueen urn scheme
- The reverse is possible using de Finetti's Theorem
  - The joint probability distribution underlying the data is invariant to permutation

$$p(\theta_1, ..., \theta_n) = \int \prod_{i=1}^{n} p(\theta_i | G) dP(G)$$

- Since θ_i are iid $\sim G$, they are exchangeable
- Thus a distribution over measures must exist making them i.i.d
- This is DP

Chinese Restaurant Process



- The first customer sits at the first table
- Assume K occupied tables, n customers, and n_k customers sit at table k
- mth subsequent customer sits at a table

$$\begin{aligned} k \text{ with prob.} &= n_k/(n-1+\alpha) \\ K+1 \text{ with prob.} &= \alpha/(n-1+\alpha) \end{aligned}$$

- Each table k has a value Φ_k drawn from a base distribution G_0
- Customer's value θ_n is assigned from his table's value

Chinese Restaurant Process

A random process in which n customers sit down in a Chinese restaurant with an infinite number of tables.

- Each table k has a value Φ_k drawn from a base distribution G_0
- Customer's value θ_n is assigned from his table's value

$$\theta_n|\theta_1, ..., \theta_{n-1}, G_0, \alpha \sim \frac{\alpha G_0}{n-1+\alpha} + \frac{\sum_{k=1}^{K} n_k \delta_{\Phi_k}}{n-1+\alpha}$$

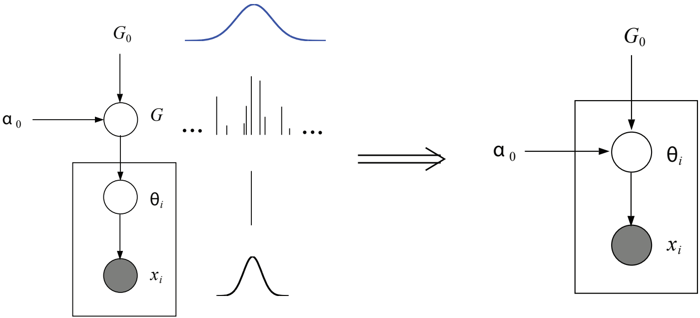
- CRP is the corresponding distribution over partitions, so CRP is exchangeable

$$p(\theta_1, ..., \theta_n) = \int \prod_{i=1}^{n} p(\theta_i|G)dP(G)$$

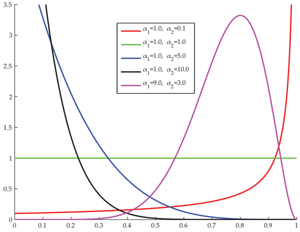
- If the DP is the prior on G, then the CRP is obtained when we integrate out G

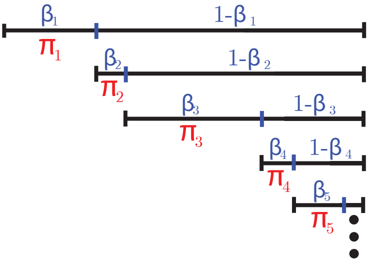
Chinese Restaurant Process

- If the DP is the prior on G, then the CRP is obtained when we integrate out G



Stick-breaking Process





- Define a sequence of Beta random variables $\beta_k \sim \text{Beta}(1, \alpha)$

- Define a sequence of mixing proportions

$$\pi_k = \beta_k \prod_{l=1}^{k=1} (1 - \beta_l)$$

Stick-breaking Process

- We can easily see $\sum_{k=1}^{\infty} \pi_k = 1$

$$
\begin{aligned}
1 - \sum_{k=1}^{K} \pi_k &= 1 - \beta_1 - \beta_2(1 - \beta_1) - \dots \\
&= (1 - \beta_1)(1 - \beta_2) - \beta_3(1 - \beta_1)(1 - \beta_2) - \dots \\
&= \prod_{k=1}^{K} (1 - \beta_k)
\end{aligned}
$$

- $G = \sum_{k=1}^{\infty} \pi_k \Phi_k$ has a clean definition of a random measure
- It is proved that G is a Dirichlet process

Stick-breaking Construction

$$G|\theta \sim \mathrm{DP}\left(\alpha+1, \frac{\alpha G_0 + \delta_\theta}{\alpha+1}\right)$$

- Given observation θ, consider a partition $(\theta, \Omega\backslash\theta)$

$$\begin{aligned}
(G(\theta), G(\Omega\backslash\theta)) &\sim \text{Dirichlet}\left((\alpha+1)\frac{\alpha G_0+\delta_\theta}{\alpha+1}(\theta), (\alpha+1)\frac{\alpha G_0+\delta_\theta}{\alpha+1}(\Omega\backslash\theta)\right) \\
&= \text{Dirichlet}(1, \alpha) \\
\Rightarrow G &= \beta\delta_\theta + (1-\beta)G' \qquad \text{with } \beta \sim \text{Beta}(1, \alpha)
\end{aligned}$$

- Agglomerative property of Dirichlet distributions implies $G' \sim \mathrm{DP}(\alpha, G_0)$
- Given observation θ'

$$G = \beta\delta_\theta + (1-\beta)(\beta'\delta_{\theta'} + (1-\beta')G'')$$

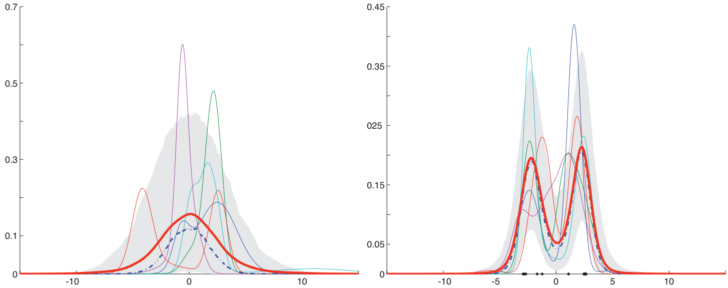
Density Estimation

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

- Problem: G is a discrete distribution; in particular it has no density!
- Solution: Convolve the DP with a smooth distribution

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, G_0) \\ F_x(\cdot) &= \int F(\cdot|\theta)dG(\theta) \\ x_i &\sim F_x \end{aligned} \quad \Rightarrow \quad \begin{aligned} G &= \sum_{k=1}^{\infty} \pi_k \delta_{\Phi_k} \\ F_x(\cdot) &= \sum_{k=1}^{\infty} \pi_k F(\cdot|\delta_{\Phi_k}) \\ x_i &\sim F_x \end{aligned}$$

Density Estimation



Dirichlet Process Mixture

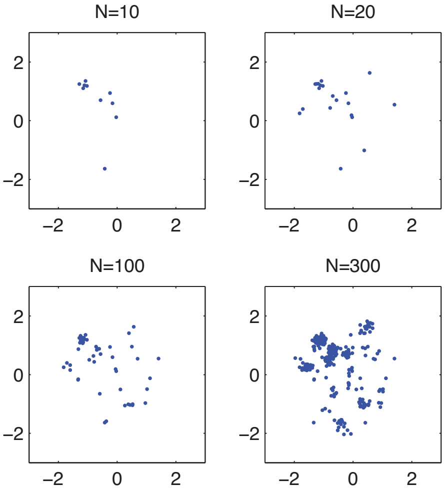
- DPs are discrete with probability one, so they are not suitable for use as a prior on continuous densities
- In a Dirichlet Process Mixture, we draw the parameters of a mixture model from a draw from a DP
- In mixture models setting, θ_i is the parameter associated with data point x_i
- Dirichlet process is prior on θ_i

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, G_0) \\ \theta_i | G &\sim G \\ x_i | \theta_i &\sim F(\cdot | \theta_i) \end{aligned}$$

- For example, if $F(\cdot|\theta_i)$ is a Gaussian density with parameters θ_i, then we have a Dirichlet Process Mixture of Gaussians

Samples from a DP Mixture of Gaussians

- More structure (clusters) appear as you draw more points



Clustering with Dirichlet Process Mixture

$$
\begin{aligned}
G &= \sum_{k=1}^{\infty} \pi_k \delta_{\Phi_k} \\
F_x(\cdot) &= \sum_{k=1}^{\infty} \pi_k F(\cdot|\delta_{\Phi_k}) \\
x_i &\sim F_x
\end{aligned}
$$

- The above model equivalent to

$$
\begin{aligned}
z_i &= \text{Multinomial}(\pi) \\
\theta_i &= \Phi_{z_i} \\
x_i|z_i &\sim F(\cdot|\Phi_{z_i})
\end{aligned}
$$

Clustering with Dirichlet Process Mixture

- DP mixture models are used in a variety of clustering applications, where the number of clusters is not known a priori
- They are also used in applications in which we believe the number of clusters grows without bound as the amount of data grows
- DPs have also found uses in applications beyond clustering, where the number of latent objects is not known or unbounded

Monte Carlo Integration

- We want to to compute the integral

$$I = \int h(x)f(x)dx$$

where $f(x)$ is a probability density function

- We can approximate this as

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N} h(X_i)dx$$

where $X_1, ..., X_N$ are sampled from f

Makov Chain Monte Carlo

- Random variable is a Markov process if the transition probabilities depends only on the random variable's current state

$$Pr(X_{t+1} = s_j | X_0 = s_k, ..., X_t = s_j) = Pr(X_{t+1} = s_j | X_t = s_j)$$

- Problem in Monte Carlo integration is sampling from some complex probability distribution $f(x)$
- Attempts to solve this problem are the roots of MCMC methods
- Metropolis-Hastings algorithm use an arbitrary transition probability function $q(X_t | X_{t-1})$, and setting the acceptance probability for a candidate point

$$\alpha = \min\left(\frac{f(X^*)q(X_{t-1}|X^*)}{f(X_{t-1})q(X^*|X_{t-1})}, 1\right)$$

where X^* is candidate point sampled from $q(X^*|X_{t-1})$

Gibbs Sampling

- A special case of Metropolis-Hastings sampling wherein the random candidate value is always accepted
- The task remains to specify how to construct a Markov Chain whose values converge to the target distribution
- The key to the Gibbs sampler is that one only considers univariate conditional distributions
- Consider a bivariate random variable (x, y), the sampler proceeds as follows

$$
\begin{aligned}
x_i &\sim f(x|y = y_{i-1}) \\
y_i &\sim f(y|x = x_i)
\end{aligned}
$$

Variational Inference

- Problem of MCMC methods are they can be slow to converge and their convergence can be difficult to diagnose
- The basic idea of variational inference is to formulate the computation of a marginal or conditional probability in terms of an optimization problem
- In Bayesian setting, we are usually interested in the posterior of $p(m|D, \theta)$
- In variational inference, we define an alternative family of distributions $q(m|\nu)$ where ν is called free variational parameters
- The optimization problem we want to solve is

$$\arg\min_{q} KL\left[q(m|\nu) || p(m|D, \theta)\right]$$

Inference for Dirichlet Process Mixtures

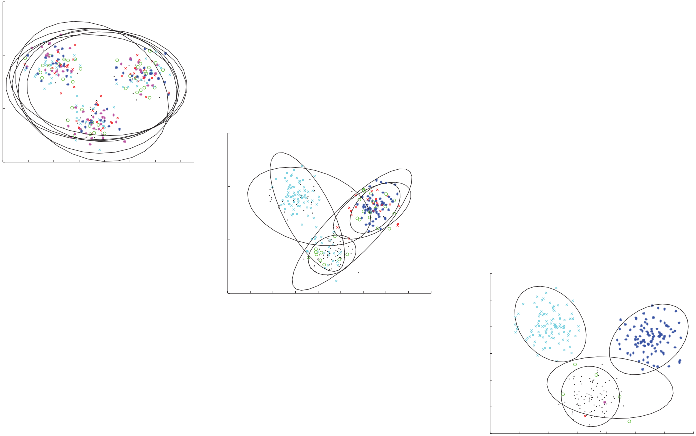
The posterior distribution under DP mixture models cannot be computed efficiently in any direct way. It must be approximated.

- Gibbs sampling (e.g. Escobar and West, 1995; Neal, 2000; Rasmussen, 2000)
- Variational approximation (Blei and Jordan, 2005)
- Expectation propagation (Minka and Ghahramani, 2003)

MCMC for Dirichlet Process Mixtures

- Key insight is to take advantage of exchangeability
- E.g. in CRP table that customer i sit is conditional on the seating choices of all other customers
- Easy when customer i is the last customer to arrive
- By exchangeability, can swap customer i with the final customer

MCMC for Dirichlet Process Mixtures



Variational Inference for Dirichlet Process Mixtures

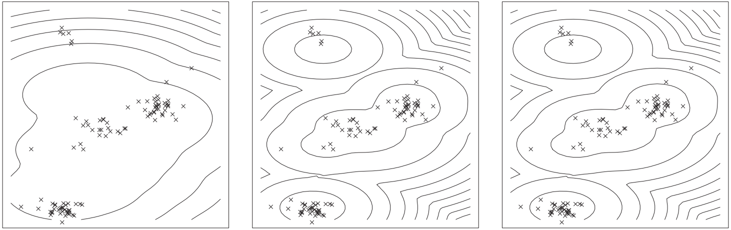
- Truncated DP: stick-breaking with fixing a value T and make $\beta_T = 1$
- Implies $\pi_k = 0$ with $k > T$

$$G_T = \sum_{k=1}^{T} \pi_k \delta_{\Phi_k}$$

is known as Truncated DP

- G_T is used to learn G

Variational Inference for Dirichlet Process Mixtures



Initial state 1st iteration 5th (and last) iteration

Summary

- Nonparametric Bayesian models allow for much flexibility, but need to place measures on measures
- The most important setting is the Dirichlet mixture model which are mixture models with countably infinite number of components
- Dirichlet process is “just” a glorified Dirichlet distribution
- Draws from a DP are probability measures consisting of a weighted sum of point masses
- Many representations: Blackwell-MacQueen urn scheme, Chinese restaurant process, stick-breaking construction
- Development of approximation for DP mixtures has enabled its application to practical data analysis problem

Thank you!