An Introduction to Bayesian Nonparametric Modelling

Yee Whye Teh

Gatsby Computational Neuroscience Unit
University College London

April 22, 2009 / MSR Cambridge

Outline

Bayesian Nonparametric Modelling

Measure Theoretic Probability Theory

Gaussian Processes

Dirichlet Processes

Extensions and Generalizations of Dirichlet Processes

Summary

Outline

Bayesian Nonparametric Modelling

Measure Theoretic Probability Theory

Gaussian Processes

Dirichlet Processes

Extensions and Generalizations of Dirichlet Processes

Summary

Modelling Data

All models are wrong, but some are useful.

—George E. P. Box, Norman R. Draper (1987).

- Models are never correct for real world data.
- How do we deal with model misfit?
  - Quantify closeness to true model, and optimality of fitted model;
  - Model selection or averaging;
  - Increase the flexibility of your model class.
- Bayesian nonparametrics are good solutions from the second and third perspectives.

Nonparametric Modelling

- What is a nonparametric model?
  - A parametric model where the number of parameters increases with data;
  - A really large parametric model;
  - A model over infinite dimensional function or measure spaces.
  - A family of distributions that is dense in some large space.
- Why nonparametric models in Bayesian theory of learning?
  - broad class of priors that allows data to “speak for itself”;
  - side-step model selection and averaging.
- How do we deal with the very large parameter spaces?
  - Marginalize out all but a finite number of parameters;
  - Define infinite space implicitly (akin to the kernel trick) using either Kolmogorov Consistency Theorem or de Finetti’s theorem.

Classification and Regression

- Learn a mapping $f : \mathbb{X} \rightarrow \mathbb{Y}$.

  Data: Pairs of data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

  Model: $y_i | x_i, w \sim F(x_i, w) + \epsilon$

  Classification: $\mathbb{Y} = \{+1, -1\}$ or $\{1, \ldots, C\}$.

  Regression: $\mathbb{Y} = \mathbb{R}$

- Prior over parameters

$$p(w)$$

- Posterior over parameters

$$p(w|\mathbf{x}, \mathbf{y}) = \frac{p(w)p(\mathbf{y}|\mathbf{x}, w)}{p(\mathbf{y}|\mathbf{x})}$$

- Prediction with posterior:

$$p(y_\star | x_\star, \mathbf{x}, \mathbf{y}) = \int p(y_\star | x_\star, w) p(w|\mathbf{x}, \mathbf{y}) dw$$

Nonparametric Classification and Regression

- Learn a mapping $f : \mathbb{X} \rightarrow \mathbb{Y}$.

  Data: Pairs of data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

  Model: $y_i | x_i, f \sim f(x_i) + \epsilon$

  Classification: $\mathbb{Y} = \{+1, -1\}$ or $\{1, \ldots, C\}$.

  Regression: $\mathbb{Y} = \mathbb{R}$

- Prior over parameters

$$p(f)$$

- Posterior over parameters

$$p(f|\mathbf{x}, \mathbf{y}) = \frac{p(f)p(\mathbf{y}|\mathbf{x}, f)}{p(\mathbf{y}|\mathbf{x})}$$

- Prediction with posterior:

$$p(y_\star|x_\star, \mathbf{x}, \mathbf{y}) = \int p(y_\star|x_\star, f)p(f|\mathbf{x}, \mathbf{y})df$$

Density Estimation

- Parametric density estimation (e.g. Gaussian, mixture models)
  Data: $\mathbf{x} = \{x_1, x_2, \dots\}$
  Model: $x_i | w \sim F(w)$
- Prior over parameters

$$p(w)$$

- Posterior over parameters

$$p(w|\mathbf{x}) = \frac{p(w)p(\mathbf{x}|w)}{p(\mathbf{x})}$$

- Prediction with posterior

$$p(x_\star|\mathbf{x}) = \int p(x_\star|w)p(w|\mathbf{x})\,dw$$

Nonparametric Density Estimation

- Nonparametric density estimation

  Data: $\mathbf{x} = \{x_1, x_2, \ldots\}$

  Model: $x_i | f \sim f$

- Prior over densities

$$p(f)$$

- Posterior over densities

$$p(f|\mathbf{x}) = \frac{p(f)p(\mathbf{x}|w)}{p(\mathbf{x})}$$

- Prediction with posterior

$$p(x_\star|\mathbf{x}) = \int f(x_\star)p(f|\mathbf{x})\,df$$

Other Tutorials on Bayesian Nonparametrics

- Zoubin Gharamani, UAI 2005.
- Michael Jordan, NIPS 2005.
- Volker Tresp, ICML nonparametric Bayes workshop 2006.
- Peter Orbanz, Foundations of Nonparametric Bayesian Methods, 2009.
- My Machine Learning Summer School 2007 tutorial and practical course.
- I have an introduction to Dirichlet processes [Teh 2007], and another to hierarchical Bayesian nonparametric models [Teh and Jordan 2009].

Outline

Bayesian Nonparametric Modelling

Measure Theoretic Probability Theory

Gaussian Processes

Dirichlet Processes

Extensions and Generalizations of Dirichlet Processes

Summary

A Tiny Bit of Measure Theoretic Probability Theory

- A σ*-algebra* Σ is a family of subsets of a set Θ such that
  - Σ is not empty;
  - If $A \in \Sigma$ then $\Theta \backslash A \in \Sigma$;
  - If $A_1, A_2, \ldots \in \Sigma$ then $\cup_{i=1}^{\infty} A_i \in \Sigma$.
- (Θ, Σ) is a *measure space* and $A \in \Sigma$ are the *measurable sets*.
- A *measure* μ over (Θ, Σ) is a function $\mu : \Sigma \to [0, \infty]$ such that
  - $\mu(\emptyset) = 0$;
  - If $A_1, A_2, \ldots \in \Sigma$ are disjoint then $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.
  - Everything we consider here will be measurable.
  - A probability measure is one where $\mu(\Theta) = 1$.
- Given two measure spaces (Θ, Σ) and (Δ, Φ), a function $f : \Theta \to \Delta$ is *measurable* if $f^{-1}(A) \in \Sigma$ for every $A \in \Phi$.

A Tiny Bit of Measure Theoretic Probability Theory

- If p is a probability measure on (Θ, Σ), a *random variable* X taking values in Δ is simply a measurable function $X : \Theta \to \Delta$.
  - Think of the probability space (Θ, Σ, p) as a black-box random number generator, and X as a function taking random samples in Θ and producing random samples in Δ.
  - The probability of an event $A \in \Phi$ is $p(X \in A) = p(X^{-1}(A))$.
- A *stochastic process* is simply a collection of random variables $\{X_i\}_{i \in \mathbb{I}}$ over the same measure space (Θ, Σ), where $\mathbb{I}$ is an index set.
  - What distinguishes a stochastic process from, say, a graphical model is that $\mathbb{I}$ can be infinite, even uncountably so.
  - This raises issues of how do you even define them and how do you ensure that they can even existence (mathematically speaking).
- Stochastic processes form the core of many Bayesian nonparametric models.
  - Gaussian processes, Poisson processes, gamma processes, Dirichlet processes, beta processes...

Outline

Bayesian Nonparametric Modelling

Measure Theoretic Probability Theory

Gaussian Processes

- Consistency
- Poisson Processes
- Gamma Processes

Dirichlet Processes

Extensions and Generalizations of Dirichlet Processes

Summary

Gaussian Processes

- A *Gaussian process* (GP) is a random function $f : \mathbb{X} \to \mathbb{R}$ such that for any finite set of input points $x_1, \ldots, x_n$,

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} c(x_1, x_1) & \ldots & c(x_1, x_n) \\ \vdots & \ddots & \vdots \\ c(x_n, x_1) & \ldots & c(x_n, x_n) \end{bmatrix} \right)$$

  where the parameters are the mean function $m(x)$ and covariance kernel $c(x, y)$.

- GPs can be visualized by iterative sampling $f(x_n)|f(x_1), \ldots, f(x_{n-1})$ on a sequence of input points $x_1, x_2, \ldots$
  - Demonstration.
- Note: a random function f is a stochastic process. It is a collection of random variables $\{f(x)\}_{x \in \mathbb{X}}$ one for each possible input value x.

[Rasmussen and Williams 2006]

Posterior and Predictive Distributions

- How do we compute the posterior and predictive distributions?
- Training set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and test input x_{n+1}.
- Out of the (uncountably infinitely) many random variables $\{f(x)\}_{x \in \mathbb{X}}$ making up the GP only $n+1$ has to do with the data:

$$f(x_1), f(x_2), \ldots, f(x_{n+1})$$

- Training data gives observations $f(x_1) = y_1, \ldots, f(x_n) = y_n$. The predictive distribution of $f(x_{n+1})$ is simply

$$p(f(x_{n+1}) | f(x_1) = y_1, \ldots, f(x_n) = y_n)$$

  which is easy to compute since $f(x_1), \ldots, f(x_{n+1})$ is Gaussian.
- This can be generalized to noisy observations $y_i = f(x_i) + \epsilon_i$ or non-linear effects $y_i \sim D(f(x_i))$ where $D(\theta)$ is a distribution parametrized by θ.

Consistency and Existence

- The definition of Gaussian processes only give finite dimensional marginal distributions of the stochastic process.
- Fortunately these marginal distributions are *consistent*.
  - For every finite set $\mathbf{x} \subset \mathbb{X}$ we have a distinct distribution $p_{\mathbf{x}}([f(x)]_{x \in \mathbf{x}})$. These distributions are said to be consistent if

    $$p_{\mathbf{x}}([f(x)]_{x \in \mathbf{x}}) = \int p_{\mathbf{x} \cup \mathbf{y}}([f(x)]_{x \in \mathbf{x} \cup \mathbf{y}}) d[f(x)]_{x \in \mathbf{y}}$$

    for disjoint and finite $\mathbf{x}, \mathbf{y} \subset \mathbb{X}$.
  - The marginal distributions for the GP are consistent because *Gaussians are closed under marginalization*.
- The *Kolmogorov Consistency Theorem* guarantees existence of GPs, i.e. the whole stochastic process $\{f(x)\}_{x \in \mathbb{X}}$.
  - For further information see [Orbanz 2009].

Poisson Processes

- A *Poisson process* (PP) is a random function $f : \Sigma \to \mathbb{R}$ such that:
  - Σ is the σ-algebra over $\mathbb{X}$.
  - For any measurable set $A \subset \mathbb{X}$,

    $$f(A) \sim \text{Poisson}(\lambda(A)),$$

    where the parameter is the rate measure λ (a function from the measurable sets of $\mathbb{X}$ to $\mathbb{R}_+$).
  - And if $A, B \subset \mathbb{X}$ are disjoint then $f(A)$ and $f(B)$ are independent.
- The above family of distributions is consistent, since the sum of two independent Poisson variables is still Poisson with the rate parameter being the sum of the individual rates.
- Note that f is also a measure, a *random measure*. It always consists of point masses:

  $$f = \sum_{i=1}^{n} \delta_{x_i}$$

  where $x_1, x_2, \ldots \in \mathbb{X}$ and $n \sim \text{Poisson}(\lambda(\mathbb{X}))$, i.e. f is a *point process*.

Gamma Processes

- A *Gamma process* (ΓP) is a random function $f : \Sigma \to \mathbb{R}$ such that:
  - For any measurable set $A \subset \mathbb{X}$,

    $$f(A) \sim \text{Gamma}(\lambda(A), 1),$$

    where the parameter is the shape measure λ.
  - And if $A, B \subset \mathbb{X}$ are disjoint then $f(A)$ and $f(B)$ are independent.
- The above family of distributions is also consistent, since the sum of two independent gamma variables (with same scale parameter 1) is still gamma with the shape parameter being the sum of the individual shape parameters.
- f is also a random measure. It always consists of weighted point masses:

  $$f = \sum_{i=1}^{\infty} w_i \delta_{x_i}$$

  with total weight $\sum_{i=1}^{\infty} w_i \sim \text{Gamma}(\lambda(\mathbb{X}))$.

Outline

Bayesian Nonparametric Modelling

Measure Theoretic Probability Theory

Gaussian Processes

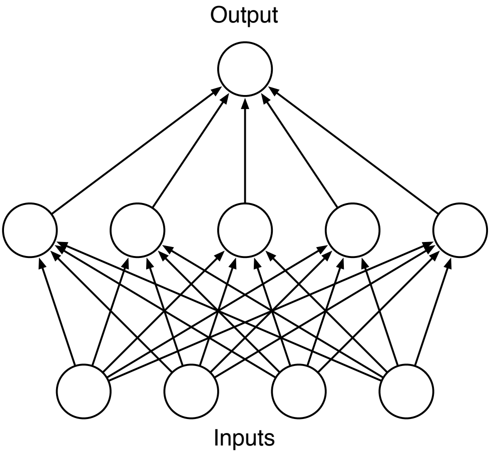
Dirichlet Processes

- Representations of Dirichlet Processes
- Applications of Dirichlet Processes
- Exchangeability

Extensions and Generalizations of Dirichlet Processes

Summary

Regression with Infinite Numbers of Features



- Bayesian neural networks with infinite numbers of features give rise to GPs.

[Neal 1994]

Infinite Mixture Models

- Mixture of K clusters:

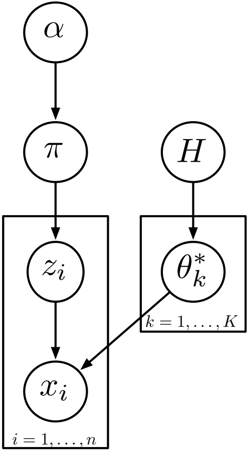
$$z_i | \boldsymbol{\pi} \sim \text{Multinomial}(\boldsymbol{\pi})$$
$$x_i | z_i, \theta^*_{z_i} \sim F(\theta_{z_i})$$

- Being Bayesian we place priors on parameters:

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\tfrac{\alpha}{K}, \ldots, \tfrac{\alpha}{K})$$
$$\theta^*_k \sim H$$

- Now we somehow take $K \to \infty$.

[Rasmussen 2000]



Infinite Mixture Models

- Assume that H is conjugate to F.
- We can integrate out parameters and Gibbs sample z_i's:

$$p(z_i = k|\mathbf{z}^{-i}) = \frac{n_k^{-i} + \frac{\alpha}{K}}{n - 1 + \alpha} f(x_i|\{x_j : j \neq i, z_j = k\})$$

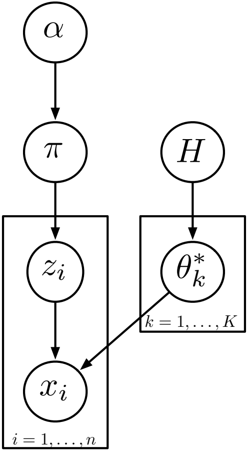
- We will assume K is very large, so many clusters will in fact be *empty*.
- We can lump these empty clusters together.

Occupied clusters:

$$p(z_i = k|\mathbf{z}^{-i}) = \frac{n_k^{-i} + \frac{\alpha}{K}}{n - 1 + \alpha} f(x_i|\{x_j : j \neq i, z_j = k\})$$

Empty clusters:

$$p(z_i = k_{\text{empty}}|\mathbf{z}^{-i}) = \frac{\alpha\frac{K-K^*}{K}}{n - 1 + \alpha} f(x_i|\{\})$$



Infinite Mixture Models

- As $K \to \infty$ the Gibbs updates simplify:

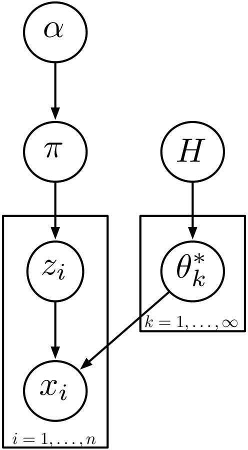
  Occupied clusters:

  $$p(z_i = k | \mathbf{z}^{-i}) = \frac{n_k^{-i}}{n - 1 + \alpha} f(x_i | \{x_j : j \neq i, z_j = k\})$$

  Empty clusters:

  $$p(z_i = k_{\text{empty}} | \mathbf{z}^{-i}) = \frac{\alpha}{n - 1 + \alpha} f(x_i | \{\})$$

- These are Gibbs updates for Dirichlet process mixture models.
- *Dirichlet processes* can be thought of as infinite dimensional Dirichlet distributions.



Infinite Mixture Models

- The actual infinite limit of finite mixture models does not make sense: any particular component will get a mixing proportion of 0.
- In the Gibbs sampler we bypassed this by lumping empty clusters together.
- Other better ways of making this infinite limit precise:
  - Look at the prior clustering structure induced by the Dirichlet prior over mixing proportions—*Chinese restaurant process*.
  - Re-order components so that those with larger mixing proportions tend to occur first, before taking the infinite limit—*stick-breaking construction*.

Dirichlet Distributions

- A *Dirichlet distribution* is a distribution over the K-dimensional probability simplex:

$$\Delta_K = \left\{(\pi_1, \dots, \pi_K) \,:\, \pi_k \geq 0, \sum_k \pi_k = 1\right\}$$

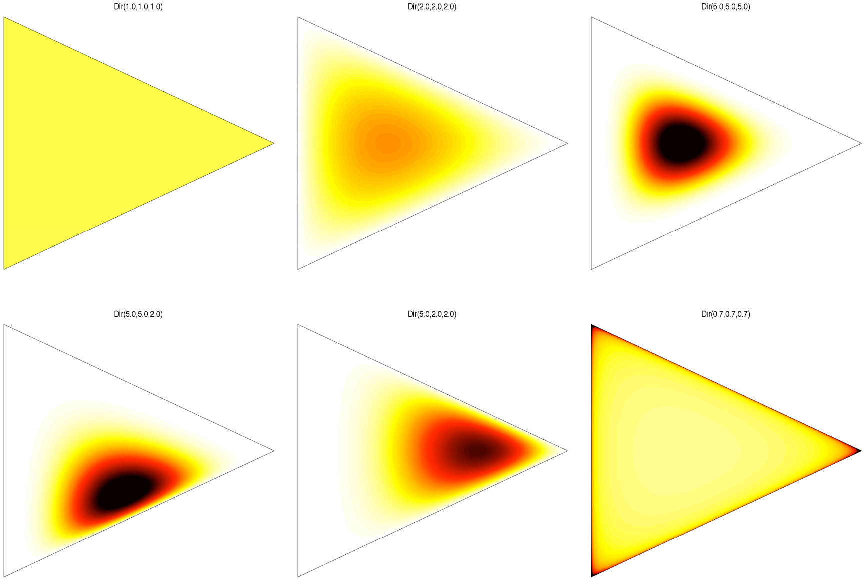
- We say $(\pi_1, \dots, \pi_K)$ is Dirichlet distributed,

$$(\pi_1, \dots, \pi_K) \sim \text{Dirichlet}(\lambda_1, \dots, \lambda_K)$$

  with parameters $(\lambda_1, \dots, \lambda_K)$, if

$$p(\pi_1, \dots, \pi_K) = \frac{\Gamma(\sum_k \lambda_k)}{\prod_k \Gamma(\lambda_k)} \prod_{k=1}^{n} \pi_k^{\lambda_k - 1}$$

Dirichlet Distributions



Dir(1.0,1.0,1.0)

Dir(2.0,2.0,2.0)

Dir(5.0,5.0,5.0)

Dir(5.0,5.0,2.0)

Dir(5.0,2.0,2.0)

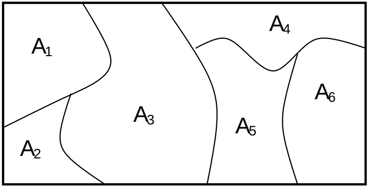
Dir(0.7,0.7,0.7)

Dirichlet Processes

- A *Dirichlet Process* (DP) is a random probability measure G over (Θ, Σ) such that for any finite set of measurable partitions $A_1 \dot\cup \ldots \dot\cup A_K = \Theta$,

$$(G(A_1), \ldots, G(A_K)) \sim \text{Dirichlet}(\lambda(A_1), \ldots, \lambda(A_K))$$

  where λ is a base measure.



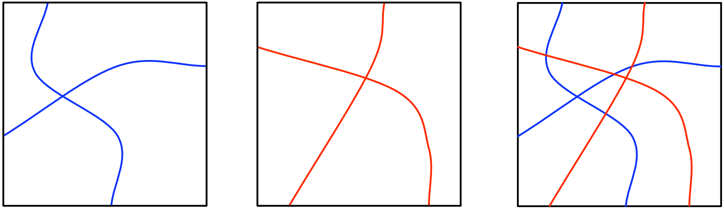
- The above family of distributions is consistent (next slide), and Kolmogorov Consistency Theorem can be applied to show existence (but there are technical conditions restricting the generality of the definition).

[Ferguson 1973, Blackwell and MacQueen 1973]

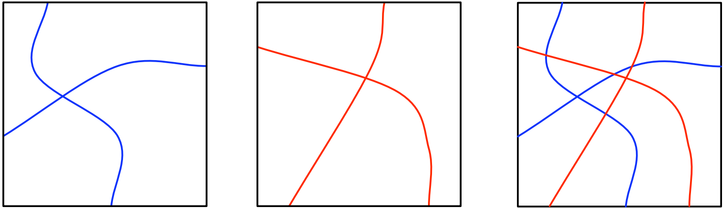
Consistency of Dirichlet Marginals

- If we have two partitions $(A_1, \ldots, A_K)$ and $(B_1, \ldots, B_J)$ of Θ, how do we see if the two Dirichlets are consistent?
- Because Dirichlet variables are normalized gamma variables and sums of gammas are gammas, if $(I_1, \ldots, I_j)$ is a partition of $(1, \ldots, K)$,

$$\left(\sum_{i \in I_1} \pi_i, \ldots, \sum_{i \in I_j} \pi_i\right) \sim \text{Dirichlet}\left(\sum_{i \in I_1} \lambda_i, \ldots, \sum_{i \in I_j} \lambda_i\right)$$



Consistency of Dirichlet Marginals



- Form the common refinement $(C_1, \ldots, C_L)$ where each C_ℓ is the intersection of some A_k with some B_j. Then:

$$\begin{aligned}
\text{By definition, } (G(C_1), \ldots, G(C_L)) &\sim \text{Dirichlet}(\lambda(C_1), \ldots, \lambda(C_L)) \\
(G(A_1), \ldots, G(A_K)) &= \left(\textstyle\sum_{C_\ell \subset A_1} G(C_\ell), \ldots, \sum_{C_\ell \subset A_K} G(C_\ell)\right) \\
&\sim \text{Dirichlet}(\lambda(A_1), \ldots, \lambda(A_K)) \\
\text{Similarly, } (G(B_1), \ldots, G(B_J)) &\sim \text{Dirichlet}(\lambda(B_1), \ldots, \lambda(B_J))
\end{aligned}$$

  so the distributions of $(G(A_1), \ldots, G(A_K))$ and $(G(B_1), \ldots, G(B_J))$ are consistent.

  - Demonstration.

Parameters of Dirichlet Processes

- Usually we split the λ base measure into two parameters $\lambda = \alpha H$:
  - *Base distribution* H, which is like the *mean* of the DP.
  - *Strength parameter* α, which is like an *inverse-variance* of the DP.
- We write:

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

if for any partition $(A_1, \ldots, A_K)$ of Θ:

$$(G(A_1), \ldots, G(A_K)) \sim \mathrm{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))$$

- The first and second moments of the DP:

$$\text{Expectation:} \qquad \mathbb{E}[G(A)] = H(A)$$

$$\text{Variance:} \qquad \mathbb{V}[G(A)] = \frac{H(A)(1 - H(A))}{\alpha + 1}$$

where A is any measurable subset of Θ.

Representations of Dirichlet Processes

- Draws from Dirichlet processes will always place all their mass on a countable set of points:

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

  where $\sum_k \pi_k = 1$ and $\theta_k^* \in \Theta$.

  - What is the joint distribution over $\pi_1, \pi_2, \ldots$ and $\theta_1^*, \theta_2^*, \ldots$?

- Since G is a (random) probability measure over Θ, we can treat it as a distribution and draw samples from it. Let

  $$\theta_1, \theta_2, \ldots \sim G$$

  be random variables with distribution G.

  - What is the marginal distribution of $\theta_1, \theta_2, \ldots$ with G integrated out?
  - There is positive probability that sets of θ_i's can take on the same value θ_k^* for some k, i.e. the θ_i's cluster together. How do these clusters look like?
  - For practical modelling purposes this is sufficient. But is this sufficient to tell us all about G?

Stick-breaking Construction

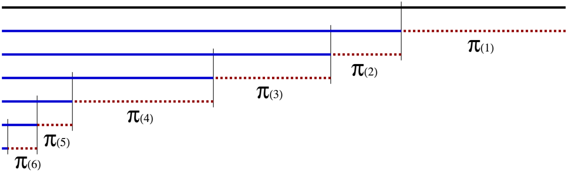
$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

- There is a simple construction giving the joint distribution of $\pi_1, \pi_2, \ldots$ and $\theta_1^*, \theta_2^*, \ldots$ called the *stick-breaking construction*.

$$\theta_k^* \sim H$$

$$\pi_k = v_k \prod_{i=1}^{k-1} (1 - v_i)$$

$$v_k \sim \text{Beta}(1, \alpha)$$



- Also known as the *GEM* distribution, write $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$.

[Sethuraman 1994]

Pólya Urn Scheme

$$\theta_1, \theta_2, \ldots \sim G$$

- The marginal distribution of $\theta_1, \theta_2, \ldots$ has a simple generative process called the *Pólya urn scheme*.

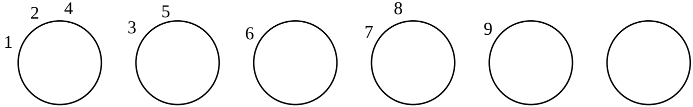
$$\theta_n | \theta_{1:n-1} \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}$$

- Picking balls of different colors from an urn:
  - Start with no balls in the urn.
  - with probability $\propto \alpha$, draw $\theta_n \sim H$, and add a ball of color θ_n into urn.
  - With probability $\propto n - 1$, pick a ball at random from the urn, record θ_n to be its color and return two balls of color θ_n into urn.
- Pólya urn scheme is like a "representer" for the DP—a finite projection of an infinite object G.
- Also known as the *Blackwell-MacQueen urn scheme*.

[Blackwell and MacQueen 1973]

Chinese Restaurant Process

- According to the Pólya urn scheme, and because G consists of weighted point masses, $\theta_1, \ldots, \theta_n$ take on $K < n$ distinct values, say $\theta_1^*, \ldots, \theta_K^*$.
- This defines a partition of $(1, \ldots, n)$ into K clusters, such that if i is in cluster k, then $\theta_i = \theta_k^*$.
- The distribution over partitions is a *Chinese restaurant process* (CRP).
- Generating from the CRP:
  - First customer sits at the first table.
  - Customer n sits at:
    - Table k with probability $\frac{n_k}{\alpha+n-1}$ where n_k is the number of customers at table k.
    - A new table $K + 1$ with probability $\frac{\alpha}{\alpha+n-1}$.
  - Customers $\Leftrightarrow$ integers, tables $\Leftrightarrow$ clusters.
- The CRP exhibits the *clustering property* of the DP.
  - *Rich-gets-richer* effect implies small number of large clusters.
  - Expected number of clusters is $K = O(\alpha \log n)$.



Density Estimation

- Parametric density estimation (e.g. Gaussian, mixture models)

  Data: $\mathbf{x} = \{x_1, x_2, \ldots\}$

  Model: $x_i | w \sim F(\cdot | w)$

- Prior over parameters

$$p(w)$$

- Posterior over parameters

$$p(w|\mathbf{x}) = \frac{p(w)p(\mathbf{x}|w)}{p(\mathbf{x})}$$

- Prediction with posteriors

$$p(x_\star|\mathbf{x}) = \int p(x_\star|w)p(w|\mathbf{x})\,dw$$

Density Estimation

- Bayesian nonparametric density estimation with Dirichlet processes

  Data: $\mathbf{x} = \{x_1, x_2, \ldots\}$

  Model: $x_i \sim G$

- Prior over distributions

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

- Posterior over distributions

$$p(G|\mathbf{x}) = \frac{p(G)p(\mathbf{x}|G)}{p(\mathbf{x})}$$

- Prediction with posteriors

$$p(x_\star|\mathbf{x}) = \int p(x_\star|G)p(G|\mathbf{x})\,dF = \int G(x_\star)p(G|\mathbf{x})\,dG$$

- *Not quite feasible, since G is a discrete distribution, in particular it has no density.*

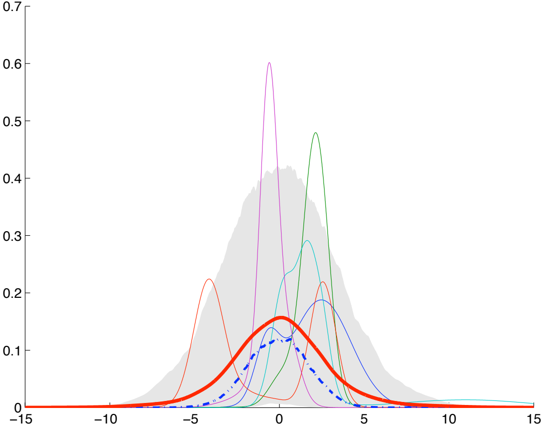
Density Estimation

- Solution: Convolve the DP with a smooth distribution:

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

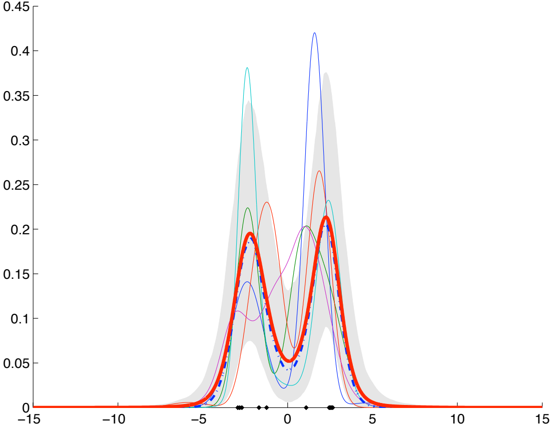
  - Demonstration.

Density Estimation



$F(\cdot|\mu,\Sigma)$ is Gaussian with mean μ, covariance Σ.
$H(\mu,\Sigma)$ is Gaussian-inverse-Wishart conjugate prior.
Red: mean density. Blue: median density. Grey: 5-95 quantile. Others: draws. Black: data points.

Density Estimation



$F(\cdot|\mu, \Sigma)$ is Gaussian with mean μ, covariance Σ.
$H(\mu, \Sigma)$ is Gaussian-inverse-Wishart conjugate prior.
Red: mean density. Blue: median density. Grey: 5-95 quantile. Others: draws. Black: data points.

Clustering

- Recall our approach to density estimation:

$$
\begin{aligned}
G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \sim \mathrm{DP}(\alpha, H) \\
F_x(\cdot) &= \sum_{k=1}^{\infty} \pi_k F(\cdot|\theta_k^*) \\
x_i &\sim F_x
\end{aligned}
$$

- Above model equivalent to:

$$
\begin{aligned}
z_i &\sim \mathrm{Discrete}(\boldsymbol{\pi}) \\
\theta_i &= \theta_{z_i}^* \\
x_i|z_i &\sim F(\cdot|\theta_i) = F(\cdot|\theta_{z_i}^*)
\end{aligned}
$$

- This is simply a mixture model with an *infinite* number of components. This is called a *DP mixture model*.

Clustering

- 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.
    - Nonparametric probabilistic context free grammars.
    - Visual scene analysis.
    - Infinite hidden Markov models/trees.
    - Haplotype inference.
    - ...
- In many such applications it is important to be able to model the same set of objects in different contexts.
- This corresponds to the problem of *grouped clustering* and can be tackled using *hierarchical Dirichlet processes*.

[Teh et al. 2006, Teh and Jordan 2009]

Semiparametric Modelling

- Example: linear regression model for inferring effectiveness of new medical treatments.

  $$y_{ij} = \beta^\top x_{ij} + b_i^\top z_{ij} + \epsilon_{ij}$$

  y_{ij} is outcome of jth trial on ith subject.

  x_{ij}, z_{ij} are predictors (treatment, dosage, age, health...).

  β are fixed-effects coefficients.

  b_i are random-effects subject-specific coefficients.

  ϵ_{ij} are noise terms.

- Care about inferring β. If x_{ij} is treatment, we want to determine $p(\beta > 0|\mathbf{x}, \mathbf{y}, \mathbf{z})$.
- Usually we assume Gaussian noise $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Is this a sensible prior? Over-dispersion, skewness,...
- May be better to model noise nonparametrically: $\epsilon_{ij} \sim F$.
- Also possible to model subject-specific random effects nonparametrically: $b_i \sim G$.

Exchangeability

- Instead of deriving the Pólya urn scheme by marginalizing out a DP, consider starting directly from the conditional distributions:

$$\theta_n|\theta_{1:n-1} \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}$$

- For any n, the joint distribution of $\theta_1, \ldots, \theta_n$ is:

$$p(\theta_1, \ldots, \theta_n) = \frac{\alpha^K \prod_{k=1}^{K} h(\theta_k^*)(m_{nk} - 1)!}{\prod_{i=1}^{n} i - 1 + \alpha}$$

  where $h(\theta)$ is density of θ under H, $\theta_1^*, \ldots, \theta_K^*$ are the unique values, and θ_k^* occurred m_{nk} times among $\theta_1, \ldots, \theta_n$.

- The joint distribution is *exchangeable* wrt permutations of $\theta_1, \ldots, \theta_n$.
- *De Finetti's Theorem* says that there must be a random probability measure G making $\theta_1, \theta_2, \ldots$ iid. This is the DP.

De Finetti's Theorem

Let $\theta_1, \theta_2, \ldots$ be an infinite sequence of random variables with joint distribution p. If for all $n \geq 1$, and all permutations $\sigma \in \Sigma_n$ on n objects,

$$p(\theta_1, \ldots, \theta_n) = p(\theta_{\sigma(1)}, \ldots, \theta_{\sigma(n)})$$

That is, the sequence is infinitely exchangeable. Then there exists a latent random parameter G such that:

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

where ρ is a joint distribution over G and θ_i's.

- θ_i's are *independent* given G.
- Sufficient to define G through the conditionals $p(\theta_n | \theta_1, \ldots, \theta_{n-1})$.
- G can be *infinite dimensional* (indeed it is often a *random measure*).

Outline

Bayesian Nonparametric Modelling

Measure Theoretic Probability Theory

Gaussian Processes

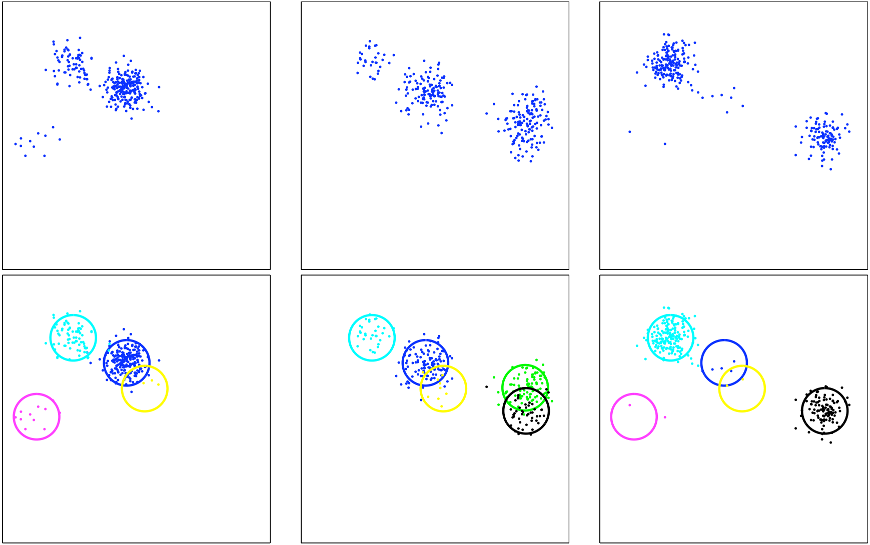
Dirichlet Processes

Extensions and Generalizations of Dirichlet Processes

- Hierarchical Dirichlet Processes
- Pitman-Yor Processes
- Indian Buffet and Beta Processes

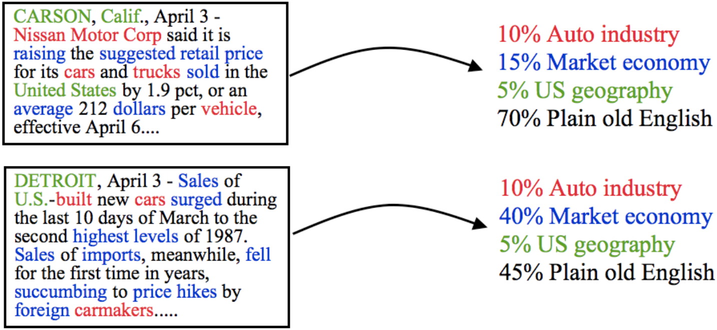
Summary

Hierarchical Dirichlet Processes



Hierarchical Dirichlet Processes

- We model documents as coming from an underlying set of topics.
  - Do not know the number of topics a priori—use DP mixtures somehow.
  - But: topics have to be shared across documents...



CARSON, Calif., April 3 - Nissan Motor Corp said it is raising the suggested retail price for its cars and trucks sold in the United States by 1.9 pct, or an average 212 dollars per vehicle, effective April 6....

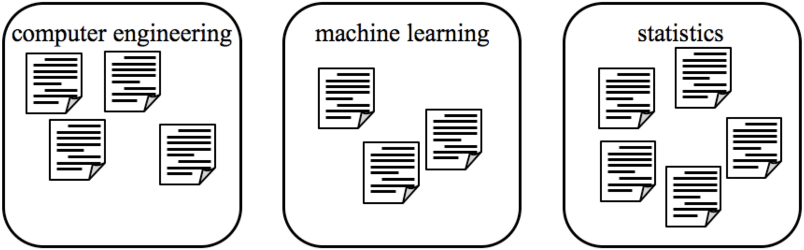
10% Auto industry
15% Market economy
5% US geography
70% Plain old English

DETROIT, April 3 - Sales of U.S.-built new cars surged during the last 10 days of March to the second highest levels of 1987. Sales of imports, meanwhile, fell for the first time in years, succumbing to price hikes by foreign carmakers.....

10% Auto industry
40% Market economy
5% US geography
45% Plain old English

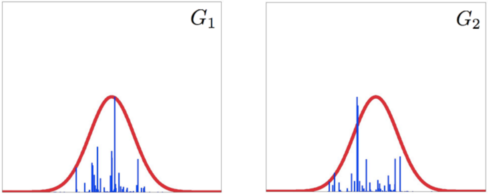
Hierarchical Dirichlet Processes

- Share topics across documents in a collection, and across different collections.
- More sharing within collections than across.
- Use DP mixture models as we do not know the number of topics a priori.



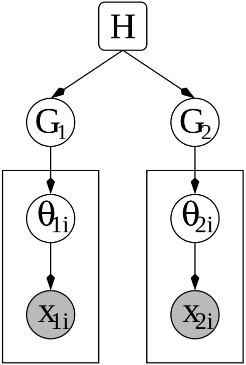
Hierarchical Dirichlet Processes

- Use a DP mixture for each group.



- Unfortunately there is no sharing of clusters across different groups because *H* is smooth.
- Solution: make the base distribution *H* discrete.
- Put a DP prior on the common base distribution.

[Teh et al. 2006]

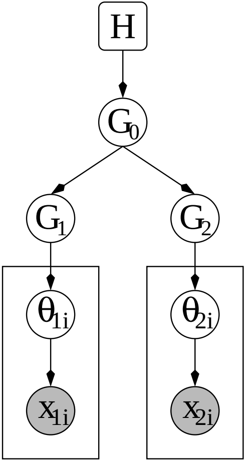


Hierarchical Dirichlet Processes

- A hierarchical Dirichlet process:

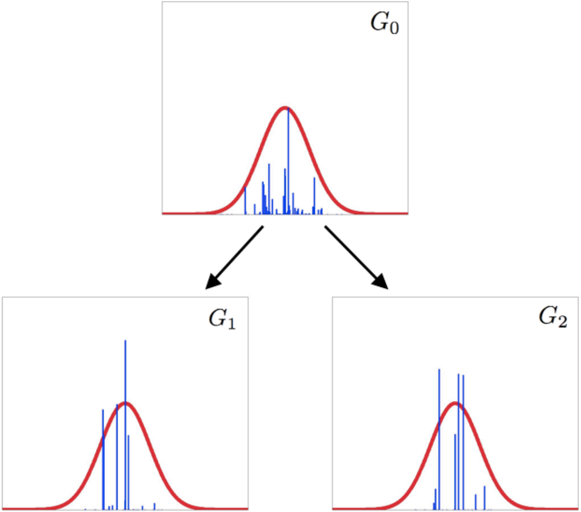
$$G_0 \sim \mathrm{DP}(\alpha_0, H)$$
$$G_1, G_2 | G_0 \sim \mathrm{DP}(\alpha, G_0) \text{ iid}$$

- Extension to larger hierarchies is straightforward.



Hierarchical Dirichlet Processes

- Making G_0 discrete forces shared cluster between G_1 and G_2.



Hierarchical Dirichlet Processes

- Document topic modelling:
    - Allows documents to be modelled with DP mixtures of topics, with topics shared across corpora.
- Infinite hidden Markov modelling:
    - Allows HMMs with an infinite number of states, with transitions from each allowable state to every other allowable state.
- Learning discrete structures from data:
    - Determining number of objects, nonterminals, states etc.
- Multi-tasking learning:
    - Allows sharing of information across tasks.

Pitman-Yor Processes

- Two-parameter generalization of the Chinese restaurant process:

$$p(\text{customer } n \text{ sat at table } k|\text{past}) = \begin{cases} \frac{n_k-\beta}{n-1+\alpha} & \text{if occupied table} \\ \frac{\alpha+\beta K}{n-1+\alpha} & \text{if new table} \end{cases}$$

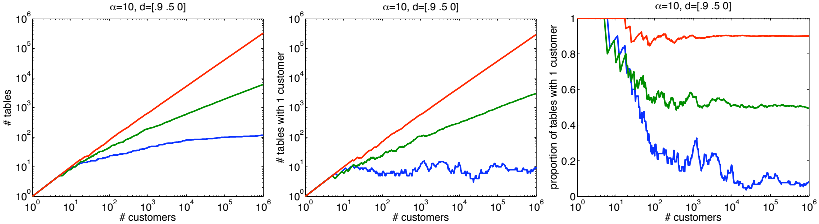
- Associating each cluster k with a unique draw $\theta_k^* \sim H$, the corresponding Pólya urn scheme is also exchangeable.
- De Finetti's Theorem states that there is a random measure underlying this two-parameter generalization.
    - This is the *Pitman-Yor process*.
- The Pitman-Yor process also has a stick-breaking construction:

$$\pi_k = v_k \prod_{i=1}^{k-1}(1 - v_i) \quad \beta_k \sim \text{Beta}(1-\beta, \alpha+\beta k) \quad \theta_k^* \sim H \quad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

[Pitman and Yor 1997, Perman et al. 1992]

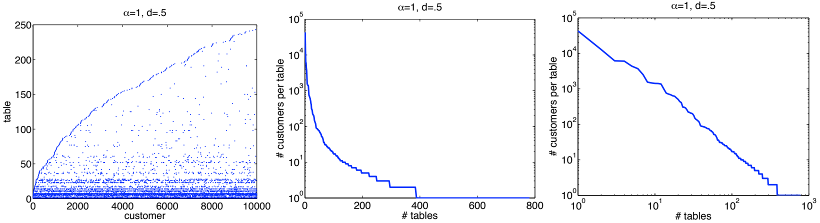
Pitman-Yor Processes

- Two salient features of the Pitman-Yor process:
  - With more occupied tables, the chance of even more tables becomes higher.
  - Tables with smaller occupancy numbers tend to have lower chance of getting new customers.
- The above means that Pitman-Yor processes produce Zipf's Law type behaviour, with $K = O(\alpha n^{\beta})$.

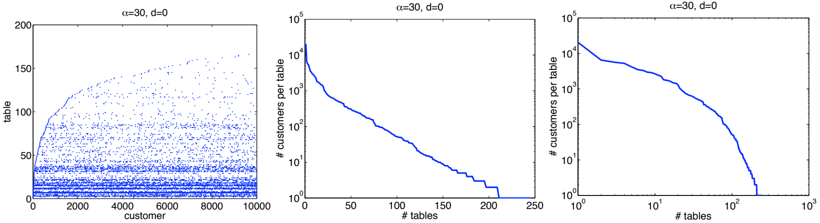


Pitman-Yor Processes

Draw from a Pitman-Yor process



Draw from a Dirichlet process



Hierarchical Pitman-Yor Language Models

- Pitman-Yor processes can be suitable models for many natural phenomena with power-law statistics.
- Language modelling with Markov assumption:

$$\begin{aligned} & p(\text{Mary has a little lamb}) \\ \approx & p(\text{Mary})p(\text{has}|\text{Mary})p(\text{a}|\text{Mary has})p(\text{little}|\text{has a})p(\text{lamb}|\text{a little}) \end{aligned}$$

- Parameterize with $p(w_3|w_1, w_2) = G_{w_1,w_2}[w_3]$ and use a hierarchical Pitman-Yor process prior:

$$\begin{aligned} G_{w_1,w_2}|G_{w_2} &\sim \text{PY}(\alpha_2, \beta_2, G_{w_2}) \\ G_{w_2}|G_\emptyset &\sim \text{PY}(\alpha_1, \beta_1, G_\emptyset) \\ G_\emptyset|U &\sim \text{PY}(\alpha_0, \beta_0, U) \end{aligned}$$

[Goldwater et al. 2006, Teh 2006]

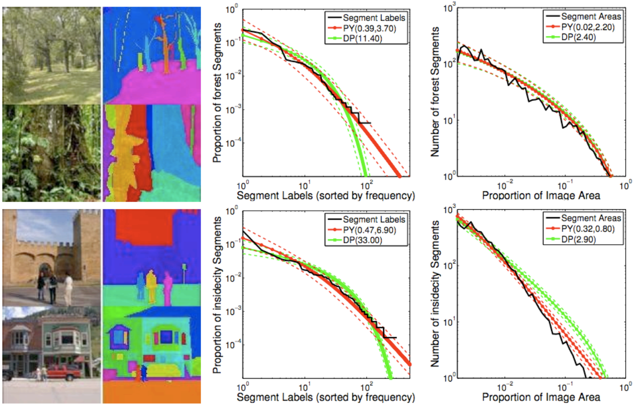
Hierarchical Pitman-Yor Language Models

T	N	IKN	MKN	HPYLM	HPYCV	HDLM
2e6	3	148.8	**144.1**	145.7	144.3	191.2
4e6	3	137.1	**132.7**	134.3	**132.7**	172.7
6e6	3	130.6	126.7	127.9	**126.4**	162.3
8e6	3	125.9	122.3	123.2	**121.9**	154.7
10e6	3	122.0	118.6	119.4	**118.2**	148.7
12e6	3	119.0	115.8	116.5	**115.4**	144.0
14e6	3	116.7	113.6	114.3	**113.2**	140.5
14e6	2	169.9	**169.2**	169.6	169.3	180.6
14e6	4	106.1	102.4	103.8	**101.9**	136.6

- Hierarchical Pitman-Yor language model produces state-of-the-art results.
- Extension to domain adaptation [Wood and Teh 2009].

T–training set size, N–context length+1, IKN–Interpolated Kneser Ney, MKN–modified Kneser-Ney, HPYLM–Hierarchical Pitman-Yor, HPYCV–HPYLM with parameters tuned by cross validation, HDLM–Hierarchical Dirichlet language model.

Image Segmentation with Pitman-Yor Processes



- Human segmentations of images also seem to follow power-law.
- An unsupervised image segmentation model based on dependent hierarchical Pitman-Yor processes achieves state-of-the-art results.

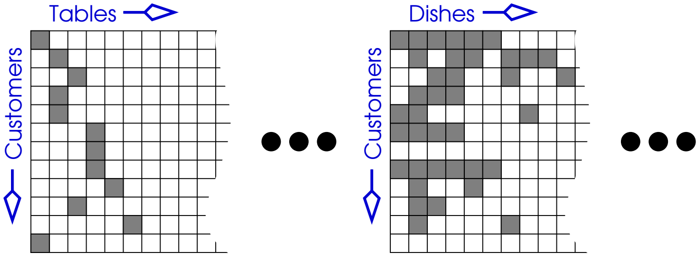
[Sudderth and Jordan 2009]

Beyond Clustering

- Dirichlet and Pitman-Yor processes are nonparametric models of clustering.
- Can nonparametric models go beyond clustering to describe data in more expressive ways?
    - Hierarchical (e.g. taxonomies)?
    - Distributed (e.g. multiple causes)?

Indian Buffet Processes

- The *Indian Buffet Process* (IBP) is akin to the Chinese restaurant process but describes each customer with a binary vector instead of cluster.
- Generating from an IBP:
  - Parameter α.
  - First customer picks Poisson(α) dishes to eat.
  - Subsequent customer i picks dish k with probability $\frac{n_k}{i}$; and picks Poisson($\frac{\alpha}{i}$) new dishes.



Infinite Independent Components Analysis

- Each image X_i is a linear combination of sparse features:

  $$X_i = \sum_k \Lambda_k y_{ik}$$

  where y_{ik} is activity of feature k with sparse prior. One possibility is a mixture of a Gaussian and a point mass at 0:

  $$y_{ik} = z_{ik} a_{ik} \qquad a_{ik} \sim \mathcal{N}(0,1) \qquad Z \sim \mathrm{IBP}(\alpha)$$

- An ICA model with infinite number of features.

[Knowles and Ghahramani 2007]

Indian Buffet Processes and Exchangeability

- The IBP is infinitely exchangeable, though this is much harder to see.
- De Finetti's Theorem again states that there is some random measure underlying the IBP.
- This random measure is the Beta process.

[Griffiths and Ghahramani 2006, Thibaux and Jordan 2007]

Beta Processes

- A *beta process* $B \sim \text{BP}(c, \alpha H)$ is a random discrete measure with form:

  $$B = \sum_{k=1}^{\infty} \mu_k \delta_{\theta_k^*}$$

  where the points $P = \{(\theta_1^*, \mu_1), (\theta_2^*, \mu_2), \ldots\}$ are spikes in a 2D Poisson process with rate measure:

  $$c\mu^{-1}(1-\mu)^{c-1} d\mu \alpha H(d\theta)$$

- The beta process with $c = 1$ is the de Finetti measure for the IBP. When $c \neq 1$ we have a two parameter generalization of the IBP.
- This is an example of a *completely random measure*.
- A beta process *does not* have Beta distributed marginals.

[Hjort 1990]

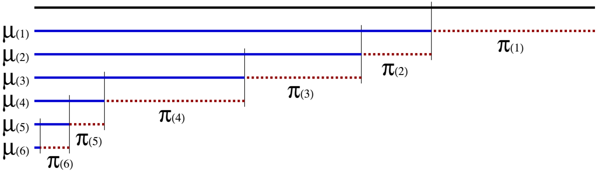
Stick-breaking Construction for Beta Processes

- When $c = 1$ it was shown that the following generates a draw of B:

$$v_k \sim \mathrm{Beta}(1, \alpha) \qquad \mu_k = (1 - v_k) \prod_{i=1}^{k-1} (1 - v_i) \qquad \theta_k^* \sim H$$

$$B = \sum_{k=1}^{\infty} \mu_k \delta_{\theta_k^*}$$

- The above is the complement of the stick-breaking construction for DPs!



[Teh et al. 2007]

Survival Analysis

- The Beta process was first proposed as a Bayesian nonparametric model for survival analysis with right-censored data.
- The hazard rate B is given a $\mathrm{BP}(c, \alpha H)$ prior. $B(\theta) d\theta$ is the chance of death in an infinitesimal interval $[\theta, \theta + d\theta)$ given that the individual has survived up to time θ.
- Data consists of a set of death times $\tau_1, \tau_2, \ldots$ and censored times $\gamma_1, \gamma_2, \ldots$, and can be summarized as:

Death measure: $$D = \sum_i \delta_{\tau_i}$$

Number-at-risk function: $$R(\theta) = D([\theta, \infty)) + \sum_i \mathbb{I}(\gamma_i \geq \theta)$$

- The posterior of B is:

$$B | D, R \sim \mathrm{BP}(c + R, \alpha H + D)$$

Note: the above is a generalization to c being a function of θ.

Outline

Bayesian Nonparametric Modelling

Measure Theoretic Probability Theory

Gaussian Processes

Dirichlet Processes

Extensions and Generalizations of Dirichlet Processes

Summary

Summary

- Motivation for Bayesian nonparametrics:
  - Allows practitioners to define and work with models with large support, sidesteps model selection.
  - New models with useful properties.
  - Large variety of applications.
- Introduced the Dirichlet process:
  - Infinite limit of finite mixture models.
  - Measure-theoretic definition.
  - Chinese restaurant process, Pólya urn scheme, stick-breaking construction.
- Touched upon two important theoretical tools:
  - Consistency and Kolmogorov's Consistency Theorem
  - Exchangeability and de Finetti's Theorem
- Described a number of applications of Bayesian nonparametrics.
- Missing: Inference methods based on MCMC, variational, and on different representations.

References I

Blackwell, D. and MacQueen, J. B. (1973).
Ferguson distributions via Pólya urn schemes.
Annals of Statistics, 1:353–355.

Ferguson, T. S. (1973).
A Bayesian analysis of some nonparametric problems.
Annals of Statistics, 1(2):209–230.

Goldwater, S., Griffiths, T., and Johnson, M. (2006).
Interpolating between types and tokens by estimating power-law generators.
In *Advances in Neural Information Processing Systems*, volume 18.

Griffiths, T. L. and Ghahramani, Z. (2006).
Infinite latent feature models and the Indian buffet process.
In *Advances in Neural Information Processing Systems*, volume 18.

Hjort, N. L. (1990).
Nonparametric Bayes estimators based on beta processes in models for life history data.
Annals of Statistics, 18(3):1259–1294.

Knowles, D. and Ghahramani, Z. (2007).
Infinite sparse factor analysis and infinite independent components analysis.
In *International Conference on Independent Component Analysis and Signal Separation*, volume 7 of *Lecture Notes in Computer Science*. Springer.

Neal, R. M. (1994).
Bayesian Learning for Neural Networks.
PhD thesis, Department of Computer Science, University of Toronto.

References II

Orbanz, P. (2009).
Foundations of nonparametric bayesian modelling.
Tutorial.

Perman, M., Pitman, J., and Yor, M. (1992).
Size-biased sampling of Poisson point processes and excursions.
Probability Theory and Related Fields, 92(1):21–39.

Pitman, J. and Yor, M. (1997).
The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator.
Annals of Probability, 25:855–900.

Rasmussen, C. E. (2000).
The infinite Gaussian mixture model.
In *Advances in Neural Information Processing Systems*, volume 12.

Rasmussen, C. E. and Williams, C. K. I. (2006).
Gaussian Processes for Machine Learning.
MIT Press.

Sethuraman, J. (1994).
A constructive definition of Dirichlet priors.
Statistica Sinica, 4:639–650.

Sudderth, E. and Jordan, M. I. (2009).
Shared segmentation of natural scenes using dependent Pitman-Yor processes.
In *Advances in Neural Information Processing Systems*, volume 21.

References III

Teh, Y. W. (2006).
A hierarchical Bayesian language model based on Pitman-Yor processes.
In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992.

Teh, Y. W. (2007).
Dirichlet processes.
Submitted to Encyclopedia of Machine Learning.

Teh, Y. W., Görür, D., and Ghahramani, Z. (2007).
Stick-breaking construction for the Indian buffet process.
In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 11.

Teh, Y. W. and Jordan, M. I. (2009).
Hierarchical Bayesian nonparametric models with applications.
In Hjort, N., Holmes, C., Müller, P., and Walker, S., editors, *To appear in Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006).
Hierarchical Dirichlet processes.
Journal of the American Statistical Association, 101(476):1566–1581.

Thibaux, R. and Jordan, M. I. (2007).
Hierarchical beta processes and the Indian buffet process.
In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, volume 11.

Wood, F. and Teh, Y. W. (2009).
A hierarchical nonparametric Bayesian approach to statistical language model domain adaptation.
In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 12.

Posterior Dirichlet Processes

- Suppose G is DP distributed, and θ is G distributed:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta | G &\sim G \end{aligned}$$

- This gives $p(G)$ and $p(\theta|G)$.
- We are interested in:

$$\begin{aligned} p(\theta) &= \int p(\theta|G) p(G)\, dG \\ p(G|\theta) &= \frac{p(\theta|G)p(G)}{p(\theta)} \end{aligned}$$

Posterior Dirichlet Processes

Conjugacy between Dirichlet Distribution and Multinomial.

- Consider:

$$\begin{aligned}(\pi_1, \ldots, \pi_K) &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \\ z|(\pi_1, \ldots, \pi_K) &\sim \text{Discrete}(\pi_1, \ldots, \pi_K)\end{aligned}$$

  z is a multinomial variate, taking on value $i \in \{1, \ldots, n\}$ with probability π_i.

- Then:

$$\begin{aligned}z &\sim \text{Discrete}\left(\tfrac{\alpha_1}{\sum_i \alpha_i}, \ldots, \tfrac{\alpha_K}{\sum_i \alpha_i}\right) \\ (\pi_1, \ldots, \pi_K)|z &\sim \text{Dirichlet}(\alpha_1 + \delta_1(z), \ldots, \alpha_K + \delta_K(z))\end{aligned}$$

  where $\delta_i(z) = 1$ if z takes on value i, 0 otherwise.

- Converse also true.

Posterior Dirichlet Processes

- Fix a partition $(A_1, \dots, A_K)$ of Θ. Then

$$(G(A_1), \dots, G(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \dots, \alpha H(A_K))$$
$$P(\theta \in A_i | G) = G(A_i)$$

- Using Dirichlet-multinomial conjugacy,

$$P(\theta \in A_i) = H(A_i)$$
$$(G(A_1), \dots, G(A_K))|\theta \sim \text{Dirichlet}(\alpha H(A_1) + \delta_\theta(A_1), \dots, \alpha H(A_K) + \delta_\theta(A_K))$$

- The above is true for every finite partition of Θ. In particular, taking a really fine partition,

$$p(d\theta) = H(d\theta)$$

- Also, the posterior $G|\theta$ is also a Dirichlet process:

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

Posterior Dirichlet Processes

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta | G &\sim G \end{aligned} \quad \Longleftrightarrow \quad \begin{aligned} \theta &\sim H \\ G|\theta &\sim \mathrm{DP}\left(\alpha + 1, \tfrac{\alpha H + \delta_\theta}{\alpha + 1}\right) \end{aligned}$$

Pólya Urn Scheme

- First sample:

$$\begin{aligned} && \theta_1 | G \sim G && G \sim \mathrm{DP}(\alpha, H) \\ \Longleftrightarrow && \theta_1 \sim H && G|\theta_1 \sim \mathrm{DP}(\alpha+1, \tfrac{\alpha H + \delta_{\theta_1}}{\alpha+1}) \end{aligned}$$

- Second sample:

$$\begin{aligned} && \theta_2 | \theta_1, G \sim G && G|\theta_1 \sim \mathrm{DP}(\alpha+1, \tfrac{\alpha H + \delta_{\theta_1}}{\alpha+1}) \\ \Longleftrightarrow && \theta_2|\theta_1 \sim \tfrac{\alpha H + \delta_{\theta_1}}{\alpha+1} && G|\theta_1, \theta_2 \sim \mathrm{DP}(\alpha+2, \tfrac{\alpha H + \delta_{\theta_1} + \delta_{\theta_2}}{\alpha+2}) \end{aligned}$$

- n^{th} sample

$$\begin{aligned} && \theta_n | \theta_{1:n-1}, G \sim G && G|\theta_{1:n-1} \sim \mathrm{DP}(\alpha+n-1, \tfrac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha+n-1}) \\ \Longleftrightarrow && \theta_n|\theta_{1:n-1} \sim \tfrac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha+n-1} && G|\theta_{1:n} \sim \mathrm{DP}(\alpha+n, \tfrac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha+n}) \end{aligned}$$

Stick-breaking Construction

- Returning to the posterior process:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta|G &\sim G \end{aligned} \qquad \Leftrightarrow \qquad \begin{aligned} \theta &\sim H \\ G|\theta &\sim \mathrm{DP}(\alpha+1, \tfrac{\alpha H+\delta_\theta}{\alpha+1}) \end{aligned}$$

- Consider a partition $(\theta, \Theta\backslash\theta)$ of Θ. We have:

$$\begin{aligned} (G(\theta), G(\Theta\backslash\theta))|\theta &\sim \mathrm{Dirichlet}((\alpha+1)\tfrac{\alpha H+\delta_\theta}{\alpha+1}(\theta), (\alpha+1)\tfrac{\alpha H+\delta_\theta}{\alpha+1}(\Theta\backslash\theta)) \\ &= \mathrm{Dirichlet}(1, \alpha) \end{aligned}$$

- G has a point mass located at θ:

$$G = \beta\delta_\theta + (1-\beta)G' \qquad \text{with} \qquad \beta \sim \mathrm{Beta}(1, \alpha)$$

  and G' is the (renormalized) probability measure with the point mass removed.

- What is G'?

Stick-breaking Construction

- Currently, we have:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta &\sim G \end{aligned} \quad \Rightarrow \quad \begin{aligned} \theta &\sim H \\ G|\theta &\sim \mathrm{DP}(\alpha + 1, \tfrac{\alpha H + \delta_\theta}{\alpha + 1}) \\ G &= \beta\delta_\theta + (1 - \beta)G' \\ \beta &\sim \mathrm{Beta}(1, \alpha) \end{aligned}$$

- Consider a further partition $(\theta, A_1, \ldots, A_K)$ of Θ:

$$\begin{aligned} &(G(\theta), G(A_1), \ldots, G(A_K)) \\ =&(\beta, (1 - \beta)G'(A_1), \ldots, (1 - \beta)G'(A_K)) \\ \sim& \mathrm{Dirichlet}(1, \alpha H(A_1), \ldots, \alpha H(A_K)) \end{aligned}$$

- The agglomerative/decimative property of Dirichlet implies:

$$(G'(A_1), \ldots, G'(A_K))|\theta \sim \mathrm{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))$$

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

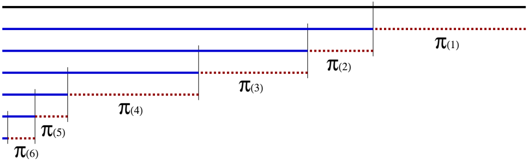
Stick-breaking Construction

- We have:

$$
\begin{aligned}
G &\sim \mathrm{DP}(\alpha, H) \\
G &= \beta_1 \delta_{\theta_1^*} + (1 - \beta_1) G_1 \\
G &= \beta_1 \delta_{\theta_1^*} + (1 - \beta_1)(\beta_2 \delta_{\theta_2^*} + (1 - \beta_2) G_2) \\
&\vdots \\
G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}
\end{aligned}
$$

  where

$$\pi_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i) \qquad \beta_k \sim \mathrm{Beta}(1, \alpha) \qquad \theta_k^* \sim H$$


$\pi_{(1)}$
$\pi_{(2)}$
$\pi_{(3)}$
$\pi_{(4)}$
$\pi_{(5)}$
$\pi_{(6)}$