Lecture 3a: Dirichlet processes

Cédric Archambeau

Centre for Computational Statistics and Machine Learning
Department of Computer Science
University College London

c.archambeau@cs.ucl.ac.uk

Advanced Topics in Machine Learning (MSc in Intelligent Systems)
January 2008

Today's plan

- The Dirichlet distribution
- The Dirichlet process
- Representations of the Dirichlet process
- Infinite mixture models
- DP mixtures of Regressors
- DP mixtures of Dynamical Systems

- Guest speakers: **Yee Whye Teh** (Gatsby unit)
  David Barber (CSML) on 05/02

(Bayesian) Nonparametric models

- Real data is complex and the parametric assumption is often wrong.
- Nonparametric models allow us to relax the parametric assumption, bringing significant flexibility to our models.
- Nonparametric models lead to model selection/averaging behaviours without the cost of actually doing model selection/averaging.
- Nonparametric models are gaining popularity, spurred by growth in computational resources and inference algorithms.
- Examples: Gaussian processes (GPs) and Dirichlet processes (DPs).
- Other models which will not be discussed include Indian buffet processes, beta processes, tree processes, ...

Gaussian processes in a nutshell

A $\mathcal{GP}$ is a distribution over (latent) **functions**.

A $\mathcal{GP}$ is defined by a **Gaussian probability measure**[1], i.e. it is characterised by a mean function $m(\cdot)$ and a covariance function $k(\cdot, \cdot)$.

The **consistency property** of the Gaussian measure preserves the Gaussian parametric form when integrating out latent function values :

- Every finite subset of latent function values is jointly Gaussian
- The marginals are Gaussian
- Gaussian conditionals can be used to sample from a $\mathcal{GP}$

[1] Let $\mathcal{A}$ be the set of subsets of a space $\mathbb{X}$. A **measure** $G : \mathcal{A} \rightarrow \Omega$ assigns a nonnegative value to any subset of $\mathbb{X}$. We say that G is a **probability measure** when $\Omega = [0, 1]$.

Dirichlet distribution

The **Dirichlet distribution** is a distribution defined over K discrete random variables $\pi_1, \ldots, \pi_K$:

$$(\pi_1, \ldots, \pi_K) \sim \mathcal{D}(\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}.$$

We say that the Dirichlet distribution is a distribution over the $(K-1)$-dimensional **probability simplex**:

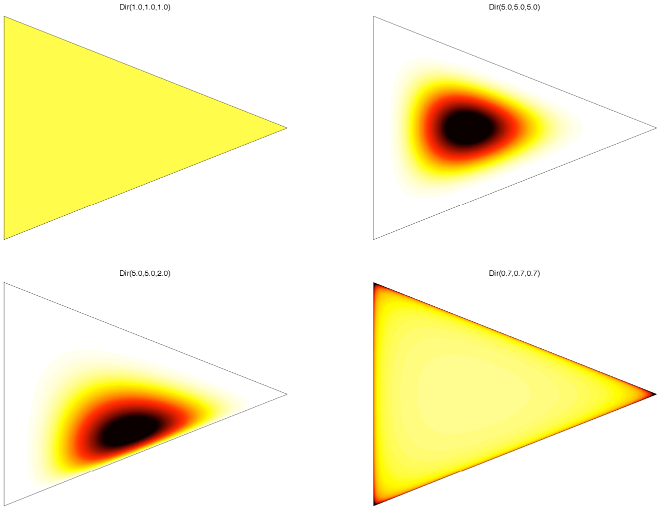
$$\Delta_K = \left\{(\pi_1, \ldots, \pi_K) : 0 \leqslant \pi_k \leqslant 1, \sum_k \pi_k = 1\right\}.$$

The **mean** and the **variance** are respectively given by

$$\langle \pi_k \rangle = \frac{\alpha_k}{\alpha}, \qquad \left\langle (\pi_k - \langle \pi_k \rangle)^2 \right\rangle = \frac{\alpha_k(\alpha - \alpha_k)}{\alpha^2(\alpha + 1)},$$

where $\alpha \equiv \sum_k \alpha_k$.

Examples



Dir(1.0,1.0,1.0)

Dir(5.0,5.0,5.0)

Dir(5.0,5.0,2.0)

Dir(0.7,0.7,0.7)

Properties of the Dirichlet distribution

- The **Beta distribution** corresponds to the Dirichlet distribution with $K = 2$:

$$\pi \sim \mathcal{B}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1}.$$

- **Agglomerative property:**
  Let $(\pi_1, \dots, \pi_K) \sim \mathcal{D}(\alpha_1, \dots, \alpha_K)$. Combining entries of probability vectors preserves the Dirichlet property:

$$(\pi_1 + \pi_2, \pi_3, \dots, \pi_K) \sim \mathcal{D}(\alpha_1 + \alpha_2, \alpha_3, \dots, \alpha_K).$$

  The property holds for any partition of $\{1, \dots, K\}$.

- **Decimative property:**
  Let $(\pi_1, \dots, \pi_K) \sim \mathcal{D}(\alpha_1, \dots, \alpha_K)$ and $(\tau_1, \tau_2) \sim \mathcal{D}(\alpha_1\beta_1, \alpha_1\beta_2)$, with $\beta_1 + \beta_2 = 1$. The converse of the agglomerative property is also true:

$$(\pi_1\tau_1, \pi_1\tau_2, \pi_2, \pi_3, \dots, \pi_K) \sim \mathcal{D}(\alpha_1\beta_1, \alpha_1\beta_2, \alpha_2, \alpha_3, \dots, \alpha_K).$$

Dirichlet process (DP)

A $\mathcal{DP}$ is a distribution over **probability measures**, such that marginals on finite partitions are Dirichlet distributed.

Let G and H be probability measures over $\mathbb{X}$. The random probability measure G is a **Dirichlet process** with **base measure** $H(\cdot)$ and **intensity parameter** α

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

if for any finite partition $\{A_1, \ldots, A_K\}$ of $\mathbb{X}$, we have

$$(G(A_1), \ldots, G(A_K)) \sim \mathcal{D}\left(\alpha H(A_1), \ldots, \alpha H(A_K)\right).$$

The **mean distribution** and its **variance** are respectively given by

$$\langle G(A) \rangle = H(A), \qquad \left\langle (G(A) - \langle G(A) \rangle)^2 \right\rangle = \frac{H(A)(1 - \alpha H(A))}{(\alpha + 1)},$$

where A is any measurable subset of $\mathbb{X}$.

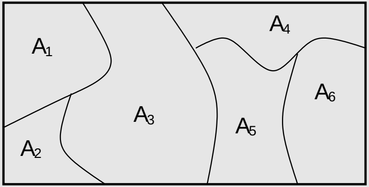
The mean measure follows directly from the expression of the mean of the Dirichlet distribution:

$$\langle G(A_k) \rangle = \frac{\alpha H(A_k)}{\sum_{k'} \alpha H(A_{k'})} = H(A_k),$$

where we used the fact that $\sum_{k'} H(A_{k'}) = 1$.

And so does the variance:

$$\begin{aligned}\left\langle (G(A_k) - H(A_k))^2 \right\rangle &= \frac{\alpha H(A_k)(\sum_{k'} \alpha H(A_{k'}) - \alpha H(A_k))}{(\sum_{k'} \alpha H(A_{k'}))^2 (\sum_{k'} \alpha H(A_{k'}) + 1)} \\ &= \frac{\alpha^2 H(A_k)(1 - \alpha H(A_k))}{\alpha^2 (\alpha + 1)} \\ &= \frac{H(A_k)(1 - \alpha H(A_k))}{(\alpha + 1)}.\end{aligned}$$



Posterior Dirichlet process

Consider a fixed partition $\{A_1, \dots, A_K\}$ of $\mathbb{X}$. If $G(\cdot) \sim \mathcal{DP}\left(\alpha, H(\cdot)\right)$, then

$$(G(A_1), \dots, G(A_K)) \sim \mathcal{D}(\alpha H(A_1), \dots, \alpha H(A_K)).$$

We may treat the random measure G as a distribution over $\mathbb{X}$, such that

$$P(\theta \in A_k | G) = G(A_k).$$

The **posterior** is again a Dirichlet distribution:

$$(G(A_1), \dots, G(A_K))\,|\theta \sim \mathcal{D}(\alpha H(A_1) + \delta_\theta(A_1), \dots, \alpha H(A_K) + \delta_\theta(A_K)),$$

where $\delta_\theta(A_k) = 1$ if $\theta \in A_k$ and 0 otherwise.

The above is true for every finite partition of $\mathbb{X}$, e.g. $H(d\theta) = p(\theta)d\theta$. Hence, this suggests that the **posterior process** is a Dirichlet process:

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

where $\delta_\theta(\cdot)$ is Dirac's delta centred at θ, which we call an **atom**.

Let us denote $G(A_k)$ by π_k and $\alpha H(A_k)$ by α_k. We introduce the **auxiliary indicator variable** z, which takes the value $k \in \{1, \ldots, K\}$ when $\theta \in A_k$.

Hence, we have

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

If z is Discrete with parameters $\pi_1 \ldots, \pi_K$, then z takes the value $k \in \{1, \ldots, K\}$ with probability π_k.

The **posterior** is given by Bayes' rule:

$$\begin{aligned}
p(\pi_1, \ldots, \pi_k|z = k) &\propto \pi_k \prod_{k'} \pi_{k'}^{\alpha_{k'}-1} \\
\Rightarrow \quad \pi_1, \ldots, \pi_k|z &\sim \mathcal{D}(\alpha_1 + \delta_z(1), \ldots, \alpha_K + \delta_z(K)),
\end{aligned}$$

where $\delta_z(\cdot)$ is Kronecker's delta centred at z.

The **marginal** is given by

$$p(z = k) = \langle \pi_k \rangle = \frac{\alpha_k}{\sum_{k'} \alpha_{k'}} \quad \Rightarrow \quad z \sim \text{Discrete}\left(\frac{\alpha_1}{\sum_k \alpha_k}, \ldots, \frac{\alpha_K}{\sum_k \alpha_k}\right).$$

Remark

If $G \sim \mathcal{DP}(\alpha, H)$ and $\theta | G \sim G$, then the **posterior process** and the **marginal** are respectively given by

$$\begin{aligned} G|\theta &\sim \mathcal{DP}\left(\alpha+1, \frac{\alpha H + \delta_\theta}{\alpha+1}\right), \\ \theta &\sim H. \end{aligned}$$

It can be shown that the **converse** is also true, such that

$$\begin{array}{r} G \sim \mathcal{DP}(\alpha, H) \\ \theta|G \sim G \end{array} \quad \Leftrightarrow \quad \begin{array}{r} \theta \sim H \\ G|\theta \sim \mathcal{DP}\left(\alpha+1, \frac{\alpha H+\delta_\theta}{\alpha+1}\right) \end{array} .$$

Blackwell-MacQueen urn scheme

Blackwell-MacQueen urn scheme tells us how to **sample sequentially** from a DP. It is like a "representer", i.e. a finite projection of an infinite object.

It produces a sequence $\theta_1, \theta_2, \ldots$ with the following conditionals:

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

This can be interpreted as picking balls of different colours from an urn:

- Start with no balls in the urn.
- Draw $\theta_N \sim H$ with probability $\propto \alpha$. Add a ball of colour θ_N in the urn.
- Pick a ball at random from the urn with probability $\propto N - 1$. Record its colour θ_N and return the ball in the urn. Place a second ball of the same colour in urn.

Hence, there is positive probability that θ_N takes the same value θ_n for some n, i.e. the $\theta_1, \ldots, \theta_N$ **cluster** together (see CRP).

The Blackwell-MacQueen urn scheme is also known as the **Pòlya urn** scheme.

- First sample:

$$\begin{aligned} G &\sim \mathcal{DP}(\alpha, H) \\ \theta_1 | G &\sim G \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} \theta_1 &\sim H \\ G|\theta_1 &\sim \mathcal{DP}\left(\alpha + 1, \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1}\right) \end{aligned} .$$

- Second sample:

$$\begin{aligned} G|\theta_1 &\sim \mathcal{DP}\left(\alpha + 1, \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1}\right) \\ \theta_2|\theta_1, G &\sim G \end{aligned}$$

$$\Leftrightarrow \quad \begin{aligned} \theta_2|\theta_1 &\sim \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1} \\ G|\theta_1, \theta_2 &\sim \mathcal{DP}\left(\alpha + 2, \frac{\alpha H + \delta_{\theta_1} + \delta_{\theta_2}}{\alpha + 2}\right) \end{aligned} .$$

$$\vdots$$

- N^{th} sample:

$$\begin{aligned} G|\theta_{1:N-1} &\sim \mathcal{DP}\left(\alpha + N - 1, \frac{\alpha H + \sum_n^{N-1} \delta_{\theta_n}}{\alpha + N - 1}\right) \\ \theta_2|\theta_{1:N-1}, G &\sim G \end{aligned}$$

$$\Leftrightarrow \quad \begin{aligned} \theta_N|\theta_{1:N-1} &\sim \frac{\alpha H + \sum_n^{N-1} \delta_{\theta_n}}{\alpha + N - 1} \\ G|\theta_{1:N} &\sim \mathcal{DP}\left(\alpha + N, \frac{\alpha H + \sum_n^{N} \delta_{\theta_n}}{\alpha + N}\right) \end{aligned} .$$

Chinese restaurant process (CRP)

Random draws $\theta_1, \dots, \theta_N$ from a Blackwell-MacQueen urn scheme induce a **random partition** of $1, \dots, N$:

- $\theta_1, \dots, \theta_N$ take $K < N$ distinct values, say $\theta_1^*, \dots, \theta_K^*$.
- Every drawn sequence defines a partition of $1, \dots, N$ into K clusters, such that if n is in cluster k, then $\theta_n = \theta_k^*$.
- The induced distribution over partitions is a CRP.

Generating from the CRP:

- First customer sits at the first table.
- Customer N sits at:
  - A new table with probability $\frac{\alpha}{\alpha+N-1}$.
  - Table k with probability $\frac{N_k}{\alpha+N-1}$, where N_k is the number of customers at table k.

CRP exhibits the **clustering property** of DP:

- Tables are clusters, i.e. distinct values $\{\theta_k^*\}_{k=1}^K$ drawn from H.
- Customers are the actual realisations, i.e. $\theta_n = \theta_{z_n}^*$ where $z_n \in \{1, \dots, K\}$ indicates at which table the customer n sat.

Stick-breaking construction

We would like to exhibit the **discrete** character of $G \sim \mathcal{DP}(\alpha, H)$.

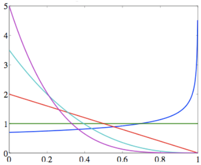
Consider $\theta \sim H$ and the partition $\{\theta, \mathbb{X}\backslash\theta\}$ of $\mathbb{X}$. The posterior process $G|\theta \sim \mathcal{DP}\left(\alpha+1, \frac{\alpha H+\delta_\theta}{\alpha+1}\right)$ implies that

$$\begin{aligned} \left(G(\theta), G(\mathbb{X}\backslash\theta)\right)|\theta &\sim \mathcal{D}\left(\alpha H(\theta)+\delta_\theta(\theta), \alpha H(\mathbb{X}\backslash\theta)+\delta_\theta(\mathbb{X}\backslash\theta)\right) \\ &= \mathcal{D}(1, \alpha) \\ \Leftrightarrow \quad G(\theta)|\theta &\sim \mathcal{B}(1, \alpha). \end{aligned}$$

Hence, G has a **point mass** located at θ, such that

$$G = \beta\delta_\theta + (1-\beta)G', \qquad \beta \sim \mathcal{B}(1, \alpha),$$

where G' is the (renormalized) probability measure without the point mass.



Stick-breaking construction (continued)

What form has the probability measure G'?

Let $\{A_1, \ldots, A_K\}$ be a partition of $\mathbb{X}\backslash\theta$. Since $G(\mathbb{X}\backslash\theta) = \sum_k G(A_k)$, the agglomerative property of the Dirichlet leads to

$$(G(\theta), G(A_1), \ldots, G(A_K)) \,|\theta \sim \mathcal{D}(1, \alpha H(A_1), \ldots, \alpha H(A_K)).$$

Hence, we see that

$$\begin{aligned} G(\theta)|\theta &= \beta \\ G(A_1)|\theta &= (1-\beta)G'(A_1) \\ &\vdots \\ G(A_K)|\theta &= (1-\beta)G'(A_K). \end{aligned}$$

From the decimative property of Dirichlet, we get

$$(G'(A_1), \ldots, G'(A_K)) \sim \mathcal{D}(\alpha H(A_1), \ldots, \alpha H(A_K)).$$

In other words, $G' \sim \mathcal{DP}(\alpha, H)$!

Stick-breaking construction (continued)

What form has the random measure G?

$$
\begin{aligned}
&G \sim \mathcal{DP}(\alpha, H) \\
&G = \beta_1 \delta_{\theta_1^*} + (1-\beta_1) G_1 && G_1 \sim \mathcal{DP}(\alpha, H) \\
&G = \beta_1 \delta_{\theta_1^*} + (1-\beta_1)(\beta_2 \delta_{\theta_2^*} + (1-\beta_2) G_2) && G_2 \sim \mathcal{DP}(\alpha, H) \\
&\quad \vdots \\
&G = \sum_{n=1}^{\infty} \pi_n \delta_{\theta_n^*}
\end{aligned}
$$

where

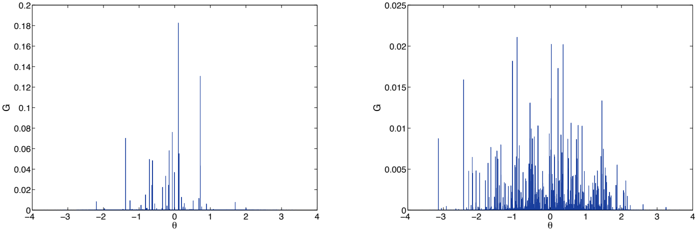
$$\pi_n = \beta_n \prod_{k=1}^{n-1} (1-\beta_k), \qquad \beta_n \sim \mathcal{B}(1, \alpha), \qquad \theta_n^* \sim H.$$

Draws from the DP look like a weighted sum of point masses, where the weights are drawn from the stick-breaking construction[2]. The steps are limited by assumptions of regularity on $\mathbb{X}$ and the smoothness on H.

[2] The construction can be viewed as repeatedly breaking the remainder, which is of length $1-\beta_n$, of a unit-length stick according to a Beta distribution.

Sampling from a DP using the stick-breaking construction (demo: stick_break)

The base measure H is a unit Gaussian with zero mean. Note how the intensity parameter α regulates the number of spikes.



(a) $\alpha = 15$.

(b) $\alpha = 150$.

Figure: The random meaasure G drawn from a Dirichlet process prior is discrete.

Exchangeability (de Finetti's theorem)

Exchangeability is a fundamental property of Dirichlet processes, more specifically for inference.

An infinite sequence $\theta_1, \theta_2, \theta_3, \ldots$ of random variables is said to be infintely exchangeable if for any finite subset $\theta_{i_1}, \ldots, \theta_{i_n}$ and any permutation $\theta_{j_1}, \ldots, \theta_{j_n}$, we have

$$P(\theta_{i_1}, \ldots, \theta_{i_n}) = P(\theta_{j_1}, \ldots, \theta_{j_n}).$$

The condition of exchangeability weaker than the assumption that they are independent and identically distributed

Exchangeable observations are conditionally independent given some (usually) unobserved quantity to which some probability distribution would be assigned.

Using de Finettis theorem, it is possible to show that draws form a $\mathcal{DP}$ are infinitely exchangeable

Density estimation with Dirichlet processes

Consider a set of realisations $\{\mathbf{x}_n\}_{n=1}^N$ of the random variable X. Based on these observations, we would like to **model** the density of X.

Would the following setting work?

$$\begin{aligned} G &\sim \mathcal{DP}(\alpha, H), \\ \mathbf{x}_n | G &\sim G. \end{aligned}$$

Obviously not, as G is a discrete distribution (not a density).

A solution is to convolve $G \sim \mathcal{DP}(\alpha, H)$ with a kernel $F(\cdot|\theta)$:

$$F(\cdot) = \int F(\cdot|\theta) G(\theta) d\theta = \sum_{n=1}^{\infty} \pi_n F(\cdot|\theta_n^*),$$

where we used the stick-breaking representation $G(\theta) = \sum_{n=1}^{\infty} \pi_n \delta_{\theta_n^*}(\theta)$.

Each data point $\mathbf{x}_n$ is modelled by a **latent parameter** θ_n, which is drawn from an unknown distribution G on which a **Dirichlet process prior** is imposed:

$$\begin{aligned} G &\sim \mathcal{DP}(\alpha, H), \\ \theta_n | G &\sim G, \\ \mathbf{x}_n | \theta_n &\sim F(\cdot|\theta_n). \end{aligned}$$

Finite mixture models

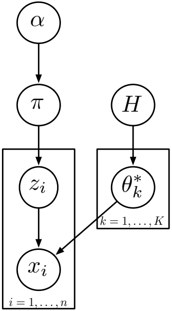
A finite mixture model makes the assumption that each data point $\mathbf{x}_n$ was generated by a single mixture component with parameters θ^*_k:

$$\mathbf{x}_n \sim \sum_k \pi_k F(\mathbf{x}_n|\theta^*_k).$$

This can be captured by introducing a latent indicator variable z_n.

The model is formalised by the following priors and likelihood:

$$\begin{aligned}
\theta^*_k &\sim H \\
\pi_1, \ldots, \pi_K &\sim \mathcal{D}(\alpha/K, \ldots, \alpha/K) \\
z_n|\pi_1, \ldots, \pi_K &\sim \text{Discrete}(\pi_1, \ldots, \pi_K) \\
\mathbf{x}_n|\theta^*_{z_n} &\sim F(\cdot|\theta^*_{z_n})
\end{aligned}$$



Inference is performed on:

- The hyperparameters of the prior H.
- Parameter α of the Dirichlet prior on the weights.
- The number of components K.

DP mixture models

If K is really **large**, there will be no overfitting if the parameters $\{\theta_k^*\}$ and mixing proportions $\{\pi_k\}$ are integrated out or at most N (but usually much less) components will be active, i.e. associated with data.

Recall the density estimation approach based on $\mathcal{DP}$:

$$F(\cdot) = \sum_{k=1}^{\infty} \pi_k F(\cdot|\theta_k^*) \qquad G \sim \mathcal{DP}(\alpha, H),$$

$$\theta_n | G \sim G = \sum_{k=1}^{\infty} \pi_k \delta^*_{\theta_k},$$

$$\mathbf{x}_n | \theta_k^* \sim F(\cdot|\theta_n = \theta_k^*).$$

This is equivalent to a mixture model with a countably **infinite** number of components:

$$\begin{aligned} \pi_1, \ldots, \pi_K &\sim \mathcal{D}(\alpha/K, \ldots, \alpha/K), \\ z_n &\sim \text{Discrete}(\pi_1, \ldots, \pi_K), \\ \mathbf{x}_n | z_n &\sim F(\cdot|\theta^*_{z_n}), \end{aligned}$$

where $K \rightarrow \infty$.

DP mixture models (continued)

- DP mixture models are used to sidestep the model selection problem by replacing it by model averaging.
- They are used in applications in which the number of clusters is not known a priori or is believed to grow without bound as the amount of data grows.
- DPs can also be used in applications, 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).
- Inference for DP mixtures depends on the representations:
    - MCMC based CRP, stick-breaking construction, etc.
    - Variational inference based on stick-breaking construction (Blei and Jordan, 2005).

References

- The slides are largely based on Y.W. Teh's Tutorial and practical course on Dirichlet processes at Machine Learning Summer School 2007.
- Tutorial on Dirichlet processes at NIPS 2005 by M. I. Jordan.

- Blackwell, D. and MacQueen, J. B. (1973), Ferguson distributions via Plya urn schemes, Annals of Statistics, 1:353355.
- Blei, D. M. and Jordan, M. I. (2006), Variational inference for Dirichlet process mixtures, Bayesian Analysis, 1(1):121144.
- Ferguson, T. S. (1973), A Bayesian analysis of some nonparametric problems, Annals of Statistics, 1(2):209230.
- Rasmussen, C. E. (2000), The infinite Gaussian mixture model, Advances in Neural Information Processing Systems, volume 12.
- Sethuraman, J. (1994), A constructive definition of Dirichlet priors, Statistica Sinica, 4:639650.
- 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):15661581.