Tutorial on Bayesian Non-parametric methods

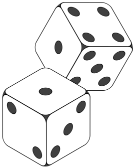
Erin Grant

Department of Computer Science, University of Toronto

November 27th, 2015

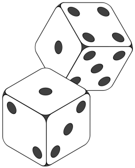
Intuition for the Dirichlet Distribution (1)

Consider a six-sided die.



Intuition for the Dirichlet Distribution (1)

Consider a six-sided die.



The possible outcomes $\{1, 2, 3, 4, 5, 6\}$ of rolling the die:

- are discrete (countable);
- are disjoint;
- have probabilities:

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}.$$

Intuition for the Dirichlet Distribution (2)

How do we model manufacturing error that causes the die outcome probabilities $\{1, 2, 3, 4, 5, 6\}$ to be skewed?

Intuition for the Dirichlet Distribution (2)

How do we model manufacturing error that causes the die outcome probabilities $\{1, 2, 3, 4, 5, 6\}$ to be skewed?

We can use the Dirichlet distribution, so that the probabilities themselves are distributed:

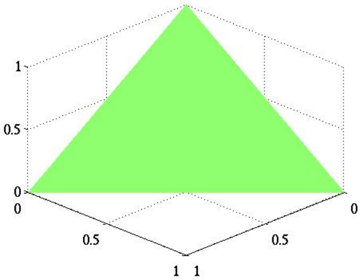
$$\big[P(1),\quad P(2),\quad P(3),\quad P(4),\quad P(5),\quad P(6)\big] \sim \text{Dir}(\vec{\alpha})$$

where $\vec{\alpha} = (\alpha_1, \ldots, \alpha_6)$ is a parameter vector of pseudocounts (prior expectations).

Math Background for the Dirichlet Distribution (1)

Probability simplex:

$$\left\{\vec{x} \in \mathbb{R}^k \mid x_1 + \cdots + x_k = 1, x_1, \ldots, x_k \geq 0\right\}$$

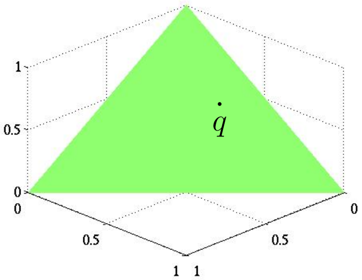


The 2-dimensional probability simplex in $\mathbb{R}^3$.

Figure adapted from Frigyik et al. [2010]

Math Background for the Dirichlet Distribution (2)

Each point q corresponds to a probability mass function (pmf) over k disjoint events.



E.g.: $q = [0.3, 0.1, 0.6]$

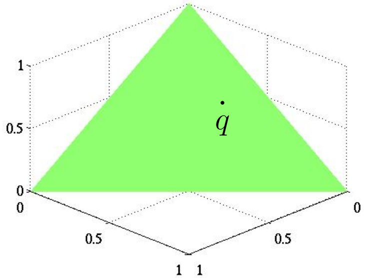
$\downarrow$

pmf over three events with probabilities $\frac{3}{10}$, $\frac{1}{10}$, and $\frac{3}{5}$.

Figure adapted from Frigyik et al. [2010]

Math Background for the Dirichlet Distribution (3)

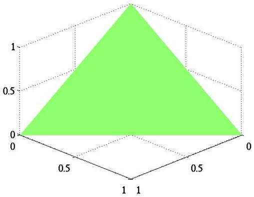
The Dirichlet distribution defines a probability for each point q in the simplex (i.e., it is a pmf over pmfs).



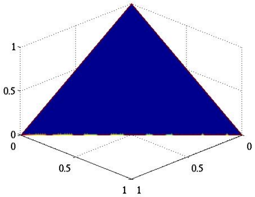
$$P(q) = \text{Dir}(\vec{\alpha}) \text{ for some } \vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3).$$

Figure adapted from Frigyik et al. [2010]

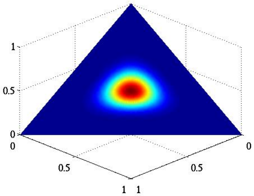
The Parameter Vector $\vec{\alpha}$



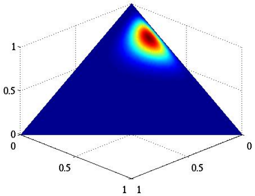
$\alpha = [1, 1, 1]$



$\alpha = [.1, .1, .1]$



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



$\alpha = [2, 5, 15]$

Figures from Frigyik et al. [2010]

Refresher: Bayesian Updating and Conjugacy (1)

We have a likelihood distribution $P(X \mid \theta)$, parametrised by some parameters θ, and a prior $P(\theta)$ over the parameters θ.

We want to infer a posterior distribution over the parameters θ after seeing some observations $X_1, \ldots X_N$.

We use Bayes' Rule:

$$P(\theta \mid X_1, \ldots, X_N) = \frac{P(X_1, \ldots, X_N \mid \theta)\, P(\theta)}{P(X_1, \ldots, X_N)}$$

Refresher: Bayesian Updating and Conjugacy (1)

We have a likelihood distribution $P(X \mid \theta)$, parametrised by some parameters θ, and a prior $P(\theta)$ over the parameters θ.

We want to infer a posterior distribution over the parameters θ after seeing some observations $X_1, \ldots X_N$.

We use Bayes' Rule:

$$P(\theta \mid X_1, \ldots, X_N) = \frac{P(X_1, \ldots, X_N \mid \theta)\, P(\theta)}{P(X_1, \ldots, X_N)}$$

But $P(X_1, \ldots, X_N) = \int_P (X_1, \ldots, X_N \mid \theta')\, P(\theta')\, \mathrm{d}\theta'$ is usually hard to compute.

Refresher: Bayesian Updating and Conjugacy (2)

$\boxed{\text{But } P(X_1, \ldots, X_N) = \int_P (X_1, \ldots, X_N \mid \theta')\, P(\theta')\, \mathrm{d}\theta' \text{ is usually hard to compute.}}$

Solution: Use a conjugate prior so that the posterior can be computed in closed form.

E.g., the Dirichlet is conjugate to the Multinomial distribution.

The Multinomial Distribution (1)

The Multinomial distribution gives the probability of N outcomes to be distributed over K categories:

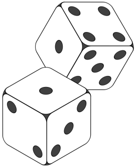
$$(X_1, \ldots, X_N) \sim \text{Multi}(n, (q_1, \ldots, q_K)),$$

where

- X_i is the number of times that the ith category occurred amongst the N events;
- $\vec{q} = (q_1, \ldots, q_K)$ gives the probabilities for each of the K categories to occur.

The Multinomial Distribution (2)

Example: Roll a die 5 times. What are the outcomes?



Let $X_i \in \{1, 2, 3, 4, 5, 6\}$ be the outcome of the ith roll. Then

$$\vec{X} = (X_1, X_2, X_3, X_4, X_5) \sim \mathsf{Multi}(n, \vec{q}),$$

with $n = 5$ and $\vec{q} = \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$.

Bayesian Updating for the Dirichlet-Multinomial

Use a Dirichlet prior over the probability vector $\vec{Q}$:

$$\vec{Q} \sim \mathrm{Dir}\left(\vec{\alpha}\right)$$

$$+$$

$$\left(\vec{X} \mid \vec{Q}\right) \sim \mathrm{Multi}\left(n, \vec{Q}\right)$$

$$\Downarrow$$

$$\left(\vec{Q} \mid \vec{X} = \vec{x}\right) \sim \mathrm{Dir}\left(\vec{\alpha} + \vec{x}\right).$$

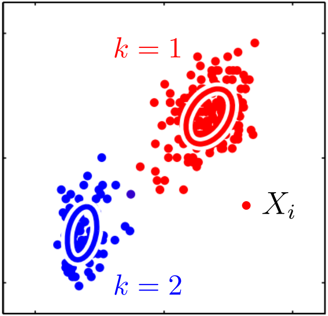
Application: Dirichlet-Multinomial Mixture Model

$\vec{Q} \mid \vec{\alpha}$	$\sim$	$\mathsf{Dir}\left(\vec{\alpha}\right)$	⬅ component assignment probability
$Z_1, \ldots, Z_N \mid \vec{Q}$	$\sim$	$\mathsf{Multi}\left(N, \vec{Q}\right)$	⬅ component assignment variables $Z_i \in \{1, \ldots, k\}$
θ_k	$\sim$	$\mathcal{G}$	⬅ prior over parameters of component distribution
$X_i \mid Z_i, \theta_{Z_i}$	$\sim$	$\mathcal{F}\left(\theta_{Z_i}\right)$	⬅ component likelihood distribution

Application: Dirichlet-Multinomial Mixture Model

$$\vec{Q} \mid \vec{\alpha} \quad \sim \quad \text{Dir}\left(\vec{\alpha}\right)$$

component assignment probability

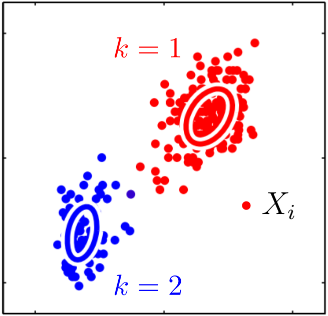


Determines $P\left(k\right)$.

Application: Dirichlet-Multinomial Mixture Model

$$Z_1, \ldots, Z_N \mid \vec{Q} \sim \text{Multi}\left(N, \vec{Q}\right)$$

component assignment variables $Z_i \in \{1, \ldots, k\}$

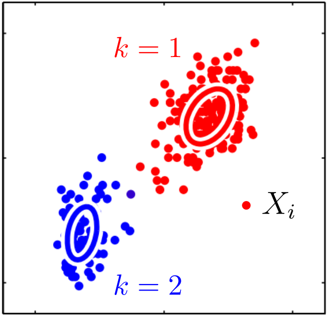


Determines $P\left(Z_i = k \mid \vec{Q}\right)$.

Application: Dirichlet-Multinomial Mixture Model

$$\theta_k \sim \mathcal{G}$$

prior over parameters of component distribution

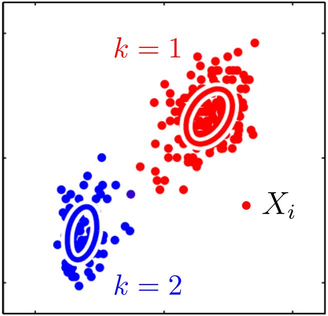


Determines $P(\theta_k)$.

Application: Dirichlet-Multinomial Mixture Model

$$X_i \mid Z_i, \theta_{Z_i} \quad \sim \quad \mathcal{F}\left(\theta_{Z_i}\right)$$

component likelihood distribution



Determines $\quad P\left(X_i \mid Z_i, \theta_{Z_i}\right)$.

Application: Dirichlet-Multinomial Mixture Model

By conjugacy, the posterior is tractable to compute:

$$P(Z_1, \ldots, Z_N, \theta_1, \ldots, \theta_K \mid X_1, \ldots, X_N)$$

Application: Dirichlet-Multinomial Mixture Model

By conjugacy, the posterior is tractable to compute:

$$P(Z_1, \ldots, Z_N, \theta_1, \ldots, \theta_K \mid X_1, \ldots, X_N)$$

- Allows us to answer:
  - What cluster do the instances belong to? $(Z_1, \ldots, Z_N)$
  - What are the properties of the clusters? $(\theta_1, \ldots, \theta_K)$

Application: Dirichlet-Multinomial Mixture Model

By conjugacy, the posterior is tractable to compute:

$$P(Z_1, \ldots, Z_N, \theta_1, \ldots, \theta_K \mid X_1, \ldots, X_N)$$

- Allows us to answer:
  - What cluster do the instances belong to? $(Z_1, \ldots, Z_N)$
  - What are the properties of the clusters? $(\theta_1, \ldots, \theta_K)$
- Applications:
  - X_i could be a document and Z_i the topic of X_i.

Background: Stochastic Processes (1)

A stochastic process is a collection of random variables indexed by some index set:

$$\{X_i\} \qquad i \in \mathcal{I} \qquad X_i \sim \mathcal{D}(\theta).$$

Background: Stochastic Processes (1)

A stochastic process is a collection of random variables indexed by some index set:

$$\{X_i\} \qquad i \in \mathcal{I} \qquad X_i \sim \mathcal{D}(\theta).$$

Any finite subset of these variables has a joint distribution; e.g.,

$$p\left(X_{j_1}, \ldots, X_{j_n}\right) \sim \mathbb{D}(\theta), \qquad (j_1, \ldots, j_n) \subset \mathbb{N}.$$

Background: Stochastic Processes (2)

Why use a stochastic process instead of a set of random variables?

Background: Stochastic Processes (2)

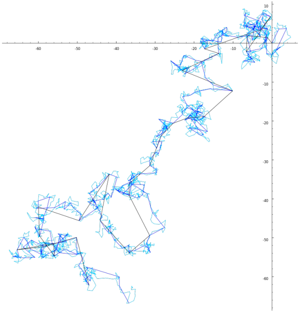
Why use a stochastic process instead of a set of random variables?

- Index set has unknown dimension
  - e.g., Topic modelling: words $w \in \{\text{corpus}\}$ are indexed by topics t, where the number of topics is not known but inferred from the data (Teh et al. [2006])

Background: Stochastic Processes (2)

Why use a stochastic process instead of a set of random variables?

- The index set is infinite-dimensional
  - e.g., Brownian motion: the random displacement $X(t)$ of a particle depend on a continuous time $t \in \mathbb{R}$



Intuition for the Dirichlet Process

We want to generalize the Dirichlet distribution to be a pmf over an infinite number of events:

$$\text{Dir}\left(\alpha_1, \ldots, \alpha_K\right) \qquad\qquad K \to \infty$$

It is a non-parametric model since the number of categories K is not fixed, and instead grows with the data.

Intuition for the Dirichlet Process

We want to generalize the Dirichlet distribution to be a pmf over an infinite number of events:

$$\text{Dir}\left(\alpha_1, \ldots, \alpha_K\right) \qquad\qquad K \to \infty$$

It is a non-parametric model since the number of categories K is not fixed, and instead grows with the data.

It becomes a Dirichlet process instead of a distribution.

Background: σ-Algebra (1)

A σ-Algebra over a set $\mathcal{B}$ is a collection of subsets of $\mathcal{B}$ that is *closed* under countably many of the following operations:

Complement: if $A \in \mathcal{B}$ then $A^C \in \mathcal{B}$;

Union: if $A_1, A_2, \cdots \in \mathcal{B}$ then $\bigcup A_i \in \mathcal{B}$;

Intersection: if $A_1, A_2, \cdots \in \mathcal{B}$ then $\bigcap A_i \in \mathcal{B}$.

Background: σ-Algebra (2)

Example: Let $\mathcal{B} = \{\, \textcolor{green}{\bullet}, \textcolor{yellow}{\bullet}, \textcolor{red}{\bullet} \,\}$ be the possible colours of a traffic light that you encounter while driving.

Then

$$\sigma = \Big\{ \textcolor{green}{\bullet}, \textcolor{yellow}{\bullet}, \textcolor{red}{\bullet}, \{\textcolor{green}{\bullet}, \textcolor{yellow}{\bullet}\}, \{\textcolor{green}{\bullet}, \textcolor{red}{\bullet}\}, \\ \{\textcolor{yellow}{\bullet}, \textcolor{red}{\bullet}\}, \{\textcolor{green}{\bullet}, \textcolor{yellow}{\bullet}, \textcolor{red}{\bullet}\}, \varnothing \Big\}$$

is a σ-algebra on $\mathcal{B}$.

Dirichlet Process: Introduction (1)

A Dirichlet process takes as its index set a σ-algebra over a space $\mathcal{B}$:

$$\forall \text{ sets } B \in \sigma\left(\mathcal{B}\right), \quad \tilde{P}(B) \in [0,1] \qquad \text{is a random variable.}$$

Dirichlet Process: Introduction (1)

A Dirichlet process takes as its index set a σ-algebra over a space $\mathcal{B}$:

$$\forall \text{ sets } B \in \sigma(\mathcal{B}), \quad \tilde{P}(B) \in [0,1] \qquad \text{is a random variable.}$$

Dirichlet Process: Introduction (2)

Marginalisation Property: For any finite partition $(B_1, \ldots, B_N)$ of the space $\mathcal{B}$, if $G \sim \text{DP}(\alpha, H)$, then

$$\begin{bmatrix} \tilde{P}(B_1) \\ \vdots \\ \tilde{P}(B_N) \end{bmatrix} \sim \text{Dir}\begin{pmatrix} \alpha H(B_1) \\ \vdots \\ \alpha H(B_N) \end{pmatrix}$$

Dirichlet Process: Introduction (2)

Marginalisation Property: For any finite partition $(B_1, \ldots, B_N)$ of the space $\mathcal{B}$, if $G \sim \mathsf{DP}(\alpha, H)$, then

$$\begin{bmatrix} \tilde{P}(B_1) \\ \vdots \\ \tilde{P}(B_N) \end{bmatrix} \sim \mathrm{Dir} \begin{pmatrix} \alpha H(B_1) \\ \vdots \\ \alpha H(B_N) \end{pmatrix}$$

Implication: Draws from the Dirichlet Process are random probability distributions.

Example of a “DP”’ ’ Indexed by a Finite σ-algebra

$$\begin{bmatrix} \tilde{P}\left(\{{\color{green}\bullet}, {\color{yellow}\bullet}\}\right) \\ \\ \tilde{P}\left({\color{red}\bullet}\right) \end{bmatrix} \sim \mathsf{Dir}\begin{pmatrix} \alpha H\left(\{{\color{green}\bullet}, {\color{yellow}\bullet}\}\right) \\ \\ \alpha H\left({\color{red}\bullet}\right) \end{pmatrix}$$

Example of a “DP”’ ’ Indexed by a Finite σ-algebra

$$\begin{bmatrix} \tilde{P}\left(\{{\color{green}\bullet}, {\color{yellow}\bullet}, {\color{red}\bullet}\}\right) \\ \\ \tilde{P}\left(\varnothing\right) \end{bmatrix} \sim \mathsf{Dir}\begin{pmatrix} \alpha H\left(\{{\color{green}\bullet}, {\color{yellow}\bullet}, {\color{red}\bullet}\}\right) \\ \\ \alpha H\left(\varnothing\right) \end{pmatrix}$$

Example of a "DP"' Indexed by a Finite σ-algebra

$$\begin{bmatrix} \tilde{P}\left(\textcolor{green}{\bullet}\right) \\ \tilde{P}\left(\textcolor{yellow}{\bullet}\right) \\ \tilde{P}\left(\textcolor{red}{\bullet}\right) \end{bmatrix} \sim \mathsf{Dir}\begin{pmatrix} \alpha H\left(\textcolor{green}{\bullet}\right) \\ \alpha H\left(\textcolor{yellow}{\bullet}\right) \\ \alpha H\left(\textcolor{red}{\bullet}\right) \end{pmatrix}$$

Dirichlet Process: Definition

$$\left[\tilde{P}\left(B_1\right), \quad \ldots, \quad \tilde{P}\left(B_N\right)\right] \sim \mathsf{Dir}\left(\alpha H(B_1), \ldots, \alpha H\left(B_N\right)\right)$$

H is the (non-random) base distribution over $\mathcal{B}$.

(Can be any probability distribution over $\mathcal{B}$.)

It determines the mean of the DP for any set B:

$$\mathbb{E}\left(\tilde{P}(B)\right) = H(B).$$

Dirichlet Process: Definition

$$\boxed{\left[\tilde{P}\left(B_1\right), \quad \ldots, \quad \tilde{P}\left(B_N\right)\right] \sim \mathrm{Dir}\left(\alpha H(B_1), \ldots, \alpha H\left(B_N\right)\right)}$$

α is a positive real concentration parameter.

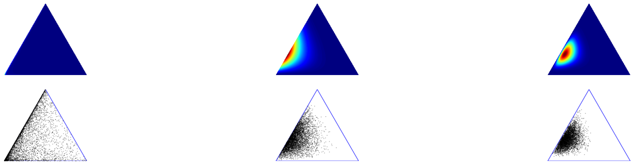
It determines the concentration of the DP about the mean:

$$\mathrm{Var}\left(\tilde{P}(B)\right) = \frac{H(B)(1 - H(B))}{\alpha + 1}.$$

Example for a Finite σ-algebra

Suppose the base distribution is given by

$$H\left({\color{green}\bullet}\right) = 0.6 \qquad H\left({\color{yellow}\bullet}\right) = 0.1 \qquad H\left({\color{red}\bullet}\right) = 0.3.$$



$\alpha = 2$ $\qquad$ $\alpha = 10$ $\qquad$ $\alpha = 20$

Generalization to Infinite σ-algebra (1)

Suppose for a space $\mathcal{B}$, $\sigma(\mathcal{B})$ is (countably) infinite.

Generalization to Infinite σ-algebra (1)

Suppose for a space $\mathcal{B}$, $\sigma(\mathcal{B})$ is (countably) infinite.

Example: We want to create a generative unigram model of a text.

Generalization to Infinite σ-algebra (1)

Suppose for a space $\mathcal{B}$, $\sigma(\mathcal{B})$ is (countably) infinite.

Example: We want to create a generative unigram model of a text.

We want to assign non-zero probability to the next word in the document.

Generalization to Infinite σ-algebra (1)

Suppose for a space $\mathcal{B}$, $\sigma(\mathcal{B})$ is (countably) infinite.

Example: We want to create a generative unigram model of a text.

We want to assign non-zero probability to the next word in the document.

But how do we distribute the probability mass when the number of words (partitions) is unknown (and possibly infinite)?

Generalization to Infinite σ-algebra (2)

Solution: Model the document as a Dirichlet Process DP(α, H) (with H initially uniform).

Then a vocabulary of N symbols corresponds to a partition into N parts.

Generalization to Infinite σ-algebra (3)

Under this model, the probability of a symbol w is given by

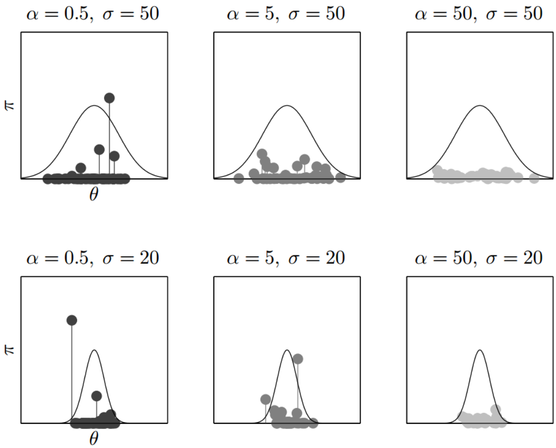
$$P(w) = \frac{F_w}{F + \alpha}$$

if w is seen, and

$$P(w) = \frac{\alpha}{F + \alpha}$$

if w is unseen.

Draws from a DP with a Gaussian H



Rows vary σ; columns vary α.

Recall: Dirichlet-Multinomial Mixture Model (1)

$\vec{Q} \mid \vec{\alpha}$	$\sim$	$\text{Dir}\left(\vec{\alpha}\right)$	⬅	cluster assignment probability
$Z_1, \ldots, Z_N \mid \vec{Q}$	$\sim$	$\text{Multi}\left(N, \vec{Q}\right)$	⬅	cluster assignment variables
θ_k	$\sim$	$\mathcal{G}$	⬅	prior over parameters of component distribution
$X_i \mid Z_i, \theta_{Z_i}$	$\sim$	$\mathcal{F}\left(\theta_{Z_i}\right)$	⬅	component distribution

Recall: Dirichlet-Multinomial Mixture Model (2)

$$\boxed{(Q_1, \ldots, Q_k) \mid \alpha_1, \ldots, \alpha_k \quad \sim \quad \text{Dir}\,(\alpha_1, \ldots, \alpha_k)}$$

What if $K \to \infty$?

Dirichlet Process Mixture Model (1)

$$\mathcal{Q} \mid \alpha, H \sim \mathsf{DP}(\alpha, H)$$

random distribution over parameters

$$\theta_i \mid \mathcal{Q} \sim \mathcal{Q}$$

latent parameter for X_i

$$X_i \mid \theta_i \sim \mathcal{F}(\theta_i)$$

likelihood distribution

Dirichlet Process Mixture Model (1)

$$
\begin{aligned}
\mathcal{Q} \mid \alpha, H &\sim \mathrm{DP}(\alpha, H) \\
\theta_i \mid \mathcal{Q} &\sim \mathcal{Q} \\
X_i \mid \theta_i &\sim \mathcal{F}(\theta_i)
\end{aligned}
$$

⬅ random distribution over parameters

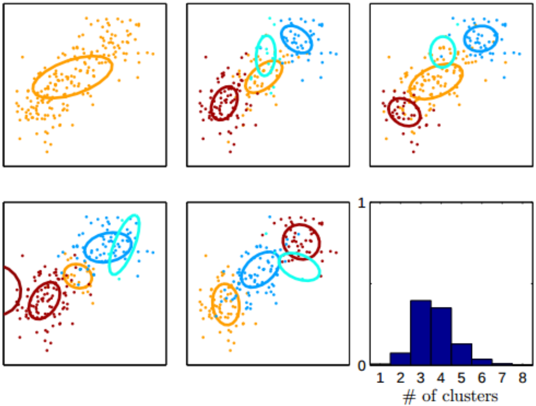
⬅ latent parameter for X_i

⬅ likelihood distribution

$\mathcal{Q}$ is discrete, so multiple θ_i can take on the same value (i.e., they cluster).

Dirichlet Process Mixture Model (2)

Posterior inference for the number of clusters can be done using Markov Chain Monte Carlo (MCMC):



Application: Dirichlet Processes in NLP

Language modelling: The number of words is unbounded.

Topic modelling: The number of topics is inferred from the data.

Discussion: Relevance to the Word Learning Model

In the present model, the meaning probability corresponds to the expected value of the posterior of a Dirichlet distribution:

$$p_t(f \mid w) = \frac{\mathsf{assoc}_t(w, f) + \gamma}{\sum_{f'} \mathsf{assoc}_t(w, f') + k \cdot \gamma}$$

However, we fix the number of features k ahead of time.

Discussion: Relevance to the Word Learning Model

In the present model, the meaning probability corresponds to the expected value of the posterior of a Dirichlet distribution:

$$p_t(f \mid w) = \frac{\text{assoc}_t(w, f) + \gamma}{\sum_{f'} \text{assoc}_t(w, f') + k \cdot \gamma}$$

However, we fix the number of features k ahead of time.

Can we effectively use a Dirichlet Process to allow the number of features to grow with the data?

References

B. A. Frigyik, A. Kapila, and M. R. Gupta. Introduction to the Dirichlet Distribution and Related Processes. Technical report, 2010. URL `http://scholar.google.com/scholar?hl=en{&}btnG=Search{&}q=intitle:Introduction+to+the+Dirichlet+Distribution+and+Related+Processes{#}0`.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476), 2006.