The Dirichlet distribution is a distribution on the set
\[ \{ (p_1, \ldots, p_r) \mid \forall j \ p_j \geq 0, \ \Sigma p_j = 1 \} \]
with pdf
\[ f (p_1, \ldots, p_{r-1}) = \frac{1}{B} \frac{\Gamma \left( \sum \alpha_j \right)}{\prod \Gamma (\alpha_j)} p_1^{\alpha_1-1} \cdots p_{r-1}^{\alpha_{r-1}-1} \]
where \( \alpha_j > 0 \) and \( B \) is a normalizing constant.

Properties

- Mean \[ E[p_j] = \frac{\alpha_j}{\alpha_0}, \quad \alpha_0 = \alpha_1 + \cdots + \alpha_r \]
- Variance \[ \text{Var} [p_j] = \frac{\alpha_j (\alpha_0 - \alpha_j)}{\alpha_0^2 (\alpha_0 + 1)} \]
- If \((n_1, \ldots, n_r) \sim \text{mult} (p_1, \ldots, p_r)\)
  and \((p_1, \ldots, p_r) \sim \text{Dir} (\alpha_1, \ldots, \alpha_r)\)
then \((p_1, \ldots, p_r) \mid (n_1, \ldots, n_r) \sim \text{Dir} (\alpha_1 + n_1, \ldots, \alpha_r + n_r)\)
Let $\mathcal{X}$ be a sample space, and

- $H$ a distribution on $\mathcal{X}$
- $\alpha > 0$

A Dirichlet process with base distribution $H$ and concentration parameter $\alpha$ is a distribution on distributions on $\mathcal{X}$, denoted $\text{DP}(\alpha, H)$, such that, for $G \sim \text{DP}(\alpha, H)$, and for any partition $A_1, \ldots, A_r$ of $\mathcal{X}$,

$$(G(A_1), \ldots, G(A_r)) \sim \text{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r)).$$

Remarks

- Existence of DP's not obvious. We'll see a construction later.
- Like Gaussian process, all finite dimensional "marginals" are Dirichlet distributed.
• For any \( A \subseteq \Theta \),

\[
E[G(A)] = H(A)
\]

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

**Posterior**

Suppose

\[
G \sim \text{DP}(\alpha, H)
\]

\( \theta_1, \ldots, \theta_n \) iid \( G \)

What is \( \text{dist. of } G/\theta_1, \ldots, \theta_n \)?

Let \( A_1, \ldots, A_r \) be a partition of \( \Theta \).

Let

\[
\eta_k = \# \{ i : \theta_i \in A_k \}
\]

By the conjugacy property,

\[
(G(A_1), \ldots, G(A_r)) | \theta_1, \ldots, \theta_n
\]

\[
\sim \text{Dir}(aH(A_1) + n_1, \ldots, a_r H(A_r) + n_r)
\]

\( \Rightarrow \) posterior also a DP. What are its base dist. and concentration parameter?
We have
\[ (G(A_1), \ldots, G(A_r)) \mid \theta_1, \ldots, \theta_n \sim \text{DP}(\beta, L) \]

where, \(\forall k,\)
\[ \beta L(A_k) = \alpha H(A_k) + \eta_k \]

Summing over \(k,\)
\[ \beta = \alpha + n \]

where \(n := n_1 + \cdots + n_r.\) Then,
\[ L(A_k) = \frac{\alpha H(A_k) + \sum_{i=1}^{n} \delta_{\theta_i}(A_k)}{\alpha + n} \]

where
\[ \delta_{\theta}(A) = \begin{cases} 1 & \text{if } \theta \in A \\ 0 & \text{if } \theta \notin A \end{cases} \]

Thus
\[ L = \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n} \]
\[ = \frac{\alpha}{\alpha + n} H + \frac{n}{\alpha + n} \cdot \frac{1}{n} \sum_{i=1}^{n} \delta_{\theta_i} \]

\(\text{empirical distribution}\)
Predictive Distribution

Once again, suppose
\[ G \sim \text{DP}(\alpha, H) \]

\[ \theta_1, \ldots, \theta_n \sim G \]

What is the "predictive" distribution of
\[ \theta_{n+1} \mid \theta_1, \ldots, \theta_n \]

with \( G \) marginalized out?

Let \( A \subseteq \Theta \). Then

\[
\Pr \{ \theta_{n+1} \in A \mid \theta_1, \ldots, \theta_n \}
\]

\[
= E_{\theta_{n+1} \mid \theta_1, \ldots, \theta_n} \left[ 1 \{ \theta_{n+1} \in A \} \right]
\]

\[
= E_{G \mid \theta_1, \ldots, \theta_n} \left[ E_{\theta_{n+1} \mid G, \theta_1, \ldots, \theta_n} \left[ 1 \{ \theta_{n+1} \in A \} \right] \right]
\]

\[
= E_{G \mid \theta_1, \ldots, \theta_n} \left[ E_{\theta_{n+1} \mid G} \left[ 1 \{ \theta_{n+1} \in A \} \right] \right]
\]

\[
= E_{G \mid \theta_1, \ldots, \theta_n} \left[ G (A) \right]
\]

\[
= \frac{1}{\alpha + n} \left( \alpha H(A) + \sum \delta_{\theta_i} (A) \right)
\]
Thus
\[
\Theta_{n+1} | \Theta_1, \ldots, \Theta_n \sim \frac{1}{a+n} \left( aH + \sum_{i=1}^{n} \Theta_i \right)
\]

Remarks

- With probability \( \frac{n}{a+n} \), \( \Theta_{n+1} \) will take on a previous value.
- If \( \Theta_i \) has been drawn once, then with positive probability it will be drawn again.

\[ \Rightarrow \text{G is discrete} \]

Clustering and the CRP

Consider \( \Theta_1, \ldots, \Theta_n \). Since \( \Theta_i \)'s are repeated, let \( \Theta_1^*, \ldots, \Theta_m^* \) be the distinct values among \( \Theta_1, \ldots, \Theta_n \), and let \( n_k = \# \) of repeats of \( \Theta_k^* \). Then
\[
\Theta_{n+1} | \Theta_1, \ldots, \Theta_n \sim \frac{1}{a+n} \left( aH + \sum n_k \Theta_k^* \right)
\]
Rich-get-richer: $\Theta_k^*$ repeated with probability proportional to $n_k$.

What is the expected number of distinct values among $\Theta_1, \ldots, \Theta_n$?

$$E[m; n] = E\left[ \sum_{i=1}^{n} 1 \{ \Theta_i \text{ is a new value} \} \right]$$

$$= \sum_{i=1}^{n} E\left[ 1 \right] 1 \{ \Theta_i \text{ is a new value} \}$$

$$= \sum_{i=1}^{n} 1 \{ \Theta_i \text{ is a new value} \}$$

$$= O(\alpha \log n)$$

Sidebar: Chinese Restaurant Process

- Chinese restaurant
- Infinite # of tables
- """" seats per table
- $(n+1)$st customer sits at a currently occupied table $k$ with prob. proportional to $n_k$, or at an unoccupied table with prob. proportional to $\alpha$.
- Induces distribution on partitions of $\{1, 2, \ldots, n\}$
**Stick Breaking Construction**

Consider the following construction

- \( \beta_k \sim \text{Beta}(1, \alpha), \quad k = 1, 2, \ldots \)

\[
\text{pdf} \propto (1 - \alpha)^{\alpha - 1}
\]

- \( \pi_k = \beta_k \cdot \frac{k-1}{\sum_{l=1}^{k-1} (1-\beta_l)} \)

- \( \theta_k \overset{\text{iid}}{\sim} \mathcal{H} \)

- \( G = \sum_{k=1}^{\infty} \pi_k \int_{\theta_k} \)

- \( 0 \quad \pi_1 \quad \pi_2 \quad \pi_3 \quad \pi_4 \quad \pi_5 \quad \pi_6 \quad \pi_7 \ldots \)

- \( 1 \)

**Fact:** \( G \sim \text{DP}(\alpha, \mathcal{H}) \)

**Notation:** \( \pi \sim \text{GEM}(\alpha) \)
Let's apply DPs to mixture modeling.
Consider $X_1, ..., X_n \in \mathbb{R}$.

Suppose
$$X_i \sim N(\mu_i, \sigma_i).$$

Let
$$\Theta_i = (\mu_i, \sigma_i)$$
$$\Theta = \mathbb{R} \times \mathbb{R}^+$$

Further suppose
$$\Theta_i \sim G$$
$$G \sim \text{DP}(\alpha, \Theta)$$
From stick-breaking perspective

\[ G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \]

- number of clusters not fixed a priori
- number of clusters may grow with \( n \)

**Inference**

Estimate \( \{ \theta_i \} \) via MCMC

**Reference**

Yee Whye Teh, "Dirichlet Processes,"
Encyclopedia of Machine Learning, 2010

**Key**

\[ A = \frac{\alpha}{\alpha + i - 1} \]