DIRICHLET PROCESSES

Dirichlet Distribution

The Dirichlet distribution is a distribution on the set

$$\{ (p_1, \dots, p_r) \mid \forall j \; p_j \geq 0, \; \Sigma p_j = 1 \}$$

with pdf

$$f(p_1, \dots, p_{r-1}) = \frac{1}{B} \prod_{j=1}^{r} p_j^{\alpha_j - 1}$$

where $\alpha_j > 0$ and B is a normalizing constant.

Properties

- Mean $E[p_j] = \dfrac{\alpha_j}{\alpha_0}$, $\alpha_0 = \alpha_1 + \dots + \alpha_r$
- Variance $Var[p_j] = \dfrac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}$
- If

$$(n_1, \dots, n_r) \sim mult(p_1, \dots, p_r)$$

and

$$(p_1, \dots, p_r) \sim Dir(\alpha_1, \dots, \alpha_r)$$

then

$$(p_1, \dots, p_r) \mid (n_1, \dots, n_r) \sim Dir(\alpha_1 + n_1, \dots, \alpha_r + n_r)$$

Dirichlet Process

Let Θ be a sample space, and

- H a distribution on Θ
- $\alpha > 0$

A Dirichlet process with <u>base distribution</u> H and <u>concentration parameter</u> α is a distribution on distributions on Θ, denoted $DP(\alpha, H)$, such that, for $G \sim DP(\alpha, H)$, and for any partition $A_1, \ldots, A_r$ of Θ,

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

<u>Remarks</u>

- Existence of DPs not obvious. We'll see a construction later
- Like Gaussian process, all finite dimensional "marginals" are Dirichlet distributed.

- For any $A \subseteq \Theta$,

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

$$Var(G(A)) = \frac{H(A)(1 - H(A))}{\alpha + 1}$$

$\boxed{\text{Posterior}}$

Suppose

$$G \sim DP(\alpha, H)$$

$$\theta_1, \dots, \theta_n \overset{iid}{\sim} G$$

What is dist. of $G \mid \theta_1, \dots, \theta_n$?

Let $A_1, \dots, A_r$ be a partition of Θ.

Let

$$n_k = \#\{i : \theta_i \in A_k\}$$

By the conjugacy property,

$$(G(A_1), \dots, G(A_r)) \mid \theta_1, \dots, \theta_n$$

$$\sim Dir(\alpha H(A_1) + n_1, \dots, \alpha_r H(A_r) + n_r)$$

$\Rightarrow$ posterior also a DP. What are its base dist. and concentration parameter?

We have

$$(G(A_1), \ldots, G(A_r)) \mid \theta_1, \ldots, \theta_n \sim DP(\beta, L)$$

where, $\forall k$,

$$\beta L(A_k) = \alpha H(A_k) + n_k$$

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

$$L = \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}$$

$$= \frac{\alpha}{\alpha+n} H + \frac{n}{\alpha+n} \cdot \frac{1}{n} \sum_{i=1}^{n} \delta_{\theta_i}$$

empirical distribution

Predictive Distribution

Once again, suppose

$$G \sim DP(\alpha, H)$$

$$\theta_1, \dots, \theta_n \overset{iid}{\sim} G$$

What is the "predictive" distribution of

$$\theta_{n+1} \mid \theta_1, \dots, \theta_n$$

with G marginalized out?

Let $A \subseteq \Theta$. Then

$$\Pr\{\theta_{n+1} \in A \mid \theta_1, \dots, \theta_n\}$$

$$= E_{\theta_{n+1} \mid \theta_1, \dots, \theta_n}\left[\mathbb{1}_{\{\theta_{n+1} \in A\}}\right]$$

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

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

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

$$= \frac{1}{\alpha + n}\left(\alpha H(A) + \sum \delta_{\theta_i}(A)\right)$$

Thus

$$\theta_{n+1} \mid \theta_1, \dots, \theta_n \sim \frac{1}{\alpha+n}\left(\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}\right)$$

Remarks

- With probability $\frac{n}{\alpha+n}$, θ_{n+1} will take on a previous value.
- If θ_i has been drawn once, then with positive probability it will be drawn again

$\Longrightarrow$ G is discrete

Clustering and the CRP

Consider $\theta_1, \dots, \theta_n$. Since θ_i s are repeated, let $\theta_1^*, \dots, \theta_m^*$ be the distinct values among $\theta_1, \dots, \theta_n$, and let $n_k = \#$ of repeats of θ_k^*. Then

$$\theta_{n+1} \mid \theta_1, \dots, \theta_n \sim \frac{1}{\alpha+n}\left(\alpha H + \sum n_k \delta_{\theta_k^*}\right)$$

Rich-get-richer : θ_k^* repeated with probability proportional to n_k.

What is the expected number of distinct values among $\theta_1, \ldots, \theta_n$?

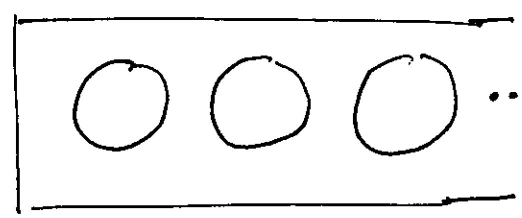
$$E[m;n] = E\left[\sum_{i=1}^{n} 1_{\{\theta_i \text{ is a new value}\}}\right]$$

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

A $$= \sum_{i=1}^{n}$$

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

Sidebar: Chinese Restaurant Process



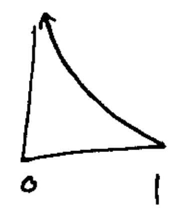
- Chinese restaurant
- Infinite # of tables
- " " " seats per table
- $(n+1)$st customer sits at a currently occupied table k with prob. proportional to n_k, or at an unoccupied table with prob. proportional to α.
- Induces distribution on partitions of $\{1, 2, \ldots, n\}$

Stick Breaking Construction

Consider the following construction

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

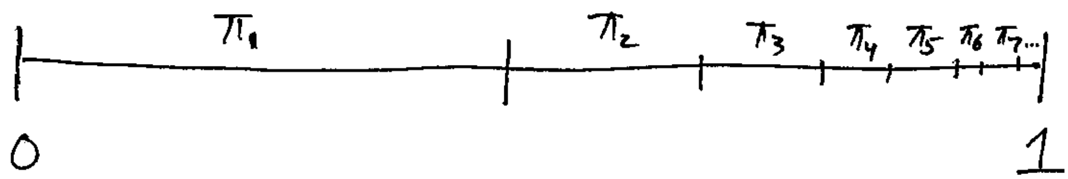
  $$\text{pdf} \propto (1-x)^{\alpha-1}$$



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

- $\theta_k^* \overset{\text{iid}}{\sim} H$

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



Fact: $G \sim DP(\alpha, H)$

Notation: $\pi \sim GEM(\alpha)$

DP Mixture Models

Let's apply DPs to mixture modeling.

Consider $x_1, \ldots, x_n \in \mathbb{R}$.

Suppose

$$x_i \sim N(\mu_i, \sigma_i).$$

Let

$$\theta_i = (\mu_i, \sigma_i)$$

$$\Theta = \mathbb{R} \times \mathbb{R}^+$$

Further suppose

$$\theta_i \sim G$$

$$G \sim DP(\alpha, H)$$

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. $\dfrac{\alpha}{\alpha + i - 1}$