Bayesian Statistics

Debdeep Pati
Florida State University

February 25, 2016

Dirichlet processes (Ferguson, 1973; 1974)

- As discussed last lecture, a simple conjugate prior for the bin probabilities corresponds to the Dirichlet distribution
- For example, we could let

$$\{P(B_1), ..., P(B_k)\} \sim Dir\{\alpha P_0(B_1), \ldots, \alpha P_0(B_k)\} \tag{1}$$

- P_0 is a "base" probability measure providing an initial guess at P & α is a prior concentration parameter
- Ferguson's idea: eliminate sensitivity to choice of $B_1, \ldots, B_k$ & induce a fully specified prior on P, through assuming (1) holds for all $B_1, \ldots, B_k$ & all k.

Dirichlet processes (Ferguson, 1973; 1974)

- For Ferguson's specification to be coherent, there must exist an RPM P such that the probs assigned to any measurable partition $B_1, \ldots, B_k$ by P is $Dir\{\alpha P_0(B_1), \ldots, \alpha P_0(B_k)\}$
- The existence of such a P can be shown by verifying the Kolmogorov consistency conditions
- The first Kolmogorov condition is automatic, since (1) is defined free of the order of the sets
- The remaining condition relates to coherence across different partitions - e.g, if we form a new partition by taking unions of some of the sets in $B_1, \ldots, B_k$ then the resulting probs assigned to this new partition must still be Dirichlet with the same form

Dirichlet process: a prior for the space of probability distributions

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

$$\Delta_K = \{(\pi_1, \pi_2, \ldots, \pi_k) : \pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1\}$$

- We say $(\pi_1, \ldots, \pi_k)$ is Dirichlet distributed $(\lambda_1, \lambda_2, \ldots, \lambda_k)$ if

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

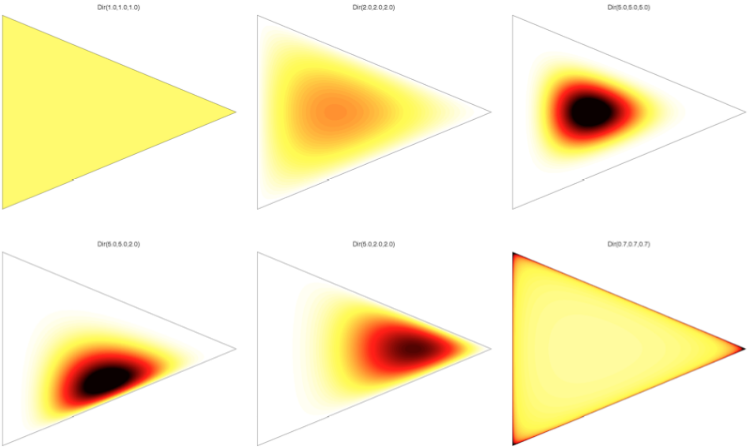
- Equivalent to normalizing a set of independent gamma variables

$$(\pi_1, \ldots, \pi_k) \stackrel{d}{=} \frac{1}{\sum_k \gamma_k}(\gamma_1, \ldots, \gamma_k)$$

$$\gamma_j \sim \text{Gamma}(\lambda_k, \beta)$$

Dirichlet distribution

Figure: Dirichlet distribution



Agglomerative & Decimative properties of DP

- Combining entries by their sum

$$\begin{aligned} (\pi_1, \ldots, \pi_K) &\sim \text{Diri}(\alpha_1, \ldots, \alpha_K) \\ (\pi_1, \ldots, \pi_i + \pi_j \ldots, \pi_K) &\sim \text{Diri}(\alpha_1, \ldots, \alpha_i + \alpha_j, \ldots \alpha_K) \end{aligned}$$

- Decimating one entry into two

$$\begin{aligned} (\pi_1, \ldots, \pi_K) &\sim \text{Diri}(\alpha_1, \ldots, \alpha_K) \\ (\tau_1, \tau_2) &\sim \text{Diri}(\alpha_i \beta_1, \alpha_i \beta_2) \\ (\pi_1, \ldots, \pi_i \tau_1, \pi_i \tau_2, \ldots, \pi_K) &\sim \text{Diri}(\alpha_1, \ldots, \alpha_i \beta_1, \alpha_i \beta_2, \ldots, \alpha_K) \end{aligned}$$

Existence of Dirichlet process

- $(B'_1, \ldots, B'_{k'})$ and $(B_1, \ldots, B_k)$ are measurable partitions
- $(B'_1, \ldots, B'_{k'})$ is a refinement of $(B_1, \ldots, B_k)$s with $B_1 = \cup_1^{r_1} B'_j, B_2 = \cup_{r_1+1}^{r_2} B'_j, \ldots B_k = \cup_{r_{k-1}+1}^{k'} B'_j$
- Then, the distribution of $P(B'_1), \ldots, P(B'_{k'})$ induces a distribution on

$$\sum_1^{r_1} P(B'_j), \sum_{r_1+1}^{r_2} P(B'_j), \cdots, \sum_{r_{k-1}+1}^{k'} P(B'_j)$$

  which is equivalent to the distribution of $P(B_1), \ldots, P(B_k)$.
- Ferguson shows this condition is sufficient for Kolmogorov consistency

Moment properties of the DP

- Let $P \sim \mathrm{DP}(\alpha, P_0)$ denote that the probability measure P on $(\Omega, \mathcal{B})$ is assigned a Dirichlet process (DP) prior with scalar precision $\alpha > 0$ and base probability measure P_0
- From the definition of the Dirichlet process & properties of the Dirichlet, we have

$$P(B) \sim beta[\alpha P_0(B), \alpha\{1 - P_0(B)\}], \text{ for all } B \in \mathcal{B}.$$

- Hence, we have $E\{P(B)\} = P_0(B)$, for all $B \in \mathcal{B}$, so that the prior for P is centered on P_0
- In addition, we have

$$V\{P(B)\} = \frac{P_0(B)\{1 - P_0(B)\}}{1 + \alpha}, \text{ for all } B \in \mathcal{B},$$

so that α is a precision parameter controlling the variance

Large Support of the DP

- Let $Q \in \mathcal{P}$ denote a fixed probability measure on $(\Omega, \mathcal{B}$
- From proposition 3 in Ferguson (1973), for any positive integer k, measurable sets $B_1, \ldots, B_k$ and $\epsilon > 0$,

$$P\{|P(B_i) - Q(B_i)| < \epsilon\} \text{ for } i = 1, \ldots, k\} > 0.$$

- The topology of pointwise convergence corresponds to $P_n \to P$,iff every $B \in \mathcal{B}$,$P_n(B) \to P(B)$.
- Under this topology, the support of the DP contains all probability measures whose support is contained in the support of P_0.

Conjugacy

- Let $P \sim \mathrm{DP}(\alpha, P_0)$ and let $y_i \sim P$ i.i.d (following standard practice in using P to denote both the probability measure and its corresponding distribution)
- For any measurable partition $B_1, \ldots, B_k$, we have

$$\{P(B_1), \ldots, P(B_k) \mid y_1, \ldots, y_n\} \sim \mathrm{Diri}\left\{\alpha P_0(B_1) + \sum_{i=1}^{n} I(y_i \in B_1), \ldots, \alpha P_0(B_k) + \sum_{i=1}^{n} I(y_i \in B_k)\right\}$$

- From this & the above development, it is straightforward to obtain

$$(P \mid y_1, \ldots, y_n) \sim \mathrm{DP}\left(\alpha P_0 + \sum_i \delta_{y_i}\right)$$

Posterior of the DP

- The updated precision parameter is $\alpha + n$, so that α is in some sense a prior sample size
- The posterior expectation of P is defined as

$$E\{P(B) \mid y\} = \frac{\alpha}{\alpha + n} P_0(B) + \frac{n}{\alpha + n} \sum_i \frac{1}{n} \delta_{y_i}$$

- Hence, the Bayes estimator of P under squared error loss is the empirical measure with equal masses at the data points shrunk towards the base measure.

Bayesian Bootstrap

- Note that in the limit as $\alpha \to 0$ we obtain the posterior,

$$(P \mid y^n) \sim DP\left(\sum_{i=1}^{n} \delta_{y_i} \right)$$

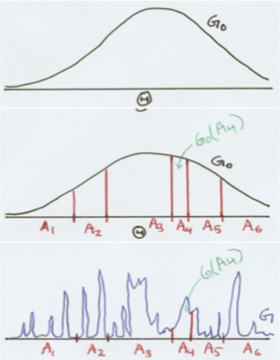
- This limiting posterior is known as the Bayesian bootstrap
- Samples from the Bayesian bootstrap correspond to discrete distributions supported at the observed data points with Dirichlet distributed weights
- Compared with the typical Efron bootstrap, the Bayesian bootstrap leads to smoothing of the weights

Dirichlet process (Ferguson 1973)

- Let Θ be a measurable space, G_0 be a probability measure (base) on Θ and $\alpha > 0$ is (precision / concentration).
- $G \sim DP(\cdot \mid G_0, \alpha)$ if for all $A_1, \ldots, A_k$ finite partitions of Θ,

$$(G(A_1), G(A_2), \ldots, G(A_K)) \sim \text{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))$$

Figure: DP



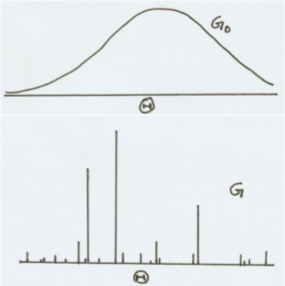
$G \sim DP(\cdot \mid G_0, \alpha)$: What does it look like

- $\{G(B) : B \in \mathcal{B}\}$ is a stochastic process
- Samples from DP are discrete with probability one. In fact,

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta), \theta_k \sim G_0$$

- $E(G) = G_0$
- As $\alpha \to \infty$, G looks more like G_0

Figure: DP realizations



Representations of Dirichlet process

- Posterior Dirichlet process

$$\left[\begin{array}{c} G \sim DP(\cdot \mid \alpha, G_0) \\ \theta \mid G \sim G \end{array}\right] \Leftrightarrow \left[\begin{array}{c} \theta \sim G_0 \\ G \mid \theta \sim DP\left(\cdot, \alpha+1, \frac{\alpha G_0 + \delta_\theta}{\alpha+1}\right) \end{array}\right]$$

- Pólya Urn Scheme

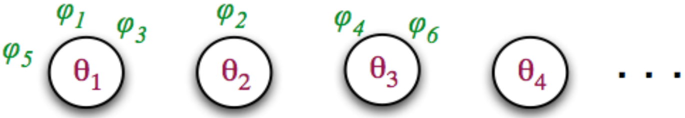
$$\begin{aligned} \theta' \mid \theta, G_0 &= \int [\theta' \mid G][G \mid \theta] dG = \int G[G \mid \theta] dG = \frac{\alpha G_0 + \delta_\theta}{\alpha + 1} \\ \theta_n \quad \mid & \quad \theta_1, \ldots, \theta_{n-1}, G_0 \sim \frac{\alpha G_0 + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1} \end{aligned}$$

Chinese Restaurant process

- This shows the clustering effect explicitly.
- Restaurant has infinitely many tables $k = 1, \ldots$.
- Tables have values θ_k drawn from G_0.
- Customers are indexed by $i = 1, \ldots$ with values ϕ_i.
- $K =$ total number of occupied tables so far.
- $n =$ total number of customers so far.
- $n_k =$ number of customers seated at table k.

Chinese Restaurant process

Figure: CRP



Generating from a CRP:

customer 1 enters the restaurant and sits at table 1.

$\phi_1 = \theta_1$ where $\theta_1 \sim G_0$, $K = 1$, $n = 1$, $n_1 = 1$

for $n = 2, \ldots,$

customer n sits at table $\begin{cases} k & \text{with prob } \frac{n_k}{n-1+\alpha} \quad \text{for } k = 1 \ldots K \\ K+1 & \text{with prob } \frac{\alpha}{n-1+\alpha} \quad \text{(new table)} \end{cases}$

if new table was chosen **then** $K \leftarrow K + 1$, $\theta_{K+1} \sim G_0$ **endif**

set ϕ_n to θ_k of the table k that customer n sat at; set $n_k \leftarrow n_k + 1$

endfor

Relationship between CRP and DP

- DP is a distribution over distributions
- DP results in discrete distributions, so if you draw n points you are likely to get repeated values
- A DP induces a partitioning of the n points e.g. $(134)(25) \Leftrightarrow \phi_1 = \phi_3 = \phi_4 \neq \phi_2 = \phi_5$
- CRP is the corresponding distribution over partitions

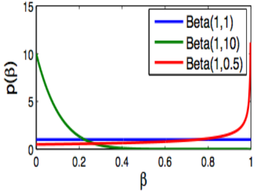
Stick Breaking construction for $G \sim DP(\cdot, \alpha, G_0)$

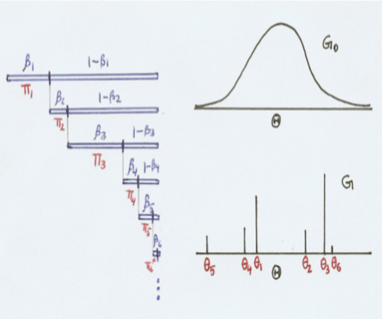
Stick-Breaking Formula

$$\pi_k = \beta_k \prod_{l=1} (1 - \beta_l), \beta_k \sim Beta(1, \alpha),$$

$$\theta_k^* \sim G_0, G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

Stick Breaking





Sketch of the proof (Sethuraman, 1994)

- Recall the posterior process

$$\begin{bmatrix} G \sim DP(\cdot \mid \alpha, G_0) \\ \theta \mid G \sim G \end{bmatrix} \Leftrightarrow \begin{bmatrix} \theta \sim G_0 \\ G \mid \theta \sim DP\left(\cdot, \alpha+1, \frac{\alpha G_0 + \delta_\theta}{\alpha+1}\right) \end{bmatrix}$$

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

$$\begin{aligned} (G(\theta), G(\Theta\backslash\theta)) &\sim \text{Diri}\left\{(\alpha+1)\frac{\alpha G_0 + \delta_\theta}{\alpha+1}(\theta), (\alpha+1)\frac{\alpha G_0 + \delta_\theta}{\alpha+1}(\Theta\backslash\theta)\right\} \\ &= \text{Beta}(1, \alpha) \end{aligned}$$

- G has a point mass located at θ:

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

  and G' is the renormalized probability measure with the point mass removed.
- What is G'?

Sketch of the proof : What is G'

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

$$\begin{aligned}(G(\theta), G(A_1), \ldots, G(A_K)) &= (\beta, (1-\beta)G'(A_1), \ldots, (1-\beta)G'(A_K)) \\ &\sim \text{Diri}(1, \alpha G_0(A_1), \ldots, \alpha G_0(A_K))\end{aligned}$$

- Renomalizing

$$\begin{aligned}(G'(A_1), \ldots, G'(A_K)) \mid \theta &= \text{Diri}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K)) \\ G' &\sim DP(\cdot, \alpha, G_0)\end{aligned}$$

Sketch of the proof

$$
\begin{aligned}
G &\sim DP(\cdot, \alpha, G_0) \\
G &= \beta_1 \delta_{\theta_1^*} + (1 - \beta_1) G_1 \\
G &= \beta_1 \delta_{\theta_1^*} + (1 - \beta_1)(\beta_2 \delta_{\theta_2^*} + (1 - \beta_2) G_2) \\
&\vdots \\
G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}
\end{aligned}
$$

where $\pi_k = \beta \prod_{l=1}(1 - \beta_l), \beta_k \sim Beta(1, \alpha), \theta_k^* \sim G_0$.