SC7/SM6 Bayes Methods HT20

Lecturer: Geoff Nicholls

Lecture 14: Bayesian non-parametrics : the Dirichlet Process

Notes and Problem sheets are available at

http://www.stats.ox.ac.uk/~nicholls/BayesMethods/
The finite Dirichlet process

We want to distribute probability randomly in $\Omega$. Let $H(d\theta)$ be some simple parameteric distribution on $\Omega$. If $\theta \in \mathbb{R}^p$ then

$$H(d\theta) = h(\theta)d\theta$$

for some simple density $h(\theta)$ on $\mathbb{R}^n$.

We can make a very simple random discrete measure as follows.

(!) Fix the number of points $K \geq 1$ in the support of $G$, and let $\alpha > 0$ be given.

1. For $k = 1, \ldots, K$, sample $\theta^*_k \sim H$.
2. Sample $w_1, \ldots, w_K \sim \text{Dirichlet}(\alpha/K)$.
3. Set $G(d\theta) = \sum_{k=1}^K w_k \delta_{\theta^*_k}(d\theta)$.

Random probability masses $w_k$ at a random locations $\theta^*_k$ in $\Omega$. 
Let $\mathcal{S}_\Omega$ be a $\sigma$-algebra of sets $A \subset \Omega$ chosen so that the following integrals exist (for example if $\Omega = \mathbb{R}^p$ at least all open sets).

For $A \in \mathcal{S}_\Omega$ let $\Pr(\theta \in A|G) = G(A)$. We have

$$G(A) = \int_\Omega \mathbb{1}_{\theta \in A} G(d\theta)$$

$$= \int_{\Omega \cap A} \sum_{k=1}^{K} w_k \delta_{\theta^*_k}(d\theta)$$

$$= \sum_{k=1}^{K} w_k \mathbb{1}_{\theta^*_k \in A}. \quad (\star)$$

In distribution notation $G = \sum_{k=1}^{K} w_k \delta_{\theta^*_k}$.

We have a simple procedure that assigns a random probability mass to measurable sets $A \subseteq \Omega$. 
By putting a prior on $w$ and $\theta^*$ we determine a prior on a probability distribution.

The choice $w \sim \text{Dirichlet}(\alpha/K, ..., \alpha/K)$, $\theta^*_k \sim H$, $k = 1, ..., K$ determines a finite Dirichlet process.

$G$ is a random distribution, but it is “centred” on the “base distribution” $H$. From Eqn (⋆), since $w$ and $\theta^*$ are independent,

$$E(G(A)) = \sum_{k=1}^{K} E(w_k)E(\mathbb{1}_{\theta^*_k \in A})$$

$$= \sum_{k} \frac{\alpha_k}{K} \sum_{j} \frac{\alpha_j}{K} H(A)$$

$$= H(A)$$

Exercise: use the Dirichlet-variance formula to calculate $\text{var}(G(A))$. 

The Dirichlet process

Definition: $G \sim DP(\alpha, H)$ if $\forall$ partitions $A_1, A_2, ..., A_r$ of $\Omega$

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

Does such a distribution exist? If it did and $G \sim DP$ then $\theta_1, ..., \theta_n \sim G$ (iid) gives an infinite exchangeable sequence.

Conversely, we give an algorithm below realising an IES $\theta_1, ..., \theta_n$ directly given $\alpha, H$. By de Finetti a random variable $G$ and $D(dG; \alpha, H)$ must exist. From the properties of the sequence we can show the random variable $G$ satisfies the definition.

The DP is also obtained as the limit $K \to \infty$ of the finite Dirichlet process (PS4). The definition above “starts afresh”.
A number of properties follow immediately from the properties of a Dirichlet Distribution. We must have

\[ G(A), G(A^c) \sim \text{Beta}(\alpha H(A), \alpha(1 - H(A))) \]

by agglomeration so \( E(G(A)) = H(A) \) as before.

If \( G \sim DP(\alpha, H) \) and \( \theta \sim G \) then marginally \( \theta \sim H \), because

\[ \Pr(\theta \in A) = E(E(\mathbb{1}_{\theta \in A}|G)) = E(G(A)) = H(A) \]

for all \( A \in \mathcal{S}_\Omega \) so \( \theta \sim H \).

If \( G \sim DP(\alpha, H) \) and \( \theta \sim G \) then for measurable \( B \subseteq A \),

\[ \Pr(\theta \in B|\theta \in A) = \frac{\Pr(\theta \in B)}{\Pr(\theta \in A)} = \frac{H(B)}{H(A)} \]

If \( H(d\theta) = h(\theta)d\theta \), so distribution \( H \) has density \( h \), then marginally \( \theta|\theta \in A \) has density \( p(\theta|\theta \in A) \) on \( A \cap \Omega \) with

\[ p(\theta|\theta \in A) = h(\theta|\theta \in A). \]
DP generative model and predictive distributions

Our generative model for \( \theta = (\theta_1, ..., \theta_n) \) is

1. \( G \sim DP(\alpha, H) \)
2. \( \theta_i \sim G, i = 1, 2, ..., n \)

Here \( G \) is a “parameter” like \( \theta \), but we want to work with the marginal \( \pi(d\theta) \) not the joint for \( G \) and \( \theta \). Use the identity

\[
\pi(d\theta) = \pi(d\theta_n|\theta_1:n-1)\pi(d\theta_{n-1}|\theta_1:n-2)\cdots\pi(d\theta_1).
\]

We know \( \pi(d\theta_1) = H(d\theta_1) \) so for \( i = 2, ..., n \) the predictive distributions \( \pi(d\theta_i|\theta_1:i-1) \) are needed.

**Claim:** The conditional distribution of \( G|\theta_1 \) is

\[
G|\theta_1 \sim DP\left(\alpha + 1, \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1}\right).
\]

and so \( \pi(d\theta_2|\theta_1) = \frac{\alpha H(d\theta_2) + \delta_{\theta_1}(d\theta_2)}{\alpha + 1} \).
Proof: let $A_1, ..., A_r$ be a partition of $\Omega$. Suppose $\theta_1 \in A_j$ and use the abbreviation $G_j \equiv G(A_j)$. Then we have

$$
\pi(G_1, ..., G_r|\theta_1) \propto p(\theta_1|G_1, ..., G_r) \pi(G_1, ..., G_r).
$$

Since $\theta_1 \in A_j$,

$$
p(\theta_1|G_1, ..., G_r) = p(\theta_1, \theta_1 \in A_j|G_1, ..., G_r)
$$

$$
= p(\theta_1|\theta_1 \in A_j, G_1, ..., G_r)p(\theta_1 \in A_j|G_1, ..., G_r)
$$

$$
= h(\theta_1|\theta_1 \in A_j)G_j,
$$

when $H(d\theta_1) = h(\theta_1)d\theta_1$ in our DP. Here $\theta_1$ is independent of $G_1, ..., G_r$ given $\theta_1 \in A_j$ as $G_j$ etc are just probability masses assigned to sets $A_1, ..., A_r$ not distributions within sets.

Recall now that

$$
G_1, ..., G_r \sim \text{Dirichlet}(\alpha H_1, ... \alpha H_r),
$$

where $H_j \equiv H(A_j)$. 
The conditional is
\[
\pi(G_1, ..., G_r|\theta_1) \propto h(\theta_1|\theta_1 \in A_j)G_j \times G_1^{\alpha H_1 - 1} \cdots G_r^{\alpha H_r - 1}
\]
\[
\propto G_1^{\alpha H_1 - 1 + \mathbb{I}_{\theta_1 \in A_1}} \times \cdots \times G_r^{\alpha H_r - 1 + \mathbb{I}_{\theta_1 \in A_r}},
\]
dropping \(h(\theta_1|\theta_1 \in A_j)\) independent of \(G_1, ..., G_r\), and giving
\[
G_1, ..., G_r|\theta_1 \sim \text{Dirichlet}(\alpha H_1 + \mathbb{I}_{\theta_1 \in A_1}, ..., \alpha H_r \mathbb{I}_{\theta_1 \in A_r}).
\]
This is a \(DP(H', \alpha')\) if we choose \(\alpha'\) and \(H'\) so that
\[
\alpha' H'(A_j) = \alpha H(A_j) + \mathbb{I}_{\theta_1 \in A_j}.
\]
for \(j = 1, ..., r\). If we take \(\alpha' = \alpha + 1\) and base distribution
\[
H' = \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1}, \quad \text{(ie } H' \text{ normalised)}
\]
then this works, so indeed \(G|\theta_1 \sim DP\left(\alpha + 1, \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1}\right)\).
The predictive distribution of $\theta_2|\theta_1, \alpha, H$ is

$$(\theta_2 \mid \theta_1, \alpha, H) \sim \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1},$$

as $G|\theta_1 \sim DP(\alpha', H')$ and $\theta_2 \sim G$ so marginally $\theta_2 \sim H'$. [EOP]

Interpretation: The updated base distribution for $G|\theta_1$ is

$$H'(d\theta) = \frac{\alpha}{\alpha + 1} h(\theta)d\theta + \frac{1}{\alpha + 1} \delta_{\theta_1}(d\theta).$$

This is a mixture.

In order to simulate $\theta_2 \sim G|\theta_1$ we would simulate $\theta \sim h$ wp $\alpha/(\alpha + 1)$ and otherwise (ie wp $1/(\alpha + 1)$) we set $\theta = \theta_1$. Notice the atom at $\theta = \theta_1$. 
Exercise (PS4): if $\theta_1, \ldots, \theta_n \sim G$ with $G \sim DP(\alpha, H)$, show

$$G|\theta_1:n \sim DP\left(\alpha + n, \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}\right).$$

Hint: induction replacing $G$ by $G|\theta_1:(n-1)$.

From the exercise, the predictive distribution of $\theta_{n+1}$ is

$$(\theta_{n+1} \mid \theta_1:n, \alpha, H) \sim \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}.$$

Sequential simulation: Generative model for $\theta = (\theta_1, \ldots, \theta_n)$

1. $\theta_1 \sim H$
2. for $j = 1, \ldots, n - 1$
   (a) With probability $\alpha/(\alpha + j)$ simulate $\theta_{j+1} \sim H$.
   (b) Otherwise simulate $\theta_{j+1} \sim U\{\theta_1, \ldots, \theta_j\}$.

No $G$ in sight!
The Dirichlet process is “atomic”, so \((\theta_1, ..., \theta_n)\) may contain repeated \(\theta\)-values (if we choose one of the “old” \(\theta\)’s). Write

\[
\sum_{i=1}^{n} \delta_{\theta_i} = \sum_{k=1}^{K} n_k \delta_{\theta_k^*}.
\]

\(\theta^* = (\theta_1^*, ..., \theta_K^*)\) are the unique \(\theta\)-values in \((\theta_1, ..., \theta_n)\).

\(K\) is random, the number of distinct values in \((\theta_1, ..., \theta_n)\).

\(n_k\) is the number of times \(\theta_k^*\) appears in \((\theta_1, ..., \theta_n)\).

For \(k = 1, ..., K\) let \(S_k = \{i; \theta_i = \theta_k^*\}\) so we have a partition \(S = \{S_1, ..., S_K\}\) with \(n_k = |S_k|\).

The mapping \(\theta \rightarrow (\theta^*, S)\) is invertible: for \(i = 1, ..., n\) let \(k_i = \{k : i \in S_k\}\); since \(S\) is a partition \(k_i\) is unique; set \(\theta_i = \theta_{k_i}^*\).
Sequential simulation: generative model for $\theta^*, S$

1. $\theta_1^* \sim H$; set $K = 1$, $S_1 = 1$ and $S = \{S_1\}$.
2. for $j = 1, \ldots, n - 1$
   
   (a) With probability $\alpha/(\alpha + j)$: simulate $\theta_{K+1}^* \sim H$; set $S_{K+1} = \{j + 1\}$ and $S \leftarrow S \cup S_K$; set $K \leftarrow K + 1$.
   
   (b) Otherwise: for $k = 1, \ldots, K$ set $n_k = |S_k|$; simulate $k \sim (n_1, \ldots, n_K)/j$; set $S_k \leftarrow S_k \cup \{j + 1\}$.

This simulates the marginal distribution $(\theta^*, S)|\alpha, H$ (so, $\theta$).

Let $P_\alpha(S)$ give the PMF for the random partition $S$.

The marginal $\pi(d\theta)$ we sought is given by

$$\pi(d\theta^*, S) = P_\alpha(S) \prod_{k=1}^K H(d\theta_k^*),$$

since $\pi(d\theta^*|S) = \prod_k H(d\theta_k^*)$. Here $S$ is a model index $S$ fixing the dimension $\dim(\theta^*) = |S| = K$. 
The Chinese Restaurant Process, AKA CRP

The sequential simulation of parameters according to a $DP(\alpha, H)$ is analogous to restaurant seating!

1) There is $j = 1$ one customer in the restaurant. They are seated at table $k = 1$.

2) Suppose that after the customer $j = 2, 3, ..., n$ arrives, there are $n_j^k$ people seated at table $k$, and $K_j$ tables in all are occupied.

3) The $j + 1$'st arrival chooses a new table wp $\alpha/(\alpha + j)$ and table $k$ with probability $n_j^k/\alpha + j$.

This divides up $n$ customers over $K$ tables. The set $S_k$ lists customers at table $k$.

Put an independent parameter $\theta_k^* \sim H$ on table $k = 1, 2, ... K$. Together, $(S, \theta^*) \sim G$ with $G \sim DP(\alpha, H)$. 