SC7/SM6 Bayes Methods HT19

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/

and via the MSc weblearn pages.
The big picture If data \( y_1, y_2, \ldots \) is an infinite exchangeable sequence then there exists a distribution \( G \) such that the prior predictive distribution \( p(y_1, \ldots, y_n) \) is given by

\[
p(y_1, \ldots, y_n) = \int_{\Omega} p(y_1, \ldots, y_n | \theta) dG(\theta).
\]

These distributions and the posterior distribution

\[
\pi(d\theta | x_1, \ldots, x_m) \propto p(x_1, \ldots, x_n | \theta) \pi(d\theta)
\]

all exist, with prior \( \pi(d\theta) = dG(\theta) \). The difficulty is that we don’t know \( G \) (like we didn’t know \( \theta \)).

The idea is to move all this up one level and infer \( G \), treating it like a parameter. Now \( G \in \mathcal{G} \) is some unknown distribution that puts probability mass on sets \( A \subset \Omega \).

Bayesian inference for \( G, \theta \) given \( y \) is inference for our \( \theta \)-prior.
Suppose the prior distribution for $G$ (a DP in our case study) is $G \sim D$. All the usual objects like $E(G(A))$, $\text{var}(G(A))$ and a posterior $G(A)|x_1, \ldots, x_n$ can be defined for sets $A$.

Suppose $\theta_1, \theta_2, \ldots$ is an IES with parameter $G$. It is convenient to define and calculate the marginal distribution,

$$
\pi(\theta_1, \ldots, \theta_n) = \int_D \prod_{i=1}^n \pi(\theta_i|G) D(dG).
$$

If the observation model is $p(y|\theta)$ then the posterior is

$$
\pi(\theta|y) \propto p(y|\theta) \pi(\theta),
$$

where now the $G$-dependence has been integrated out of the full posterior “$\pi(dG, \theta|y)$”. Our $G$ will leave its mark on the analysis, as the parameter $\theta = (\theta_1, \ldots, \theta_n)$ has dimension $K \leq n$.

*This hides alot of detail which we will unpack in the next two lectures!*
Bayesian Non-parametrics

With enough data you reject any parametric model. Small model violations (skew, correlation, heavy tails, number of mixture components) built-in by parametric model assumptions may become glaring as the data set grows large.

Non parametric models allow fitting with an unbounded number of parameters. NP models adapt themselves to data as more data is added - they are able to model data with much greater complexity.

Often, the “scientific model” is parametric, but the noise has some unknown complex structure - so express the science with a parametric model and model the noise non-parametrically. NP models commonly have parametric elements.
Prior elicitation and careful modeling of the observation process still matter. The NP model you specify will have built in structure, it won't adapt to every form of model mispecification.

The model contains information which may bias the inference. This is good if the model biases towards the truth!

We will look at one example of a NP model and BNP fitting: the Dirichelet process mixture for density estimation and clustering.

The model changes but the methodological framework we have developed (model averaging, marginal likelihoods...) is the same.
The Dirichlet Distribution - reference slide

Let \( w = (w_1, \ldots, w_K) \) with \( \sum_k w_k = 1 \).
Suppose \( w \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \). The density of \( w \) is

\[
\pi(w_1, w_2, \ldots, w_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} w_1^{\alpha_1-1} w_2^{\alpha_2-1} \ldots w_K^{\alpha_K-1}.
\]

Agglomeration: if \( w_1, \ldots, w_K \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \) then

\[
w_1 + w_2, w_3, \ldots, w_K \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K).
\]

Conjugate prior to multinomial: if \( (n_1, \ldots, n_K) \sim \text{Multinom}(n, w) \) with \( n = \sum_k n_k \) and \( w \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \) then

\[
\pi(w|n_1, \ldots, n_K) \propto w_1^{\alpha_1+n_1-1} w_2^{\alpha_2+n_2-1} \ldots w_K^{\alpha_K+n_K-1}
\]so \( w|n_1, \ldots, n_K \sim \text{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_K + n_K) \).
The finite Dirichlet process

We want to distribute probability randomly in $\Omega$. Let $H(d\theta)$ be some simple parameteric distribution on $\Omega$, for example

$$H(d\theta) = \pi(\theta)d\theta.$$  

Fix the number of points $K \geq 1$ in the support of $G$, and let $\alpha = (\alpha_1, \ldots, \alpha_K)$ be given.

We can make a random discrete measure as follows.

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

We put probability masses $w_k$ at a random locations $\theta^*_k$ in $\Omega$. 

Now given $G$ (ie, given $w, \theta^*$) we have

$$\Pr(\theta = \theta^*_k|G) = w_k,$$

or in standard notation, if $\theta \sim G$ then $\theta = \theta^*_k \mathrm{wp} w_k$. If $A \subseteq \Omega$ then $\Pr(\theta \in A|G)$ is the chance to draw one of the $\theta^*$'s in $A$,

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

We can write this random discrete distribution using our delta function notation

$$dG(\theta) = \sum_{k=1}^{K} w_k \delta_{\theta^*_k}(\theta)d\theta.$$

Now $\Pr(\theta \in A|G) = G(A)$ by definition.
For $A \subset \Omega$,

$$G(A) = \int_{\Omega} \mathbb{I}_{\theta \in A} dG(\theta)$$

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

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

We sometimes use the notation

$$G = \sum_{k=1}^{K} w_k \delta_{\theta_k^*},$$

for writing distributions of this sort since $\delta_{\theta_k^*}(A) = \mathbb{I}_{\theta \in A}$ but $\delta_{\theta_k^*}$ refers to the distribution.
We have a simple procedure that assigns a random probability mass to sets $A \subseteq \Omega$.

By putting a prior on $w$ and $\theta^*$ we determine a prior on a probability distribution. The choice $w \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$, $\theta^*_k \sim H, \ k = 1, \ldots, K$ determines a finite Dirichlet process.

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

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

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

$$= H(A)$$

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

The Dirichlet process

The Dirichlet process is obtained from the finite Dirichlet process in the limit $K \to \infty$ (See PS4). The definition we give below “starts anew” but the two approaches are equivalent.

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

\[
G(A_1), \ldots, G(A_r) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_r))
\]

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

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

so $Pr(\theta \in A) = H(A)$ integrating $dG$.

**Stick breaking representation - outside course**

The equivalence to the distribution obtained in the limit may be shown using the “stick breaking” construction of the DP.

If $\beta_k \sim \text{Beta}(1, \alpha)$ and for each $k = 1, 2, ..., K, ...$ we set

$$w_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_j) \quad \theta^*_k \sim H \quad \text{and} \quad G(A) = \sum_{k=1}^{\infty} w_k \mathbb{1}_{\theta^*_k \in A}$$

then we recover the Dirichlet process as the number of terms in the sum defining $G$ goes from $K$ to $\infty$. 
Updating the DP

Recall that $G$ is the “parameter”, the thing we don’t know and are trying to infer. Suppose we observe a sequence of $\theta$-values drawn from $G$. The process determining the $\theta$’s is

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

Here $DP(\alpha, H)$ is a prior for $G$ and $\theta_1, ..., \theta_n$ are data.

Claim: The posterior distribution of $G|\theta_1$ is

$$G|\theta_1 \sim DP \left(\alpha + 1, \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1}\right).$$
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 $A_1, ..., A_r$ is a partition, $\theta_1 \in A_j$ for some unique $j$, so

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

assuming $H(d\theta) = h(\theta)d\theta$ in our DP. The conditioning is legitimate as $\theta_1$ is independent of $G_1, ..., G_r$ given $\theta_1 \in A_j$ (a tricky point - $\theta_1$ would depend on $G$ but we are conditioning on less information, just $G(A_1), ..., G(A_r)$). Recall now that

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

where $H_j \equiv H(A_j)$. 
The posterior is
\[
\pi(G_1, \ldots, G_r | \theta_1) \propto h(\theta_1 | \theta_1 \in A_j) G_j \times G_1^{\alpha H_1 - 1} \ldots G_r^{\alpha H_r - 1}
\]
\[
\propto G_1^{\alpha H_1 - 1 + \mathbb{1}_{\theta_1 \in A_1}} \times \ldots \times G_r^{\alpha H_r - 1 + \mathbb{1}_{\theta_1 \in A_r}},
\]
dropping \(h(\theta_1 | \theta_1 \in A_j)\) independent of \(G_1, \ldots, G_r\), and giving
\[
G_1, \ldots, G_r | \theta_1 \sim \text{Dirichlet}(\alpha H_1 + \mathbb{1}_{\theta_1 \in A_1}, \ldots, \alpha H_r \mathbb{1}_{\theta_1 \in A_r}).
\]
This is a \(DP(H', \alpha')\) if we choose the right \(\alpha'\) and \(H'\). If we take as our new \textit{normalised} base distribution
\[
H' = \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1}
\]
and our new \(\alpha' = \alpha + 1\) then for \(j = 1, \ldots, r\),
\[
\alpha' H'(A_j) = \alpha H(A_j) + \mathbb{1}_{\theta_1 \in A_j}
\]
and so indeed \(G | \theta_1 \sim DP\left(\alpha + 1, \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1}\right)\). \[\text{EOP}\]
Interpretation The new base “density” for $G|\theta_1$,

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

is defined by its action in integrals, as $H'(A) = \int_A h'(\theta) d\theta$ gives

$$H'(A) = \frac{\alpha}{\alpha + 1} H(A) + \frac{1}{\alpha + 1} \mathbb{I}_{\theta_1 \in A}$$

as above. This is a mixture: to simulate $\theta \sim G|\theta_1$ 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): show that if $\theta_1, ..., \theta_n \sim G$ with $G \sim DP$ then

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

[Hint: induction replacing $G$ by $G|\theta_1:(n-1)$]
The predictive distribution of $\theta_{n+1}|\theta_n, \alpha, H$ is

$$(\theta_{n+1} \mid \theta_{1:n}, \alpha, H) \sim \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}.$$ because if $G \sim DP(\alpha, H)$ and $\theta \sim G$ then marginally $\theta \sim H$.

Again, this is a mixture of the continuous distribution $H$ and the atoms at $\theta_i, i = 1,\ldots,n$. To simulate this distribution we simulate $\theta \sim H$ wp $\alpha/(\alpha + n)$ and otherwise (ie wp $1/(\alpha + n)$) we sample $\theta \sim U\{\theta_1,\ldots,\theta_n\}$.

[see R code for this lecture]