23: Dirichlet Process Models

Taylor

University of Virginia

Introduction

We'll take a look at **Dirichlet Processes** now, and see how they're useful for mixture modeling.

Bayesian histograms

A probability model for the density analagous to the histogram

$$f(y \mid \pi_1, \dots, \pi_k) = \sum_{h=1}^{k} 1_{\xi_{h-1} < y \leq \xi_h} \frac{\pi_h}{(\xi_h - \xi_{h-1})}$$

where $\xi_0 < \xi_1 < \cdots < \xi_k$ are your **knot points**, and $(\pi_1, \dots, \pi_k)$ is an unknown probability vector.

Bayesian histograms

A probability model for the density analagous to the histogram

$$f(y \mid \pi_1, \ldots, \pi_k) = \sum_{h=1}^{k} 1_{\xi_{h-1} < y \le \xi_h} \frac{\pi_h}{(\xi_h - \xi_{h-1})}$$

where $\xi_0 < \xi_1 < \cdots < \xi_k$ are your **knot points**, and $(\pi_1, \ldots, \pi_k)$ is an unknown probability vector.

Note

$$\int f(y) \mathrm{d}y = \sum_{h=1}^{k} \text{base}_h \times \text{height}_h = \sum_{h=1}^{k} (\xi_h - \xi_{h-1}) \frac{\pi_h}{(\xi_h - \xi_{h-1})} = 1.$$

Bayesian histograms

$$f(y \mid \pi) = \sum_{h=1}^{k} 1_{\xi_{h-1} < y \leq \xi_h} \frac{\pi_h}{(\xi_h - \xi_{h-1})}$$

We can put a Dirichlet$(\alpha_1, \ldots, \alpha_k)$ prior on the parameters $\pi = (\pi_1, \ldots, \pi_k)$:

$$p(\pi) = \frac{\Gamma\left(\sum_{h=1}^{k} a_h\right)}{\prod_{h=1}^{k} \Gamma(\alpha_h)} \prod_{h=1}^{k} \pi_h^{a_h - 1}$$

where $a = (a_1, \ldots, a_k)$ are the chosen parameters of the prior.

Bayesian histograms

Note that $f(y \mid \pi)$ was for one data point y. Let $\sigma(i) = \{k : \xi_{k-1} < y_i \leq \xi_k\}$. Notice that this function is many-to-one. Then

$$
\begin{aligned}
p(y \mid \pi) &= \prod_{i=1}^{n} f(y_i \mid \pi) \\
&= \prod_{i=1}^{n} \left[\sum_{h=1}^{k} 1_{\xi_{h-1} < y_i \leq \xi_h} \frac{\pi_h}{(\xi_h - \xi_{h-1})} \right] \\
&= \prod_{i=1}^{n} \left[\frac{\pi_{\sigma(i)}}{(\xi_{\sigma(i)} - \xi_{\sigma(i)-1})} \right] \\
&= \prod_{h=1}^{k} \left[\frac{\pi_h}{(\xi_h - \xi_{h-1})} \right]^{n_h}
\end{aligned}
$$

where $n_h = \sum_{i=1}^{n} 1_{\xi_{h-1} < y_i \leq \xi_h}$.

Bayesian histograms

Bayes' rule:

$$\begin{aligned} p(\pi \mid y) &\propto p(y \mid \pi) p(\pi) \\ &= \left[\prod_{h=1}^{k} \frac{\pi_h}{(\xi_h - \xi_{h-1})} \right]^{n_h} \prod_{h=1}^{k} \pi_h^{a_h - 1} \\ &\propto \prod_{h=1}^{k} \pi_h^{a_h + n_h - 1} \end{aligned}$$

So $p(\pi \mid y) = \text{Dirichlet}(a_1 + n_1, \ldots, a_k + n_k)$.

Bayesian histograms

Bayes' rule:

$$
\begin{aligned}
p(\pi \mid y) &\propto p(y \mid \pi) p(\pi) \\
&= \left[\prod_{h=1}^{k} \frac{\pi_h}{(\xi_h - \xi_{h-1})} \right]^{n_h} \prod_{h=1}^{k} \pi_h^{a_h - 1} \\
&\propto \prod_{h=1}^{k} \pi_h^{a_h + n_h - 1}
\end{aligned}
$$

So $p(\pi \mid y) = \text{Dirichlet}(a_1 + n_1, \ldots, a_k + n_k)$.

But bin specification is annoying!

A quick note

If $\pi \sim \text{Dirichlet}(a_1, \ldots, a_k)$ then

$$E[\pi] = \left(\frac{a_1}{\alpha}, \cdots, \frac{a_k}{\alpha}\right)$$

where

$$\alpha = a_1 + \cdots + a_k.$$

So we can write

$$\pi \sim \text{Dirichlet}(\alpha E[\pi])$$

as well. α has the interpretation of a sample size.

Definitions

A **random probability measure** assigns probabilities to sets, and they still satisfy the three probability axioms, but these probabilities are random.

Definitions

A **random probability measure** assigns probabilities to sets, and they still satisfy the three probability axioms, but these probabilities are random.

The random probability measure P is a **Dirichlet process** if for **any** measurable partition $B_1, \ldots, B_k$ of a sample space Ω, the vector

$$P(B_1), \ldots, P(B_k) \sim \text{Dirichlet}(\alpha P_0(B_1), \ldots, \alpha P_0(B_k)).$$

Definitions

1. P_0 is a P_0 is a **baseline probability measure** we pick (e.g. normal)
2. shorthand: $P \sim \text{DP}(\alpha P_0)$
3. $P(B) \sim \text{Beta}(\alpha P_0(B), \alpha(1 - P_0(B)))$ for any measurable B
4. $E[P(B)] = P_0(B)$
5. $V[P(B)] = P_0(B)[1 - P_0(B)]/(1 + \alpha)$

Returning to the Bayesian histogram

Assume $y_i \overset{iid}{\sim} P$ and $P \sim \mathrm{DP}(\alpha P_0)$. Then, for any partition $B_1, \ldots, B_k$,

$$\begin{aligned} &P(B_1), \ldots, P(B_k) \mid y_1, \ldots, y_n \\ &\sim \text{Dirichlet}\left(\alpha P_0(B_1) + \sum_{i=1}^{n} 1_{y_i \in B_1}, \ldots, \alpha P_0(B_k) + \sum_{i=1}^{n} 1_{y_i \in B_k}\right) \end{aligned}$$

using the same reasoning as in slide 5.

Returning to the Bayesian histogram

Assume $y_i \overset{iid}{\sim} P$ and $P \sim \mathrm{DP}(\alpha P_0)$. Then, for any partition $B_1, \ldots, B_k$,

$$\begin{aligned} &P(B_1), \ldots, P(B_k) \mid y_1, \ldots, y_n \\ &\sim \text{Dirichlet}\left(\alpha P_0(B_1) + \sum_{i=1}^{n} 1_{y_i \in B_1}, \ldots, \alpha P_0(B_k) + \sum_{i=1}^{n} 1_{y_i \in B_k}\right) \end{aligned}$$

using the same reasoning as in slide 5.

What happens when we take the noninformative prior $\alpha \downarrow 0$?

Returning to the Bayesian histogram

In particular, for any measurable B, $P(B) \mid y_1, \ldots, y_n$ follows a

$$\text{Beta}\left(\alpha P_0(B) + \sum_{i=1}^{n} 1_{y_i \in B}, \alpha + n - \left[\alpha P_0(B) + \sum_{i=1}^{n} 1_{y_i \in B}\right]\right)$$

and

$$\begin{aligned} E[P(B) \mid y_1, \ldots, y_n] &= \frac{\alpha P_0(B) + \sum_{i=1}^{n} 1_{y_i \in B}}{\alpha + n} \\ &= \frac{\alpha}{\alpha + n} P_0(B) + \frac{n}{\alpha + n} \sum_{i=1}^{n} \frac{1_{y_i \in B}}{n} \end{aligned}$$

Returning to the Bayesian histogram

In particular, for any measurable B, $P(B) \mid y_1, \ldots, y_n$ follows a

$$\text{Beta}\left(\alpha P_0(B) + \sum_{i=1}^{n} 1_{y_i \in B}, \alpha + n - \left[\alpha P_0(B) + \sum_{i=1}^{n} 1_{y_i \in B}\right]\right)$$

and

$$\begin{aligned} E[P(B) \mid y_1, \ldots, y_n] &= \frac{\alpha P_0(B) + \sum_{i=1}^{n} 1_{y_i \in B}}{\alpha + n} \\ &= \frac{\alpha}{\alpha + n} P_0(B) + \frac{n}{\alpha + n} \sum_{i=1}^{n} \frac{1_{y_i \in B}}{n} \end{aligned}$$

not "smooth" even if P_0 was!

The stick-breaking representation

We can write a DP as a countably infinite mixture of point masses:

$$P(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}(\cdot).$$

Round 1

1. sample location $\theta_1 \sim P_0$
2. sample associated probability $V_1 \sim \text{Uniform}(0, 1)$
3. we have $1 - V_1$ probability left over...

The stick-breaking representation

We can write a DP as a countably infinite mixture of point masses:

$$P(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}(\cdot).$$

Round 1

1. sample location $\theta_1 \sim P_0$
2. sample associated probability $V_1 \sim \text{Uniform}(0, 1)$
3. we have $1 - V_1$ probability left over...

Round 2

1. sample location $\theta_2 \sim P_0$
2. sample $V_2 \sim \text{Uniform}(0, 1)$
3. probability at second location is now $(1 - V_1)V_2$
4. we have $1 - (V_1 + V_2(1 - V_1)) = 1 - V_1 - V_2(1 - V_1) = (1 - V_1)(1 - V_2)$ probability left over...

The stick-breaking representation

We can write a DP as a countably infinite mixture of point masses:

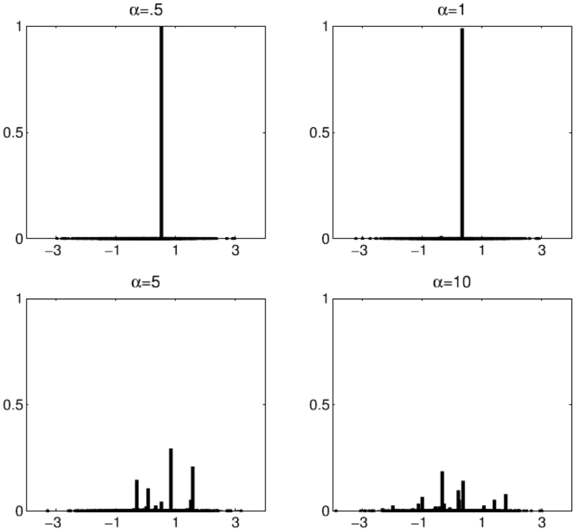
$$P(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}(\cdot).$$

with

$$\pi_h = V_h \prod_{l<h} (1 - V_l)$$

and $V_h \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$, $\theta_h \overset{\text{iid}}{\sim} P_0$

The stick-breaking representation



The stick-breaking representation

$$
\begin{aligned}
E[P(B)] &= \sum_{h=1}^{\infty} E[\pi_h 1_{\theta_h \in B}] \\
&= \sum_{h=1}^{\infty} E[\pi_h] E[1_{\theta_h \in B}] \\
&= \sum_{h=1}^{\infty} E[V_h \prod_{l<h} (1 - V_l)] P_0(B) \\
&= P_0(B) \sum_{h=1}^{\infty} E[V_h] \prod_{l<h} E[(1 - V_l)] \\
&= P_0(B) \sum_{h=1}^{\infty} \frac{1}{1+\alpha} \left(\frac{\alpha}{1+\alpha} \right)^{h-1} \\
&= P_0(B)
\end{aligned}
$$

Dirichlet process mixtures

Sampling $P \sim \mathrm{DP}(\alpha P_0)$ and then using that random histogram as the distribution for a sample of continuous y_i random variables is problematic because we would like P to be smooth.

Dirichlet process mixtures

Sampling $P \sim \mathrm{DP}(\alpha P_0)$ and then using that random histogram as the distribution for a sample of continuous y_i random variables is problematic because we would like P to be smooth.

We can use DPs for **general kernel mixture models** though:

$$f(y \mid P) = \int \mathcal{K}(y \mid \theta) P(\mathrm{d}\theta)$$
$$= \sum_{h=1}^{\infty} \pi_h \mathcal{K}(y \mid \theta_h)$$

This is a mixture model, but there are an infinite number of mixands!