THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

Some notes on the Dirichlet

Sinead Williamson
The University of Texas Department of Statistics and Data Science

Dirichlet distribution

Dirichlet prior over probability vector:

$$p(\theta) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_i)}{\prod_{k=1}^{K} \Gamma(\alpha_i)} \prod_{k=1}^{K} \theta_i^{\alpha_i - 1}$$

Discrete prior on observation category:

$$p(Z_i = k|\theta) = \theta_i$$

Posterior over θ:

$$p(\theta|z_i) \propto p(z_i|\theta)p(\theta) \propto \theta_{z_i} \prod_{k=1}^{K} \theta_i^{\alpha_i - 1} = \prod_{k=1}^{K} \theta_i^{\alpha_i + \mathbb{I}(z_i = k) - 1}$$

so $(\theta|z_i) \sim \text{Dirichlet}\left((\alpha_k + \mathbb{I}(z_i = k))_{k=1}^{K}\right)$

Dirichlet distribution

If we have multiple samples $z_i \sim \theta$, then

$$p(\theta|z_1, \ldots, z_n) = \text{Dirichlet}((\alpha_k + m_k)_{k=1}^{K})$$

where $m_k = \sum_{i=1}^{n} \mathbb{I}(z_i = k)$

Posterior predictive

Let's start with $p(z_{i+1}|z_{1:i})$:

$$\begin{aligned} p(z_{i+1} = k|z_{1:i}) &= \int_{\mathcal{M}} p(z_{i+1}|\theta)p(\theta|z_{1:i})d\theta \\ &= \int_{\mathcal{M}} \theta_k \text{Dirichlet}(\theta|(\alpha_k + m_k)_{k=1}^K))d\theta \\ &= \frac{\alpha_k + m_k}{\sum_{j=1}^K \alpha_j + m_j} \end{aligned}$$

Posterior predictive: multiple observations

How about $p(z_{i+1:i+j}|z_{1:i})$?

$$
\begin{aligned}
p(z_{i+1:i+j}|z_{1:i}) =& p(z_{i+j}|z_{1:i+j-1})p(z_{i+j-1}|z_{1:i+j-2})\cdots p(z_{i+1}|z_{1:i}) \\
=& \frac{\alpha_{z_{i+j}} + m_{z_{i+j}}^{(1:i+j-1)}}{i+j-1+\sum_k \alpha_k} \times \frac{\alpha_{z_{i+j-1}} + m_{z_{i+j-1}}^{(1:i+j-2)}}{i+j-2+\sum_k \alpha_k} \\
& \times \cdots \times \frac{\alpha_{z_{i+1}} + m_{z_{i+1}}^{(1:i)}}{i+\sum_k \alpha_k} \\
=& \frac{\Gamma(i+\sum_k \alpha_k)}{\Gamma(i+j+\sum_k \alpha_k)} \prod_{k=1}^{K} \frac{\Gamma(m_k^{1:i+j}+\alpha_k)}{\Gamma(m_k^{1:i}+\alpha_k)}
\end{aligned}
$$

where $m_k^{1:j} = \sum_{i=1}^{j} \mathbb{I}(z_i = k)$. (If you look up the Dirichlet-Multinomial distribution, you will find this with some constant combinatorics terms, and $i = 0$).

Mixture models

General format:

$$\begin{aligned}
\pi &\sim \text{Dirichlet}(\alpha) \\
\theta_k &\sim p(\theta), \qquad k = 1, \dots, K \\
z_i &\sim \pi, \qquad i = 1, \dots, n \\
x_i &\sim f(\theta_{z_i})
\end{aligned}$$

Conditional distributions:

$$\begin{aligned}
p(z_i = k | z_{-i}) &\propto m_k^{-i} + \alpha_k \\
p(z_i = k | z_{-i}, x_i, \theta_k) &\propto (m_k^{-i} + \alpha_k) f(x_i; \theta_k)
\end{aligned}$$

- We can sample $z_i | z_{-i}, x_i, \theta_k$ by calculating each unnormalized probability, normalizing, and using them to parametrize a multinomial.
- If f and θ are conjugate, we may be able to integrate out θ_k and instead use $p(x_i | \{x_j, j \neq i, z_j = k\})$

Concrete example: Mixture of multinomials

$$\begin{aligned}
\pi &\sim \text{Dirichlet}(\alpha) \\
\eta_k &\sim \text{Dirichlet}(\beta), \qquad k = 1, \ldots, K \\
z_i &\sim \pi, \qquad i = 1, \ldots, n \\
x_i &\sim \text{Multinomial}(M_i, \eta_{z_i})
\end{aligned}$$

If we don't integrate out π and η:

$$\begin{aligned}
p(z_i = k | \pi, x_i, \eta_k) &\propto \pi_i \prod_{v=1}^{V} \eta_{k,v}^{\sum_{j=1}^{M_i} x_{i,j} = v} \\
\pi | z_{1:n} &\sim \text{Dirichlet}\left(\left(\alpha_k + \sum_i z_i = k \right)_{k=1}^{K} \right) \\
\eta_k | \{x_i : z_i = k\} &\sim \text{Dirichlet}\left(\left(\beta_v + \sum_{i:z_i=k} \sum_{j=1}^{M_i} x_{i,j} = v \right)_{v=1}^{V} \right)
\end{aligned}$$

Concrete example: Mixture of multinomials

$$p(z_i = k|\pi, x_i, \eta_k) \propto \pi_i \prod_{v=1}^{V} \eta_{k,v}^{\sum_{j=1}^{M_i} x_{i,j}=v}$$

$$\pi|z_{1:n} \sim \text{Dirichlet}\left(\left(\alpha_k + \sum_i z_i = k\right)_{k=1}^{K}\right)$$

$$\eta_k|\{x_i : z_i = k\} \sim \text{Dirichlet}\left(\left(\beta_v + \sum_{i:z_i=k} \sum_{j=1}^{M_i} x_{i,j} = v\right)_{v=1}^{V}\right)$$

Integrating out π:

$$p(z_i = k|z_{-i}x_i, \eta_k) \propto (m_k^{-i} + \alpha_k) \prod_{v=1}^{V} \eta_{k,v}^{\sum_{j=1}^{M_i} \mathbb{I}(x_{i,j}=v)}$$

$$\eta_k|\{x_i : z_i = k\} \sim \text{Dirichlet}\left(\left(\beta_v + \sum_{i:z_i=k} \sum_{j=1}^{M_i} \mathbb{I}(x_{i,j} = v)\right)_{v=1}^{V}\right)$$

Concrete example: Mixture of multinomials

$$p(z_i = k | z_{-i} x_i, \eta_k) \propto (m_k^{-i} + \alpha_k) \prod_{v=1}^{V} \eta_{k,v}^{\sum_{j=1}^{M_i} \mathbb{I}(x_{i,j}=v)}$$

$$\eta_k | \{x_i : z_i = k\} \sim \text{Dirichlet}\left(\left(\beta_v + \sum_{i:z_i=k} \sum_{j=1}^{M_i} \mathbb{I}(x_{i,j} = v) \right)_{v=1}^{V} \right)$$

Integrating out both π and η_k:

$$p(z_i = k | z_{-i}, x, \eta_k)$$
$$\propto (m_k^{-i} + \alpha_k) \frac{\Gamma(\sum_v \rho_{k,v}^{-i} + \sum_v \beta_v)}{\Gamma(\sum_v \rho_{k,v}^{-i} + M_i + \sum_v \beta_v)} \prod_{v=1}^{V} \frac{\Gamma(\rho_{k,v}^{-i} + \sum_{j=1}^{M_i} \mathbb{I}(x_{i,j} = v) + \beta_v)}{\Gamma(\rho_{k,v}^{-i} + \beta_v)}$$

where $\rho_{k,v} = \sum_{i:z_i=k} \sum_{j=1}^{M_i} \mathbb{I}(x_{i,j} = v)$ is the number of times we've seen token v in cluster k.

Latent Dirichlet allocation

Let's assume we use the the mixture model above to model documents. In such a model, a document is associated with a single cluster, or topic. It might be more reasonable to associate each document with a mixture over topics, so that

$$
\begin{aligned}
\theta_i &\sim \mathrm{Dirichlet}_K(\alpha), && i = 1, \ldots, N \\
\eta_k &\sim \mathrm{Dirichlet}_V(\beta), && k = 1, \ldots, K \\
z_{i,j} &\sim \mathrm{Discrete}(\theta_i), && j = 1, \ldots, M_i \\
w_{i,j} &\sim \mathrm{Discrete}(\eta_{z_{i,j}}),
\end{aligned}
$$

An uncollapsed Gibbs sampler

$$
\begin{aligned}
p(z_{i,j} = k|\theta_i, \eta_k, w_{i,j} = v) \propto& \theta_{i,k}\eta_{k,v} \\
p(\theta_i|\{z_{i,j}\}_{j=1}^{M_i}) \sim& \text{Dirichlet}\left((m_{i,k} + \alpha)_{k=1}^{K}\right) \\
p(\eta_k|\{w_{i,j} : z_{i,j} = k\}_{j=1}^{M_i}) \sim& \text{Dirichlet}\left((\rho_{v,k} + \eta)_{k=1}^{K}\right)
\end{aligned}
$$

where m_k is the number of times we've seen topic k in document i, and $\rho_{k,v}$ is the number of times we've seen word v in topic k.

A collapsed Gibbs sampler

Integrating out θ_i:

$$\begin{aligned}p(z_{i,j}=k|z_{i,-j},\eta_k,w_{i,j}=v) \propto& p(z_{i,j}=k|z_{i,-j})p(w_{i,j}=v|\eta_k)\\ =& \frac{m_{i,k}^{-j}+\alpha_k}{M_i-1+\sum_k \alpha_k}\eta_{k,v}\end{aligned}$$

Integrating out η_k:

$$\begin{aligned}p(z_{i,j}=k|z_{i,-j},\eta_k,w_{i,j}=v) \propto& p(z_{i,j}=k|z_{i,-j})p(w_{i,j}=v|\{w_{i,j}:z_{i,j}=k\})\\ =& \frac{m_{i,k}^{-j}+\alpha_k}{M_i-1+\sum_k \alpha_k}\cdot\frac{\rho_{k,v}^{-w_{i,j}}+\beta_v}{\sum_{v'}(\rho_{k,v'}^{-w_{i,j}}+\beta_{v'})}\\ \propto& (m_{i,k}^{-j}+\alpha_k)\cdot\frac{\rho_{k,v}^{-w_{i,j}}+\beta_v}{\sum_{v'}(\rho_{k,v'}^{-w_{i,j}}+\beta_{v'})}\end{aligned}$$