The Dirichlet-Multinomial and Dirichlet-Categorical models for Bayesian inference

Stephen Tu
tu.stephenl@gmail.com

1 Introduction

This document collects in one place various results for both the Dirichlet-multinomial and Dirichlet-categorical likelihood model. Both models, while simple, are actually a source of confusion because the terminology has been very sloppily overloaded on the internet. We'll clear the confusion in this writeup.

2 Preliminaries

Let's standardize terminology and outline a few preliminaries.

2.1 Dirichlet distribution

The Dirichlet distribution, which we denote $\text{Dir}(\alpha_1, ..., \alpha_K)$, is parameterized by positive scalars $\alpha_i > 0$ for i=$1, ..., K$, where $K \geq 2$. The support of the Dirichlet distribution is the $(K-1)$-dimensional simplex $\mathcal{S}_K$; that is, all K dimensional vectors which form a valid probability distribution. The probability density of $\mathbf{x} = (x_1, ..., x_K)$ when $\mathbf{x} \in \mathcal{S}_K$ is

$$f(x_1, ..., x_K; \alpha_1, ..., \alpha_K) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K x_i^{\alpha_i - 1}$$

There are some pedagogical concerns to be noted. First, the pdf $f(\mathbf{x})$ is technically defined only in $(K-1)$-dimensional space and not K; this is because $\mathcal{S}_K$ has zero measure in K dimensions[1]. Second, the support is actually the *open* simplex, meaning all $x_i > 0$ (which implies no $x_i = 1$).

The Dirichlet distribution is a generalization of the Beta distribution, which is the conjugate prior for coin flipping.

[1] If you read the Wikipedia entry on the Dirichlet distribution, this is what the mathematical jargon referring to the Lebesgue measure is saying.

2.2 Multinomial distribution

The Multinomial distribution, which we denote $\mathrm{Mult}(p_1, ..., p_K, n)$, is a discrete distribution over K dimensional non-negative integer vectors $\mathbf{x} \in \mathbb{Z}_+^K$ where $\sum_{i=1}^K x_i = n$. Here, $\mathbf{p} = (p_1, ..., p_K)$ is a element of $\mathcal{S}_K$ and $n \geq 1$. The probability mass function is given as

$$f(x_1, ..., x_K; p_1, ..., p_K, n) = \frac{\Gamma(n+1)}{\prod_{i=1}^K \Gamma(x_i+1)} \prod_{i=1}^K p_i^{x_i}$$

The interpretation is simple; a draw $\mathbf{x}$ from $\mathrm{Mult}(p_1, ..., p_K; n)$ can be intepreted as drawing n iid values from a Categorical distribution with pmf $f(X{=}i) = p_i$. Each entry x_i counts the number of times value i was drawn. The strange looking gamma functions in the pmf account for the combinatorics of a draw; it is simply the number of ways of placing n balls in K bins. Indeed, since $\Gamma(n+1) = n!$ for non-negative integer n,

$$\frac{\Gamma(n+1)}{\prod_{i=1}^K \Gamma(x_i+1)} = \frac{n!}{x_1! \cdots x_K!}$$

The Multinomial distribution is a generalization of the Binomial distribution.

2.3 Categorical distribution

The Categorical distribution, which we denote as $\mathrm{Cat}(p_1, ..., p_K)$, is a discrete distribution with support $\{1, ..., K\}$. Once again, $\mathbf{p} = (p_1, ..., p_K) \in \mathcal{S}_K$. It has a probability mass function given as

$$f(x{=}i) = p_i$$

The source of confusion with the Multinomial distribution is because often the popular 1-of-K encoding is used to encode a value drawn from the Categorical distribution, and in this case, we can actually see the Categorical distribution is just a Multinomial with $n = 1$.

In the scalar form, the Categorical distribution is a generalization of the Bernoulli distribution (coin flipping).

2.4 Conjugate priors

Much ink has been spilled on conjugate priors, so we won't attempt to provide any lengthy discussion. What we will mention now is that, when doing Bayesian inference, we are often interested in a few key distributions. Below, D refers to the dataset we have at hand, and we typically assume each $y_i \in D$ is drawn iid from the same distribution $f(y; \theta)$, where θ parameterizes the likelihood model. When people say they are "being Bayesian", what this often means is they are treating θ as an unknown, but postulating that θ follows some prior distribution $f(\theta; \alpha)$, where α parameterizes the prior distribution (and is often called the *hyper-parameter*). Of course you can keep playing this game and put a prior on α (called the *hyper-prior*), but we won't go there.

While so far this seems reasonable, often Bayesian analysis is intractable since we have to integrate out the unknown θ, and for many problems the integral cannot be solved analytically (and one has to resort to numerical techniques). Conjugate priors, however, are the (tiny[2]) class of models where we *can* analytically compute distributions of interest. These distributions are:

Posterior predictive. This is the distribution denoted notationally by $f(y|D)$, where y is a new datapoint of interest. By the iid assumption, we have

$$f(y|D) = \int f(y,\theta|D)\ d\theta = \int f(y|\theta)f(\theta|D)\ d\theta$$

The distribution $f(\theta|D)$, often called the posterior distribution, can be broken down further

$$f(\theta|D) = \frac{f(\theta,D)}{f(D)} \propto f(\theta,D) = f(\theta|\alpha)\prod_{y_i\in D} f(y_i|\theta)$$

Marginal distribution of the data. This is denoted $f(D)$, and is derived by integrating out the model parameter

$$f(D) = \int f(D,\theta)\ d\theta = \int f(\theta|\alpha)\prod_{y_i\in D} f(y_i|\theta)\ d\theta$$

Dirichlet distribution as a prior. It turns out (to further the confusion), that the Dirichlet distribution is the conjugate prior for both the Categorical and Multinomial distributions! For the remainder of this document, we will list the results of both the posterior predictive and marginal distribution on both Dirichlet-Categorical and Dirichlet-Multinomial.

3 Dirichlet-Multinomial

Model.

$$\begin{aligned} p_1,...,p_K &\sim \text{Dir}(\alpha_1,...,\alpha_K) \\ y_1,...y_K &\sim \text{Mult}(p_1,...,p_K) \end{aligned}$$

[2] This class is essentially just the exponential family of distributions.

Posterior.

$$\begin{aligned}
f(\theta|D) &\propto f(\theta, D) \\
&= f(p_1, ..., p_K|\alpha_1, ..., \alpha_K) \prod_{y_i \in D} f(y_i|p_1, ...p_K) \\
&\propto \prod_{j=1}^{K} p_j^{\alpha_j - 1} \prod_{y_i \in D} \prod_{j=1}^{K} p_j^{y_i^{(j)}} \\
&= \prod_{j=1}^{K} p_j^{\alpha_j - 1 + \sum_{y_i \in D} y_i^{(j)}}
\end{aligned}$$

This density is exactly that of a Dirichlet distribution, except we have

$$\alpha_j' = \alpha_j + \sum_{y_i \in D} y_i^{(j)}$$

That is, $f(\theta|D) = \mathrm{Dir}(\alpha_1', ..., \alpha_K')$.

Posterior Predictive.

$$\begin{aligned}
f(y|D) &= \int f(y|\theta) f(\theta|D)\, d\theta \\
&= \int f(y|p_1, ..., p_K) f(p_1, ..., p_K|D)\, d\mathcal{S}_K \\
&= \int \frac{\Gamma(n+1)}{\prod_{j=1}^{K} \Gamma(y^{(j)}+1)} \prod_{j=1}^{K} p_j^{y^{(j)}} \frac{\Gamma(\sum_{j=1}^{K} \alpha_j')}{\prod_{j=1}^{K} \Gamma(\alpha_j')} \prod_{j=1}^{K} p_j^{\alpha_j' - 1}\, d\mathcal{S}_K \\
&= \frac{\Gamma(n+1)}{\prod_{j=1}^{K} \Gamma(y^{(j)}+1)} \frac{\Gamma(\sum_{j=1}^{K} \alpha_j')}{\prod_{j=1}^{K} \Gamma(\alpha_j')} \int \prod_{j=1}^{K} p_j^{y^{(j)} + \alpha_j' - 1}\, d\mathcal{S}_K \\
&= \frac{\Gamma(n+1)}{\prod_{j=1}^{K} \Gamma(y^{(j)}+1)} \frac{\Gamma(\sum_{j=1}^{K} \alpha_j')}{\prod_{j=1}^{K} \Gamma(\alpha_j')} \frac{\prod_{j=1}^{K} \Gamma(y^{(j)} + \alpha_j')}{\Gamma(n + \sum_{j=1}^{K} \alpha_j')} \qquad (1)
\end{aligned}$$

where $d\mathcal{S}_K$ denotes integrating $(p_1, ..., p_K)$ with respect to the $(K-1)$ simplex.

Marignal. This derivation is almost the same as the posterior predictive.

$$
\begin{aligned}
f(D) &= \int f(\theta|\alpha) \prod_{y_i \in D} f(y_i|\theta)\; d\theta \\
&= \int f(p_1, ..., p_K|\alpha_1, ..., \alpha_K) \prod_{y_i \in D} f(y_i|p_1, ..., p_K)\; d\mathcal{S}_K \\
&= \int \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K p_j^{\alpha_j - 1} \prod_{y_i \in D} \frac{\Gamma(n+1)}{\prod_{j=1}^K \Gamma(y_i^{(j)}+1)} \prod_{j=1}^K p_j^{y_i^{(j)}}\; d\mathcal{S}_K \\
&= \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \left[\prod_{y_i \in D} \frac{\Gamma(n+1)}{\prod_{j=1}^K \Gamma(y_i^{(j)}+1)} \right] \int \prod_{j=1}^K p_j^{\sum_{y_i \in D} y_i^{(j)} + \alpha_j - 1}\; d\mathcal{S}_K \\
&= \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \left[\prod_{y_i \in D} \frac{\Gamma(n+1)}{\prod_{j=1}^K \Gamma(y_i^{(j)}+1)} \right] \frac{\prod_{j=1}^K \Gamma(\sum_{y_i \in D} y_i^{(j)} + \alpha_j)}{\Gamma(|D|\, n + \sum_{j=1}^K \alpha_j)} \qquad (2)
\end{aligned}
$$

4 Dirichlet-Categorical

The derivations here are almost identical to before (with some minor syntatic differences).

Model.

$$
\begin{aligned}
p_1, ..., p_K &\sim \mathrm{Dir}(\alpha_1, ..., \alpha_K) \\
y &\sim \mathrm{Cat}(p_1, ..., p_K)
\end{aligned}
$$

Posterior.

$$
\begin{aligned}
f(\theta|D) &\propto f(\theta, D) \\
&= f(p_1, ..., p_K|\alpha_1, ..., \alpha_K) \prod_{y_i \in D} f(y_i|p_1, ...p_K) \\
&\propto \prod_{j=1}^K p_j^{\alpha_j - 1} \prod_{y_i \in D} \prod_{j=1}^K p_j^{\mathbb{1}\{y_i = j\}} \\
&= \prod_{j=1}^K p_j^{\alpha_j - 1 + \sum_{y_i \in D} \mathbb{1}\{y_i = j\}}
\end{aligned}
$$

This density is exactly that of a Dirichlet distribution, except we have

$$
\alpha_j' = \alpha_j + \sum_{y_i \in D} \mathbb{1}\{y_i = j\}
$$

That is, $f(\theta|D) = \mathrm{Dir}(\alpha_1', ..., \alpha_K')$.

Posterior Predictive.

$$\begin{aligned}
f(y{=}x|D) &= \int f(y{=}x|\theta)f(\theta|D)\; d\theta \\
&= \int f(y{=}x|p_1,...,p_K)f(p_1,...,p_K|D)\; d\mathcal{S}_K \\
&= \int p_x \frac{\Gamma(\sum_{j=1}^K \alpha'_j)}{\prod_{j=1}^K \Gamma(\alpha'_j)} \prod_{j=1}^K p_j^{\alpha'_j-1}\; d\mathcal{S}_K \\
&= \frac{\Gamma(\sum_{j=1}^K \alpha'_j)}{\prod_{j=1}^K \Gamma(\alpha'_j)} \int \prod_{j=1}^K p_j^{\mathbb{1}\{x=j\}+\alpha'_j-1}\; d\mathcal{S}_K \\
&= \frac{\Gamma(\sum_{j=1}^K \alpha'_j)}{\prod_{j=1}^K \Gamma(\alpha'_j)} \frac{\prod_{j=1}^K \Gamma(\mathbb{1}\{x=j\}+\alpha'_j)}{\Gamma(1+\sum_{j=1}^K \alpha'_j)} \\
&= \frac{\alpha'_x}{\sum_{j=1}^K \alpha'_j} \qquad (3)
\end{aligned}$$

where we used the fact that $\Gamma(n+1) = n\Gamma(n)$ to simplify the second to last line.

Marignal.

$$\begin{aligned}
f(D) &= \int f(\theta|\alpha) \prod_{y_i\in D} f(y_i|\theta)\; d\theta \\
&= \int f(p_1,...,p_K|\alpha_1,...,\alpha_K) \prod_{y_i\in D} f(y_i|p_1,...,p_K)\; d\mathcal{S}_K \\
&= \int \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K p_j^{\alpha_j-1} \prod_{y_i\in D}\prod_{j=1}^K p_j^{\mathbb{1}\{y_i=j\}}\; d\mathcal{S}_K \\
&= \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \int \prod_{j=1}^K p_j^{\sum_{y_i\in D}\mathbb{1}\{y_i=j\}+\alpha_j-1}\; d\mathcal{S}_K \\
&= \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \frac{\prod_{j=1}^K \Gamma(\sum_{y_i\in D}\mathbb{1}\{y_i=j\}+\alpha_j)}{\Gamma(|D|+\sum_{j=1}^K \alpha_j)} \qquad (4)
\end{aligned}$$