The Dirichlet-multinomial distribution

David Mimno

Let's say we have observations in the form of a sequence of variables $x_1, ..., x_N$ where each x_i is a number from 1 to K. We can summarize this sequence as a vector of K count variables $n_1, ..., n_K$, such that $n_k = \sum_i^N \mathbb{I}[x_i = k]$. We want to estimate the probability that the next observation, x_{N+1} is some value k, $P(k|\boldsymbol{x})$.

The maximum likelihood estimate of this probability is exactly what we would expect, $P(k|\boldsymbol{x}) = \frac{n_k}{N}$. This estimator assigns zero probability to events that haven't occurred in the training data $\boldsymbol{x}$. The Dirichlet-multinomial model provides a useful way of adding "smoothing" to this predictive distribution.

The Dirichlet distribution by itself is a density over K positive numbers $\theta_1, ..., \theta_K$ that sum to one, so we can use it to draw parameters for a multinomial distribution. The parameters of the Dirichlet distribution are positive real numbers $\alpha_1, ..., \alpha_K$. These do not need to sum to one, and in fact their sum has an important effect on the density. Its probability function is

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}. \tag{1}$$

Let's say we draw a distribution $\boldsymbol{\theta}$ from a Dirichlet with parameters $\boldsymbol{\alpha}$, and then sample a sequence of N discrete variables $x_1, ..., x_N$. The probability of $\boldsymbol{x}$ given $\boldsymbol{\theta}$ is $\prod_k \theta_k^{n_k}$. Combining this term with Eq. 1 we get

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{\alpha}) &= p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) && (2) \\ &= \prod_k \theta_k^{n_k} \times \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1} && (3) \\ &= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{n_k + \alpha_k - 1}. && (4) \end{aligned}$$

This form works out so nicely because the Dirichlet and the multinomial are a conjugate pair. We'll talk more about conjugacy later in the course.

We can simplify Eq. 4 by integrating over the distribution $\boldsymbol{\theta}$ to get the marginal probability $p(\boldsymbol{x}|\boldsymbol{\alpha})$. There's a useful trick for doing this kind of integration. The probability density function over the variables $\boldsymbol{\theta}$ has to integrate to one when we integrate over all possible values of $\boldsymbol{\theta}$.

$$\int \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1} d\boldsymbol{\theta} = 1. \tag{5}$$

We can divide a density function into parts that don't involve the variable we're integrating over, and therefore pop outside the integral, and parts that have to stay inside the integral. Using this fact, we get a "cheat sheet" that tells us what scary-looking functions integrate to.

$$\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \int \prod_k \theta_k^{\alpha_k - 1} d\boldsymbol{\theta} = 1 \tag{6}$$

$$\int \prod_k \theta_k^{\alpha_k - 1} d\boldsymbol{\theta} = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}. \tag{7}$$

The joint distribution over $\boldsymbol{x}$ and $\boldsymbol{\theta}$ had just this form, but with parameters "shifted" by the observations: $n_k + \alpha_k$.

$$p(\boldsymbol{x} \mid \boldsymbol{\alpha}) = \int \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{n_k + \alpha_k - 1} d\boldsymbol{\theta} \tag{8}$$

$$= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \int \prod_k \theta_k^{n_k + \alpha_k - 1} d\boldsymbol{\theta} \tag{9}$$

$$= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(\alpha_k + n_k)}{\Gamma(\sum_k \alpha_k + n_k)}. \tag{10}$$

This is the Dirichlet-multinomial distribution, also known as the Dirichlet Compound Multinomial (DCM) or the Pólya distribution. The giant blob of gamma functions is a distribution over a set of K count variables, conditioned on some parameters $\boldsymbol{\alpha}$. We can now get back to our original question: given that you've seen $x_1, ..., x_N$, what is the probability that x_{N+1} is k? By the definition of conditional probability this value is $P(x_{N+1}|x_1, ..., x_N) = P(x_1, ..., x_N, x_{N+1})/P(x_1, ..., x_N)$. Let's define $n_1, ..., n_K$ as the count variables for all the observations up to x_N. Adding one more observation of type k to the N previous observations means that the total number of instances of type k is now $n_k + 1$, and the total number of observations is $N + 1$. The last fact to remember is that $\Gamma(x + 1) = x\Gamma(x)$.

Use Eq. 10 combined with the expression for conditional probability and the gamma recursion.

$$P(x_{N+1}|x_1, ..., x_N) = \quad ??? \tag{11}$$