Bayesian inference of a multinomial distribution

Thomas P. Minka

Abstract

This note derives the posterior, the evidence, and the predictive density for a multinomial distribution, given a conjugate parameter prior.

1 Introduction

The multinomial distribution is

$$p(x|\mathbf{p}) = \prod_{k=1}^{K} p_k^{\delta(x=k)} \tag{1}$$

where $\delta(x = k)$ is an indicator function. The joint probability of N IID samples $\mathbf{X} = \{x_1, ..., x_N\}$ is therefore

$$\begin{aligned} p(\mathbf{X}|\mathbf{p}) &= \prod_{k=1}^{K} p_k^{N_k} & (2) \\ N_k &= \sum_i \delta(x_i = k) & (3) \end{aligned}$$

The joint probability of a set of counts N_k, on the other hand, is

$$p(N_1..N_K|\mathbf{p}) = \binom{N}{N_1..N_K} \prod_{k=1}^{K} p_k^{N_k} \tag{4}$$

since we must account for all possible permutations.

The conjugate prior for $\mathbf{p}$ is the Dirichlet distribution:

$$\begin{aligned} p(\mathbf{p}) &\sim \mathcal{D}(\alpha_1, ..., \alpha_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k p_k^{\alpha_k - 1} & (5) \\ \text{where } p_k &> 0 & (6) \\ \sum_k p_k &= 1 & (7) \end{aligned}$$

This density is proper when all $\alpha_k > 1$. The Beta distribution is the special case when $K = 2$.

The Dirichlet distribution has the properties

$$p(p_1) \sim \mathcal{D}(\alpha_1, \alpha_2 + ... + \alpha_K) \tag{8}$$

$$E[p_1] = \frac{\alpha_1}{\sum_k \alpha_k} \tag{9}$$

$$E[\log p_1] = \Psi(\alpha_1) - \Psi(\sum_k \alpha_k) \tag{10}$$

$$\text{where } \Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)} \tag{11}$$

The maximum of the density is at $p_k = (\alpha_k - 1)/((\sum_k \alpha_k) - K)$, when this value is valid. Otherwise the maximum is on the boundary of the simplex and must be found by nonlinear programming.

Given a Dirichlet prior, the joint distribution is

$$p(\mathbf{X}, \mathbf{p}|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k p_k^{N_k+\alpha_k-1} \tag{12}$$

so the posterior is

$$p(\mathbf{p}|\mathbf{X}, \alpha) \sim \mathcal{D}(N_k + \alpha_k) \tag{13}$$

and the evidence is

$$\begin{aligned} p(\mathbf{X}|\alpha) &= \int_{\mathbf{p}} p(\mathbf{X}, \mathbf{p}|\alpha) && (14) \\ &= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(N_k + \alpha_k)}{\Gamma(\sum_k N_k + \alpha_k)} \int_{\mathbf{p}} \mathcal{D}(\mathbf{p}; N_k + \alpha_k) && (15) \\ &= \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N + \sum_k \alpha_k)} \prod_k \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)} && (16) \end{aligned}$$

which is the probability that this data all came from one multinomial distribution.

For example:

$$\begin{aligned} p(\mathbf{X}|\alpha_k = 1/2) &= \frac{\Gamma(K/2)}{\Gamma(N + K/2)\Gamma(1/2)^K} \prod_k \Gamma(N_k + 1/2) \text{ (Jeffreys' prior)} && (17) \\ p(\mathbf{X}|\alpha_k = 1) &= \frac{\Gamma(K)}{\Gamma(N + K)} \prod_k \Gamma(N_k + 1) \text{ (Uniform prior)} && (18) \\ &= \frac{1}{\binom{N}{N_1..N_K}\binom{N+K-1}{K-1}} && (19) \end{aligned}$$

The posterior predictive distribution is

$$\begin{aligned} p(x = k|\mathbf{X}, \alpha) &= E[p_k|\mathbf{X}] && (20) \\ &= \frac{N_k + \alpha_k}{N + \sum_k \alpha_k} && (21) \end{aligned}$$

$$= \frac{N_k + 1/2}{N + K/2} \text{ (Jeffreys' prior)} \quad (22)$$

$$= \frac{N_k + 1}{N + K} \text{ (Uniform prior)} \quad (23)$$

By contrast, the maximum of the posterior is

$$\hat{p}_k = \frac{N_k + \alpha_k - 1}{N - K + \sum_k \alpha_k} \quad (24)$$

$$= \frac{N_k - 1/2}{N - K/2} \text{ (Jeffreys' prior)} \quad (25)$$

$$= \frac{N_k}{N} \text{ (Uniform prior)} \quad (26)$$

Using the posterior predictive distribution to represent our knowledge of $\mathbf{p}$ was the main argument of Bayes (1763).

For predicting multiple samples, we let M_k be the counts for the new sample and get

$$p(\mathbf{Y}|\mathbf{X}, \alpha) = p(\mathbf{Y}, \mathbf{X}|\alpha)/p(\mathbf{X}|\alpha) \quad (27)$$

$$= \frac{\Gamma(N + \sum_k \alpha_k)}{\Gamma(M + N + \sum_k \alpha_k)} \prod_k \frac{\Gamma(M_k + N_k + \alpha_k)}{\Gamma(N_k + \alpha_k)} \quad (28)$$

i.e. it is just as if the prior were $\mathcal{D}(N_k + \alpha_k)$. Note that the new samples are not independent: we cannot use the posterior predictive distribution for one sample as though it was the true distribution.

2 Samples vs. Counts

There is some confusion in the literature about when to use (2) versus (4). The preceding text used (2), which is the correct choice for most tasks. The confusion stems from the misconception that "the samples are equivalent to their counts." The samples are not equivalent to their counts because the samples occurred in some particular order. Every possible ordering has the same probability but that doesn't mean we should sum over orderings: the data only had one ordering.

For example, consider tossing a fair coin twice. The sample $< H, T >$ has probability 1/4, as does $< H, H >$. However, the count event $\{N_H = 1, N_T = 1\}$ has probability 1/2, which is greater than the probability of the count event $\{N_H = 2\}$. The discrepancy arises because we are considering a different event space. If someone told you that the sample had one head and one tail and asked for the probability of *that sample* (not its counts), then the answer would be 1/4, since every sample with those counts has probability 1/4.

As another example, suppose we have N samples from a univariate Gaussian distribution. The N-dimensional joint density of the samples only depends on the sample mean and sample variance of the sample; these are sufficient statistics. Every sample with the same sufficient statistics has the same probability. However, the joint density of the sufficient statistics has only two dimensions, with a different normalizing term. Asking for the probability of a sample with given sufficient statistics is not the same as asking for the probability *of those statistics*, which concerns a different event space.

3 Evidence vs. Entropy

When the parameter vector $\mathbf{p}$ is uniform, it is well-known (Bishop, 1995) that the log-probability of a set of counts can be approximated by the entropy of the ML estimate, i.e.

$$\begin{aligned}
\log \Gamma(x+1) &\approx x \log(x) - x && (29)\\
\log p(N_1..N_K|p_k = 1/K) &= \log \binom{N}{N_1..N_K} - N \log K && (30)\\
&\approx N \log N - N - \sum_k (N_k \log N_k - N_k) - N \log K && (31)\\
&= -\sum_k N_k \log \frac{N_k}{N} - N \log K && (32)\\
&= N\mathcal{H}(N_k/N) - N \log K && (33)
\end{aligned}$$

supporting the intuition that more entropic counts are more probable. Here it is crucial that we compute the probability of counts; the probability of any particular sample is constant.

There is a corresponding result for the *evidence* of a sample, when the parameter *prior* is uniform. For a large sample, the evidence is approximately the *negative* entropy of the ML estimate:

$$\begin{aligned}
\Gamma(K)/\Gamma(N+K) &\approx 1/\Gamma(N+1) && (34)\\
\log p(\mathbf{X}|\alpha_k = 1) &\approx -N \log N + N + \sum_k N_k \log N_k - N_k && (35)\\
&= \sum_k N_k \log \frac{N_k}{N} && (36)\\
&= -N\mathcal{H}(N_k/N) && (37)
\end{aligned}$$

That is, a uniform prior over parameters prefers less entropic samples. Here it is crucial that we compute the evidence of the particular sample; the counts have constant evidence.

Another information-theoretic quantity is mutual information, which arises in the following way. Suppose we want to know the probability that two samples $\mathbf{X}$ and $\mathbf{Y}$ came from the same multinomial distribution vs. different ones. That is,

$$p(\text{same}|\mathbf{X},\mathbf{Y}) = \frac{p(\mathbf{X},\mathbf{Y}|\text{same})p(\text{same})}{p(\mathbf{X},\mathbf{Y}|\text{same})p(\text{same}) + p(\mathbf{X},\mathbf{Y}|\text{different})p(\text{different})} \tag{38}$$

$$= \frac{1}{1 + \frac{p(\mathbf{X},\mathbf{Y}|\text{different})}{p(\mathbf{X},\mathbf{Y}|\text{same})}\frac{p(\text{different})}{p(\text{same})}} \tag{39}$$

The quantity $\frac{p(\mathbf{X},\mathbf{Y}|\text{different})}{p(\mathbf{X},\mathbf{Y}|\text{same})}$ is the evidence ratio in favor of independence. It can also be written as

$$\frac{p(\mathbf{X}|\alpha)p(\mathbf{Y}|\alpha)}{p(\mathbf{X},\mathbf{Y}|\alpha)} = \frac{\Gamma(M + \sum_k \alpha_k)\Gamma(N + \sum_k \alpha_k)}{\Gamma(\sum_k \alpha_k)\Gamma(M + N + \sum_k \alpha_k)} \prod_k \frac{\Gamma(\alpha_k)\Gamma(M_k + N_k + \alpha_k)}{\Gamma(M_k + \alpha_k)\Gamma(N_k + \alpha_k)} \tag{40}$$

Using the entropy approximation, we can approximate the logarithm of this ratio as

$$-M\mathcal{H}(\frac{M_k}{M}) - N\mathcal{H}(\frac{N_k}{N}) + (M+N)\mathcal{H}(\frac{M_k + N_k}{M+N}) = \tag{41}$$

$$M\mathcal{D}(\frac{M_k}{M} \mid\mid \frac{M_k + N_k}{M+N}) + N\mathcal{D}(\frac{N_k}{N} \mid\mid \frac{M_k + N_k}{M+N}) \tag{42}$$

$$\text{where } \mathcal{D}(p \mid\mid q) = \sum_k p_k \log \frac{p_k}{q_k} \tag{43}$$

i.e. the sum of two Kullback-Liebler divergences. This "average divergence to the mean" can also be written as a mutual information:

$$p(x|c=1) = N_k/N \tag{44}$$

$$p(x|c=2) = M_k/M \tag{45}$$

$$p(c=1) = N/(N+M) \tag{46}$$

$$p(x) = \sum_c p(x|c)p(c) \tag{47}$$

$$\log \frac{p(\mathbf{X}|\alpha)p(\mathbf{Y}|\alpha)}{p(\mathbf{X},\mathbf{Y}|\alpha)} \approx (M+N)\mathcal{H}(x) - N\mathcal{H}(x|c=1) - M\mathcal{H}(x|c=2) \tag{48}$$

$$= (N+M)\mathcal{H}(x) - (N+M)\sum_j p(c=j)\mathcal{H}(x|c=j) \tag{49}$$

$$= (N+M)\mathcal{H}(x) - (N+M)\mathcal{H}(x|c) \tag{50}$$

$$= (N+M)\mathcal{I}(x,c) \tag{51}$$

That is, it measures the information we gain about x by knowing c (or about c by knowing x).

4 Discrete Markov Chains

A discrete Markov Chain is a sequence of random variables x_i such that the conditional distribution $p(x_i|x_{i-1}..x_1) = p(x_i|x_{i-1})$ is multinomial. The probability of a sequence $\mathbf{X} =< x_1, ..., x_N >$

is therefore

$$p(\mathbf{X}|\mathbf{P}) = p(x_1)\prod_{i=2}^{N} p(x_i|x_{i-1}) \tag{52}$$

$$= p(x_1)\prod_{jk} p_{jk}^{N_{jk}} \tag{53}$$

$$N_{jk} = \sum_{i=2}^{N} \delta(x_i = j)\delta(x_{i-1} = k) \tag{54}$$

The conjugate prior for p_{jk} is

$$p(\mathbf{p}_k) \sim \mathcal{D}(\alpha_{1k}..\alpha_{Kk}) \tag{55}$$

where $\mathbf{p}_k$ is the k-th column of $\mathbf{P}$. A shorthand for the latter equation is

$$p(\mathbf{P}) \sim \mathcal{D}(\alpha_{jk}) \tag{56}$$

$$= \frac{\prod_k \Gamma(\sum_j \alpha_{jk})}{\prod_{jk} \Gamma(\alpha_{jk})} \prod_{jk} p_{jk}^{\alpha_{jk}-1} \tag{57}$$

The joint distribution is

$$p(\mathbf{X}, \mathbf{P}|\alpha) = p(x_1)\frac{\prod_k \Gamma(\sum_j \alpha_{jk})}{\prod_{jk} \Gamma(\alpha_{jk})} \prod_{jk} p_{jk}^{N_{jk}+\alpha_{jk}-1} \tag{58}$$

so the posterior is

$$p(\mathbf{P}|\mathbf{X}, \alpha) \sim \mathcal{D}(N_{jk} + \alpha_{jk}) \tag{59}$$

and the evidence is

$$p(\mathbf{X}|\alpha) = \int_{\mathbf{P}} p(\mathbf{X}, \mathbf{P}|\alpha) \tag{60}$$

$$= p(x_1)\frac{\prod_k \Gamma(\sum_j \alpha_{jk})}{\prod_{jk} \Gamma(\alpha_{jk})} \frac{\prod_{jk} \Gamma(N_{jk} + \alpha_{jk})}{\prod_k \Gamma(\sum_j N_{jk} + \alpha_{jk})} \int_{\mathbf{P}} \mathcal{D}(\mathbf{P}; N_{jk} + \alpha_{jk}) \tag{61}$$

$$= p(x_1)\prod_k \frac{\Gamma(\sum_j \alpha_{jk})}{\Gamma(\sum_j N_{jk} + \alpha_{jk})} \prod_j \frac{\Gamma(N_{jk} + \alpha_{jk})}{\Gamma(\alpha_{jk})} \tag{62}$$

The posterior predictive distribution is

$$p(x_{N+1} = j|x_N = k, \mathbf{X}, \alpha) = E[p_{jk}|\mathbf{X}] \tag{63}$$

$$= \frac{N_{jk} + \alpha_{jk}}{\sum_j N_{jk} + \alpha_{jk}} \tag{64}$$

For predicting multiple samples, we let M_{jk} be the counts for the new sample and get

$$p(\mathbf{Y}|\mathbf{X}, \alpha) = p(\mathbf{Y}, \mathbf{X}|\alpha)/p(\mathbf{X}|\alpha) \tag{65}$$

$$= \prod_k \frac{\Gamma(\sum_j N_{jk} + \alpha_{jk})}{\Gamma(\sum_j M_{jk} + N_{jk} + \alpha_{jk})} \prod_j \frac{\Gamma(M_{jk} + N_{jk} + \alpha_{jk})}{\Gamma(N_{jk} + \alpha_{jk})} \tag{66}$$

References

[1] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc. London*, 53:370–418, 1763. Reprinted in *Biometrika* 45, 293–315, and in Press (1989).

[2] C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.