The Dirichlet-tree distribution

Tom Minka
Justsystem Pittsburgh Research Center
July, 1999 (revised Oct, 2004)

Abstract

This note further explores the Dirichlet-tree distribution developed by Dennis (1991). An inclusion relationship is given for different tree structures and an independence relationship is given for probabilities in the tree. The posterior, evidence, and predictive density are derived for multinomial observations.

1 Introduction

The Dirichlet distribution has enjoyed considerable popularity as a prior distribution for multinomial parameters. This is mainly because the Dirichlet is conjugate to the multinomial under the conventional parameterization. But the Dirichlet distribution has key limitations:

1. Each variable has its own mean, but they must all share a common variance parameter.
2. Aside from the constraint that they sum to one, the variables are mutually independent (Mosimann, 1962).

This note describes a new distribution which overcomes these limitations while preserving computational simplicity. The new distribution is simply the conjugate distribution to the multinomial under a different parameterization.

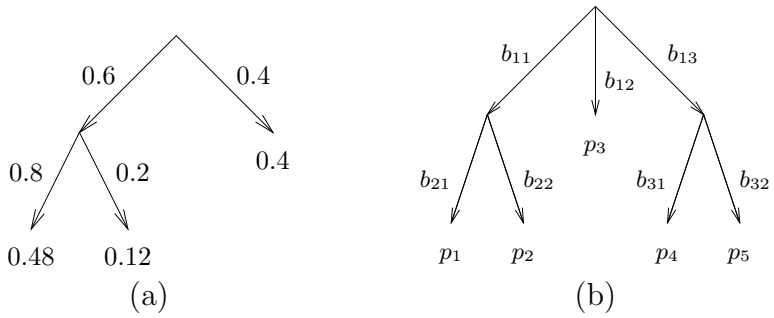


Figure 1: (a) A finite stochastic process (b) Notation for a general process

Instead of representing a multinomial sample as the outcome of a K-sided die, we can represent it as the outcome of a finite stochastic process, such as the tree shown in figure 1(a). The probability of a leaf is the product of branch probabilities leading to that leaf. To represent an arbitrary tree,

we can use the notation in figure 1(b). Before, the parameters were the leaf probabilities $[p_1 ... p_K]$, so that the probability of a sample x was

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

Under the tree parameterization, it is instead written

$$p(x|\mathbf{B}, T) = \prod_{\text{(nodes } j\text{)}} \prod_{\text{(branches } c\text{)}} b_{jc}^{\delta_{jc}(x)} \tag{2}$$

$$\delta_{jc}(x) = \begin{cases} 1 & \text{if branch } jc \text{ leads to } x \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The probability of reaching any interior node can also be computed this way. The conjugate prior for this parameterization is no longer a single Dirichlet density but rather a product of Dirichlet densities, one for each node:

$$p(\mathbf{B}|\alpha) = \prod_{\text{(nodes } j\text{)}} p(\mathbf{b}_j|\alpha) \tag{4}$$

$$p(\mathbf{b}_j|\alpha) \sim \mathcal{D}(\alpha_{jc}) \tag{5}$$

The "Dirichlet-tree distribution" is the distribution over leaf probabilities $[p_1 ... p_K]$ that results from this prior on branch probabilities. The distribution is a function of the tree structure T as well as α. The explicit density over $[p_1 ... p_K]$ can be computed by noting that

$$b_{jc} = \frac{\sum_k \delta_{jc}(k) p_k}{\sum_{kc'} \delta_{jc'}(k) p_k} \tag{6}$$

i.e. b_{jc} is proportional to the probability mass in its subtree. Combining this with (4) and multiplying by the Jacobian gives (Dennis, 1991)

$$p(\mathbf{p}|\alpha, T) = \prod_k p_k^{\alpha_{\text{parent}(k)}-1} \prod_j \frac{\Gamma(\sum_c \alpha_{jc})}{\prod_c \Gamma(\alpha_{jc})} \left(\sum_{kc} \delta_{jc}(k) p_k \right)^{\beta_j} \tag{7}$$

$$\beta_j = \alpha_{\text{parent}(j)} - \sum_c \alpha_{jc} \qquad \text{(or 0 if } j \text{ is the root node)} \tag{8}$$

where $\alpha_{\text{parent}(j)}$ means the α parameter for the branch immediately leading to j.

Since the Dirichlet distribution at a node can be arbitrarily broad or sharp, the Dirichlet-tree distribution can give an independent variance to each p_k. Also, the leaves in a subtree are correlated since they all depend on the ancestors of that subtree.

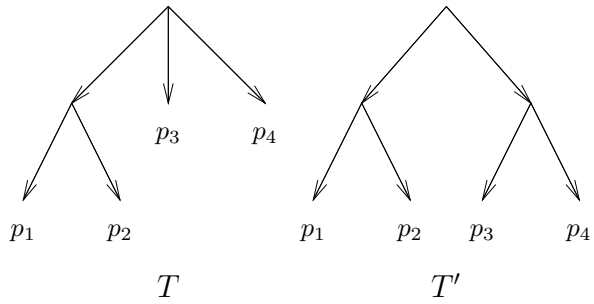
2 Properties

The Dirichlet-tree distribution obviously generalizes the Dirichlet distribution, since we can always have a tree of depth 1. Surprisingly, a depth 1 tree is not necessary: it is possible for any tree

structure to realize a Dirichlet distribution on $[p_1 ... p_K]$, if the α's are chosen carefully. In particular, if the β_j in (7) are all zero, then the distribution is Dirichlet. This result is generalized in the following theorem:

Theorem 1 Let $S(T)$ be the set of Dirichlet-tree distributions realizable by tree structure T. If T' is identical to T except for an additional interior node, then $S(T')$ is a proper superset of $S(T)$.

Proof T' adds a new α to the tree and a new β_j to (7). By setting $\beta_j = 0$, we can realize anything in $S(T)$. If $\beta_j \neq 0$, then we can get distributions not in $S(T)$.



The marginal distribution of any p_k is difficult to compute, since it is a multiplicative convolution of Dirichlet densities. But the moments of p_k are easy. The first two moments are (Dennis, 1991):

$$E[p_k] = \prod_{jc} E[b_{jc}]^{\delta_{jc}(k)} = \prod_{jc} \left(\frac{\alpha_{jc}}{\sum_{c'} \alpha_{jc'}} \right)^{\delta_{jc}(k)} \tag{9}$$

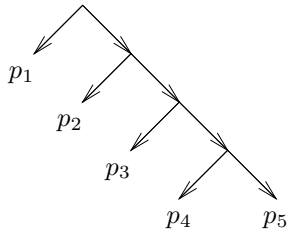
$$E[p_k^2] = \prod_{jc} E[b_{jc}^2]^{\delta_{jc}(k)} = E[p_k] \prod_{jc} \left(\frac{1 + \alpha_{jc}}{1 + \sum_{c'} \alpha_{jc'}} \right)^{\delta_{jc}(k)} \tag{10}$$

The most probable $\mathbf{p}$, given by maximizing (7), corresponds to setting the branch probabilities to

$$b_{jc} \propto \alpha_{jc} - \sum_k \delta_{jc}(k) \tag{11}$$

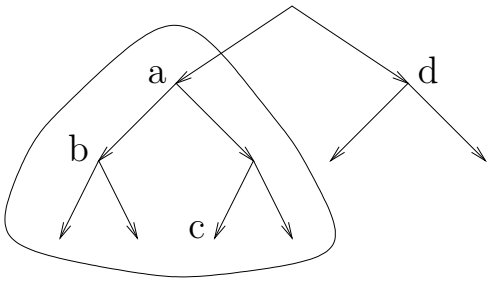
The subtracted term is simply the total number of leaves under branch jc. In the Dirichlet case, this reduces to the usual formula $p_k \propto \alpha_k - 1$.

A special case of the Dirichlet-tree distribution was developed by Connor & Mosimann (1969). It restricts the tree structure to be a binary cascade, as shown to the right.



Connor & Mosimann showed that the cascade-tree distribution satisfies a certain independence property. The following theorem generalizes this property to arbitrary trees.

Theorem 2 Let a be a node in the tree such that nodes b and c are descendants of a while node d is not. Then the ratio p_b/p_c is independent of p_d. That is, relative probabilities in a subtree are independent of anything outside the subtree.



Proof The probability of a node is the product of branch probabilities leading to it. The ratio p_b/p_c only involves branch probabilities below a, so the theorem follows.

3 Bayesian inference

Since the Dirichlet-tree distribution is a conjugate distribution, it is straightforward to compute the posterior distribution given data $D = \{x_1 ... x_N\}$:

$$\begin{aligned} p(\mathbf{B}|D,\alpha,T) &\propto p(\mathbf{B}|\alpha)\prod_i p(x_i|\mathbf{B},T) && (12)\\ &\propto \prod_{jc} b_{jc}^{\alpha_{jc}-1} b_{jc}^{\sum_i \delta_{jc}(x_i)} && (13)\\ &\sim \prod_j \mathcal{D}(\alpha_{jc} + \sum_i \delta_{jc}(x_i)) && (14)\end{aligned}$$

We simply add 1 to α_{jc} for every data point reachable from branch jc. Each observation x is equivalent to a boolean matrix of decision outcomes, given by $\delta_{jc}(x)$. So from the perspective of node j, it is just as if its children were leaves. The same result holds for the evidence:

$$\begin{aligned} p(D|\alpha,T) &= \int_{\mathbf{B}} p(\mathbf{B}|\alpha)\prod_i p(x_i|\mathbf{B},T) && (15)\\ &= \prod_j \left(\frac{\Gamma(\sum_c \alpha_{jc})}{\Gamma(\sum_c \alpha_{jc} + \sum_c n_{jc})} \prod_c \frac{\Gamma(\alpha_{jc}+n_{jc})}{\Gamma(\alpha_{jc})} \right) && (16)\\ n_{jc} &= \sum_i \delta_{jc}(x_i) && (17)\end{aligned}$$

The predictive density for a new sample $D' = \{y_1 ... y_M\}$ is therefore

$$\begin{aligned} p(D'|D,\alpha,T) &= p(D',D|\alpha,T)/p(D|\alpha,T) && (18)\\ &= \prod_j \left(\frac{\Gamma(\sum_c \alpha_{jc} + \sum_c n_{jc})}{\Gamma(\sum_c \alpha_{jc} + \sum_c n_{jc} + \sum_c m_{jc})} \prod_c \frac{\Gamma(\alpha_{jc}+n_{jc}+m_{jc})}{\Gamma(\alpha_{jc}+n_{jc})} \right) && (19)\\ m_{jc} &= \sum_i \delta_{jc}(y_i) && (20)\end{aligned}$$

These results also appear in Dennis (1996).

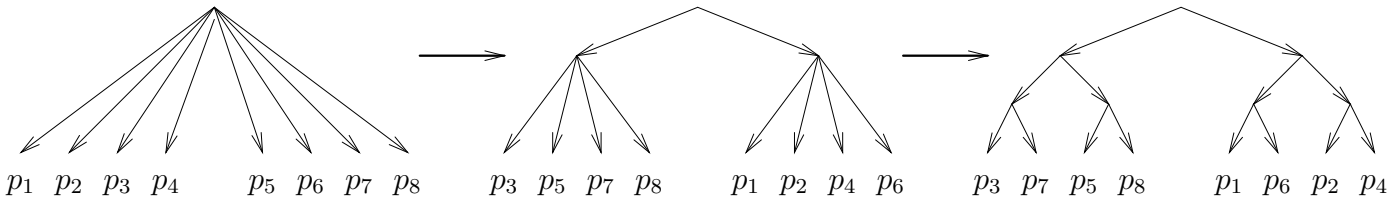


Figure 2: Greedy construction of tree structure

4 Learning the tree structure

Suppose we want to find the tree structure which best fits a data set. The data set is a matrix of counts $D = \{\mathbf{x}_1 ... \mathbf{x}_N\}$ where each column $\mathbf{x}_s$ is a set of samples. For any particular tree structure, we can convert this dataset into a matrix of counts n_{jc}, which is the number of times branch jc was taken. Define $D_j = \{n_{j1}, ..., n_{jC}\}$ to be the dataset from the perspective of node j.

Formally, we want to do model selection by maximizing the probability of the data, with parameters integrated out (the "structure evidence"):

$$p(D|T) = \int_\alpha p(D|\alpha, T)p(\alpha|T) \quad (21)$$

$$= \prod_j \int_{\alpha_j} p(D_j|\alpha_j)p(\alpha_j) \quad (22)$$

Because each interior node has separate α parameters, this decouples into a product of integrals, one for each interior node. Intuitively, the structure evidence for an interior node measures how well the data for that node matches an ordinary Dirichlet distribution.

To evaluate this criterion efficiently, one could use Laplace's method at the maximum-likelihood estimate of α, or a variational method based on the lower bounds derived by Minka (2000).

Given an efficient way to evaluate the evidence at each interior node, one way to optimize it is to greedily build the tree structure top-down, as illustrated in figure 2. By Theorem 1, we only need to consider binary trees. Start with one interior node. At each step, pick an interior node r with > 2 children (which must all be leaves), and introduce two new interior nodes, j_1 and j_2, which will parent the children. Each possible division of the children is scored using the evidence. Note that when we make a local change to the tree structure, most of the terms in (22) are unchanged. In particular, when we introduce j_1 and j_2, we change the term for r and add terms for j_1 and j_2. All the other terms are unchanged.

Further development of these ideas is left to the reader.

Acknowledgement

This work was performed under the supervision and encouragement of Andrew McCallum at Justsystem Pittsburgh Research Center.

References

[1] R. J. Connor and J. E. Mosimann. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association* 64: pp 194–206, 1969.

[2] S. Y. Dennis III. On the Hyper-Dirichlet Type 1 and Hyper-Liouville distributions. *Communications in Statistics—Theory and Methods* 20(12): pp 4069–4081, 1991.

[3] S. Y. Dennis III. A Bayesian analysis of tree-structured statistical decision problems. *Journal of Statistical Planning and Inference* 53(3): pp 323–344, 1996.

[4] T. P. Minka. Estimating a Dirichlet distribution. `research.microsoft.com/~minka/papers/dirichlet/`, 2000.

[5] J. E. Mosimann. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* 49(1,2): pp 65–82, 1962.