Machine Learning and Pattern Recognition: Note on Dirichlet Multinomial

Course Lecturer:Amos J Storkey
Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
10 Crichton Street, Edinburgh UK
a.storkey@ed.ac.uk
Course page : `http://www.inf.ed.ac.uk/teaching/courses/mlpr/`

1 The Multinomial Distribution

These are intended to be helpful pointers. They are highly concise, and you will need to look parts up elsewhere.

If our data is a multivariate variable, we typically represent it by a multivariate distribution.

If we have many IID data items then the many multivariate distributions of each and every data item can be summarised using the multinomial distribution for the counts c_i of those items. I.e. we will have

$$P(D|\boldsymbol{\theta}) = \frac{N!}{\prod c_i!} \prod \theta_i^{c_i}$$

where $N = \sum_i c_i$. Note this multinomial distribution has one parameter θ_i for each class i the data can take. These parameters are in fact probabilities. For example in document analysis, there would be one θ value for each possible word, and the c_i would be the number of occurrences of the ith word.

2 The Dirichlet Distribution

The conjugate prior for the multinomial distribution is the Dirichlet distribution.

$$P(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod \theta_i^{\alpha_i - 1}.$$

This should be obvious: as a function of $\boldsymbol{\theta}$ they both have the same form. If you want to know more about the normalisation function B, then look this up elsewhere.

3 The Bayesian Posterior

So in a conjugate Bayesian analysis we will have the multinomial likelihood and the Dirichlet prior. After observing data with counts $\{c_i\}$. we have the posterior distribution for the parameters as being

$$\begin{aligned} P(\boldsymbol{\theta}|D, \boldsymbol{\alpha}) &\propto P(D|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{\alpha}) \\ &= \frac{N!}{\prod c_i!} \prod \theta_i^{c_i} \frac{1}{B(\alpha)} \prod \theta_i^{\alpha_i - 1} \propto \prod \theta_i^{\alpha_i + c_i - 1} \end{aligned} \quad (1)$$

which we recognise as having the form of a Dirichlet distribution. Hence we know the constant of proportionality immediately from the form of Dirichlet distribution and so we have

$$P(\boldsymbol{\theta}|D, \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha} + \mathbf{c})} \prod \theta_i^{\alpha_i + c_i - 1}.$$

4 Bayesian Inference

Fine. But suppose we now want to infer the probability of some new data point x. What would that be? We we have a multivariate distribution for x. Let $\mathbf{x}$ be the indicator vector for x, with a one in the row of $\mathbf{x}$ corresponding to the class x. Now the posterior distribution for the parameters of that multivariate distribution.

$$P(x|D, \boldsymbol{\alpha}) = \int d\boldsymbol{\theta} \frac{N!}{\prod c_i!} \prod_i \theta_i^{x_i} \frac{1}{B(\boldsymbol{\alpha} + \mathbf{c})} \prod \theta_i^{\alpha_i + c_i - 1}$$

We can do this integral (to work through this you need to know what B is) and we get

$$P(x = j|D, \boldsymbol{\alpha}) = \frac{\alpha_j + c_j}{\sum_j (\alpha_j + c_j)}.$$

We note that a Dirichlet multinomial model is equivalent to a regularised maximum posterior model for a different Dirichlet prior! The result for the regularised maximum posterior model with our prior would be

$$P(x = j|D, \boldsymbol{\theta}_{max}\boldsymbol{\alpha}) = \frac{\alpha_j + c_j - 1}{\sum_j (\alpha_j + c_j - 1)}.$$