Computational Cognitive Science 2019-2020 Tutorial 7: Dirichlet distributions

The Dirichlet Distribution and the Dirichlet-Categorical Distribution

The Dirichlet distribution is a distribution over probability distributions; that is, draws from a Dirichlet are a probability distribution. In the lecture on overhypotheses, the Dirichlet distribution was parameterised with two values: Dirichlet($\alpha\boldsymbol{\beta}$), where $\alpha > 0$ is a scalar and $\boldsymbol{\beta}$ is a probability distribution of size K, where all $\beta_k > 0$ and $\sum_k \beta_k = 1$.

The two parameters play different roles: $\boldsymbol{\beta}$ is the **base distribution** and α is the **concentration parameter**, governing how much draws from Dirichlet($\alpha\boldsymbol{\beta}$) will diverge from the base distribution.

This exercise is meant to strengthen your intuitions about the role of the α and $\boldsymbol{\beta}$ hyperparameters in the context of Dirichlet priors with categorical likelihoods, by examining how the hyperparameters influence the prediction of the next draw.

Dirichlet priors with categorical likelihoods have a convenient closed form for the predictive posterior, integrating over all possible draws from the Dirichlet (θ in the lectures). With the Dirichlet($\alpha\boldsymbol{\beta}$) parameterisation, the predictive posterior is:

$$p(y = k|D, \alpha, \boldsymbol{\beta}) = \frac{\alpha\beta_k + N_k}{\sum_{k'} \alpha\beta_{k'} + N_{k'}}$$

where N_k refers to the number of items of category k in D.

Exercise: Consider a dataset consisting of ten marbles, two of which are black (marbles can only be black and white; so $K = 2; N = 10; N_{\textbf{black}} = 2$). Given this dataset, calculate the probability of the next marble being black, using the following hyperparameter settings (9 different settings; preferably vary α while keeping $\boldsymbol{\beta}$ stable):

- $\boldsymbol{\beta} = [0.5, 0.5], \boldsymbol{\beta} = [0.2, 0.8], \boldsymbol{\beta} = [0.8, 0.2]$
- $\alpha = 0.01, 1, 100$

You can calculate these by hand (try this for 1-2 to see what the numerator and denominator look like) or use R/Matlab or any language you prefer.

Question: What is the effect of a small α? A large α?

Question: The empirical likelihood of black is 0.2. When and why do the predictive posteriors match from the empirical likelihood?

In this example, the posterior predictive probability always matches or increases the probability of black, as compared to the empirical likelihood.

Question: What kind of hyperparameters would result in the predictive probability of black being *lower* than the empirical likelihood?

Kemp et al.'s hierarchical model does not specify α directly, but instead specifies a prior over possible values of α.

Question: Give an example where this leads to very different predictions than we'd see from a model that uses a fixed value of α.

Question: What prior over α did Kemp et al. use? Describe one or more features that any prior over alpha should have.

Finally, Kemp et al.'s model can discover that some features (like shape) are important for some kinds of things (like rigid objects), whereas other features (like color) are important for other kinds of objects (like substances).

Question: Do we need to specify the numbers of kinds of things in advance, in this model? Why or why not?

Question: Can you give a high level summary of how a "Chinese restaurant process" works?