The Dirichlet Distribution

Matthew Stephens

2017-02-19

Last updated: 2017-03-06

Code version: c7339fc

Overview

The purpose of this vignette is to introduce the Dirichlet distribution. You should be familiar with the Beta distribution (beta.html) since the Dirichlet can be thought of as a generalization of the Beta distribution.

If you want more details you could look at Wikipedia (https://en.wikipedia.org/wiki/Dirichlet_distribution).

The Dirichlet Distribution

You can think of the J-dimensional Dirichlet distribution as a distribution on probability vectors, $q = (q_1, \dots, q_J)$, whose elements are non-negative and sum to 1. It is perhaps the most commonly-used distribution for probability vectors, and plays a central role in Bayesian inference from multinomial data.

The Dirichlet distribution has J parameters, $\alpha_1, \dots, \alpha_J$ that control the mean and variance of the distribution. If $q \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_J)$ then:

- The expectation of q_j is $\alpha_j/(\alpha_1 + \cdots + \alpha_J)$.
- The variance of q_j becomes smaller as the sum $\sum_j \alpha_j$ increases.

As a generalization of the Beta distribution

The 2-dimensional Dirichlet distribution is essentially the Beta distribution. Specifically, let $q = (q_1, q_2)$. Then $q \sim Dirichlet(\alpha_1, \alpha_2)$ implies that

$$q_1 \sim \text{Beta}(\alpha_1, \alpha_2)$$

and $q_2 = 1 - q_1$.

Other connections to the Beta distribution

More generally, the marginals of the Dirichlet distribution are also beta distributions.

That is, if $q \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_J)$ then $q_j \sim \text{Beta}(\alpha_j, \sum_{j' \neq j} \alpha_{j'})$.

Density

The density of the Dirichlet distribution is most conveniently written as

$$p(q|\alpha) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_J)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_J)} \prod_{j=1}^{J} q_j^{\alpha_j - 1} \qquad (q_j \geq 0; \quad \sum_j q_j = 1).$$

where $Gamma$ here denotes the gamma function.

Actually when writing the density this way, a little care needs to be taken to make things formally correct. Specifically, if you perform standard (Lebesgue) integration of this “density” over the J dimensional space $q_1, \dots, q_J$ it integrates to 0, and not 1 as a density should. This problem is caused by the constraint that the qs must sum to 1, which means that the Dirichlet distribution is effectively a $J-1$-dimensional distribution and not a J dimensional distribution.

The simplest resolution to this is to think of the J dimensional Dirichlet distribution as a distribution on the $J-1$ numbers $(q_1, \dots, q_{J-1})$, satisfying $\sum_{j=1}^{J-1} q_j \leq 1$, and then define $q_J := (1 - q_1 - q_2 - \cdots - q_{J-1})$. Then, if we integrate the density

$$p(q_1, \dots, q_{J-1}|\alpha) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_J)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_J)} \prod_{j=1}^{J-1} q_j^{\alpha_j - 1} (1 - q_1 - \cdots - q_{J-1})^{\alpha_J} \qquad (q_j \geq 0; \quad \sum_{j=1}^{J-1} q_j \leq$$

over $(q_1, \dots, q_{J-1})$, it integrates to 1 as a density should.

Examples

Session information

```
sessionInfo()
```

```
R version 3.3.2 (2016-10-31)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.5 LTS

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] workflowr_0.4.0    rmarkdown_1.3.9004

loaded via a namespace (and not attached):
 [1] backports_1.0.5 magrittr_1.5    rprojroot_1.2   htmltools_0.3.5
 [5] tools_3.3.2     yaml_2.1.14     Rcpp_0.12.9     stringi_1.1.2
 [9] knitr_1.15.1    git2r_0.18.0    stringr_1.2.0   digest_0.6.12
[13] evaluate_0.10
```

This site was created with R Markdown (http://rmarkdown.rstudio.com)