The Dirichlet Distribution

Selected topics from Part 1 of Data Mining Intro

2025-01-26 12:54 GMT

Abstract

The Dirichlet and Multinomial distributions are introduced in preparation for a discussion of Latent Dirichlet Allocation (LDA).

Table of Contents

## Background

Latent Dirichlet Allocation (LDA) was proposed as a method of topic modeling in 2003 in a paper by Blei, Ng, and Jordan. The method is briefly mentioned in Part 1 of the course. Several course participants requested a more detailed description. This note prepares for the requested response by introducing the Dirichlet and Multinomial distributions.

## The Multinomial Distribution

Consider a categorical (qualitative) random variable that randomly selects one of distinct categories .

Probability vector is thus the probability distribution of on .

If the categories are ordered, we might represent numerically as a random index . But for present purposes we suppose the categories are not ordered, and we represent as a vector of random indicator variables[[1]](#footnote-21).

That is, the random indicator vector selects one of the Euclidean basis vectors with probability . Therefore the expected value of equals the vector of category probabilities .

Now suppose that are independent random variables all having the same distribution as . Corresponding to we have an indicator vector that we’ll denote as . Let denote the sum over the indicator vectors .

For any given set of possible counts , that is, of non-negative integers summing to , and for a given probability vector , the probability that is as follows.

where

gives the number of assignments of objects to the categories such that each category receives the prescribed number of objects.

The probability distribution of is called the *multinomial distribution*.

For , this simplifies to the binomial distribution, with parameters , where . If denotes the observed number of succeses in Bernoulli trials, then .

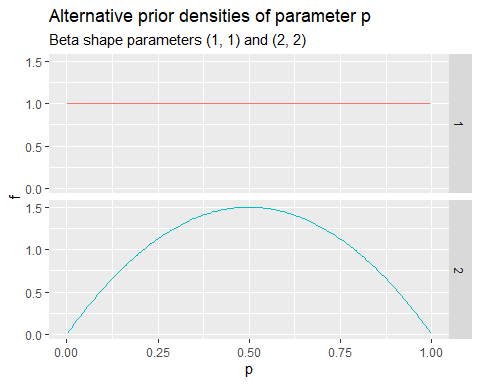
## Bayesian Inference

The Bayesian approach to statistical estimation provides a framework for representing the state of knowledge, or degree of uncertainty, about a model parameter before and after collecting relevant data.

As an example consider the binomial distribution mentioned above, where we are estimating the probability of success based on a sequence of independent Bernoulli trials. Following the notation above, we use two dependent indicator random variables to represent success and failure respectively, with , and with .

Prior to observing the sequence of Bernoulli trials, we might represent the state of information about as a uniform distribution, so that each possible value of is deemed equally likely. Alternatively, if the Bernoulli sequence represents the outcomes of tossing a coin that is presumed fair, or nearly fair, we might represent that information as a probability distribution having a mode at the value .

More specifically, it turns out to be mathematically convenient to represent prior information about probability as a member of the beta family of probability distributions over the unit interval. The uniform distribution is the special case of setting beta shape parameters to the values . Alternatively setting the shape parameters to gives a density function symmetric about the mode at .



In general, a beta distribution having shape parameters has the following density function.

Suppose now that we adopt the distribution just mentioned, , as the prior distribution of success probability , for some specified positive parameter values . We then observe successes and failures from independent Bernoulli trials. Based on these observations we update the prior distribution to form the posterior distribution of as follows.

Note that

so that

Consequently, the ratio of the last two expressions gives

That is, the posterior distribution is .

This is the “mathematical convenience” previously alluded to: the prior and posterior probability distributions of belong to the same family of parametric probability distributions, namely the family. For this reason the family is said to be *conjugate* to the binomial family.

## The Dirichlet Probability Distribution

### Definition

The binomial distribution is a special case of the multinomial distribution in which the number of categories is equal to 2. (In the discussion above we referred to the categories as success and failue, respectively.) More generally, for a fixed integer , the family of distributions conjugate to the multinomial family of distributions over categories is the following *Dirichlet* family.

This is the distribution over , where probability vector ranges over the dimesional simplex such that each component is non-negative and all the components together sum to unity.

### Special Case:

We will denote the sum of the components of as .[[2]](#footnote-28)

Consider the special case in which all the components of have the same value

The probability distribution is then symmetric in the components of , and is referred to as the *concentration parameter*. The uniform distribution over the domain of (that is, over the dimensional simplex) is obtained by setting . Setting concentrates the distribution around the centroid

For example the previous figure shows two alternative concentrations of the symmetric beta density function of scalar parameter , where , and .

Alternatively, setting concentrates the distribution away from the centroid toward the corners of the simplex.

### Posterior Distribution

Recall that constructed above has a multinomial distribution if it can be represented as the sum of independent trials, each trial yielding one of the Euclidean basis vectors with probability . Then , the component of , counts the number of trials yielding .

Now suppose that has prior distribution , for some specification of , and that the observed outcome of the trials is a given vector of counts. What is the posterior distribution of given the observation ? This turns out to be

That is, the Dirichlet family is conjugate to the multinomial family, just as the beta family is conjugate to the binomial family.

## Resources

[Dirichlet distribution - Wikipedia](https://en.wikipedia.org/wiki/Dirichlet_distribution)

1. In machine learning this mapping of a categorical variable to an indicator vector is called “one-hot encoding”. [↑](#footnote-ref-21)
2. An alternative notation for the sum of is . [↑](#footnote-ref-28)