Conjugacy for Categorical Distributions

Carlo Tomasi

Learning amounts to estimating the parameters θ of a probabilistic model for variable z from the training data $z_1, \ldots, z_I$. If the model is a prior or a posterior, then z is the world state w. If the model is a likelihood, then z is the data feature x. If the model is a joint distribution, then z is the concatenation of x and w.

An important question in machine learning and more generally in parameter estimation is whether the parameter θ is viewed as a random variable[1] or a deterministic unknown. Is there a "true value" for θ out there and estimation needs to find it (possibly approximately), or is θ inherently unknowable, and all we can do is to determine its distribution given the data?

In the first approach, parameter estimation finds what is called a *point estimate* $\hat{\theta}$ of θ. The measure of fit is then called Maximum Likelihood (ML) if no prior knowledge of θ is assumed, or Maximum a Posteriori (MAP) if one assumes to *know* a prior $p(\theta|\pi)$ for the parameter. The prior itself is of course instantiated by a set of parameters (π), which are assumed to be known. The second approach, in which θ is a random variable, is called the *Bayesian approach* to parameter estimation.[2]

Carefully note the distinction between MAP and the Bayesian approach: In both approaches, the prior $p(\theta|\pi)$ is known, including all its parameters. However, the MAP approach finds a point estimate $\hat{\theta}(z_1, \ldots, z_I, \pi)$ given the data. This point estimate is then simply replaced in the likelihood function $p(z|\theta)$ to yield the predictive distribution $p(z|z_1, \ldots, z_I, \pi) = p(z|\hat{\theta}(z_1, \ldots, z_I, \pi))$. The Bayesian approach uses the data to compute a *posterior distribution* $p(\theta|z_1, \ldots, z_I, \pi)$ for θ given the data. Because of this, the parameter θ must be marginalized out to find

[1] "Variable" here is used generically to denote a scalar, a vector, or a matrix.

[2] Do not confuse the use of "Bayesian" for parameter estimation with the use of "Bayesian" in "Bayesian inference." We use Bayesian inference regardless of whether we estimate the parameters of the posterior with ML, MAP, or the Bayesian approach.

the predictive distribution:

$$p(z \mid z_1, \ldots, z_I, \pi) = \int p(z|\theta)\, p(\theta|z_1, \ldots, z_I, \pi)\, d\theta \; . \tag{1}$$

Conjugacy

The integral in equation (1) can be difficult to compute either analytically or, when the dimensionality of θ is large, numerically. Because of this difficulty, the forms of data distribution $p(z|\theta)$ and parameter prior $p(\theta|\pi)$ are carefully designed so that the integral becomes easy to compute. Specifically, suppose that these distributions can be chosen so that

$$p(z|\theta)\, p(\theta|\pi) = c(z, \pi)\, p(\theta|\pi') \; .$$

In words, the product of data distribution and parameter prior (the left-hand side) is proportional to a distribution $p(\theta|\pi')$ that has the same analytic form as the parameter prior $p(\theta|\pi)$, although possibly with different parameters ($\pi' \neq \pi$). Crucially, the constant of proportionality $c(z, \pi)$ is independent of θ. In that case, the parameter prior $p(\theta|\pi)$ is said to be the *conjugate* to the data distribution $p(z|\theta)$.

The reason why conjugacy is useful is that it makes the integral disappear:

$$\begin{aligned} \int p(z|\theta)\, p(\theta|\pi)\, d\theta &= \int c(z, \pi)\, p(\theta|\pi')\, d\theta & (2) \\ &= c(z, \pi) \int p(\theta|\pi')\, d\theta & (3) \\ &= c(z, \pi) \; . & (4) \end{aligned}$$

Look at these equalities carefully:

(2) Conjugacy replaces the product of two functions of θ with the product of a probability distribution on θ and one that does not depend on this variable.

(3) The function $c(z, \pi)$ can be taken out of the integral because it does not depend on θ.

(4) The integral of what is left is one, because $p(\theta|\pi')$ is a distribution.

Clearly, this magic requires very special forms for the distributions involved. This note explores conjugacy for discrete distributions on z, with a continuous vector $\theta = \boldsymbol{\lambda}$ of parameters.

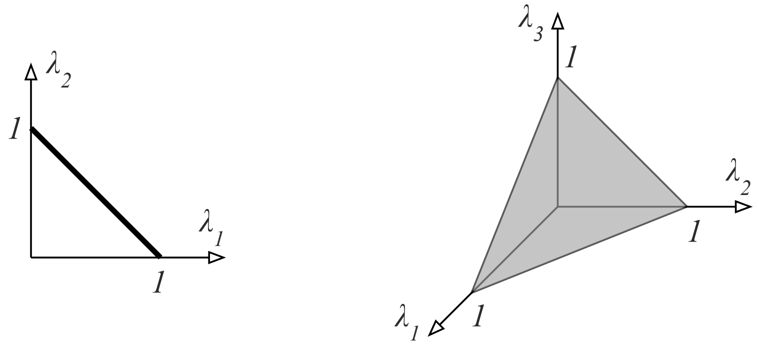

Figure 1:. The 1-simplex (left) and the 2-simplex (right).

The Categorical Distribution

When θ is defined on a finite universe with no further constraints, its distribution is a categorical distribution. This is always the case, as the categorical is the most general distribution on finite universes, although it may be possible to parameterize this distribution more succinctly (for example, with a binomial or other finite distribution). If Z is a discrete random variable over K values, the categorical distribution is

$$p(z) = \lambda_z = \mathbb{P}[Z = z] \quad \text{for} \quad z = 1, \ldots, K \ .$$

Since $p(z)$ is a distribution, we have

$$\lambda_k \geq 0 \quad \text{for} \quad l = 1, \ldots, K \quad \text{and} \quad \sum_{k=1}^{K} \lambda_k = 1 \ ,$$

This is a distribution with k parameters, which can be collected into a vector parameter $\boldsymbol{\lambda} = \left[\begin{array}{ccc} \lambda_1, \ldots, \lambda_K \end{array} \right]^T$. The constraints on $\boldsymbol{\lambda}$ define what is called the *unit* $(K{-}1)$*-simplex*. Figure 1 shows the 1-simplex and the 2-simplex.

The Dirichlet Distribution

One possible probability measure on the unit $(K-1)$-simplex is the *Dirichlet distribution* with parameter $\boldsymbol{\alpha}$, a vector with K strictly positive entries:

$$p(\boldsymbol{\lambda}|\boldsymbol{\alpha}) = c(\boldsymbol{\alpha}) \prod_{k=1}^{K} \lambda_k^{\alpha_k - 1} \quad \text{with} \quad \alpha_k > 0 \quad \text{for} \quad k = 1, \ldots, K \ ,$$

where

$$c(\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \ . \tag{5}$$

In this expression, $\Gamma(t)$ is the so-called *gamma function*:

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} \, dx \ . \tag{6}$$

When t is an integer n, integration by parts shows that this formula simplifies to a factorial, $\Gamma(n) = (n-1)!$. However, the entries of $\boldsymbol{\alpha}$ need not be integers. The property

$$\Gamma(n+1) = n\Gamma(n)$$

for integer arguments follows immediately from the definition of factorial. This property also holds for real-valued t, as we now show.

From the definition (6) of the gamma function we obtain

$$\Gamma(t+1) = \int_0^\infty x^t e^{-x} \, dx$$

and integration by parts yields

$$\Gamma(t+1) = \int_0^\infty x^t e^{-x} \, dx = -x^t e^{-x} \big|_{x=0}^{\infty} + \int t x^{t-1} e^{-x} \, dx = t \int x^{t-1} e^{-x} \, dx$$

that is,

$$\Gamma(t+1) = t\,\Gamma(t) \ . \tag{7}$$

Categorical, Dirichlet, and Conjugacy

We now show that

- The Dirichlet distribution is the conjugate prior of the categorical distribution
- Bayes training of the Dirichlet parameter is greatly simplified by conjugacy: The posterior of $\boldsymbol{\lambda}$ given the training data is still Dirichlet.
- The Bayes predictive distribution for the categorical distribution given a Dirichlet parameter and the training data is a categorical distribution whose parameter vector $\boldsymbol{\lambda}'$ can be easily computed thanks to conjugacy.

Dirichlet is Conjugate to Categorical

Let

$$p(z|\boldsymbol{\lambda}) = \lambda_z$$

be a categorical distribution over K values with parameter

$$\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_K]^T$$

on the unit simplex:

$$\lambda_k \geq 0 \quad \text{for} \quad k = 1, \ldots, K \quad \text{and} \quad \lambda_1 + \ldots + \lambda_K = 1 \;,$$

and let $\boldsymbol{\lambda}$ have prior Dirichlet distribution with hyper-parameter $\boldsymbol{\alpha}$. So $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$ play the roles of θ and π in the previous sections. Conjugacy is immediate:

$$\begin{aligned}
\mathrm{Cat}_z[\boldsymbol{\lambda}]\,\mathrm{Dir}_{\boldsymbol{\lambda}}[\boldsymbol{\alpha}] &= p(z|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\boldsymbol{\alpha}) \\
&= \lambda_z c(\boldsymbol{\alpha}) \prod_{k=1}^{K} \lambda_k^{\alpha_k - 1} = c(\boldsymbol{\alpha}) \prod_{k=1}^{K} \lambda_k^{\tilde{\alpha}_k - 1} \\
&= \frac{c(\boldsymbol{\alpha})}{c(\tilde{\boldsymbol{\alpha}})} c(\tilde{\boldsymbol{\alpha}}) \prod_{k=1}^{K} \lambda_k^{\tilde{\alpha}_k - 1} = \frac{c(\boldsymbol{\alpha})}{c(\tilde{\boldsymbol{\alpha}})} \mathrm{Dir}_{\boldsymbol{\lambda}}[\tilde{\boldsymbol{\alpha}}]
\end{aligned}$$

where

$$\tilde{\boldsymbol{\alpha}} = [\tilde{\alpha}_1, \ldots, \tilde{\alpha}_K] \quad \text{with} \quad \tilde{\alpha}_k = \alpha_k + \delta_{kz} \;.$$

In this expression, δ_{kz} is the Kronecker delta:

$$\delta_{kz} = \begin{cases} 1 & \text{for } k = z \\ 0 & \text{otherwise} \end{cases} \;.$$

The Posterior of $\boldsymbol{\lambda}$ is Dirichlet

The posterior distribution on $\boldsymbol{\lambda}$ after seeing training samples $z_1, \ldots, z_I$ is

$$p(\boldsymbol{\lambda}|z_1, \ldots, z_I, \boldsymbol{\alpha}) = \frac{p(\boldsymbol{\lambda}|\boldsymbol{\alpha}) \prod_{i=1}^{I} p(z_i|\boldsymbol{\lambda})}{p(z_1, \ldots, z_I, \boldsymbol{\alpha})} \; .$$

and applying conjugacy I times yields

$$\begin{aligned} p(\boldsymbol{\lambda}|\boldsymbol{\alpha}) \prod_{i=1}^{I} p(z_i|\boldsymbol{\lambda}) &= \mathrm{Dir}_{\boldsymbol{\lambda}}[\boldsymbol{\alpha}] \prod_{i=1}^{I} \lambda_{z_i} \\ &= \mathrm{Dir}_{\boldsymbol{\lambda}}[\boldsymbol{\alpha}] \prod_{k=1}^{K} \lambda_k^{\sum_{i=1}^{I} \delta_{kz_i}} \\ &= \frac{c(\boldsymbol{\alpha})}{c(\boldsymbol{\alpha}')} \mathrm{Dir}_{\boldsymbol{\lambda}}[\boldsymbol{\alpha}'] \end{aligned}$$

where

$$\boldsymbol{\alpha}' = [\alpha'_1, \ldots, \alpha'_K]^T \quad \text{with} \quad \alpha'_k = \alpha_k + \sum_{i=1}^{I} \delta_{kz_i} = \alpha_k + h(k \,|\, z_1, \ldots, z_I) \; . \qquad (8)$$

In this expression,

$$h(k \,|\, z_1, \ldots, z_I) = \sum_{i=1}^{I} \delta_{kz_i} \quad \text{for} \quad k = 1, \ldots, K \qquad (9)$$

is the *histogram* of the data, that is, the count of data points z_i that are equal to k.

The *evidence*, that is, the denominator in the posterior distribution on $\boldsymbol{\lambda}$ is then

$$\begin{aligned} p(z_1, \ldots, z_I, \boldsymbol{\alpha}) &= \int p(\boldsymbol{\lambda}|\boldsymbol{\alpha}) \prod_{i=1}^{I} p(z_i|\boldsymbol{\lambda}) \, d\boldsymbol{\lambda} \\ &= \int \frac{c(\boldsymbol{\alpha})}{c(\boldsymbol{\alpha}')} \mathrm{Dir}_{\boldsymbol{\lambda}}[\boldsymbol{\alpha}'] \, d\boldsymbol{\lambda} \\ &= \frac{c(\boldsymbol{\alpha})}{c(\boldsymbol{\alpha}')} \int \mathrm{Dir}_{\boldsymbol{\lambda}}[\boldsymbol{\alpha}'] \, d\boldsymbol{\lambda} \\ &= \frac{c(\boldsymbol{\alpha})}{c(\boldsymbol{\alpha}')} \end{aligned}$$

thanks to conjugacy again, so that the posterior on $\boldsymbol{\lambda}$ is simply

$$p(\boldsymbol{\lambda}|z_1,\ldots,z_I,\boldsymbol{\alpha}) = \mathrm{Dir}_{\boldsymbol{\lambda}}[\boldsymbol{\alpha}']$$

(of course, the constant $c(\boldsymbol{\alpha})/c(\boldsymbol{\alpha}')$ *had* to cancel out, because $p(\boldsymbol{\lambda}|z_1,\ldots,z_I,\boldsymbol{\alpha})$ is a probability distribution, so the last manipulation is merely a confirmation).

The Predictive Distribution

The Bayesian predictive distribution of z is

$$\begin{aligned}
p(z|z_1,\ldots,z_I,\boldsymbol{\alpha}) &= \int p(z|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|z_1,\ldots,z_I,\boldsymbol{\alpha})\,d\boldsymbol{\lambda} \\
&= \int \mathrm{Cat}_z[\boldsymbol{\lambda}]\,\mathrm{Dir}_{\boldsymbol{\lambda}}[\boldsymbol{\alpha}']\,d\boldsymbol{\lambda} \\
&= \int \lambda_z\,\mathrm{Dir}_{\boldsymbol{\lambda}}[\boldsymbol{\alpha}']\,d\boldsymbol{\lambda} \\
&= \int c(\boldsymbol{\alpha}')\prod_{k=1}^{K}\lambda_k^{\alpha'_k+\delta_{kz}}\,d\boldsymbol{\lambda} \\
&= \frac{c(\boldsymbol{\alpha}')}{c(\boldsymbol{\alpha}'')}\int \mathrm{Dir}_{\boldsymbol{\lambda}}[\boldsymbol{\alpha}'']\,d\boldsymbol{\lambda} \\
&= \frac{c(\boldsymbol{\alpha}')}{c(\boldsymbol{\alpha}'')}\ ,
\end{aligned}$$

where conjugacy came to the rescue once again, and where

$$\boldsymbol{\alpha}'' = [\alpha''_1,\ldots,\alpha''_K]^T \quad \text{with} \quad \alpha''_k = \alpha'_k+\delta_{kz} = \alpha_k+\delta_{kz}+h(k\,|\,z_1,\ldots,z_I)\ . \tag{10}$$

Since z is a discrete variable over K values, its distribution is categorical. So if

$$\boldsymbol{\lambda}' = [\lambda'_1,\ldots,\lambda'_K]^T$$

is its parameter, then

$$p(z|z_1,\ldots,z_I,\boldsymbol{\alpha}) = \lambda'_z = \frac{c(\boldsymbol{\alpha}')}{c(\boldsymbol{\alpha}'')}\ .$$

We now rewrite the last fraction to determine the values in $\boldsymbol{\lambda}'$ and to verify that they satisfy the unit-simplex constraints

$$\lambda'_z \geq 0 \quad \text{for} \quad z = 1,\ldots,K \quad \text{and} \quad \lambda'_1+\ldots+\lambda'_K = 1\ .$$

From the definition (10) of $\boldsymbol{\alpha}''$ we see that

$$\alpha''_z = 1 + \alpha'_z \quad \text{and} \quad \alpha''_k = \alpha'_k \quad \text{for} \quad k \neq z \tag{11}$$

and therefore

$$\sum_{k=1}^{K} \alpha''_k = 1 + \sum_{k=1}^{K} \alpha'_k$$

so that equation (7) with $t = \sum_{k=1}^{K} \alpha'_k$ yields

$$\Gamma\left(\sum_{k=1}^{K} \alpha''_k\right) = \left(\sum_{k=1}^{K} \alpha'_k\right) \, \Gamma\left(\sum_{k=1}^{K} \alpha'_k\right) \; .$$

Therefore,

$$\begin{aligned}
\lambda'_z &= \frac{c(\boldsymbol{\alpha}')}{c(\boldsymbol{\alpha}'')} = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha'_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha'_k)} \, \frac{\prod_{k=1}^{K} \Gamma(\alpha''_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha''_k\right)} \\
&= \frac{\prod_{k=1}^{K} \Gamma(\alpha''_k)}{\prod_{k=1}^{K} \Gamma(\alpha'_k)} \, \frac{1}{\sum_{k=1}^{K} \alpha'_k} \\
&= \frac{\alpha'_z}{\sum_{k=1}^{K} \alpha'_k}
\end{aligned}$$

where the last passage follows from (11).

This expression shows that the values λ'_z in $\boldsymbol{\lambda}$ are properly normalized:

$$\lambda'_1 + \ldots + \lambda'_K = 1 \; .$$

From the definition (8) of α'_k we obtain the desired result:

The Bayesian predictive distribution of a categorical distribution with Dirichlet prior on its parameter $\boldsymbol{\lambda}$ and hyper-parameter vector $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$ after seeing data $z_1, \ldots, z_I$ with histogram

$$h(k \,|\, z_1, \ldots, z_I) = \sum_{i=1}^{I} \delta_{kz_i} \quad \text{for} \quad k = 1, \ldots, K$$

is a categorical distribution whose vector $\boldsymbol{\lambda}'$ of posterior parameters has entries defined as follows:

$$\lambda'_z = \frac{\alpha_z + h(z \,|\, z_1, \ldots, z_I)}{\sum_{k=1}^{K} [\alpha_k + h(k \,|\, z_1, \ldots, z_I)]} \quad \text{for} \quad z = 1, \ldots, K \, .$$

Since the α_ks are positive real numbers and the histograms are integer counts, the entries of $\boldsymbol{\lambda}'$ are nonnegative.

The result above is also given (without proof) in equation (4.35) of the textbook. Comparison with equations (4.30) in the textbook shows that the ML point estimate

$$\hat{\lambda}_z^{(ML)} = \frac{h(z \,|\, z_1, \ldots, z_I)}{\sum_{k=1}^{K} h(k \,|\, z_1, \ldots, z_I)}$$

of λ_z yields the same predictive distribution for z as the Bayesian estimation approach with the uninformative prior $\boldsymbol{\alpha} = [0, \ldots, 0]$. Comparison with equation (4.32) in the textbook shows that the MAP point estimate

$$\hat{\lambda}_z^{(MAP)} = \frac{\alpha_z + h(z \,|\, z_1, \ldots, z_I) - 1}{\sum_{k=1}^{K} [\alpha_k + h(k \,|\, z_1, \ldots, z_I) - 1]}$$

of λ_z yields a less “diffuse” predictive distribution than the Bayesian estimate, in the sense that the MAP parameters can be obtained from the Bayes ones by de-normalizing (that is, multiply then by the sum in their common denominator), subtracting 1 from each parameter, and re-normalizing. This transformation increases the ratios between larger and smaller values, because

$$v > u > 1 \quad \Rightarrow \quad \frac{v-1}{u-1} > \frac{v}{u} \, .$$

The ratios between larger and smaller values of the categorical parameters are even greater for the ML estimate, because the α_z are replaced by zeros.

In particular, since the α_z are strictly positive, the Bayesian estimate $\boldsymbol{\lambda}'$ cannot have any zero entries, while the MAP estimate can, and even more so can the ML estimate. In this sense, the ML estimate is more “peaked” than the MAP estimate and the MAP estimate is more “peaked” than the Bayes estimate.