Binomial and multinomial distributions

Kevin P. Murphy

Last updated October 24, 2006

* Denotes more advanced sections

1 Introduction

In this chapter, we study probability distributions that are suitable for modelling discrete data, like letters and words. This will be useful later when we consider such tasks as classifying and clustering documents, recognizing and segmenting languages and DNA sequences, data compression, etc. Formally, we will be concerned with density models for $X \in \{1, \ldots, K\}$, where K is the number of possible values for X; we can think of this as the number of symbols/ letters in our alphabet/ language. We assume that K is known, and that the values of X are unordered: this is called **categorical** data, as opposed to **ordinal** data, in which the discrete states can be ranked (e.g., low, medium and high). If $K = 2$ (e.g., X represents heads or tails), we will use a **binomial** distribution. If $K > 2$, we will use a **multinomial** distribution.

2 Bernoullis and Binomials

Let $X \in \{0, 1\}$ be a binary random variable (e.g., a coin toss). Suppose $p(X = 1) = \theta$. Then

$$p(X|\theta) = \text{Be}(X|\theta) = \theta^X (1-\theta)^{1-X} \tag{1}$$

is called a **Bernoulli** distribution. It is easy to show that

$$E[X] = p(X = 1) = \theta \tag{2}$$

$$\text{Var}[X] = \theta(1-\theta) \tag{3}$$

The likelihood for a *sequence* $\mathcal{D} = (x_1, \ldots, x_N)$ of coin tosses is

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} \theta^{x_n} (1-\theta)^{1-x_n} = \theta^{N_1} (1-\theta)^{N_0} \tag{4}$$

where $N_1 = \sum_{n=1}^{N} x_n$ is the number of heads ($X = 1$) and $N_0 = \sum_{n=1}^{N} (1 - x_n) = N - N_1$ is the number of tails ($X = 0$).

The likelihood of a given *number* of heads N_1 out of N trials is

$$p(N_1|N) = \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N-N_1} = Bi(\theta, N) \tag{5}$$

where

$$\binom{N}{N_1} = \frac{N!}{(N-N_1)!N_1!} \tag{6}$$

is the number of ways to choose N_1 items from N. This is called a **Binomial** distribution. Note that it defines a distribution on counts, not on sequences.

To understand the difference between counts and sequences, it is helpful to consider the following analogy [Min03]: the sufficient statistics of N samples from a univariate Gaussian are the sample mean and sample variance; all samples with the same statistics have the same probability; however, the probability *of those statistics* is not the same as the probability of the samples that generated the statistics. (For one thing, the statistics define a 2D joint distribution.) Since data is usually samples, not counts, we will use the Bernoulli rather than the binomial.

2.1 Maximum likelihood parameter estimation

In this section, we discuss one popular approach to estimating the parameters of a probability density function. Imagine that we have N samples x_n drawn independently from the same distribution, $x_n \sim p(x|\theta)$; this is called an **iid (independent, identical distributed)** sample, which we will call D (since it represents our "training" data). One intutively appealing approach to parameter estimation is to find the parameter setting that makes the data as likely as possible:

$$\hat{\theta}^{MLE} = \arg\max_{\theta} p(D|\theta) \tag{7}$$

where $p(\mathcal{D}|\theta)$ is called the **likelihood of the parameters given the data**; it is viewed as a function of *theta*, since $\mathcal{D}$ is fixed, so is often written $L(theta)$. $\hat{\theta}^{MLE}$ is called a **maximum likelihood estimate (mle)**. There are various theoretical reasons why this is a reasonable thing to do, which we will discuss later.

Since the data $\mathcal{D} = \{x_1, \ldots, x_N\}$, is iid, the likelihood factorizes

$$L(\theta) = \prod_{n=1}^{N} p(x_n|\theta) \tag{8}$$

It is often more convenient to work with log probabilities; this will not change $\arg\max L(\theta)$, since log is a monotonic function. Hence we define the **log likelihood** as $\ell(\theta) = \log p(\mathcal{D}|\theta)$. For iid data this becomes

$$\ell(\theta) = \sum_{n=1}^{N} \log p(x_n|\theta) \tag{9}$$

The mle then maximizes $\ell(\theta)$.

Now consider the case of MLE for the Bernoulli distirbution. Let $X \in \{0, 1\}$. Given $\mathcal{D} = (x_1, \ldots, x_N)$, the likelihood is

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} p(x_n|\theta) \tag{10}$$

$$= \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{1-x_n} \tag{11}$$

$$= \theta^{N_1}(1-\theta)^{N_0} \tag{12}$$

where $N_1 = \sum_n x_n$ is the number of heads and $N_0 = \sum_n (1 - x_n)$ is the number of tails (these counts are the **sufficient statistics**). The log-likelihood is

$$\ell(\theta) = \log p(\mathcal{D}|\theta) = N_1 \log\theta + N_0 \log(1-\theta) \tag{13}$$

Solving for $\frac{d\ell}{d\theta} = 0$ yields

$$\hat{\theta}^{ML} = \frac{N_1}{N} \tag{14}$$

the empirical fraction of heads.

Suppose we have seen 3 tails out of 3 trials. Then we predict that the probability of heads is zero:

$$\hat{\theta}^{ML} = \frac{N_1}{N_1 + N_0} = \frac{0}{0+3} \tag{15}$$

This is an example of **overfitting** and is a result of using maximum likelihood estimation. In this context, this problem is called the **sparse data problem**: if we fail to see something in the training set, we predict that it can never happen in the future, which seems a little extreme. To consider another example, suppose we have seen 3 white swans and 0 black swans; can we infer all swans are white? No! On visiting Australia, we may encounter a black swan. This is called the **black swan paradox**, and is an example of the famous **problem of induction** in philosophy. Below we will see how a Bayesian approach solves this problem.

2.2 Bayesian parameter estimation

In the Bayesian approach to statistics all uncertainty (including uncertainty about parameter values) is modelled with probability theory. Our belief about the value of the parameters θ after having seen data D can be computed using Bayes rule:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{16}$$

Here $p(D|\theta)$ is called the **likelihood of data given the parameters**, $p(\theta)$ is our **prior belief** and $p(\theta|D)$ is our **posterior belief**. The normalization constant

$$p(D) = \int p(D, \theta) d\theta \tag{17}$$

is called the **marginal likelihood** or **evidence**; this quantity depends on the model that we are using. To make this explicit (when performing model comparison), we write $p(D|M)$ etc. (All our probabilities are implicitly conditioned on M, but we drop it from the notation except where necessary.)

The appropriate prior $p(\theta)$ to use depends on the nature of the model, and the nature of your prior knowledge. Somtetimes we choose a **conjugate prior**, which is a mathematically convenient prior we will define below. Othertimes we choose a **non-informative/ objecive/ reference prior**, to let the data "speak for itself"; often we can approximatge a non-informative prior using a conjugate prior. We will see examples below.

2.3 Conjugate prior

Since the probability of heads, θ, is uncertain, let us put a ("second order") probability distribution on this probability. A Bayesian estimate of θ requires a prior $p(\theta)$. It is common to use a **conjugate prior**, to simplify the math. A prior is called **conjugate** if, when multiplied by the likelihood $p(D|\theta)$, the resulting posterior is in the same parametric family as the prior. (We say the model is **closed under Bayesian updating**.) Since the Bernoulli likelihood has the form

$$p(D|\theta) \propto [\theta^{N_1}(1-\theta)^{N_0}] \tag{18}$$

we seek a prior of the form

$$p(\theta) = \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1} \tag{19}$$

since then the posterior can be gotten by adding the exponents:

$$\begin{aligned} p(\theta|D) &\propto p(D|\theta)p(\theta) && (20)\\ &\propto [\theta^{N_1}(1-\theta)^{N_0}][\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}] && (21)\\ &= \theta^{N_1+\alpha_1-1}(1-\theta)^{N_0+\alpha_0-1} && (22) \end{aligned}$$

α_1	α_0	$E\theta$	mode θ	Var θ
1/2	1/2	1/2	NA	∞
1	1	1/2	NA	0.25
2	2	1/2	1/2	0.08
10	10	1/2	1/2	0.017

Table 1: The mean, mode and variance of various beta distributions. As the strength of the prior, $\alpha_0 = \alpha_1 + \alpha_0$, increases, the variance decreases. Note that the mode is not defined if $\alpha_0 \leq 2$: see Figure 1 for why.

where N_1 is the number of heads and N_0 is the number of tails. α_1, α_0 are **hyperparameters** (parameters of the prior) and correspond to the number of "virtual" heads/tails (**pseudo counts**). $\alpha = \alpha_1 + \alpha_0$ is called the **effective sample size** (strength) of the prior. This prior is called a **beta prior**:

$$p(\theta|\alpha_1, \alpha_0) = \text{Beta}(\theta|\alpha_1, \alpha_0) = \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)}\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1} \tag{23}$$

where the gamma function is defined as

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du \tag{24}$$

Note that $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(1) = 1$. Also, for integers, $\Gamma(x+1) = x!$. Note that **Stirling's approximation** gives

$$\log\Gamma(x+1) \approx x\log x - x \tag{25}$$

Also, the **beta function** is define as

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{26}$$

The normalization constant

$$1/Z(\alpha_1, \alpha_0) = \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \tag{27}$$

ensures

$$\int_0^1 \text{Beta}(\theta|\alpha_1, \alpha_0)d\theta = 1 \tag{28}$$

If $\theta \sim Beta(\alpha_1, \alpha_0)$, then we have the following properties

$$\text{mean} = \frac{\alpha_1}{\alpha_1 + \alpha_0} \tag{29}$$

$$\text{mode} = \frac{\alpha_1 - 1}{\alpha_1 + \alpha_0 - 2} \tag{30}$$

$$\text{Var} = \frac{\alpha_1\alpha_0}{(\alpha_1 + \alpha_0)^2(\alpha_1 + \alpha_0 - 1)} \tag{31}$$

See Figure 1 for some plots of some beta distributions, and Table 1 for some typical values.

$\alpha_1 = \alpha_0 = 1$ is a uniform prior, also called the **Laplace prior**, and is often considered an uniformative prior, since then $\hat{\theta}^{MAP} = \hat{\theta}^{MLE}$ (see below). However, $\alpha_1 = \alpha_0 = 1/2$ is also an uniformative prior (called **Jeffrey's prior**), since the prior variance is infinite.

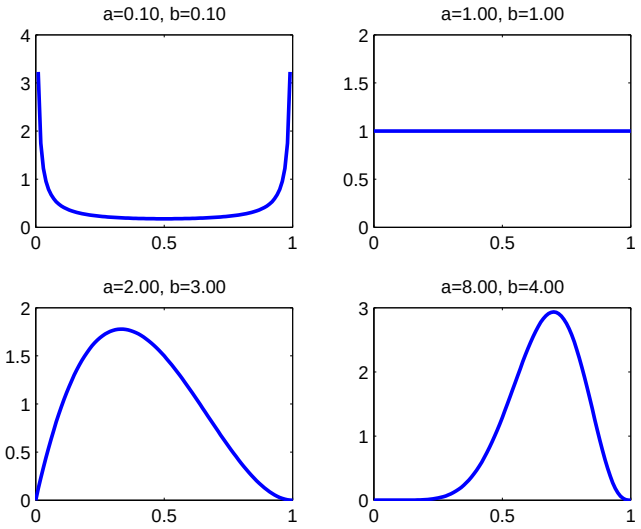

Figure 1: Some beta $Be(a, b)$ distributions.

2.4 Updating a beta distribution

If we start with a beta prior $Beta(\theta|\alpha_1, \alpha_0)$ and see N_1 heads and N_0 tails, we end up with a beta posterior $Beta(\theta|\alpha_1 + N_1, \alpha_0 + N_0)$:

$$\begin{aligned} p(\theta|D) &= \frac{1}{p(D)} p(D|\theta) p(\theta|\alpha_1, \alpha_0) && (32) \\ &= \frac{1}{p(D)} [\theta^{N_1}(1-\theta)^{N_0}] \frac{1}{Z(\alpha_1, \alpha_0)} [\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}] && (33) \\ &= Beta(\theta|N_1 + \alpha_1, N_0 + \alpha_0) && (34) \end{aligned}$$

For example, suppose we start with $Beta(\theta|\alpha_1 = 2, \alpha_0 = 2)$ and observe $x = 1$, so $N_1 = 1, N_0 = 0$; then the posterior is $Beta(\theta|\alpha_1 = 3, \alpha_0 = 2)$. So the mean shifts from $E[\theta] = 2/4$ to $E[\theta|D] = 3/5$. We can plot the prior and posterior, as in Figure 2.

We can also perform the updating sequentially. Suppose we start with beta prior $Beta(\theta|\alpha_1, \alpha_0)$. If we observe N trials with N_1 heads and N_0 tails, then the posterior becomes

$$p(\theta|\alpha_1, \alpha_0, N_1, N_0) = Beta(\theta; \alpha_1 + N_1, \alpha_0 + N_0) = Beta(\theta; \alpha_1', \alpha_0') \quad (35)$$

where α_1', α_0' are the parameters of the new prior. If we observe another N' trials with N_1' heads and N_0' tails, then the posterior becomes

$$\begin{aligned} p(\theta|\alpha_1', \alpha_0', N_1', N_0') &= Beta(\theta; \alpha_1' + N_1', \alpha_0' + N_0') && (36) \\ &= Beta(\theta; \alpha_1 + N_1 + N_1', \alpha_0 + N_0 + N_0') && (37) \end{aligned}$$

So we see that sequentially absorbing data in any order is equivalent to batch updating (assuming iid data and exact Bayesian updating). This is useful for **online learning** and for processing large datasets, since we don't need to store the original data.

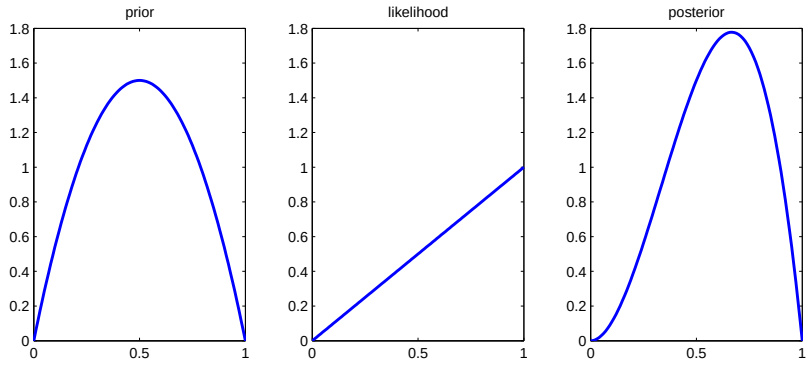

Figure 2: Updating a $Beta(2,2)$ prior with sufficient statistics $N_1 = 1$,$N_0 = 0$ to yield a $Beta(3,2)$ posterior. The mean shifts from 0.5 to 0.6. Note that the likelihood is just a line, since $p(X|\theta) = \theta$.

2.5 Marginal likelihood *

The marginal likelihood, $p(D)$, is the normalization constant in the denominator of Bayes rule:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)} \tag{38}$$

Hence

$$p(D) = \int p(\theta)p(D|\theta)d\theta \tag{39}$$

If we are only interested in parameter estimation, we do not need to compute this constant, but it is necessary for choosing between different models, as we will see later.

Since the posterior is normalized,

$$\frac{1}{Z(\alpha_1 + N_1, \alpha_0 + N_0)} = \frac{1}{P(D)} \frac{1}{Z(\alpha_1, \alpha_0)} \tag{40}$$

Hence the marginal likelihood is the ratio of the normalizing constants:

$$p(D) = \frac{Z(\alpha_1 + N_1, \alpha_0 + N_0)}{Z(\alpha_1, \alpha_0)} \tag{41}$$

$$= \frac{\Gamma(\alpha_1 + N_1)\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_1 + \alpha_0 + N)} \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \tag{42}$$

Here is an alternative derivation of this fact. By the chain rule of probability,

$$p(x_{1:N}) = p(x_1)p(x_2|x_1)p(x_3|x_{1:2})\ldots \tag{43}$$

Also, as we show in Section 2.6, after N data cases, we have

$$p(X = k|D_{1:N}) = \frac{N_k + \alpha_k}{\sum_i N_i + \alpha_i} \stackrel{\text{def}}{=} \frac{N_k + \alpha_k}{N + \alpha} \tag{44}$$

Now suppose $D = H, T, T, H, H, H$ or $D = 1, 0, 0, 1, 1, 1$. Then

$$p(D) = \frac{\alpha_1}{\alpha} \cdot \frac{\alpha_0}{\alpha + 1} \cdot \frac{\alpha_0 + 1}{\alpha + 2} \cdot \frac{\alpha_1 + 1}{\alpha + 3} \cdot \frac{\alpha_1 + 2}{\alpha + 4} \tag{45}$$

$$= \frac{[\alpha_1(\alpha_1 + 1)(\alpha_1 + 2)]\,[\alpha_0(\alpha_0 + 1)]}{\alpha(\alpha + 1)\cdots(\alpha + 4)} \tag{46}$$

$$= \frac{[(\alpha_1)\cdots(\alpha_1 + N_1 - 1)]\,[(\alpha_0)\cdots(\alpha_0 + N_0 - 1)]}{(\alpha)\cdots(\alpha + N)} \tag{47}$$

For integers,

$$(\alpha)(\alpha+1)\cdots(\alpha+M-1) \tag{48}$$

$$= \frac{(\alpha+M-1)!}{(\alpha-1)!} \tag{49}$$

$$= \frac{(\alpha+M-1)(\alpha+M-2)\cdots(\alpha+M-M)(\alpha+M-M-1)\cdots 2\cdot 1}{(\alpha-1)(\alpha-2)\cdots 2\cdot 1} \tag{50}$$

$$= \frac{(\alpha+M-1)(\alpha+M-2)\cdots(\alpha)(\alpha-1)\cdots 2\cdot 1}{(\alpha-1)(\alpha-2)\cdots 2\cdot 1} \tag{51}$$

For reals, we replace $(\alpha-1)!$ with $\Gamma(\alpha)$. Hence

$$p(D) = \frac{[(\alpha_1)\cdots(\alpha_1+N_1-1)]\,[(\alpha_0)\cdots(\alpha_0+N_0-1)]}{(\alpha)\cdots(\alpha+N)} \tag{52}$$

$$= \frac{\Gamma(\alpha)}{\Gamma(\alpha+N)}\cdot\frac{\Gamma(\alpha_1+N_1)}{\Gamma(\alpha_1)}\cdot\frac{\Gamma(\alpha_0+N_0)}{\Gamma(\alpha_0)} \tag{53}$$

2.6 Posterior predictive distribution

The prior predicted probability of heads can be computed as follows:

$$p(X=1) = \int_0^1 p(X=1|\theta)p(\theta)d\theta \tag{54}$$

$$= \int_0^1 \theta\, p(\theta)d\theta = E[\theta] = \frac{\alpha_1}{\alpha_0+\alpha_1} \tag{55}$$

The posterior predicted probability of heads, given the data, can be computed as follows:

$$p(X=1|D) = \int_0^1 p(X=1|\theta)p(\theta|D)d\theta \tag{56}$$

$$= \int_0^1 \theta\, p(\theta|D)d\theta = E[\theta|D] = \frac{N_1+\alpha_1}{\alpha_0+\alpha_1+N} \tag{57}$$

With a uniform prior $\alpha_1=\alpha_0=1$, we get **Laplace's rule of succession**

$$p(X=1|N_1,N_0) = \frac{N_1+1}{N_1+N_0+2} \tag{58}$$

This avoids the sparse data problem we encountered earlier.

2.7 Predictive distribution for Binomials *

If you lookup the predictive distribution for the binomial distribution with beta prior in a textbook (such as [BS94]), you will find it called the **binomial-beta** distribution, defined as

$$p(x) = \int_0^1 Bi(x|\theta,n)Beta(\theta|\alpha_1,\alpha_0)d\theta \tag{59}$$

$$\stackrel{\text{def}}{=} Bb(x|\alpha_0,\alpha_1,n) \tag{60}$$

$$= \frac{B(x+\alpha_1,n-x+\alpha_0)}{B(\alpha_1,\alpha_0)}\binom{n}{x} \tag{61}$$

where $B(\alpha_1,\alpha_0)$ is the beta function

$$B(\alpha_1,\alpha_0) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_0)}{\Gamma(\alpha_1+\alpha_0)} \tag{62}$$

The posterior predictive is the same, but with updated hyper-parameters: $\alpha_0' = \alpha_0 + N_0$, $\alpha_1' = \alpha_1 + N_1$. The mean and variance of the binomial-beta are given by

$$E[x] = n\frac{\alpha_1}{\alpha_0 + \alpha_1} \tag{63}$$

$$\text{Var}[x] = \frac{n\alpha_0\alpha_1}{(\alpha_0 + \alpha_1)^2}\frac{(\alpha_0 + \alpha_1 + n)}{\alpha_0 + \alpha_1 + 1} \tag{64}$$

Let us verify this gives the same result as the (prior) predictive derived above for the Bernoulli distribution (in which case $n = 1$):

$$p(X = 1) = Bb(1|\alpha_1, \alpha_0, 1) \tag{65}$$

$$= \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)\Gamma(\alpha_1 + \alpha_0 + 1)}\binom{1}{1}\Gamma(\alpha_1 + 1)\Gamma(\alpha_0 + 1 - 1) \tag{66}$$

$$= \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)(\alpha_1 + \alpha_0)\Gamma(\alpha_1 + \alpha_0)}\alpha\Gamma(\alpha_1)\Gamma(\alpha_0) \tag{67}$$

$$= \frac{\alpha_1}{\alpha_1 + \alpha_0} \tag{68}$$

where we used the fact thta $\Gamma(\alpha_0 + \alpha_1 + 1) = (\alpha_0 + \alpha + 1)\Gamma(\alpha_0 + \alpha_1)$.

2.8 Point estimation

The MAP estimate is given by

$$\hat{\theta}^{MAP} = \arg\max_\theta p(D|\theta)p(\theta) \tag{69}$$

$$= \arg\max_\theta Be(\alpha_1 + N_1, \alpha_0 + N_0) \tag{70}$$

$$= \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_0 + N - 2} \tag{71}$$

So if $\alpha_1 = \alpha_0 = 1$, then $\hat{\theta}^{MAP} = \hat{\theta}^{MLE}$. Hence using a $Beta(1, 1)$ prior with a plug-in estimate

$$p(X = 1|\hat{\theta}^{MAP}) = \hat{\theta}^{MAP} \tag{72}$$

will suffer from the sparse data problem. However, if we use the posterior mean

$$\hat{\theta}^{mean} = \int \theta p(\theta|D) \tag{73}$$

$$= \frac{\alpha_1 + N_1}{\alpha_1 + \alpha_0 + N} \tag{74}$$

then we get the same result as the the predictive distribution, which does not suffer from the sparse data problem.

2.9 Effect of prior strength

Let $N = N_1 + N_0$ be number of samples (observations). Let N' be the number of pseudo observations (strength of prior) and define the prior means as fractions of N':

$$\alpha_1 = N'\alpha_1', \;\; \alpha_0 = N'\alpha_2', \tag{75}$$

where

$$0 < \alpha_1', \alpha_0' < 1 \tag{76}$$

$$\alpha_1' + \alpha_0' = 1 \tag{77}$$

Then posterior mean is a **convex combination** of the prior mean and the MLE

$$\begin{aligned} p(X=1|\alpha_1,\alpha_0,N_1,N_0) &= \frac{\alpha_1+N_1}{\alpha_1+N_1+\alpha_0+N_0} && (78)\\ &= \frac{N'\alpha'_1+N_1}{N+N'} && (79)\\ &= \frac{N'}{N+N'}\alpha'_1+\frac{N}{N+N'}\frac{N_1}{N} && (80)\\ &= \lambda\alpha'_1+(1-\lambda)\frac{N_1}{N} && (81)\end{aligned}$$

where $\lambda = N'/(N+N')$.

Suppose we have a uniform prior $\alpha'_1 = \alpha'_0 = 0.5$, and we observe $N_1 = 3$, $N_0 = 7$. If we have a weak prior $N' = 2$, the posterior prediction is

$$p(X=1|\alpha_1=1,\alpha_0=1,N_1=3,N_0=7) = \frac{3+1}{3+1+7+1} = \frac{1}{3} \approx 0.33 \quad (82)$$

If we have a strong prior $N' = 20$, the posterior prediction is prediction:

$$p(X=1|\alpha_1=10,\alpha_0=10,N_1=3,N_0=7) = \frac{3+10}{3+10+7+10} = \frac{13}{30} \approx 0.43 \quad (83)$$

So with a strong prior, new data moves us less far away from the prior mean.

Now imagine we are performing sequential updating of our estimate. The MLE can change dramatically with small sample sizes. The Bayesian estimate changes much more smoothly, depending on the strength of the prior. See Figure 3.

Eventually, if we have enough data, it washes away the prior. For example, if we observe $N_1 = 300$, $N_0 = 700$, then it doesn't matter if our prior strenght is $N' = 2$ or $N' = 20$, since

$$\frac{300+1}{1000+2} \approx \frac{300+10}{1000+20} \approx 0.3 \quad (84)$$

As $N \to \infty$, $p(\theta|D) \to \delta(\theta - \hat{\theta}^{MLE})$, so $E[\theta|D] \to \hat{\theta}^{MLE}$. So the posterior mode converges to the MLE.

2.10 Setting the hyperparameters*

To set the hyper parameters, suppose your prior is that the probability of heads should be about p, and you believe this prior with strength equivalent to about $N = 1$ samples (see below). Then you just solve the following equations for α_1, α_0:

$$\begin{aligned} p &= \frac{\alpha_1}{\alpha_1+\alpha_0} && (85)\\ N &= \alpha_1+\alpha_0 && (86)\end{aligned}$$

So we find the very intuitive result that we set α_1 to the expected number of heads, $\alpha_1 = Np$, and α_0 to the expected number of tails, $\alpha_0 = N - Np$.

If we do not know the values of α_1, α_0, or some equivalent information, we may either put priors on the hyperparameters, or one may adopt an **empirical Bayes** approach, also called **type II maximum likelihood**, in which we optimize the marginal likelihood wrt the hyperparameters:

$$\hat{\vec{\alpha}} = \arg\max_{\vec{\alpha}} p(D|\vec{\alpha}) \quad (87)$$

For details on how to do this, see [Min00].

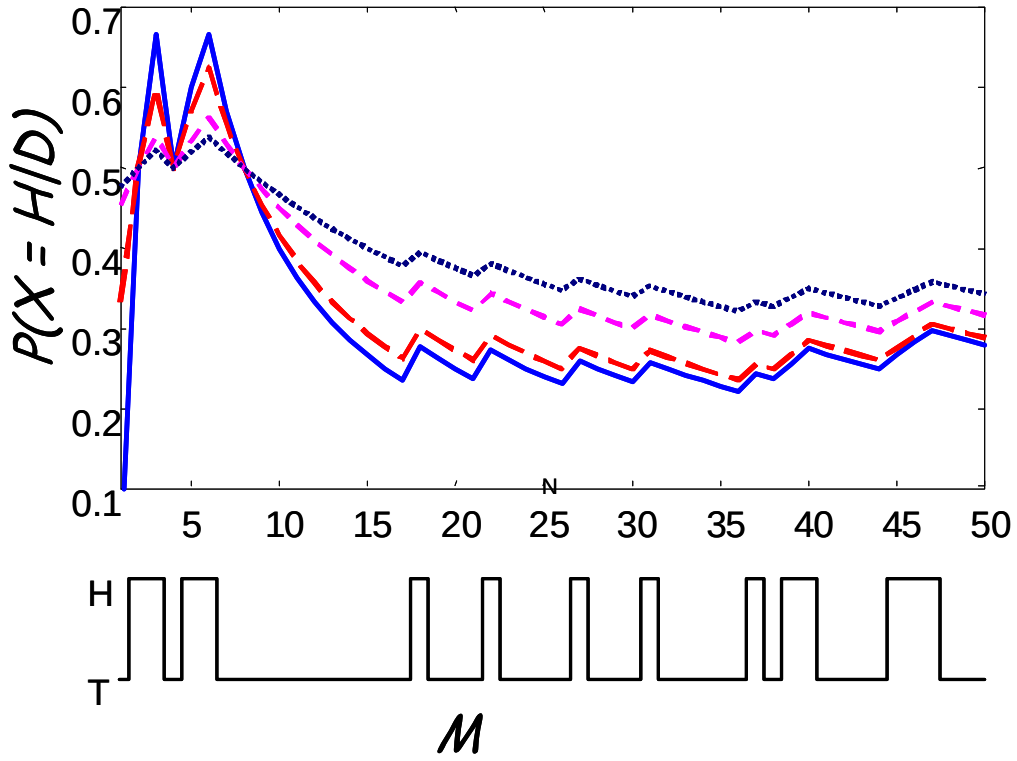

Figure 3: Posterior update as data streams in for different prior strenghts. Lower blue=MLE, red = beta(1,1), pink = beta(5,5), upper blue = beta(10,10).. Source: [KF06].

Prior

$$p(\theta|\alpha_1,\alpha_0) = Be(\theta|\alpha_1,\alpha_0) = \frac{1}{Z(\alpha_1,\alpha_0)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1} \quad (90)$$

$$Z(\alpha_1,\alpha_0) = \frac{\Gamma(\alpha_1+\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \quad (91)$$

Likelihood (where N_0, N_1 are the counts derived from the data D)

$$p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0} \quad (92)$$

Posterior

$$p(\theta|D,\alpha_1,\alpha_0) = \frac{1}{Z(\alpha_1,\alpha_0)}p(D|\alpha_1,\alpha_0)\theta^{\alpha_1+N_1-1}(1-\theta)^{\alpha_0+N_0-1} \quad (93)$$

$$= Be(\alpha_1+N_1,\alpha_0+N_0) = Be(\alpha_1',\alpha_0') \quad (94)$$

Marginal likelihood (evidence):

$$p(D|\alpha_1,\alpha_0) = \frac{Z(\alpha_1',\alpha_0')}{Z(\alpha_1,\alpha_0)} = \frac{\Gamma(\alpha_1')\Gamma(\alpha_0')}{\Gamma(\alpha_1'+\alpha_0')}\frac{\Gamma(\alpha_1+\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \quad (95)$$

Posterior predictive

$$p(X=1|D,\alpha_1,\alpha_0) = \frac{\alpha_1'}{\alpha_1'+\alpha_0'} = \frac{\alpha_1+N_1}{\alpha_1+\alpha_0+N} \quad (96)$$

Figure 4: Summary of results for the Beta-Bernoulli model.

2.11 Non-informative prior

One might think that, since $Beta(1,1)$ is a uniform distribution, that this would be an uniformative prior. But the posterior mean in this case is

$$E[\theta|D] = \frac{N_1+1}{N_1+N_0+2} \quad (88)$$

whereas the MLE is $\frac{N_1}{N_1+N_0}$. Clearly by decreasing the magnitude of the pseudo counts, we we can lessen the impact of the prior. Jeffreys' prior in this case is $\alpha_1 = \alpha_0 = \frac{1}{2}$. However, by the above argument, the most non-informative prior is

$$\lim_{c\longrightarrow 0} Beta(c,c) \quad (89)$$

which is a mixture of two equal point masses at 0 and 1 (see [ZL04] for a proof). This is also called the **Haldane prior**. Note that it is an improper priorm in the sense that $\int Be(\theta|0,0d\theta = \infty$. For a Gaussian, the maximum variance distribution is flattest, but for a Beta (because of its compact support in 0:1), the maximum variance distribution is this mixture of spikes.

2.12 Summary

We can summarize all our results for the Beta-Bernoulli model as shown in Figure 4.

3 Multinomials

Let $X \in \{1, \ldots, K\}$ be a **categorical random variable** with K states, and $p(X = j) = \theta_j$. Then

$$p(X|\theta) = \text{Mu}(X|\theta) = \prod_{j=1}^{K} \theta_j^{I(X=j)} \tag{97}$$

is called a **multinomial** distribution, where $I(x = j) = 1$ if $x = j$ and $I(x = j) = 0$ otherwise. (Sometimes we will use the notation $\delta(x - j)$ instead.) Sometimes we will represent categorical variables as bit vectors, using a **1-of-K encoding**: if $x = j$ then $x_j = 1$ and $x_{j'} = 0$ for $j' \neq j$ (e.g., if $K = 3$, $x = 3$ maps to $(0, 0, 1)$). Then we can write

$$p(x|\theta) = \prod_{j=1}^{K} \theta_j^{x_j} \tag{98}$$

The likelihood for a *sequence* $\mathcal{D} = (x_1, \ldots, x_N)$ of data is

$$p(D|\theta) = \prod_{n=1}^{N} \prod_{j=1}^{K} \theta_j^{I(x_n=j)} = \prod_j \theta_j^{N_j} \tag{99}$$

where $N_j = \sum_n I(x_n = j)$ is the number of times $X = j$. The likelihood of a given number of *counts* $N_1, \ldots, N_K$ out of N trials is

$$p(N_1, \ldots, N_K|N) = Mu(\theta, N) = \binom{N}{N_1 \cdots N_K} \prod_j \theta_j^{N_j} \tag{100}$$

This is *also* called a **multinomial** distribution. Note that it defines a distribution on counts, not on sequences. We shall not need this distribution in this book.

3.1 Maximum likelihood estimation

The log-likelihood is

$$\ell(\theta) \quad = \log p(D|\theta) = \textstyle\sum_k N_k \log \theta_k \tag{101}$$

We need to maximize this subject to the constraint $\sum_k \theta_k = 1$, so we use a **Lagrange multiplier**. The constrained cost function becomes

$$\tilde{\ell} \quad = \quad \sum_k N_k \log \theta_k + \lambda \left(1 - \sum_k \theta_k\right) \tag{102}$$

Taking derivatives wrt θ_k yields

$$\frac{\partial \tilde{\ell}}{\partial \theta_k} \quad = \quad \frac{N_k}{\theta_k} - \lambda = 0 \tag{103}$$

Taking derivatives wrt λ yields the original constraint:

$$\frac{\partial \tilde{\ell}}{\partial \lambda} \quad = \quad \left(1 - \sum_k \theta_k\right) = 0 \tag{104}$$

Using this sum-to-one constraint we have

$$N_k = \lambda\theta_k \tag{105}$$

$$\sum_k N_k = \lambda \sum_k \theta_k \tag{106}$$

$$N = \lambda \tag{107}$$

$$\hat{\theta}_k^{ML} = \frac{N_k}{N} \tag{108}$$

Hence $\hat{\theta}_k^{ML}$ is the fraction of times k occurs. If we did not observe $X = k$ in the training data, we set $\hat{\theta}_k^{ML} = 0$, so we have the same sparse data problem as in the Bernoulli case (in fact it is worse, since K may be large, so it is quite likely that we didn't see some of the symbols, especially if our data set is small).

3.2 Conjugate Bayesian analysis

Let $X \in \{1, \ldots, K\}$ have a multinomial distribution

$$p(X|\theta) = \theta_1^{I(X=1)}\theta_2^{I(X=2)} \cdots \theta_K^{I(X=k)} \tag{109}$$

In other words, $p(X = j|\theta) = \theta_j$. For a set of data $X_{1:N}$, the sufficient statistics are the counts $N_j = \sum_n I(X_n = j)$. The conjugate prior is called **Dirichlet**, and has parameters α

$$p(\theta|\alpha) = \mathcal{D}(\theta|\alpha) = \frac{1}{Z(\alpha)} \cdot \theta_1^{\alpha_1 - 1} \cdot \theta_2^{\alpha_0 - 1} \cdots \theta_K^{\alpha_K - 1} I(\sum_{k=1}^{K} \theta_k - 1) \tag{110}$$

where the I term ensures the parameters sum to one, and $Z(\alpha)$ is the normalizing constant

$$Z(\alpha) = \int \cdots \int \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_K - 1} d\theta_1 \cdots d\theta_K \tag{111}$$

$$= \frac{\prod_{j=1}^{K} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{K} \alpha_j)} \tag{112}$$

See Figure 5 for some examples of Dirichlet distributions.

α_j can be thought of as a **pseudo-count** for state j, and $\alpha \stackrel{\text{def}}{=} \sum_{k=1}^{K} \alpha_k$ is the total strenght of the prior. If $\theta \sim Dir(\theta|\alpha_1, \ldots, \alpha_K)$, then we have these properties

$$E[\theta_k] = \frac{\alpha_k}{\alpha} \tag{113}$$

$$\text{mode}[\theta_k] = \frac{\alpha_k - 1}{\alpha - K} \tag{114}$$

$$\text{Var}\,[\theta_k] = \frac{\alpha_k(\alpha - \alpha_k)}{\alpha^2(\alpha + 1)} \tag{115}$$

Bayesian analysis with a Dirichlet prior is analogous to the beta-bernoulli case: we summarize the results in Figure 6. It is common to use the prior

$$\vec{\alpha} = (\alpha', \ldots, \alpha') \tag{116}$$

where $K\alpha'$ is the prior strength. Often we set $\alpha' = 1/K$ (Jeffrey's prior) or $\alpha' = 1$ (uniform prior).

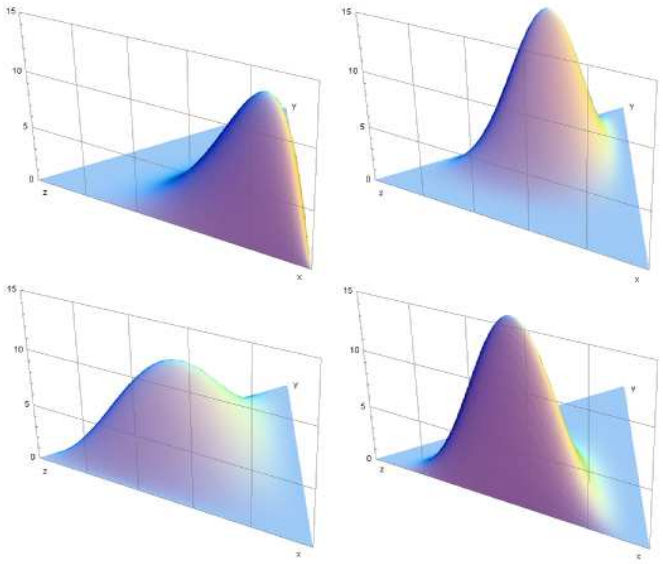

Figure 5: The Dirichlet distribution when $K = 3$ defines a distribution over the 2-simplex, which can be represented by a triangle. Parameter settings, clockwise from top-left: $\alpha = (6, 2, 2)$, $\alpha = (3, 7, 5)$, $\alpha(6, 2, 6)$, $\alpha = (2, 3, 4)$. Source: `http://en.wikipedia.org/wiki/Dirichlet_distribution`.

4 Mixtures of conjugate priors *

In addition, we have assumed conjugate priors for mathematical convenience. A simple way to extend the flexibility of the prior, while maintaining tractability, is to use a **mixture of Dirichlets**, since mixtures of conjugate priors are still conjugate. This has been used in DNA sequence alignment [BHK+93], for example. In this case, the prior is that each location is likely to be a "pure" distribution over A, C, G or T, but we don't know which; we can express this prior as follows:

$$p(\vec{\alpha}) = w_1 Dir(\vec{\alpha}|1, 0, 0, 0) + w_2 Dir(\vec{\alpha}|0, 1, 0, 0) + w_3 Dir(\vec{\alpha}|0, 0, 1, 0) + w_4 Dir(\vec{\alpha}|0, 0, 0, 1) \tag{117}$$

where w_i are the mixing weights. Initially $w_i = \frac{1}{4}$, but we can update the posterior distribution over these weights, when we discover what kind of base is present in this sequence location.

In general, we are free to use any kind of prior we want, but if it is not conjugate, we have to resort to numerical methods to compute the posterior.

References

[BHK+93] M. P. Brown, R. Hughey, A. Krogh, I. S. Mian, K. Sjölander, and D. Haussler. Using dirichlet mixtures priors to derive hidden Markov models for protein families. In *Intl. Conf. on Intelligent Systems for Molecular Biology*, pages 47–55, 1993.

[BS94] J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley, 1994.

[KF06] D. Koller and N. Friedman. *Bayesian networks and beyond*. 2006. To appear.

[Min00] T. Minka. Estimating a Dirichlet distribution. Technical report, MIT, 2000.

[Min03] Tom Minka. Bayesian infernece, entropy and the multinomial distribution. Technical report, CMU, 2003.

Prior

$$p(\theta|\vec{\alpha}) = Dir(\theta|\vec{\alpha}) = \frac{1}{Z(\alpha)} \cdot \theta_1^{\alpha_1 - 1} \cdot \theta_2^{\alpha - 1} \cdots \theta_K^{\alpha_K - 1} \tag{118}$$

$$Z(\vec{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)} \tag{119}$$

Likelihood (where $N_j = \sum_n I(X_n = j)$ are the counts)

$$p(D|\vec{\theta}) = \prod_{j=1}^{K} \theta_j^{N_j} \tag{120}$$

Posterior

$$p(\theta|D, \vec{\alpha}) = \frac{1}{Z(\alpha)p(\vec{N}|\alpha)} \theta_1^{\alpha_1 + N_1 - 1} \cdots \theta_K^{\alpha_K + N_k - 1} \tag{121}$$

$$= Dir(\alpha_1 + N_1, \ldots, \alpha_K + N_K) = Dir(\alpha'_1, \ldots, \alpha'_K) \tag{122}$$

Marginal likelihood (evidence):

$$p(D|\vec{\alpha}) = \frac{Z(\vec{N} + \vec{\alpha})}{Z(\vec{\alpha})} = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N + \sum_k \alpha_k)} \prod_k \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)} \tag{123}$$

Posterior predictive

$$p(x = j|D, \vec{\alpha}) = \frac{\alpha'_j}{\sum_k \alpha'_k} = \frac{\alpha_j + N_j}{N + \sum_k \alpha_k} \tag{124}$$

Figure 6: Summary of results for the Dirichlet-multinomial model.

[ZL04] M. Zhu and A. Lu. The counter-intuitive non-informative prior for the bernoulli family. *J. Statistics Education*, 2004.