Statistic

for machine learning

Tran Trong Khiem

AI lab tranning

2024/05/26

1. Bayesian statistics *
2. The Dirichlet-multinomial model
3. The Gaussian-Gaussian model

Introduction

So far, we have discussed several ways to estimate parameters from data. However, **these approaches ignore any uncertainty in the estimates**, which can be **important** for some applications, such as active learning, or avoiding overfitting, or just knowing how much to trust the estimate of some scientifically meaningful quantity.

In statistics, modeling uncertainty about parameters using a probability distribution (as opposed to just computing a point estimate) is known as inference.

We use the **posterior distribution** to represent our uncertainty. This is the approach adopted in the field of **Bayesian statistics**.

Step 1: We start with a prior distribution $p(\theta)$, which reflects what we know before seeing the data.

Step 2 : We then define a likelihood function $p(D|\theta)$, which reflects the data we expect to see for each setting of the parameters.

Step 3 : We then use Bayes' rule to condition the prior on the observed data to compute the posterior $p(\theta|D)$ as follows:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)} = \frac{p(\theta)p(D|\theta)}{\int p(\theta_0)p(D|\theta_0)d\theta_0}$$

Introduction

$p(D)$ is called the marginal likelihood, since it is computed by marginalizing over (or integrating out) the unknown θ.

This can be interpreted as the average probability of the data, where the average is with respect to the prior.

However, $p(D)$ is a constant, independent of θ, so we will often ignore it when we just want to infer the relative probabilities of θ values.

Once we have computed **the posterior over the parameters**, we can compute **the posterior predictive distribution** over outputs given inputs by marginalizing out the unknown parameters.

$$p(y|x,D) = \int p(y|x,\theta)p(\theta|D)d\theta$$

This can be viewed as a form of **Bayes model averaging (BMA)**, since we are making predictions using an infinite set of models (parameter values), each one weighted by how likely it is.

Conjugate priors

We consider a set of (prior, likelihood) pairs for which we can compute the posterior in closed form.

A prior $p(\theta) \in \mathcal{F}$ is a conjugate prior for a likelihood function $p(D|\theta)$ if the posterior is in the same parameterized family as the prior, i.e., $p(\theta|D) \in \mathcal{F}$.

If the family $\mathcal{F}$ corresponds to the exponential family, then the computations can be performed in closed form.

The beta-binomial model

Suppose we toss a coin N times, and want to infer the probability of heads. Let $y_n = 1$ denote the event that the n-th trial was heads, $y_n = 0$ represent the event that the n-th trial was tails, and let $D = \{y_n : n = 1 : N\}$ be all the data. We assume $y_n \sim \text{Ber}(\theta)$, where $\theta \in [0, 1]$ is the rate parameter (probability of heads).

Bernoulli likelihood

We assume the data are iid :

$$p(D|\theta) = \prod_{n=1}^{N} \theta^{y_n} (1-\theta)^{1-y_n} = \theta^{N_1}(1-\theta)^{N_0}$$

Binomial likelihood

Binomial likelihood model, in which we perform N trials and observe the number of heads, y, rather than observing a sequence of coin tosses

$$p(D|\theta) = \text{Bin}(y|N,\theta) = \binom{N}{y} \theta^{y}(1-\theta)^{N-y}$$

Prior

To simplify the computations, we will assume that the prior $p(\theta) \in \mathcal{F}$ is a conjugate prior for the likelihood function $p(y|\theta)$. This means that the posterior is in the same parameterized family as the prior, i.e., $p(\theta|D) \in \mathcal{F}$.
To ensure this property when using the Bernoulli (or Binomial) likelihood, we should use a prior of the following form:

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \text{Beta}(\theta|\alpha, \beta)$$

We recognize this as the probability density function (pdf) of a beta distribution.

Posterior

If we multiply the Bernoulli likelihood in Equation with the beta prior in Equation we get a beta posterior:

$$p(\theta|D) \propto \theta^{N_1}(1-\theta)^{N_0}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\propto \text{Beta}(\theta|\alpha+N_1, \beta+N_0) = \text{Beta}(\theta|\hat{\alpha}, \hat{\beta})$$

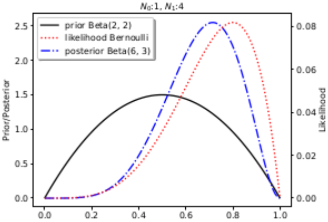
Where $\hat{\alpha} = \alpha + N_1$ and $\hat{\beta} = \beta + N_0$ are posterior parameter.
Since the posterior has the same functional form as the prior, we say that the beta distribution is a conjugate prior for the Bernoulli likelihood.
The strength of the prior is controlled by $N = \alpha + \beta$; this is called the equivalent sample size, since it plays a role analogous to the observed sample size, $N = N_0 + N_1$.

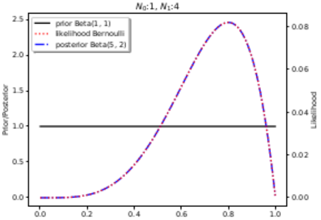
Example

Suppose we set $\alpha = \beta = 2$. This is like saying we believe we have already seen two heads and two tails before we see the actual data; this is a very weak preference for the value of $\lambda = 0.5$.

If we set $\alpha = \beta = 1$ the corresponding prior becomes the uniform distribution:

$$p(\theta) = \text{Beta}(\theta \mid 1, 1) \propto \theta^0 (1-\theta)^0 = \text{Unif}(\theta \mid 0, 1)$$





Posterior mode (MAP estimate)

The most probable value of the parameter is given by the MAP estimate

$$\begin{aligned} \hat{\theta}_{\text{MAP}} &= \arg\max_{\theta} p(\theta|D) \qquad (1) \\ &= \arg\max_{\theta} \log p(\theta|D) \qquad (2) \\ &= \arg\max_{\theta} \left(\log p(\theta) + \log p(D|\theta)\right) \qquad (3) \end{aligned}$$

Using calculus, one can show that this is given by

$$\hat{\theta}_{\text{MAP}} = \frac{\alpha + N_1 - 1}{\alpha + N_1 - 1 + \beta + N_0 - 1}$$

If we use a $\text{Beta}(\theta \mid 2, 2)$ prior, this amounts to add-one smoothing.

$$\begin{aligned} \hat{\theta}_{\text{MAP}} &= \frac{N_1 + 1}{N_1 + 1 + N_0 + 1} \\ &= \frac{N_1 + 1}{N + 2} \end{aligned}$$

Posterior mode (MAP estimate)

If we use a uniform prior, $p(\theta) \propto 1$, the MAP estimate becomes the MLE, since $\log p(\theta) = 0$.

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \log p(D \mid \theta)$$

When we use a Beta prior, the uniform distribution is $\alpha = \beta = 1$. In this case, the MAP estimate reduces to the MLE:

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \frac{N_1}{N_1 + N_0} \\ &= \frac{N_1}{N}\end{aligned}$$

Posterior mean

If $p(\theta|D) = \text{Beta}(\theta|\hat{\alpha}, \hat{\beta})$, then the posterior mean is given by

$$E[\theta|D] = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}} = \frac{\hat{\alpha}}{\hat{N}}$$

where $\hat{N} = \hat{\beta} + \hat{\alpha}$ is the strength (equivalent sample size) of the posterior.
We will now show that the posterior mean is a convex combination of the prior mean, $m = \frac{\alpha}{N}$ (where $N \hat{=} \alpha + \beta$ is the prior strength), and MLE $\theta_{mle} = \frac{N_1}{N}$

$$\begin{aligned} E[\theta|D] &= \frac{\alpha + N_1}{\alpha + \beta + N_1 + N_0} \\ &= \lambda m + (1 - \lambda)\hat{\theta}_{\text{mle}} \end{aligned}$$

Where $\lambda = \frac{N}{\hat{N}}$ is the ratio of the prior to posterior equivalent sample size. The weaker prior , the smaller λ, the closer mean of posterior to mean MLE.

Posterior variance

To capture some notion of uncertainty in our estimate, a common approach is to compute the standard error of our estimate, which is just the posterior standard deviation:

$$\mathrm{se}(\theta) = V[\theta|D]$$

In the case of the Bernoulli model, we showed that the posterior is a beta distribution.

$$\begin{aligned} V[\theta|D] &= \frac{\hat{\alpha}\hat{\beta}}{(\hat{\alpha}+\hat{\beta})^2(\hat{\alpha}+\hat{\beta}+1)} \\ &= E[\theta|D]^2 \frac{\hat{\beta}}{\hat{\alpha}(1+\hat{\alpha}+\hat{\beta})} \end{aligned}$$

Where $\hat{\alpha} = \alpha + N_1$, $\hat{\beta} = \beta + N_0$. If $N >> \alpha + \beta$, this simply to :

$$\begin{aligned} V[\theta|D] &\approx \frac{\hat{\theta}(1-\hat{\theta})}{N} \\ &= \frac{N_1 N_2}{N_3} \end{aligned}$$

Posterior variance

where $\hat{\theta}$ is the MLE. Hence the standard error is given by

$$\sigma = \sqrt{V[\theta|D]} \approx \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{N}} \quad (4.128) \qquad (4)$$

We see that the uncertainty goes down at a rate of $\frac{1}{\sqrt{N}}$. We also see that the uncertainty (variance) is maximized when $\hat{\theta} = 0.5$, and is minimized when $\hat{\theta}$ is close to 0 or 1.

Posterior predictive

Suppose we want to predict future observations.
Step 1 : Compute an estimate of the parameters based on training data, $\hat{\theta}(D)$
Step 2: Plug that parameter back into the model and use $p(y|\hat{\theta})$ to predict the future.
$=>$ this is called a **plug-in approximation**
However, this can result in overfitting.
As an extreme example, suppose we have seen $N = 3$ heads in a row. The MLE is $\hat{\theta} = \frac{3}{3} = 1$. However, if we use this estimate, we would predict that tails are impossible.
Bernoulli model
For the Bernoulli model, the resulting posterior predictive distribution has the form :

$$p(y = 1|D) = \int_0^1 p(y = 1|\theta)p(\theta|D)\,d\theta \quad = \int_0^1 \theta\,\text{Beta}(\theta|\hat{\alpha},\hat{\beta})\,d\theta \tag{5}$$

$$= E[\theta|D] \qquad = \frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}} \tag{6}$$

Bernoulli model

For the Bernoulli model, the resulting posterior predictive distribution has the form :

$$p(y=1|D) = \int_0^1 p(y=1|\theta)p(\theta|D)\,d\theta \quad = \int_0^1 \theta\,\text{Beta}(\theta|\hat{\alpha},\hat{\beta})\,d\theta \qquad (7)$$

$$= E[\theta|D] \quad = \frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}} \qquad (8)$$

In the Bayesian approach, we can get effect ”recover add-one smoothing,” using a uniform prior, $p(\theta) = \text{Beta}(\theta|1,1)$, since the predictive distribution becomes:

$$p(y=1|D) = \frac{N_1+1}{N_1+N_0+2}$$

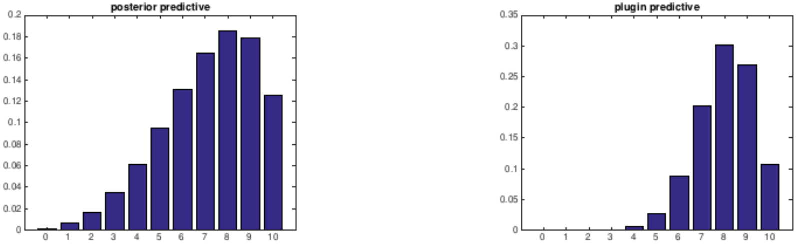
This is known as Laplace’s rule of succession.

Binomial model

Now suppose we were interested in predicting the number of heads in $M > 1$ future coin tossing trials, i.e., we are using the binomial model instead of the Bernoulli model.

$$
\begin{aligned}
p(y|D,M) &= \int_0^1 \mathrm{Bin}(y|M,\theta)\,\mathrm{Beta}(\theta|\alpha,\beta)\,d\theta \\
&= \binom{M}{y}\frac{1}{B(\alpha,\beta)}\int_0^1 \theta^y(1-\theta)^{M-y}\theta^{\alpha-1}(1-\theta)^{\beta-1}\,d\theta \\
&= \binom{M}{y}\frac{1}{B(\alpha,\beta)}\int_0^1 \theta^{y+\alpha-1}(1-\theta)^{M-y+\beta-1}\,d\theta \\
&= \binom{M}{y}\frac{B(y+\alpha,M-y+\beta)}{B(\alpha,\beta)}
\end{aligned}
$$

Binomial model



(a) Posterior predictive distributions for 10 future trials after seeing $N_1 = 4$ heads and $N_0 = 1$ tails.
(b) Plug-in approximation based on the same data. In both cases, we use a uniform prior

Marginal likelihood

The marginal likelihood or evidence for a model M is defined as :

$$p(D|M) = \int p(\theta|M)p(D|\theta,M)\,d\theta$$

When performing inference for the parameters of a specific model, we can ignore this term, since it is constant wrt θ.
1,This quantity plays a vital role when choosing between different models.
2,Useful for estimating the hyperparameters from data (an approach known as empirical Bayes)
In general, computing the marginal likelihood can be hard. However, in the case of the beta- Bernoulli model:
marginal likelihood is $\propto \frac{posteriornormalizer}{priornormalizer}$

Marginal likelihood

The posterior for the beta-binomial model is given by $p(\theta|D) = \text{Beta}(\theta|a_0, b_0)$, where $a_0 = a + N_1$ and $b_0 = b + N_0$.The normalization constant of the posterior is $B(a, b)$.

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} = \frac{1}{p(D)}[\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}][\binom{N}{N_1}\theta^{N_1}(1-\theta)^{N_0}]$$

$$= \binom{N}{N_1}\frac{1}{p(D)}\frac{1}{B(a,b)}\theta^{N_1+a-1}(1-\theta)^{N_0+b-1}$$

So, we have :

$$p(D) = \binom{N}{N_1}\frac{1}{B(a,b)}B(N_1+a, N_0+b)$$

Mixtures of conjugate priors

We can represent a mixture by introducing a latent indicator variable h, where h = k means that θ comes from mixture component k. The prior has the form

$$p(\theta) = \sum_k p(h = k) p(\theta | h = k)$$

Where each $p(\theta|h = k)$ is conjugate, and $p(h = k)$ are called the (prior) mixing weights. The posterior given by :

$$p(\theta|D) = \sum_k p(h = k|D) p(\theta|D, h = k)$$

where $p(h = k|D)$ are the posterior mixing weights given by

$$p(h = k|D) = \frac{p(h = k) p(D|h = k)}{\sum_{k'} p(h = k') p(D|h = k')}$$

The quantity $p(D|h = k)$ is the marginal likelihood for mixture component k.

1. Bayesian statistics *
2. The Dirichlet-multinomial model
3. The Gaussian-Gaussian model

Likelihood

Let $Y \sim \mathrm{Cat}(\theta)$ be a discrete random variable drawn from a categorical distribution. The likelihood has the form

$$p(D|\theta) = \prod_{n=1}^{N} \mathrm{Cat}(y_n|\theta) = \prod_{n=1}^{N}\prod_{c=1}^{C} \theta_c^{I(y_n=c)} = \prod_{c=1}^{C} \theta_c^{N_c},$$

where $N_c = \sum_n I(y_n = c)$.

Prior

The conjugate prior for a categorical distribution is the Dirichlet distribution.This has support over the probability simplex, defined by

$$S_K = \left\{ \theta : 0 \leq \theta_k \leq 1, \sum_{k=1}^{K} \theta_k = 1 \right\}$$

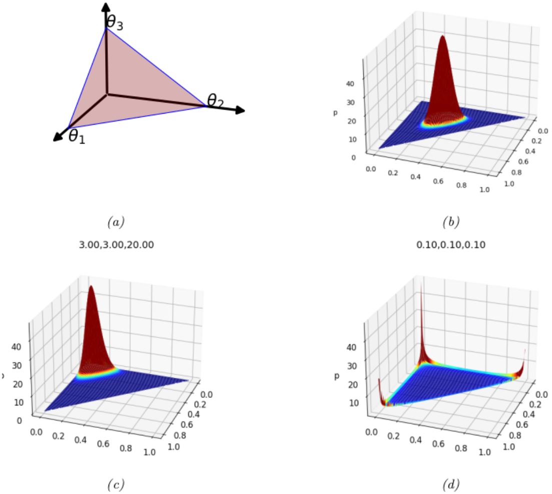
The pdf of the Dirichlet is defined as follows:

$$Dir(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} I(\theta \in S_K)$$

Where $B(\alpha)$ is multivariate beta function :

$$B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}$$

Prior



(a)

(b)

3.00,3.00,20.00

0.10,0.10,0.10

(c)

(d)

(b) Plot of the Dirichlet density for $\alpha = (20, 20, 20)$.
(c) Plot of the Dirichlet density for $\alpha = (3, 3, 20)$.

Posterior

We can combine the multinomial likelihood and Dirichlet prior to compute the posterior

$$p(\theta|D) \propto p(D|\theta)\mathrm{Dir}(\theta|\alpha) = [\prod_k \theta_k^{N_k}][\prod_k \theta_k^{\alpha_k - 1}]$$

$$= \mathrm{Dir}(\theta|\alpha_1 + N_1, \ldots, \alpha_K + N_K) = \mathrm{Dir}(\theta|\hat{\alpha})$$

Where $\hat{\alpha_k} = \alpha_k + N_k$, are the parameters of the posterior.So we see that the posterior can be computed by adding the empirical counts to the prior counts.

The posterior mean is given by :

$$\theta_k = \frac{\hat{\alpha_k}}{\sum_{k'=1}^{K} \hat{\alpha_{k'}}}$$

The posterior mode, which corresponds to the MAP estimate, is given by :

$$\hat{\theta}_k = \frac{\alpha_k - 1}{\sum_{k'}^{K} \alpha_{k'} - 1}$$

Posterior predictive

The posterior predictive distribution is given by :

$$p(y=k|D)=\int p(y=k|\theta)p(\theta|D)d\theta=\int \theta_k p(\theta_k|D)d\theta_k$$

$$=E[\theta_k|D]=\frac{\hat{\alpha_k}}{\sum_{k'=1}^{K}\hat{\alpha_{k'}}}$$

In others words, the posterior predictive distribution is given by

$$p(y|D)=\text{Cat}(y|\bar{\theta})$$

Where $\bar{\theta}\hat{=}E[\theta|D]$ are the posterior mean parameters. If instead we plug-in the MAP estimate, we will suffer from the zero-count problem.
The probability of a single future event, conditioned on past observations $\mathbf{y}=(y_1,\ldots,y_N)$, In some cases, we want to know the probability of observing a batch of future data, denoted by $\dot{\mathbf{y}}=(\dot{y}_1,\ldots,\dot{y}_M)$. We can compute this as follows:

$$p(\dot{\mathbf{y}}|\mathbf{y})=\frac{p(\dot{\mathbf{y}},\mathbf{y})}{p(\mathbf{y})}$$

Marginal likelihood

The marginal likelihood for the Dirichlet-categorical model is given by

$$p(D) = \frac{B(N+\alpha)}{B(\alpha)} \tag{9}$$

where

$$B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)} \tag{10}$$

Hence we can rewrite the above result in the following form, which is what is usually presented in the literature:

$$p(D) = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k N + \sum_k \Gamma(\alpha_k)} \prod_k \frac{\Gamma(N_k + \alpha_k)}{\Gamma\left(\sum_k (\alpha_k)\right)} \tag{11}$$

1. Bayesian statistics *
2. The Dirichlet-multinomial model
3. The Gaussian-Gaussian model

Univariate case

If σ is a known constant, the likelihood for μ has the form

$$p(D|\mu) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \mu)^2\right)$$