Bayesian Statistics

Intelligent Systems and Control

2019

Sepehr Maleki

University of Lincoln
School of Engineering

Bayes' Theorem for Point Probabilities

Let $A_1, ..., A_k$ be a partition of a sample space Ω such that $p(A_i) > 0$ for each i. If $p(B) > 0$, for event B, then:

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{\sum_j p(B|A_j)p(A_j)} \ .$$

- $p(A_i)$ is called the prior probability of A_i.
- $p(B|A_i)$ is called the likelihood of event B, given A_i.
- $p(A_i|B)$ is called the posterior probability of A_i.

(S. Maleki 2019)

Bayes' Theorem Applied to Probability Distributions

Bayesian statistics typically involves using probability distributions rather than point probabilities.

We observe $X_1, ..., X_n \sim F.$ We want to infer (estimate or learn) F.

- In Bayesian statistics the prior uncertainty about model parameters are represented with a probability distribution.
- This prior uncertainty is updated with current data to produce a posterior probability distribution for the parameter that contains less uncertainty.

(S. Maleki 2019)

Bayesian Procedure

- We choose a probability density $p(\theta)$ that describes our beliefs about a parameter θ, before seeing any data (prior).

- We choose a statistical model $p(data|\theta)$ that reflects our beliefs about $data$, given θ (likelihood).

- After observing data $\mathcal{D}_n = \{X_1, \ldots, X_n\}$, we update our beliefs and calculate the posterior distribution $p(\theta|\mathcal{D}_n)$.

$$p(\theta|data) = \frac{p(data|\theta)p(\theta)}{p(data)} \ .$$

Posterior $\propto$ Likelihood $\times$ Prior

(S. Maleki 2019)

Conjugate Priors

In Bayesian probability theory, if the posterior distributions $p(\theta|data)$ are in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function.

Consider the general problem of inferring a distribution for a parameter θ given some $data$.

$$p(\theta|data) = \frac{p(data|\theta)p(\theta)}{p(data)} \ .$$

The likelihood function is usually well-determined from a statement of the data-generating process. It is clear that different choices of the prior distribution $p(\theta)$ may make the denominator, $p(data)$, more or less difficult to calculate.

(S. Maleki 2019)

Conjugate Priors

For certain choices of the prior, the posterior has the same algebraic form as the prior (generally with different parameter values). Consider the hypothesis θ and data x. Then some of the conjugate priors are as follows:

Prior	**Likelihood**	**Posterior**
Beta(α, β)	Binomial(N, θ)	Beta$(\alpha + x, \beta + N - x)$
Beta(α, β)	Bernoulli(θ)	Beta$(\alpha + 1, \beta)$ or Beta$(\alpha, \beta + 1)$
$\mathcal{N}(\mu_{prior}, \sigma^2_{prior})$	$\mathcal{N}(\theta, \sigma^2)$	$\mathcal{N}(\mu_{post}, \sigma^2_{post})$ $\mu_{post} = \frac{\sigma^2 \mu_{prior} + n\bar{x}\sigma^2_{prior}}{\sigma^2 + n\sigma^2_{prior}}$ $\sigma^2_{post} = \frac{\sigma^2 \sigma^2_{prior}}{\sigma^2 + n\sigma^2_{prior}}$

(S. Maleki 2019)

Bayes’ Theorem With Distributions

A Beta Prior-Binomial Likelihood Example

Suppose we’re given a population where a fraction of people are suffering from a specific disease. We draw a sample of size $N = 10$ (select 10 people from the population) to perform some blood tests. The results reveal that only 1 person has the disease.

We would like to know what is the probability that any person from the population has the disease. In fact, if we denote the parameter that governs this probability with θ, we want to find what posterior distribution we can assign to θ after observing our sample of 10 people.

Likelihood: It is easy to see that a binomial distribution is an appropriate choice for likelihood:

$$p(X|\theta) = \binom{10}{1} \theta^1 \, (1-\theta)^9 \, .$$

(S. Maleki 2019)

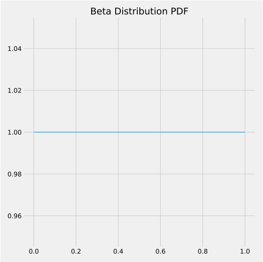
Bayes' Theorem With Distributions

A Beta Prior-Binomial Likelihood Example

Prior: Let's assume that we don't really have any idea as to what percentage of the population is suffering from the disease. In this case a uniform distribution seems to be an appropriate choice to represent our "non-informative" prior.

However, instead, we use a Beta distribution which can be equivalent to a uniform distribution if $Beta(1,1)$. Therefore:

$$p(\theta) \sim Beta(1,1)\ ,$$



(S. Maleki 2019)

Bayes' Theorem With Distributions

A Beta Prior-Binomial Likelihood Example

Posterior $$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \ .$$

One way to find the posterior distribution, i.e., $p(\theta|X)$ is to find the above numerator and then integrate it to find the denominator. However, we don't need to! Since we have a Beta prior, which is conjugate to the binomial distribution for likelihood, our posterior is also a Beta distribution with the following parameters:

$$p(\theta|X) \sim Beta(1+1, 1+10-1) \ .$$

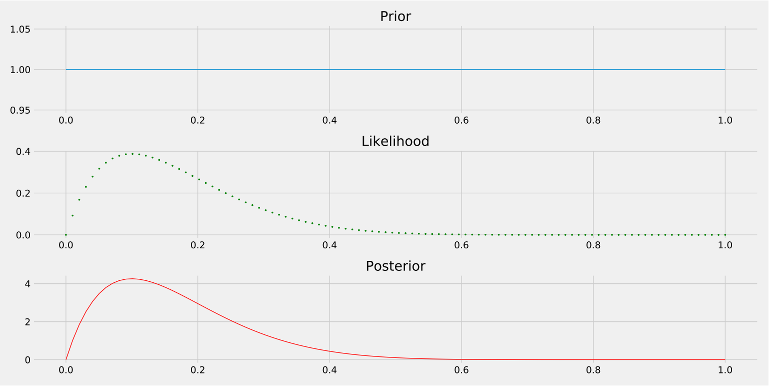
In general if $p(\theta) \sim Beta(a, b)$, and $p(X|\theta)$ has the PMF: $\binom{N}{k}\theta^k(1-\theta)^{N-k}$, Then:

$$p(\theta|X) \sim Beta(a+b, b+N-X) \ .$$

(S. Maleki 2019)

Bayes’ Theorem With Distributions

A Beta Prior-Binomial Likelihood Example



(S. Maleki 2019)

Bayes' Theorem With Distributions

A Normal Prior-Normal Likelihood Example

Consider an analytical chemist whose balance produces measurements that are normally distributed with mean equal to the true mass of the sample and standard deviation that has been estimated by the manufacturer balance and confirmed against calibration standards provided by the National Institute of Standards and Technology.

Assume the prior on the unknown mean follows a normal distribution, i.e. $\mu \sim \mathcal{N}(\nu, \tau^2)$.

We also assume that the data $x_1, x_2, ..., x_n$ are independent and come from a normal distribution with variance σ^2.

(S. Maleki 2019)

Bayes' Theorem With Distributions

A Normal Prior-Normal Likelihood Example

Suppose the chemist wants to measure the mass of a sample of ammonium nitrate.

Her balance has a known standard deviation of 0.2 milligrams. By looking at the sample, she thinks this mass is about 10 milligrams and based on her previous experience in estimating masses, her guess has the standard deviation of 2. So she decides that her prior for the mass of the sample is a normal distribution with mean, 10 milligrams, and standard deviation, 2 milligrams.

Now she collects five measurements on the sample and finds that the average of those is 10.5.

By conjugacy of the normal-normal family, our posterior belief about the mass of the sample has the normal distribution.

(S. Maleki 2019)

Bayes' Theorem With Distributions

A Normal Prior-Normal Likelihood Example

The new mean of that posterior normal is found by plugging into the formula:

$$\mu \sim \mathcal{N}(\nu = 10, \tau^2 = 2^2)$$

$$\mu_{post} = \frac{\nu\sigma^2 + n\bar{x}\tau^2}{\sigma^2 + n\tau^2} = \frac{10 \times (0.2)^2 + 5 \times 10.5 \times 2^2}{\times (0.2)^2 + 5 \times 2^2} = 10.499$$

$$\sigma_{post} = \sqrt{\frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}} = \sqrt{\frac{(0.2)^2 \times 2^2}{(0.2)^2 + 5 \times 2^2}} = 0.089$$

(S. Maleki 2019)

Bayes' Theorem With Distributions

A Normal Prior-Normal Likelihood Example

Before seeing the data, the Bayesian analytical chemist thinks the ammonium nitrate has mass 10 mg and uncertainty (standard deviation) 2 mg.

After seeing the data, she thinks the mass is 10.499 mg and standard deviation 0.089 mg.

Her posterior mean has shifted quite a bit and her uncertainty has dropped by a lot. That's exactly what an analytical chemist wants.

Prior Predictive Distributions

Suppose we assume a parametric sampling distribution $p(data|\theta)$ and express our uncertainty about the parameter θ as a distribution $p(\theta)$. Before we observe any data, what do we expect the distribution of observations to be?

In the continuous case:

$$p(data_{new}) = \int_{\Theta} p(data_{new}, \theta) d\theta = \int_{\Theta} p(data_{new}|\theta) p(\theta) d\theta \ .$$

In the discrete case:

$$p(data_{new}) = \sum_{i} p(data_{new}|\theta_i) p(\theta_i) \ .$$

- What we would predict for $data_{new}$ given no data.
- Useful for assessing whether choice of prior distribution does capture prior beliefs.

(S. Maleki 2019)

Posterior Predictive Distributions

After taking the sample, we have a better representation of the uncertainty in θ via our posterior $p(\theta|data)$. So the posterior predictive distribution for a new data is:

$$\begin{aligned} p(data_{new}|data) &= \int_{\Theta} p(data_{new}|\theta, data)p(\theta|data)d\theta \\ &= \int_{\Theta} p(data_{new}|\theta)p(\theta|data)d\theta \ , \end{aligned}$$

and in the discrete case:

$$p(data_{new}|data) = \sum_i p(data_{new}|\theta_i)p(\theta_i|data) \ .$$

- This reflects how we would predict new data to behave / vary.
- If the data we did observe follow this pattern closely, it indicates we have chosen our model and prior well.

(S. Maleki 2019)

Bayes Factors

Model Comparison:

$$p(Model_1|data) \quad \text{vs} \quad p(Model_2|data)$$

Bayes' rule:

$$p(model_1|data) = \frac{\overbrace{p(data|M_1)}^{\text{marginal likelihood of } M_1} \times \overbrace{p(M_1)}^{\text{prior}}}{\underbrace{p(data)}_{\text{marginal likelihood of both models}}}$$

From posterior distribution:

$$p(\theta|data, \theta, M_1) = \frac{p(data|\theta, M_1) \times p(\theta|M_1)}{p(data|M_1)}$$

(S. Maleki 2019)

Bayes Factors

So

$$p(data|M_1) = \int p(data|\theta, M_1)\, p(\theta|M_1)\, d\theta \ ,$$

or:

$$p(data|M_1) = \sum_{\theta} p(data|\theta, M_1)\, p(\theta|M_1) \ .$$

On the other hand:

$$p(data) = p(data|M_1) \times p(M_1) + p(data|M_2) \times p(M_2) \ .$$

(S. Maleki 2019)

Bayes Factors

$$BF_{1,2} := \underbrace{\frac{p(M_1|data)}{p(M_2|data)}}_{\text{posterior odds}} = \underbrace{\frac{p(data|M_1)}{p(data|M_2)}}_{\text{Bayes factor}} \times \underbrace{\frac{p(M_1)}{p(M_2)}}_{\text{prior odds}} \; .$$

Bayes factor ($BF_{1,2}$)	Interpretation
$BF < \frac{1}{100}$	Decisive evidence for M_2
$BF < \frac{1}{10}$	Strong evidence for M_2
$\frac{1}{10} < BF < \frac{1}{3}$	Moderate evidence for M_2
$\frac{1}{3} < BF < 1$	Weak evidence for M_2
$1 < BF < 3$	Weak evidence for M_1
$3 < BF < 10$	Moderate evidence for M_1
$BF > 10$	Strong evidence for M_1
$BF > 100$	Decisive evidence for M_1

Table: Jeffrey's scale of evidence for interpreting Bayes factors.

(S. Maleki 2019)

Bayes Factors – Example

Suppose we toss a coin 10 times and observer 9 heads. Is this a fair coin?

Let $\mathcal{H}_1$ be the hypothesis (model) that the coin is fair and $\mathcal{H}_2$ the hypothesis that the coin is unfair. In the case where we, have a coin fair, we assign the prior probability of $\frac{1}{2}$ to each outcome. However, we're not sure of the prior probability for an unfair coin. Therefore we assign a uniform prior p on $[0, 1]$.

$$p(9 \text{ H out of 10 trials}|\mathcal{H}_1) = \binom{10}{9}\frac{1}{2}^9\frac{1}{2}^{10-9} = 10 \times \frac{1}{2}^{10}$$

$$p(9 \text{ H out of 10 trials}|\mathcal{H}_2) = \int_0^1 \binom{10}{9} p^9(1-p)^{10-9}dp = 10 \times \frac{100}{110}$$

$$BF_{1,2} = \frac{\frac{1}{2}^{10}}{\frac{100}{110}} = 0.00107 \longrightarrow unfair$$

(S. Maleki 2019)

Credible Intervals

For a random variable θ (parameter), a Bayesian credible interval of size $1-\alpha$ is an interval (a,b) such that:

$$p(a \leq \theta \leq b|data) = 1-\alpha \ .$$

That is:

$$\int_a^b p(\theta|data)d\theta = 1-\alpha \ .$$

These intervals are not unique, since there will be many intervals with the correct probability coverage for a given posterior distribution.

A $100(1-\alpha)\%$ Highest Density Interval (HDI) for θ is the region $C_\alpha = \{\theta : p(\theta|data) \geq \gamma\}$ where γ is chosen so that:

$$p(\theta \in C_\alpha|data) = 1-\alpha \ .$$

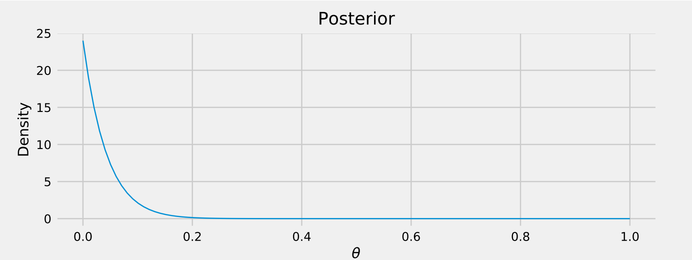
(S. Maleki 2019)

Credible Intervals – Example

Suppose that the posterior distribution for θ is a $\mathrm{Beta}(1, 24)$ distribution, with probability density function:

$$p(\theta|data) = 24(1-\theta)^{23}, \quad 0 < \theta < 1 \ .$$

Determine the $100(1-\alpha)\%$ HDI for θ.



(S. Maleki 2019)

Credible Intervals – Example

The HDI must include those values of θ with highest posterior density and so must take the form $C_\alpha = (0, b)$. The end–point b must satisfy:

$$\int_0^b 24(1-\theta)^{23} d\theta = 1 - \alpha \ .$$

Now

$$\int_0^b 24(1-\theta)^{23} d\theta = \Big[-(1-\theta)^{24} \Big]_0^b = 1 - (1-b)^{24} \ .$$

Therefore

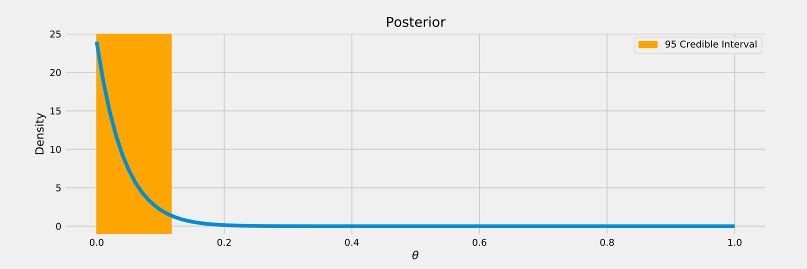
$$1 - (1-b)^{24} = 1 - \alpha \Longrightarrow 1 - b = \alpha^{1/24} \Longrightarrow b = 1 - \alpha^{1/24} \ .$$

Hence a $100(1-\alpha)\%$ HDI for θ is $(0, 1 - \alpha^{1/24})$.

(S. Maleki 2019)

Credible Intervals – Example

For example a 95% HDI for θ is $(0, 0.117)$.



Which means the probability that C_α contains θ is 0.95.

(S. Maleki 2019)