Lecture Notes

BAYESIAN ANALYSIS

DSA 8505

Strathmore University

Lecturer: Prof. Jacob Ong'ala

Contents

3 Posterior Distribution

3.1 Introduction

In Bayesian statistics, the **posterior distribution** represents the updated belief about an unknown parameter θ after observing data D. Unlike frequentist statistics where parameters are treated as fixed constants, Bayesian inference treats θ as a random variable whose uncertainty is described using probability distributions.

Before observing data, our uncertainty about θ is described by a **prior distribution** $p(\theta)$. Once data are observed, we use Bayes' theorem to combine the prior information with the likelihood of the observed data. This yields the posterior distribution:

$$p(\theta \mid D).$$

The posterior is the most important object in Bayesian inference because it contains all updated knowledge about θ after observing D. From it we can compute Bayesian estimates (posterior mean, MAP estimate), credible intervals, and predictive distributions.

Bayes' theorem provides the mathematical rule for updating beliefs:

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)}.$$

3.2 Deriving posterior distribution

3.2.1 Beta prior and binomial likelihood

We will derive the posterior distribution by combining the Beta prior with the Binomial likelihood using Bayes' Theorem.

1. Beta Prior Distribution

The Beta distribution is defined for a probability parameter $\theta \in [0, 1]$. Its probability density function (PDF) is:

$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

Where:

- θ is the unknown probability parameter (e.g., the probability of heads),
- α and β are the shape parameters of the Beta distribution,

- $B(\alpha, \beta)$ is the Beta function:

$$B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

2. Binomial Likelihood Function

The likelihood function for k successes out of n Bernoulli trials with success probability θ is given by the Binomial distribution:

$$P(D|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

Where:

- k is the number of successes (e.g., heads),
- n is the total number of trials (e.g., tosses),
- θ is the probability of success.

This likelihood function gives the probability of observing k successes in n trials, given a specific value of θ.

3. Bayes' Theorem

Bayes' Theorem provides the posterior distribution by updating the prior distribution using the likelihood of the observed data:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

Where:

- $P(\theta|D)$ is the posterior distribution, representing the updated belief about θ after observing the data D,
- $P(D|\theta)$ is the likelihood, the probability of observing data D given θ,
- $P(\theta)$ is the prior distribution of θ,
- $P(D)$ is the marginal likelihood, a normalizing constant.

4. Deriving the Posterior Distribution

Now, we multiply the prior and likelihood to get the unnormalized posterior:

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

Substitute the Binomial likelihood and Beta prior:

$$P(\theta|D) \propto \left(\binom{n}{k} \theta^k (1-\theta)^{n-k} \right) \times \left(\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} \right)$$

Since $\binom{n}{k}$ and $B(\alpha, \beta)$ are constants with respect to θ, we can drop them for proportionality:

$$P(\theta|D) \propto \theta^k (1-\theta)^{n-k} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Simplifying the exponents:

$$P(\theta|D) \propto \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$$

5. Recognizing the Posterior as a Beta Distribution

The expression $\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$ corresponds to the form of a Beta distribution. Therefore, the posterior distribution is a Beta distribution with updated parameters:

$$\theta|D \sim \text{Beta}(\alpha + k, \beta + n - k)$$

Where:

- $\alpha + k$ is the updated shape parameter for successes,
- $\beta + n - k$ is the updated shape parameter for failures.

Example: Beta-Binomial Model for Coin Tosses

As an example, consider estimating the probability θ of heads in a series of coin tosses. Suppose we observe 10 tosses, out of which 7 are heads. We can use Bayesian inference to estimate θ, the probability of heads.

- **Prior Distribution**: Assume we use a **Beta prior**, specifically $\theta \sim \text{Beta}(2, 2)$. The Beta distribution is a common prior for probabilities, as it is conjugate to the Binomial distribution.
- **Likelihood**: The data follows a **Binomial distribution**. For 7 heads out of 10 tosses, the likelihood is proportional to $\theta^7(1-\theta)^3$.
- **Posterior Distribution**: Using Bayes' Theorem, the posterior distribution is updated by combining the prior and the likelihood. Since the Beta distribution is conjugate to the Binomial, the posterior is also a Beta distribution. Specifically:

  $$\theta|D \sim \text{Beta}(2 + 7, 2 + 3) = \text{Beta}(9, 5)$$

  This posterior distribution reflects our updated belief about θ after observing the data.

The posterior Beta(9,5) distribution suggests that the probability of heads is now most likely around $\frac{9}{14} \approx 0.64$, which represents a more informed belief after observing 7 heads out of 10 tosses.

3.2.2 Uniform prior and normal likelihood

To derive the posterior distribution with a uniform prior and a normal likelihood, we can use Bayes' theorem. Let's denote the following:

- θ: the parameter we want to estimate. - x: the observed data. - $p(\theta)$: the prior distribution of θ. - $p(x \mid \theta)$: the likelihood of the data given the parameter θ. - $p(\theta \mid x)$: the posterior distribution of θ given the data x.

1. Set the Prior

Assume a uniform prior for θ:

$$p(\theta) = c$$

where c is a constant. The uniform prior does not provide information about the parameter, so it can be considered constant over the parameter's support.

Assume a uniform prior for θ on the interval $[a, b]$:

$$p(\theta) = \begin{cases} \dfrac{1}{b-a}, & a \leq \theta \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

$$p(\theta) = c, \quad a \leq \theta \leq b,$$

where the normalization constant is

$$c = \frac{1}{b-a}.$$

2. Set the Likelihood

Assume the likelihood is normally distributed:

$$p(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$$

where σ^2 is the variance of the normal distribution.

$$p(x \mid \theta) \propto \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$$

where $\exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$ is the kernel of a normal distribution.

3. Apply Bayes' Theorem

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)}$$

Since $p(x)$ does not depend on θ, we can ignore it for the purpose of deriving the form of the posterior distribution.

4. Substitute the Prior and Likelihood

$$p(\theta \mid x) \propto p(x \mid \theta)p(\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)\right) \cdot c$$

$$p(\theta \mid x) \propto \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$$

5. Identify the Form of the Posterior

The expression $\exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$ is the kernel of a normal distribution. Thus, the posterior distribution can be recognized as a normal distribution.

6. Normalize the Posterior

The posterior distribution is normally distributed:

$$p(\theta \mid x) \sim \mathcal{N}(x, \sigma^2)$$

This indicates that the posterior distribution of θ given the observed data x is a normal distribution centered at the observed value x with the same variance σ^2.

When using a uniform prior with a normal likelihood, the posterior distribution of the parameter θ is a normal distribution:

$$\theta \mid x \sim \mathcal{N}(x, \sigma^2)$$

This reflects the fact that observing x gives us the best estimate of θ while retaining the variance inherent in the likelihood.

Example

A public health researcher wants to estimate the average number of daily visits (θ) to a local health clinic. The researcher believes that the average could be anywhere between 0 and 100 visits (uniform prior), and they have data from 10 days showing the following number of visits:

$$x = \{20, 25, 30, 35, 40, 45, 50, 55, 60, 65\}$$

Solution

1. Calculate the Sample Mean and Variance:
First, we compute the sample mean $\bar{x}$ and sample variance s^2:

$$\bar{x} = \frac{20 + 25 + 30 + 35 + 40 + 45 + 50 + 55 + 60 + 65}{10} = 42.5$$

To calculate the sample variance s^2:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Where $n = 10$. - Now calculate the variance:

$$s^2 = \frac{2500}{10-1} = \frac{2500}{9} \approx 277.78$$

Thus, $\sigma^2 = s^2 \approx 277.78$.

2. Set the Prior:
Assume a uniform prior for θ:

$$p(\theta) = c \quad \text{for } \theta \in [0, 100]$$

3. Set the Likelihood:
The likelihood is:

$$p(x \mid \theta) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{(x-\theta)^2}{2s^2}\right)$$

For our calculations, we'll use the sample mean and variance in the likelihood.

4. Apply Bayes' Theorem:
The posterior is proportional to the product of the likelihood and the prior:

$$p(\theta \mid x) \propto p(x \mid \theta)p(\theta)$$

Ignoring constants not dependent on θ:

$$p(\theta \mid x) \propto \exp\left(-\frac{(42.5-\theta)^2}{2 \cdot 277.78}\right)$$

5. Identify the Form of the Posterior:
The posterior distribution is:

$$p(\theta \mid x) \sim \mathcal{N}\left(\bar{x}, s^2\right) \text{ where } \bar{x} = 42.5 \text{ and } s^2 \approx 277.78$$

Thus, the posterior distribution of θ is:

$$\theta \mid x \sim \mathcal{N}(42.5, 277.78)$$

Normal-Normal Model

1. Prior Distribution Assume that the prior distribution for the parameter μ is normally distributed:

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

where μ_0 is the mean and σ_0^2 is the variance of the prior.

2. Likelihood Assume we have observed data x which follows a Normal distribution centered at μ:

$$x|\mu \sim \mathcal{N}(\mu, \sigma^2)$$

where σ^2 is the variance of the likelihood.

Deriving the Posterior Distribution

Using Bayes' theorem, the posterior distribution $P(\mu|x)$ is proportional to the product of the likelihood and the prior:

$$P(\mu|x) \propto P(x|\mu)P(\mu)$$

Substituting in the expressions for the likelihood and the prior:

1. Prior:

$$P(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

2. Likelihood:

$$P(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Combining the Terms

Now we can combine these to find the posterior:

$$P(\mu|x) \propto \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right) \left(\frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)\right)$$

Simplifying the Expression

Combining the exponentials:

$$P(\mu|x) \propto \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

This can be rearranged by combining the terms in the exponent. We want to complete the square for the term:

$$-\frac{(x-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}$$

Completing the Square

Completing the square in the expression:

$$P(\mu|x) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\left(\mu - \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}\right)^2 + \text{constant}\right)$$

$$-\frac{(x-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}$$

Completing the Square

Step 1: Expand both quadratic terms

$$-\frac{1}{2\sigma^2}\left(\mu^2 - 2x\mu + x^2\right) - \frac{1}{2\sigma_0^2}\left(\mu^2 - 2\mu_0\mu + \mu_0^2\right)$$

Step 2: Collect terms in μ

$$-\frac{1}{2}\left[\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)\mu + \left(\frac{x^2}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2}\right)\right]$$

Step 3: Factor the coefficient of μ^2

$$-\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\left[\mu^2 - 2\frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}\mu\right] + \text{constant}$$

Step 4: Complete the square

Let

$$a = \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

Using

$$\mu^2 - 2a\mu = (\mu - a)^2 - a^2,$$

we obtain

$$-\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)(\mu - a)^2 + \text{constant}$$

Step 5: Final form

$$P(\mu \mid x) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\left(\mu - \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}\right)^2\right)$$

Posterior Mean and Variance

From the completed square form, the posterior distribution is also normal, with mean and variance given by:

$$\mu_n = \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

$$\sigma_n^2 = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

Thus, the posterior distribution is:

$$\mu|x \sim \mathcal{N}\left(\mu_n, \sigma_n^2\right)$$

Example:

Suppose we have a sample of $n = 10$ measurements with sample mean $\bar{X} = 5$ and known variance $\sigma^2 = 4$. If we use a prior $\mu \sim \mathcal{N}(4, 1)$, the posterior mean and variance are computed as:

$$\mu_n = \frac{\frac{10}{4} \cdot 5 + \frac{1}{1} \cdot 4}{\frac{10}{4} + 1} = \frac{12.5 + 4}{3.5} = 4.71$$

and

$$\tau_n^2 = \left(\frac{10}{4} + 1\right)^{-1} = \frac{1}{3.5} = 0.286$$

Thus, the posterior distribution is $\mu|X \sim \mathcal{N}(4.71, 0.286)$.

Gamma-Poisson Model

For a Poisson likelihood, where the parameter λ represents the rate of occurrence, the conjugate prior is the Gamma distribution.

Model Setup

1. **Data Model:** Let Y be the number of events (counts), and assume that:

$$Y|\lambda \sim \text{Poisson}(\lambda)$$

where λ is the rate of the Poisson distribution.

2. **Prior Distribution**: Assume a Gamma prior for the rate parameter λ:

$$\lambda \sim \Gamma(\alpha, \beta)$$

Here, α is the shape parameter and β is the rate parameter of the Gamma distribution.

Derivation of the Posterior Distribution

Using Bayes' theorem, the posterior distribution of λ given the observed data $Y = y$ can be computed:

$$P(\lambda|Y) \propto P(Y|\lambda)P(\lambda)$$

Step 1: Likelihood Function

The likelihood of the data given λ is:

$$P(Y = y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

Step 2: Prior Distribution

The prior distribution is given by:

$$P(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

Step 3: Posterior Distribution

Now, substituting the likelihood and prior into Bayes' theorem:

$$P(\lambda|Y) \propto \left(\frac{\lambda^y e^{-\lambda}}{y!}\right)\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}\right)$$

Ignoring constants that do not depend on λ, we have:

$$P(\lambda|Y) \propto \lambda^{y+\alpha-1}e^{-(1+\beta)\lambda}$$

Step 4: Identify the Posterior Distribution

The above expression is recognized as the kernel of a Gamma distribution:

$$P(\lambda|Y) \sim \Gamma(y + \alpha, 1 + \beta)$$

3.2.3 Predictive and Marginal Distributions

The **predictive distribution** allows us to make predictions about future observations based on the current model and the posterior distribution. It represents the distribution of future data given the observed data and is obtained by averaging the likelihood over the posterior distribution of the parameter:

Derivation of the Predictive Distribution

We want to derive the predictive distribution:

$$P(x_{\text{new}}|D) = \int P(x_{\text{new}}|\theta)P(\theta|D)d\theta$$

where $P(x_{\text{new}}|D)$ is the probability of a new data point x_{new} given the observed data D.

Derivation

1. **Start with the definition of the predictive distribution** :

The predictive distribution of a new data point x_{new} given the observed data D can be expressed as the marginal probability over the parameter θ:

$$P(x_{\text{new}}|D) = \int P(x_{\text{new}}, \theta|D)d\theta$$

2. **Apply the chain rule of probability**:

The joint probability $P(x_{\text{new}}, \theta|D)$ can be decomposed using the chain rule:

$$P(x_{\text{new}}, \theta|D) = P(x_{\text{new}}|\theta, D)P(\theta|D)$$

3. **Simplify $P(x_{\text{new}}|\theta, D)$**:

Since x_{new} depends only on θ and not directly on the dataset D once θ is known, we can simplify:

$$P(x_{\text{new}}|\theta, D) = P(x_{\text{new}}|\theta)$$

4. **Substitute back into the integral**:

Substituting the decomposed expression into the integral, we get:

$$P(x_{\text{new}}|D) = \int P(x_{\text{new}}|\theta)P(\theta|D)d\theta$$

Explanation of Components

- $P(x_{\text{new}}|\theta)$: The likelihood of the new data point x_{new} given the parameter θ.
- $P(\theta|D)$: The posterior distribution of the parameter θ after observing the data D.
- $\int P(x_{\text{new}}|\theta)P(\theta|D)d\theta$: The integral sums over all possible values of θ, weighted by the posterior probability $P(\theta|D)$.

Coin Toss Example: Predictive Distribution

We are interested in computing the probability of getting heads on the next toss, given that we have already observed 9 heads and 5 tails. We'll use Bayesian inference to update our belief about the probability of heads, θ, and compute the predictive distribution.

Step 1: Prior Distribution

We start by specifying a prior distribution for θ, the probability of getting heads. We use a Beta distribution as the prior, $\text{Beta}(\alpha_0, \beta_0)$, which is conjugate to the binomial likelihood (coin tosses).

Assuming a uniform prior, $\text{Beta}(1, 1)$, which corresponds to no strong prior belief about θ, we have:

$$P(\theta) = \text{Beta}(1, 1)$$

This is equivalent to the uniform distribution over $[0, 1]$.

Step 2: Likelihood

The likelihood function represents the probability of observing the data D (9 heads and 5 tails) given θ. Since the data follows a binomial distribution, the likelihood is:

$$P(D|\theta) = \theta^9(1-\theta)^5$$

Step 3: Posterior Distribution

Using Bayes' theorem, we update the prior distribution based on the observed data to obtain the posterior distribution for θ:

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

Substituting the expressions for the likelihood and the prior, we get:

$$P(\theta|D) \propto \theta^9(1-\theta)^5 \cdot \theta^{1-1}(1-\theta)^{1-1}$$

Simplifying this expression:

$$P(\theta|D) \propto \theta^9(1-\theta)^5$$

This is the kernel of a Beta distribution. Specifically:

$$P(\theta|D) = \text{Beta}(9+1, 5+1) = \text{Beta}(10, 6)$$

Thus, the posterior distribution for θ after observing 9 heads and 5 tails is $\text{Beta}(10, 6)$.

Step 4: Predictive Distribution

The predictive distribution for the next toss being heads is:

$$P(\text{heads}|9, 5) = \int_0^1 P(\text{heads}|\theta)P(\theta|9, 5)d\theta$$

where:

$$P(\text{heads}|\theta) = \theta$$

is derived from the definition of θ in the context of a Bernoulli process (such as a coin toss), where θ represents the probability of observing a "head" in a single trial (coin toss)

and:

$$P(\theta|9, 5) = \text{Beta}(10, 6)$$

Step 5. Compute the Predictive Probability

The mean of the Beta distribution Beta(α, β) is:

$$\mathbb{E}[\theta] = \frac{\alpha}{\alpha + \beta}$$

For $\alpha = 10$ and $\beta = 6$, the mean is:

$$\mathbb{E}[\theta] = \frac{10}{10 + 6} = \frac{10}{16} = 0.625$$