An Introduction to Bayesian Methods

Rishabh Iyer

March 19, 2024

1 Introduction

Bayesian methods stand at the intersection of statistics and probability theory, offering a robust framework for quantifying uncertainty in a wide array of domains, from scientific research to financial analysis. At the heart of Bayesian inference is the concept of updating our beliefs or knowledge about uncertain events in light of new evidence or data. This approach contrasts with traditional statistical methods that often rely on fixed datasets to make inferences. Bayesian techniques, by utilizing the mathematical language of probability, allow for a more nuanced understanding of uncertainty, providing a mechanism to systematically incorporate both prior knowledge and new information.

One of the most compelling aspects of Bayesian methods is their ability to incorporate prior knowledge into the inferential process. This prior knowledge, or 'prior', can be based on previous studies, expert opinion, or any relevant information available before examining the current data. As new data becomes available, Bayesian inference uses the likelihood of observing this data, given different hypotheses, to update the prior beliefs and form a 'posterior' distribution. This posterior distribution reflects a new level of understanding, melding prior knowledge with the latest evidence. Such a dynamic approach to making inferences allows Bayesian methods to adapt and refine predictions as more information is gathered, making it an invaluable tool in fields where data is continuously updated or where uncertainty is a constant companion.

2 Fundamentals of Probability

2.1 Random Variables and Distributions

A random variable is a variable whose possible values are numerical outcomes of a stochastic process. The distribution of a random variable describes the probabilities of its possible outcomes.

2.1.1 Bernoulli and Binomial Distributions

The Bernoulli distribution is a discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability

$1 - p$, symbolically:

$$P(X = x) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0. \end{cases}$$

The Binomial distribution represents the number of successes in n independent Bernoulli trials, each with success probability p. Its probability mass function (PMF) is given by:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

where k is the number of successes, n is the number of trials, and $\binom{n}{k}$ is the binomial coefficient.

2.1.2 Dice Rolls

Dice rolls can be modeled using the discrete uniform distribution, which assumes that all outcomes are equally likely. If a dice has m faces, the probability of any specific outcome is:

$$P(X = x) = \frac{1}{m}$$

For a standard six-sided dice, $m = 6$, and the probability of any side coming up is $\frac{1}{6}$. A more general case is where $P(X = x) = p_x$ where $\sum_{i=1}^{m} p_i = 1$.

2.1.3 Distribution of N Dice Rolls

When rolling N dice, the distribution of counts for each side can be modeled using the multinomial distribution. For a m-sided die, the probability mass function (PMF) for observing a specific combination of counts $(c_1, c_2, \ldots, c_6)$, where c_i represents the count of side i, is given by:

$$P(X_1 = c_1, X_2 = c_2, \ldots, X_m = c_m) = \frac{N!}{c_1! c_2! \cdots c_m!} p_1^{c_1} p_2^{c_2} \cdots p_m^{c_m}$$

where N is the total number of dice rolls, and the sum of all counts equals N: $c_1 + c_2 + \cdots + c_m = N$ and $p_1, \cdots p_m$ is the probability for the dice roll appearing at $1, \cdots, m$ (in a standard dice, $m = 6$).

This distribution captures the probabilities of all possible outcomes for the counts of each side when N fair dice are rolled.

2.1.4 Gaussian Distribution

The Gaussian or normal distribution is a continuous distribution that is fully characterized by its mean (μ) and variance (σ^2). The probability density function (PDF) of a Gaussian distribution is:

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

where x is the variable, μ is the mean, and σ^2 is the variance. This distribution models phenomena in the natural and social sciences where most observations cluster around a central value with symmetry in the distribution of values to either side.

These mathematical formulations provide a foundation for understanding how different types of data and experiments can be modeled using probability distributions, setting the stage for further exploration into statistical inference and Bayesian methods.

2.2 Expectation and Variance

The expectation (or expected value) of a random variable provides a measure of the central tendency of its distribution, essentially representing the long-run average of its outcomes. Mathematically, the expectation of a discrete random variable X is defined as:

$$E[X] = \sum_x x \cdot P(X = x)$$

where x are the possible values of X, and $P(X = x)$ is the probability of X taking the value x. For a continuous random variable, the expectation is defined as:

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

where $f(x)$ is the probability density function of X.

Variance measures the spread of the distribution around its mean, quantifying the variability of the outcomes. The variance of a random variable X is defined as:

$$\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

2.2.1 Examples

Bernoulli Distribution: For a Bernoulli random variable X with success probability p, the expectation and variance are:

$$E[X] = p, \quad \text{Var}(X) = p(1-p)$$

Binomial Distribution: For a Binomial random variable X representing the number of successes in n independent Bernoulli trials with success probability p, the expectation and variance are:

$$E[X] = np, \quad \text{Var}(X) = np(1-p)$$

Gaussian Distribution: For a Gaussian (normal) random variable X with mean μ and variance σ^2, the expectation and variance are directly the parameters of the distribution:

$$E[X] = \mu, \quad \text{Var}(X) = \sigma^2$$

These examples illustrate how the expectation and variance are computed for different distributions, reflecting the central tendency and variability of the outcomes. The Bernoulli and Binomial distributions provide simple models for binary and count data, respectively, while the Gaussian distribution models continuous data with a symmetric distribution around the mean.

3 Maximum Likelihood Estimation (MLE)

MLE is a method used in statistics to estimate the parameters of a statistical model. It identifies the parameter values that maximize the likelihood function, which measures how likely it is to observe the given data D under different parameter values.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} p(D|\theta)$$

where $\hat{\theta}_{MLE}$ is the estimate that maximizes the likelihood function $L(\theta; D)$ for the parameter θ given the observed data. The likelihood function $L(\theta; data) = P(D|\theta)$, i.e., the probability of observing the dataset given the parameters θ. Given a dataset $D = \{X_1, \cdots, X_k\}$, the MLE formulation is basically:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \prod_{i=1}^{k} p(X_i|\theta)$$

3.1 Coin Flips Example

Consider a series of n coin flips with k heads. Assuming each flip is independent, the likelihood of observing k heads with a probability p of heads is given by the binomial distribution:

$$L(p; data) = P(D|\theta) = p^k(1-p)^{n-k}$$

To find the MLE of p, we take the derivative of the log-likelihood with respect to p, set it to zero, and solve for p:

$$\frac{d}{dp} \log L(p; D) = \frac{d}{dp} \left(k \log p + (n-k) \log(1-p)\right) = 0$$

Solving this equation gives the MLE of p:

$$\hat{p}_{MLE} = \frac{k}{n}$$

3.2 Dice Rolls Example

For n rolls of a fair die, let k_i be the count of times face i appears, with $i = 1, \ldots, 6$. The likelihood function, assuming each roll is independent, is:

$$L(p_1, \ldots, p_6; D) = p(D|\theta) = \prod_{i=1}^{6} p_i^{k_i}$$

Note that there is a constraint here that $\sum_i p_i = 1$. The MLE estimates are:

$$p_i = k_i/n$$

note that $\sum_i k_i = n$ so the probabilities all add up to 1.

3.3 Normal Distribution Example

For data modeled by a normal distribution with unknown mean μ and variance σ^2, given n observations $x_1, x_2, \ldots, x_n$, the likelihood function is:

$$L(\mu, \sigma^2; D) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right)$$

Taking the derivative of the log-likelihood with respect to μ and σ^2, setting them to zero, and solving gives the MLE estimates:

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_{MLE})^2$$

These derivations illustrate how MLE is used to estimate parameters that make the observed data most probable under the assumed model.

4 Maximum A-Posteriori Estimation (MAP)

MAP estimation is a Bayesian inference technique that combines prior knowledge about a parameter with the likelihood of observing the given data to produce a posterior distribution. This method seeks the parameter value that maximizes the posterior distribution.

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|D) = \arg\max_{\theta} \frac{P(D|\theta)P(\theta)}{P(D)}$$

Since $P(D)$ is constant for all θ, the MAP estimate simplifies to:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(D|\theta)P(\theta)$$

4.1 Prior and Posterior

The prior distribution $P(\theta)$ represents our initial beliefs about the parameter θ before observing any data. The likelihood $P(D|\theta)$ quantifies how probable the observed data is for different values of θ. The posterior distribution $P(\theta|D)$ updates our beliefs about θ in light of the observed data.

4.2 Why MAP is Useful in Low Data Settings

In settings with sparse data, the prior plays a crucial role in shaping the posterior, providing a way to incorporate external knowledge into the estimation process. This can lead to more robust estimates than those obtained by MLE, particularly when the observed data are not sufficiently informative on their own.

4.3 MAP Estimation Examples

4.3.1 Coin Flips

Given a series of coin flips with k heads observed out of n flips and a Beta prior $Beta(\alpha, \beta)$ for the probability of heads p, the posterior distribution for p is:

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

The MAP estimate for p is:

$$\hat{p}_{MAP} = \frac{k+\alpha-1}{n+\alpha+\beta-2}$$

4.3.2 Dice Rolls

Consider rolling a die N times, resulting in a set of outcomes where each face i appears c_i times, and the total number of rolls is $N = \sum_{i=1}^{6} c_i$. Assuming a Dirichlet prior for the probabilities of each face $p = (p_1, \ldots, p_6)$, $Dir(\alpha_1, \ldots, \alpha_6)$, where α_i represents the prior belief in the probability of face i appearing.

The Dirichlet distribution is given by:

$$P(p; \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{6} p_i^{\alpha_i - 1}$$

where $B(\alpha)$ is the Dirichlet normalization constant, and $\alpha = (\alpha_1, \ldots, \alpha_6)$.

The likelihood of observing the given data (counts of each face) assuming a multinomial distribution is:

$$P(D|p) = p_1^{c_1} \cdots p_6^{c_6}$$

The posterior distribution for p after observing the data, combining the Dirichlet prior and the multinomial likelihood, is also a Dirichlet distribution:

$$P(p|D) \propto P(D|p)P(p) = \frac{1}{B(\alpha')} \prod_{i=1}^{6} p_i^{c_i+\alpha_i-1}$$

where $\alpha' = (\alpha_1 + c_1, \ldots, \alpha_6 + c_6)$. The MAP estimate for p maximizes this posterior distribution. Given the properties of the Dirichlet distribution, the mode (or MAP estimate) for each p_i is:

$$\hat{p}_{i,MAP} = \frac{\alpha_i + c_i - 1}{\sum_{j=1}^{6}(\alpha_j + c_j - 1)}$$

This formulation shows how the prior beliefs (α_i) are updated with the observed data (c_i) to estimate the probabilities of each face of the die, reflecting both the prior knowledge and the empirical evidence from the rolls.

4.3.3 Normal Distribution

Given a set of exam scores modeled by a normal distribution with unknown mean μ and known variance σ^2, and assuming a normal prior for μ, $N(\mu_0, \tau^2)$, the posterior for μ after observing data X is also normal, with:

$$\mu_{posterior} = \frac{\frac{\mu_0}{\tau^2} + \frac{\sum x_i}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}, \quad \sigma^2_{posterior} = \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)^{-1}$$

The MAP estimate for μ is the mode of this posterior distribution, which, for a normal distribution, coincides with the mean $\mu_{posterior}$.

These examples illustrate how MAP estimation leverages both prior knowledge and observed data to make inferences about unknown parameters, offering a principled way to incorporate external information into the estimation process.

5 Conclusion

Bayesian methods offer a nuanced approach to statistical inference, blending prior knowledge with observed data. Through examples of coin flips, dice rolls, and normal distributions, we've seen how Bayesian techniques, particularly MLE and MAP, provide a robust framework for making inferences about the world.