Bayesian Inference

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available.

Frequentist View

- **Parameters are Fixed**: In the frequentist view, parameters (e.g., the mean or variance of a population) are considered fixed but unknown values. These parameters are not random; they are fixed properties of the population.
- **Estimate Using Data**: We use data from samples to estimate these fixed parameters. For example, we might use the sample mean to estimate the population mean.
- **One True Value**: There is only one true value for a given parameter. However, since we can have many different datasets (samples), our estimates will vary depending on the sample.
- **Uncertainty in Estimation**: Because our sample might not perfectly represent the population, our estimate of the parameter will have some uncertainty. This uncertainty is often quantified using confidence intervals.

Bayesian View

- **Parameters are Random Variables**: In the Bayesian view, parameters are not fixed. Instead, they are treated as random variables with their own probability distributions.
- **Distributions**: These distributions (called prior distributions) represent our initial beliefs about the parameters before seeing the data.
- **One Dataset**: There is only one dataset, and we use this data to update our knowledge about the parameters' distributions.
- **Updating Knowledge**: Bayes' theorem allows us to update our prior distributions with the observed data to obtain posterior distributions. These posterior distributions reflect our updated beliefs about the parameters after considering the data.

Comparing the Two Views

- **Frequentist**:
    - Parameters are fixed.
    - There is one true value for each parameter.
    - Estimates have uncertainty due to sampling variability.
- **Bayesian**:
    - Parameters are random variables with distributions. ○ We use the data to update these distributions. ○ Posterior distributions represent our updated knowledge about the parameters.

Example

Consider estimating the mean of a population:

Frequentist Approach:

- Assume the mean is a fixed value.
- Collect a sample and calculate the sample mean.
- Construct a confidence interval around the sample mean to express the uncertainty.

Bayesian Approach:

- Assume the mean has a prior distribution (e.g., normal distribution with some mean and variance).
- Collect a sample and update the prior distribution using Bayes' theorem to get the posterior distribution.
- The posterior distribution provides a full description of our uncertainty about the mean after considering the data.

Conclusion

Both approaches aim to make inferences about unknown parameters based on data, but they differ fundamentally in their treatment of parameters and the interpretation of probability. Frequentists view probability as the long-run frequency of events, while Bayesians view it as a measure of belief or certainty about events.

Key Concepts in Bayesian Inference

Likelihood Function

- **Definition**: P(Data∣θ)P(\text{Data} \mid \theta)P(Data∣θ) viewed as a function of θ\thetaθ
- **Purpose**: Measures the probability of the observed data given a specific value of the parameter θ\thetaθ.

Maximum Likelihood Estimate (MLE)

- **Definition**: The θ\thetaθ value that maximizes the likelihood function.
- **Purpose**: Provides the parameter estimate that makes the observed data most probable.

Prior Distribution

- **Definition**: Prior (before seeing data) belief about the distribution of θ\thetaθ.

- **Purpose**: Incorporates any existing knowledge or assumptions about θ\thetaθ.

Posterior Distribution

- **Definition**: Posterior (after seeing data) belief about the distribution of θ\thetaθ.
- **Formula**: Posterior∝Likelihood×Prior\text{Posterior} \propto \text{Likelihood} \times \text{Prior}Posterior∝Likelihood×Prior This means the posterior distribution is proportional to the product of the likelihood function and the prior distribution.

Maximum a Posteriori Probability (MAP)

- **Definition**: The θ\thetaθ value that maximizes the posterior distribution.
- **Purpose**: Provides the parameter estimate that is most probable given both the observed data and the prior information.

Detailed Explanation

Likelihood Function

The likelihood function P(Data∣θ)P(\text{Data} \mid \theta)P(Data∣θ) measures the probability of observing the given data for different values of θ\thetaθ. For example, if you have data points from a normal distribution, the likelihood function would tell you how likely it is to observe those data points for different values of the mean μ\muμ and variance σ2\sigma^2σ2.

Maximum Likelihood Estimate (MLE)

The MLE is the value of θ\thetaθ that maximizes the likelihood function. This means we are finding the parameter value that makes the observed data most likely. Mathematically, it is:

θ^MLE=arg⁡max⁡θP(Data∣θ)\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} P(\text{Data} \mid \theta)θ^MLE=argθmaxP(Data∣θ)

Prior Distribution

Before seeing any data, we might have some beliefs about the parameter θ\thetaθ. This belief is encoded in the prior distribution P(θ)P(\theta)P(θ). For example, if you believe θ\thetaθ is likely to be around a certain value, you would choose a prior distribution centered around that value.

Posterior Distribution

After observing the data, we update our prior beliefs using Bayes' theorem to get the posterior distribution. The posterior distribution combines the prior distribution and the likelihood of the observed data. Mathematically, it is:

P(θ|Data)∝P(Data|θ)×P(θ)P(\theta \mid \text{Data}) \propto P(\text{Data} \mid \theta) \times P(\theta)P(θ|Data)∝P(Data|θ)×P(θ)

This means the posterior distribution is proportional to the likelihood function multiplied by the prior distribution.

Maximum a Posteriori Probability (MAP)

The MAP estimate is the value of θ\thetaθ that maximizes the posterior distribution. This is similar to the MLE but takes into account the prior distribution as well. Mathematically, it is:

θ^MAP=arg⁡maxθP(θ|Data)\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} P(\theta \mid \text{Data})θ^MAP=argθmaxP(θ|Data)

Using the relationship between the posterior, likelihood, and prior, this can be rewritten as:

θ^MAP=arg⁡maxθ(P(Data|θ)×P(θ))\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} \left(P(\text{Data} \mid \theta) \times P(\theta) \right)θ^MAP=argθmax(P(Data|θ)×P(θ))

Example

Consider a simple case where we want to estimate the bias θ\thetaθ of a coin (the probability of heads):

1. **Likelihood Function**: Suppose we flip the coin 10 times and observe 7 heads. The likelihood function for θ\thetaθ is:

   P(Data|θ)=θ7(1−θ)3P(\text{Data} \mid \theta) = \theta^7 (1 - \theta)^3P(Data|θ)=θ7(1−θ)3

2. **MLE**: The MLE would be the value of θ\thetaθ that maximizes this likelihood. In this case, it would be: θ^MLE=0.7\hat{\theta}_{\text{MLE}} = 0.7θ^MLE=0.7

3. **Prior Distribution**: Suppose our prior belief is that the coin is fair, so we use a Beta distribution Beta(1,1)\text{Beta}(1, 1)Beta(1,1) (which is uniform).
4. **Posterior Distribution**: Using Bayes' theorem, we combine the prior and the likelihood to get the posterior distribution, which is also a Beta distribution but with updated parameters:

   P(θ|Data)∝θ7(1−θ)3×1P(\theta \mid \text{Data}) \propto \theta^7 (1 - \theta)^3 \times 1P(θ|Data)∝θ7(1−θ)3×1

   This results in a Beta(8,4)\text{Beta}(8, 4)Beta(8,4) distribution.

5. **MAP**: The MAP estimate would be the value of θ\thetaθ that maximizes this posterior distribution. For a Beta distribution Beta(a,b)\text{Beta}(a, b)Beta(a,b), the MAP estimate is:

   θ^MAP=a−1a+b−2=8−18+4−2=710=0.7\hat{\theta}_{\text{MAP}} = \frac{a - 1}{a + b - 2} = \frac{8 - 1}{8 + 4 - 2} = \frac{7}{10} = 0.7θ^MAP=a+b−2a−1=8+4−28−1=107=0.7

In this simple example, the MLE and MAP estimates coincide, but this is not always the case, especially when the prior distribution significantly influences the posterior distribution.

Understanding Standard Normal Distribution

The standard normal distribution N(0,1)N(0, 1)N(0,1) has a mean μ=0\mu = 0μ=0 and a standard deviation σ=1\sigma = 1σ=1. When you generate samples from this distribution using `np.random.randn(n)`, you get `n` samples from N(0,1)N(0, 1)N(0,1).

Scaling and Shifting

To transform these samples to follow a different normal distribution N(μ,σ2)N(\mu, \sigma^2)N(μ,σ2), you need to scale and shift the standard normal samples. Here's how:

1. **Scaling**:
   - When you multiply the standard normal samples by σ\sigmaσ, the standard deviation of the resulting distribution changes from 1 to σ\sigmaσ. This is because the standard deviation is a measure of spread, and multiplying by σ\sigmaσ stretches or compresses the distribution accordingly.
2. **Shifting**:
   - When you add μ\muμ to the scaled samples, the mean of the distribution shifts from 0 to μ\muμ. This is because adding a constant shifts all values by that constant.

Example: Transforming N(0,1)N(0, 1)N(0,1) to N(μ2,σ22)N(\mu_2, \sigma_2^2)N(μ2,σ22)

Let's go through the steps with your parameters μ2=10\mu_2 = 10μ2=10 and σ2=1\sigma_2 = 1σ2=1:

1. **Generate samples from N(0,1)N(0, 1)N(0,1)**:

   ```
   python
   Copy code
   normal_samples_2 = np.random.randn(n)
   ```

   This gives you `n` samples from the standard normal distribution N(0,1)N(0, 1)N(0,1).

2. **Scale the samples**:

```
python
Copy code
scaled_samples_2 = normal_samples_2 * sigma2
```

Since σ2=1\sigma_2 = 1σ2=1, the scaled samples are still N(0,1)N(0, 1)N(0,1). However, if σ2\sigma_2σ2 were different, this step would adjust the standard deviation of the samples to σ2\sigma_2σ2.

3. **Shift the samples**:

```
python
Copy code
shifted_samples_2 = scaled_samples_2 + mu2
```

This shifts the mean of the samples from 0 to μ2=10\mu_2 = 10μ2=10.

Combining these steps:

```
python
Copy code
normal_samples_2 = np.random.randn(n) * sigma2 + mu2
```

This line generates `n` samples from N(μ2,σ22)N(\mu_2, \sigma_2^2)N(μ2,σ22) by first generating samples from N(0,1)N(0, 1)N(0,1), then scaling by σ2\sigma_2σ2, and finally shifting by μ2\mu_2μ2.

Summary

To generate samples from N(μ2,σ22)N(\mu_2, \sigma_2^2)N(μ2,σ22) using `np.random.randn(n)`:

- **Scale** by multiplying by σ2\sigma_2σ2.
- **Shift** by adding μ2\mu_2μ2.

This method transforms the standard normal samples to follow the desired normal distribution.