

Bayesian Neural Networks (BNN)

General Principles

To model complex, non-linear relationships between variables, we can use multiple approaches including, splines, polynomials, gaussian processes, and neural networks. Here, we will focus on a Bayesian Neural Network (BNN). Think of a neural network as a highly flexible function made of interconnected layers of “neurons.” Each connection between neurons has a weight, and each neuron has a bias. These weights and biases are like a vast set of adjustable knobs. In a standard network, the goal is to find the single best setting for all these knobs to map inputs to outputs. Unlike a standard neural network which learns a single set of optimal weights, a BNN learns distributions over its weights and biases. This allows it to capture not just the relationship in the data, but also its own uncertainty about that relationship. For this, we need to define:

- 1) **A Network Architecture**, which specifies the number of layers, the number of neurons in each layer, and the activation functions (e.g., ReLU, tanh) that introduce non-linearity. This defines the structure of our “knobs.”
- 2) **Priors for Arrays of Weights and Biases.** In a simple model like linear regression, we define a prior for each individual parameter (e.g., one prior for the slope β). In a neural network, which can have thousands or millions of weights, we don’t define a unique prior for every single one. Instead, we define a prior that acts as a template for an entire **array of parameters**. For example, we might declare that all weights in a specific layer are drawn from the same $\text{Normal}(0, 1)$ distribution. This allows us to efficiently specify our beliefs about the entire set of network parameters.
- 3) **An Output Distribution (Likelihood)**, which defines the probability of the data given the network’s predictions. For a continuous variable (regression), this is often a Normal distribution with a variance term σ that quantifies the data’s noise around the model’s predictions.

Considerations

🔥 Caution

- Like all Bayesian models, BNNs consider model parameter uncertainty . The parameters here are the network's **weights (W)** and **biases (b)**. We quantify our uncertainty about them through their posterior distribution . Therefore, we must declare prior distributions for all weights and biases, as well as for the output variance σ .
- Unlike in a linear regression where the coefficient has a direct interpretation (e.g., the effect of weight on height), the individual weights and biases in a BNN are not directly interpretable. A single weight's influence is entangled with thousands of other parameters through non-linear functions. Consequently, BNNs are best viewed as powerful **predictive tools** rather than explanatory ones. They excel at learning complex patterns and quantifying predictive uncertainty, but if the goal is to isolate and interpret the effect of a specific variable, a simpler model is often more appropriate.
- Prior distributions are built following these considerations:
 - As the data is typically scaled (see introduction), we can use a standard Normal distribution (mean 0, standard deviation 1) as a weakly-informative prior for all weights and biases. This acts as a form of regularization.
 - Since the output variance σ must be positive, we can use a positively-defined distribution, such as the Exponential or Half-Normal.
- BNNs can be used for both regression and classification. The final layer's activation and the chosen likelihood distribution depend on the task. For binary classification, a sigmoid activation is paired with a Bernoulli likelihood, which requires a link function (logit) to connect the linear output of the network to the probability space [0, 1]. For regression, the identity activation is often used with a Gaussian likelihood.

Example

Below is an example code snippet demonstrating a Bayesian Neural Network for regression using the Bayesian Inference (BI) package. Data consist of two continuous variables (height and weight), and the goal is to predict height from weight using a non-linear model.

Python

```
from BI import bi

# Setup device-----
m = bi(platform='cpu')

# Import Data & Data Manipulation -----
# Import
from importlib.resources import files
data_path = files('BI.resources.data') / 'Howell1.csv'
m.data(data_path, sep=';')
m.df = m.df[m.df.age > 18] # Manipulate
m.scale(['weight']) # Scale

# Define model -----
def model(weight, height):
    # Define the BNN architecture and get its output (mu)
    # 1 input -> 10 hidden neurons (tanh) -> 1 output neuron (identity)
    # Priors for weights/biases are Normal(0,1) by default
    mu = m.bnn(x=weight, n_neurons=[10, 1], activations=['tanh', 'identity'], name='bnn')

    # Prior for the output standard deviation
    s = m.dist.exponential(1, name='s')

    # Likelihood
    m.normal(mu, s, obs=height)

# Run mcmc -----
m.fit(model) # Approximate posterior distributions for weights, biases, and sigma

# Summary -----
m.summary() # Get posterior distributions
```

R

```
library(BI)
m=importbi(platform='cpu')

# Load csv file
m$data(paste(system.file(package = "BI"),"/data/Howell1.csv", sep = ''), sep=';')
```

```

# Filter data frame
m$df = m$df[m$df$age > 18,]

# Scale
m$scale(list('weight'))

# Convert data to JAX arrays
m$data_to_model(list('weight', 'height'))

# Define model -----
model <- function(height, weight){
  # Define the BNN architecture
  # 1 input -> 10 hidden neurons (tanh) -> 1 output neuron (identity)
  # Priors for weights/biases are Normal(0,1) by default
  mu <- bi$bnn(x = weight, n_neurons = list(10, 1), activations = list('tanh', 'identity'), 

  # Prior for the output standard deviation
  s = bi$dist$exponential(1, name = 's')

  # Likelihood
  m$normal(mu, s, obs = height)
}

# Run mcmc -----
m$run(model) # Approximate posterior distributions

# Summary -----
m$summary()

```

Mathematical Details

Frequentist Formulation

A standard (non-Bayesian) neural network with one hidden layer is defined by forward propagation:

$$h_i = f(X_i W_1 + b_1)$$

$$\hat{Y}_i = g(h_i W_2 + b_2)$$

Where: - Y_i is the predicted output for observation i . - X_i is the input vector for observation i . - W_1, b_1 are the weight matrix and bias vector for the hidden layer. - W_2, b_2 are the weight matrix and bias vector for the output layer. - h_i is the activation of the hidden layer. - f and g are activation functions (e.g., ReLU, tanh, sigmoid, identity). - Parameters W_1, b_1, W_2, b_2 are learned as single optimal values via optimization.

Bayesian Formulation

In the Bayesian formulation, we place priors on all weights and biases and define a likelihood for the output. For a regression task with a one-layer BNN:

$$Y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = g(h_i W_2 + b_2)$$

$$h_i = f(X_i W_1 + b_1)$$

The parameters are now distributions:

$$W_1 \sim \text{Normal}(0, 1)$$

$$b_1 \sim \text{Normal}(0, 1)$$

$$W_2 \sim \text{Normal}(0, 1)$$

$$b_2 \sim \text{Normal}(0, 1)$$

$$\sigma \sim \text{Exponential}(1)$$

Where: - Y_i is the observed dependent variable for observation i . - μ_i is the mean predicted by the network, which is itself a distribution because it is a function of the distributions of weights and biases. - W_1, b_1, W_2, b_2 are the weights and biases, treated as random variables. - σ is the standard deviation of the normal distribution, quantifying observation noise.

Notes

Note

- The primary difference between a *Frequentist* and *Bayesian* neural network lies in how parameters are treated. In the frequentist approach, weights and biases are point estimates found by minimizing a loss function (e.g., via gradient descent). Techniques like *Dropout* or *L2 regularization* are often used to prevent overfitting, which can be interpreted as approximations to a Bayesian treatment. In contrast, the *Bayesian* formulation does not seek a single best set of weights. Instead, it uses methods like MCMC or Variational Inference to approximate the entire posterior distribution for every weight and bias. This provides a principled and direct way to quantify model uncertainty.
- While present an example of non-linear regression, the Bayesian Neural Network can be used for linear regressions as well (keeping in mind that interpretation of the weights are impossible).

```

from BI import bi

# Setup device-----
m = bi(platform='cpu')

# Import Data & Data Manipulation -----
# Import
from importlib.resources import files
data_path = files('BI.resources.data') / 'Howell1.csv'
m.data(data_path, sep=';')
m.df = m.df[m.df.age > 18] # Manipulate
m.scale(['weight']) # Scale

# Define model -----
def model(weight, height):
    # Define the BNN architecture and get its output (mu)
    # 1 input -> 10 hidden neurons (tanh) -> 1 output neuron (identity)
    # Priors for weights/biases are Normal(0,1) by default
    mu = m.bnn(x=weight, n_neurons=[10, 1], activations=['tanh', 'identity'], name='bnn')

    # Prior for the output standard deviation
    s = m.dist.exponential(1, name='s')

    # Likelihood
    m.normal(mu, s, obs=height)

# Run mcmc -----
m.fit(model) # Approximate posterior distributions for weights, biases, and sigma

# Summary -----
m.summary() # Get posterior distributions

```

Reference(s)

(neal1995bayesian?)