Measuring Uncertainty through Bayesian Neural Networks:

An Application in Insurance Claim Cost Prediction

Mai Tra My

Data Science in Economics and Business
Faculty of Mathematical Economics
National Economics University

June 13, 2024

Overview

1. Introduction
2. Bayesian Neural Networks
3. Application in Insurance Claim Cost
4. Conclusion and Perspectives
5. Appendix

Introduction

Uncertainty

”Uncertainty: The Importance of Knowing What We Don’t Know” [Gal, 2016]

1. **Aleatoric uncertainty** arises from the inherent noise in the data, represents the variability in the data that cannot be reduced, even with more data or better models.
2. **Epistemic uncertainty** represents the model’s lack of knowledge.

Aleatoric uncertainty cannot be reduced only identified and quantified, while epistemic uncertainty can be reduced through more comprehensive study.

Measuring Epistemic Uncertainty

- **Bayesian approaches:** Incorporate prior distributions over model parameters and update these distributions in light of the observed data.
- **Recent advanced models**:
  - Uncertainty representation through the joint distribution of predictions [Osband et al., 2021]
  - Conformal prediction
    Straightforward way to generate prediction sets for any model.

Bayesian Neural Networks

Neural Networks

A feed-forward neural network

$x \in \mathbb{R}^p$, $y \in Y$,
m hidden layers, weight w, bias b and a non-linear activation function ϕ:

$$
\begin{aligned}
h^{(1)} &= \phi(w^{(1)}x + b^{(1)}) \\
h^{(2)} &= \phi(w^{(2)}h^{(1)} + b^{(2)}) \\
&\vdots \\
h^{(m)} &= \phi(w^{(m)}h^{(m-1)} + b^{(m)}) \\
y^* &= g(w^{(m+1)}h^{(m)} + b^{(m+1)}).
\end{aligned}
$$

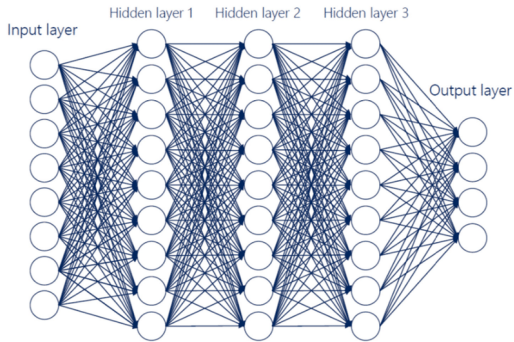


Figure: An Artificial Neural Network

Limitations of Neural Networks

- A lot of labeled data is necessary
- Optimization is difficult and time-consuming (many parameters, initialization, etc.)
- Not transparent: Black boxes
- Sensitive to noisy data
- No guarantee for predictions

Sources of uncertainty

- **Noisy data**: The observed labels might be noisy, leading to uncertainty in predictions.
- **Model parameter uncertainty**: When multiple models can effectively explain a dataset, choosing the optimal model parameters for prediction becomes uncertain (uncertainty in cognitive aspect).
- **Uncertainty about model architecture**: This uncertainty is related to the choice of model architecture itself.

Bayesian Neural Networks

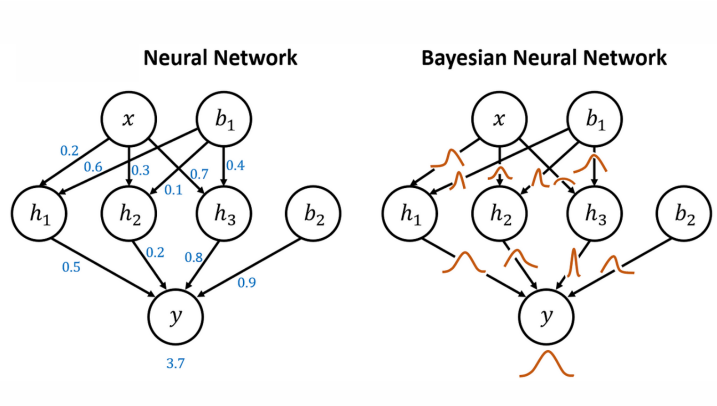


Figure: Traditional Neural Network vs. Bayesian Neural Network
(Source: Internet)

The weights W are represented as distributions rather than fixed (unknown) values, and the biases B are also represented as distributions:

$$h^{(1)} = \phi(W^{(1)}x + B^{(1)})$$
$$h^{(2)} = \phi(W^{(2)}h^{(1)} + B^{(2)})$$
$$\vdots$$
$$h^{(m)} = \phi(W^{(m)}h^{(m-1)} + B^{(m)})$$
$$y^* = g(W^{(m+1)}h^{(m)} + B^{(m+1)}).$$

Bayesian Learning

- (Training) Given data $D = (X, Y)$, prior distribution $p(\theta)$ of the model parameter, and the likelihood distribution $p(D|\theta)$:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)} = \frac{p(\theta)\prod_i p(y_i|x_i,\theta)}{p(D)} \tag{1}$$

- (Inference) Predict a new data point (x^*, y^*) following the posterior predictive distribution:

$$\begin{aligned} p(y^*|x^*, D) &\approx E_{p(\theta|D)}[p(y^*|x^*,\theta)] \\ &= \int_\Theta p(y^*|x^*,\theta)P(\theta|D)d\theta \end{aligned} \tag{2}$$

Computation Methods

	MCMC	Variational Inference
Examples	Gibbs Sampling Metropolis Hasting Hamiltonian MC	Stochastic Variational Inference Bayes by Backprop

Table: Some Computation Bayesian methods

Variational Inference

KL divergence and ELBO

Use KL divergence for continuous distributions - the posterior p and the variational q:

$$
\begin{aligned}
KL(q(\theta) \,||\, p(\theta|D)) &= \mathbb{E}_q \left[log \frac{q(\theta)}{p(\theta|D)} \right] \\
&= -\mathbb{E}_q \left[log \frac{p(\theta|D)}{q(\theta)} \right] \\
&= -\left(\mathbb{E}_q \left[log \frac{p(\theta, D)}{q(\theta)} \right] - \mathbb{E}_q [log p(D)] \right) \\
&= -(\mathbb{E}_q \left[log p(\theta, D) - log q(\theta) \right]) + log p(D)
\end{aligned}
$$

i.e, $\boxed{KL\left(q(\theta) \,||\, p(\theta|D)\right) = log p(D) - ELBO}$

Regression and Curve Fitting

$$y(x) = x + 0.3 * \sin\left[2 * \pi * (x + \epsilon_1)\right] + 0.3 * \sin\left[4 * \pi * (x + \epsilon_1)\right] + \epsilon_2$$
$$= f(x, \epsilon_1) + \epsilon_2, \quad \text{where } \epsilon_1, \epsilon_2 \sim N(0, 0.02).$$

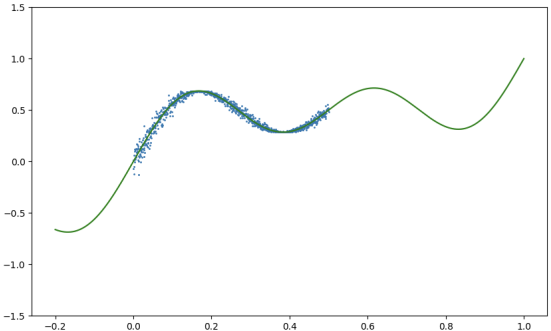


Figure: Simulated Data

Regression and Curve Fitting

$$y(x) = x + 0.3 * \sin\left[2 * \pi * (x + \epsilon_1)\right] + 0.3 * \sin\left[4 * \pi * (x + \epsilon_1)\right] + \epsilon_2$$
$$= f(x, \epsilon_1) + \epsilon_2, \quad \text{where } \epsilon_1, \epsilon_2 \sim N(0, 0.02).$$

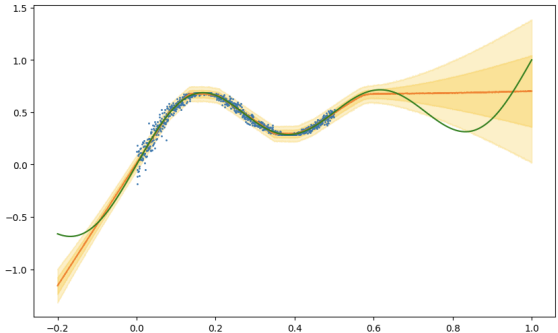


Figure: BNN Predictions

Classification

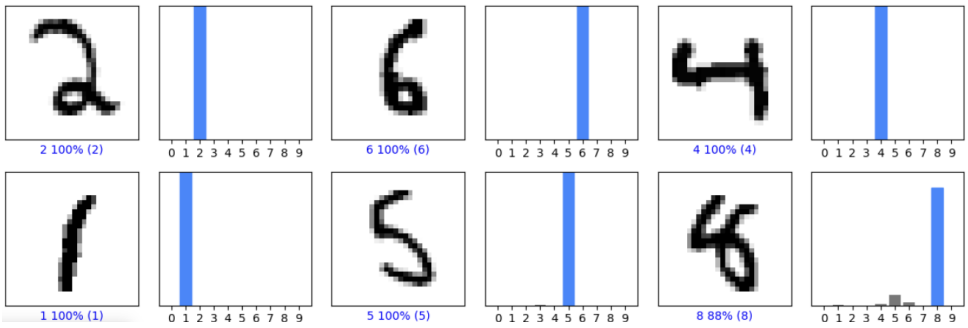


Figure: MNIST: CNN Prediction

Classification

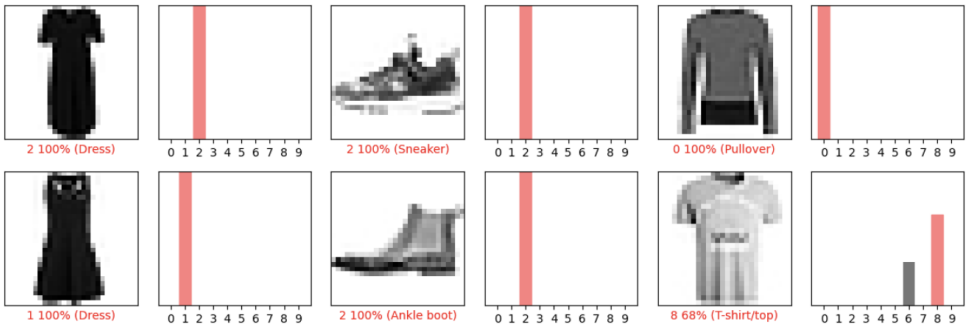
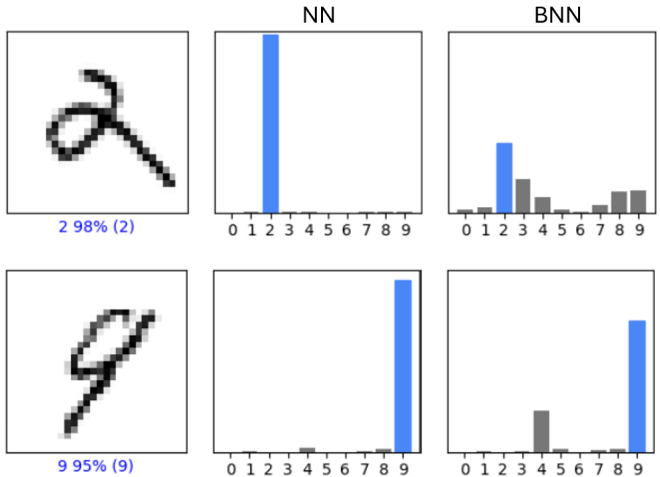
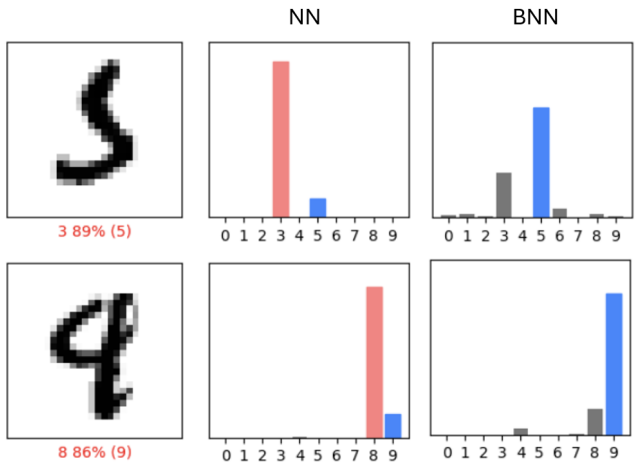


Figure: Fashion-MNIST: CNN Prediction

Classification



Classification



Classification

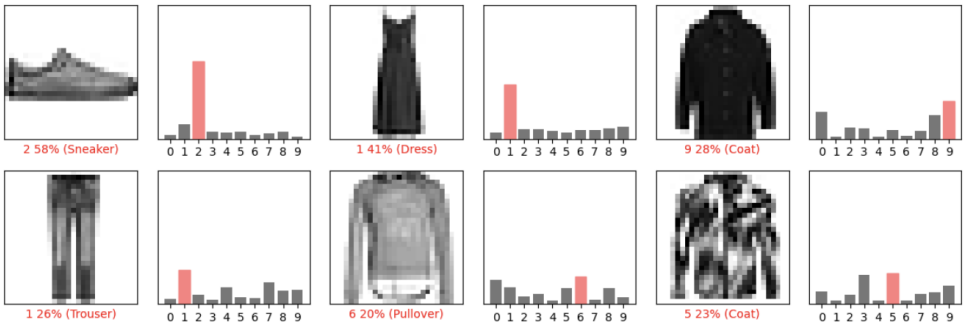
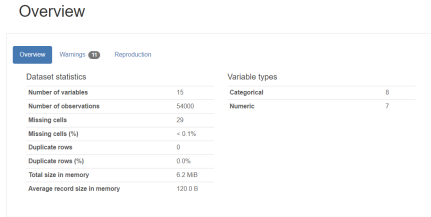


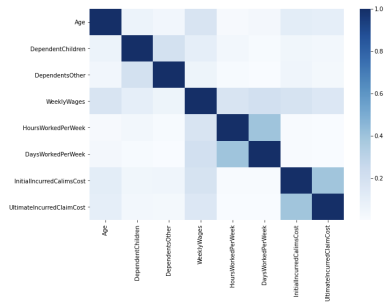
Figure: Fashion MNIST: Bayesian CNN Prediction

Application in Insurance Claim Cost

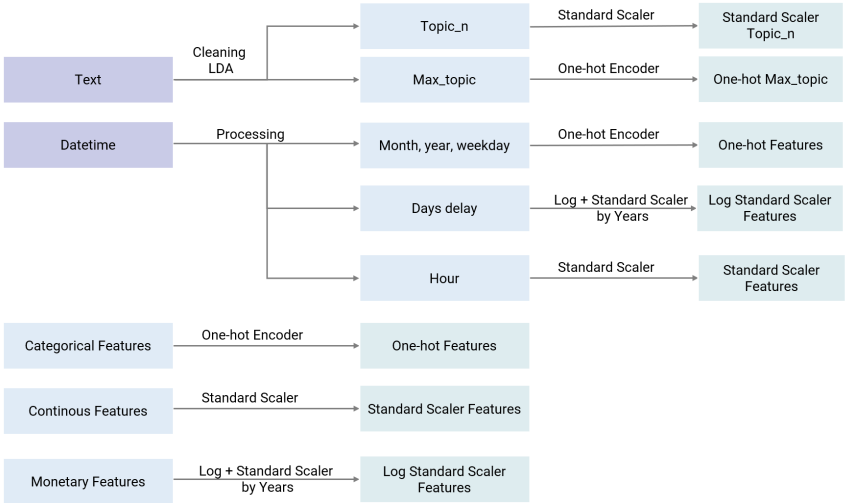
Context

- The Actuarial Loss Prediction Competition 2020/21.
- 15 columns with 54000 records from 1988 to 2005.
  Target: Ultimate Claim Cost
- Data includes categorical, numerical, datetime and text features.
- Skewed: Initial & Ultimate Claim Cost, Hours Worked Per Week





Feature Engineering



Approaches to the Claim Cost Prediction

- *Traditional Statistical Methods:* GLMs, GAMs, etc.
  Traditional statistical methods need many assumptions about the distribution of the data.
- *Machine Learning:* Decision trees and their ensemble forms
- *Deep Learning:* Feedforward Neural Networks, RNNs and LSTMs

Results

- Out-of-distribution observations have a much wider IQR, with more outliers.
- Normal inputs have a much smaller range.

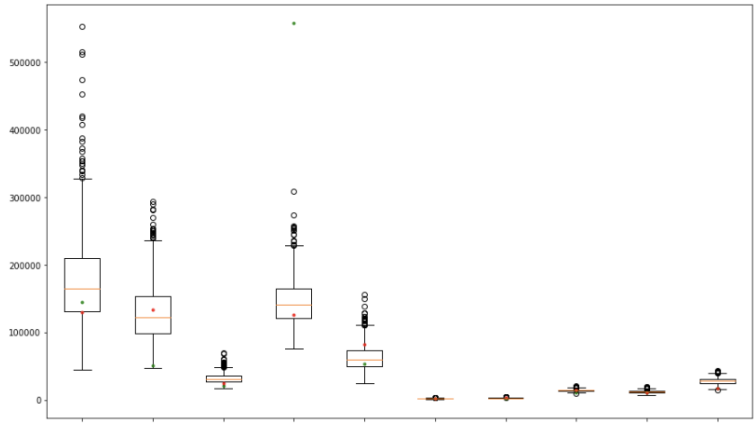


Figure: Boxplots for the Sample's Result

Robustness

- BNN and NN provide better results and common GLM (not specifically adapted) lag behind.

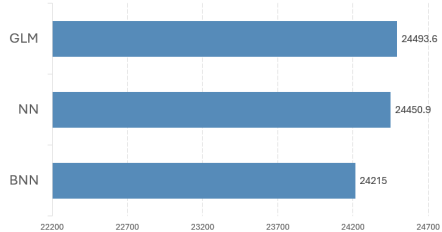


Figure: Comparing RMSE

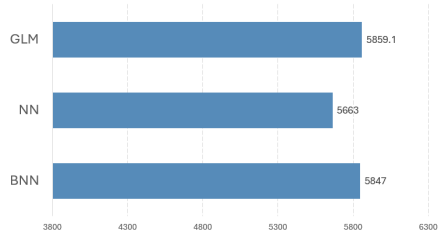


Figure: Comparing MAE

Time Consuming

- BNNs require a much longer time to converge for training
- GLM has the fastest training time among the three models

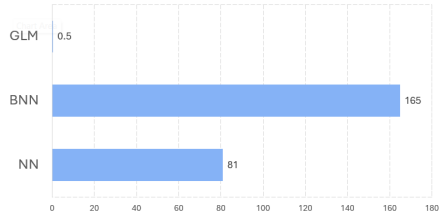


Figure: Comparing training time(s)

Conclusion and Perspectives

Conclusion and Perspectives

It can be seen that:

- BNNs offer robust uncertainty estimates.
- Results are promising in terms of performance.
- We also observe BNNs drawbacks in training/inference time.

Several perspectives can be discussed:

- Tune the model in order to have better performance.
- The analysis of feature importance.
- Experiments with other approaches to uncertainty quantification
- Integration such uncertainty measures within daily processes.

References

Yarin Gal (2016)
Uncertainty in Deep Learning
PhD. Thesis of Department of Engineering, University of Cambridge

Osband (2021)
Epistemic Neural Networks
NeurIPS 2023

Angelopoulos & Bates (2021)
A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification
Journal Name 12(3), 45 – 678.

Thank you for your attention

Mai Tra My

Data Science in Economics and Business
Faculty of Mathematical Economics
National Economics University

June 13, 2024

Appendix

NN - Activation Functions

- **Sigmoid function**

$$\phi(x) = \frac{1}{1 + e^{-x}}$$

- **Hyperbolic tangent**

$$\phi(x) = tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-1}}$$

→ Vanishing gradients

- **Rectified Linear Units (ReLU)**

$$\phi(x) = ReLU(x) = max(0, x)$$

NN - Regularizations

- **L^1-regularization** (Lasso)
  Penalize large weights in terms of absolute value with parameter $\lambda \in R^+$.

$$L(y^*|X^*, w) + \lambda \sum_i^m \sum_j^{h_i} |w_{i,j}|$$

- **L^2-regularization** (Ridge regression)
  Penalize large weights by adding L^2-penalty to the model loss function.

$$L(y^*|X^*, w) + \lambda \sum_i^m \sum_j^{h_i} (w_{i,j}^2)$$

- **Dropout**

NN - Regularizations

- L^1-**regularization** (Lasso)
- L^2-**regularization** (Ridge regression)
- **Dropout**
  A special case of Probabilistic Deep Learning
  Monte Carlo dropout is widely popular for its simplicity and effectiveness.

Bayesian Learning

The Law of Total Variance

$$Var(X) = E[Var(X|Y)] + Var(E[X|Y])$$

Factoring the variance into two terms:

$$Var(p(y^*|x^*, D)) = E[Var(p(y^*|x^*, D, \theta)] + Var(E[p(y^*|x^*, D, \theta])$$

Variational Inference

Complicated posterior distributions of the weights ≈ a simple distribution (variational distribution).

Evidence Lower Bound (ELBO)

Also called variational lower bound. Starting with log of evidence, we have:

$$
\begin{aligned}
log p(D) = log \int_{\theta} p(D, \theta) &= log \int_{\theta} p(D, \theta) \frac{q(\theta)}{q(\theta)} \\
&= log \mathbb{E}_q \left[\frac{p(D, \theta)}{q(\theta)} \right] \\
&\geq \mathbb{E}_q \left[log \frac{p(D, \theta)}{q(\theta)} \right] = \mathbb{E}_q \left[log p(D, \theta) \right] - \mathbb{E}_q \left[q(\theta) \right]
\end{aligned}
$$

Results of BNN vs NN - Data points near mean of distribution

- More consistent data with fewer outliers and lower variance.