Performance analysis of Bayesian neural networks with different a priori weights on different datasets

ШуСяохэн
Мех-Мат
Магистра -1
группа: МаТИС
2023-05-03

Код:https://github.com/MachineShu/porject-mldl---2023/tree/main

Тескт:https://github.com/MachineShu/porject-mldl---2023/blob/main/README.md

CONTENTS

01

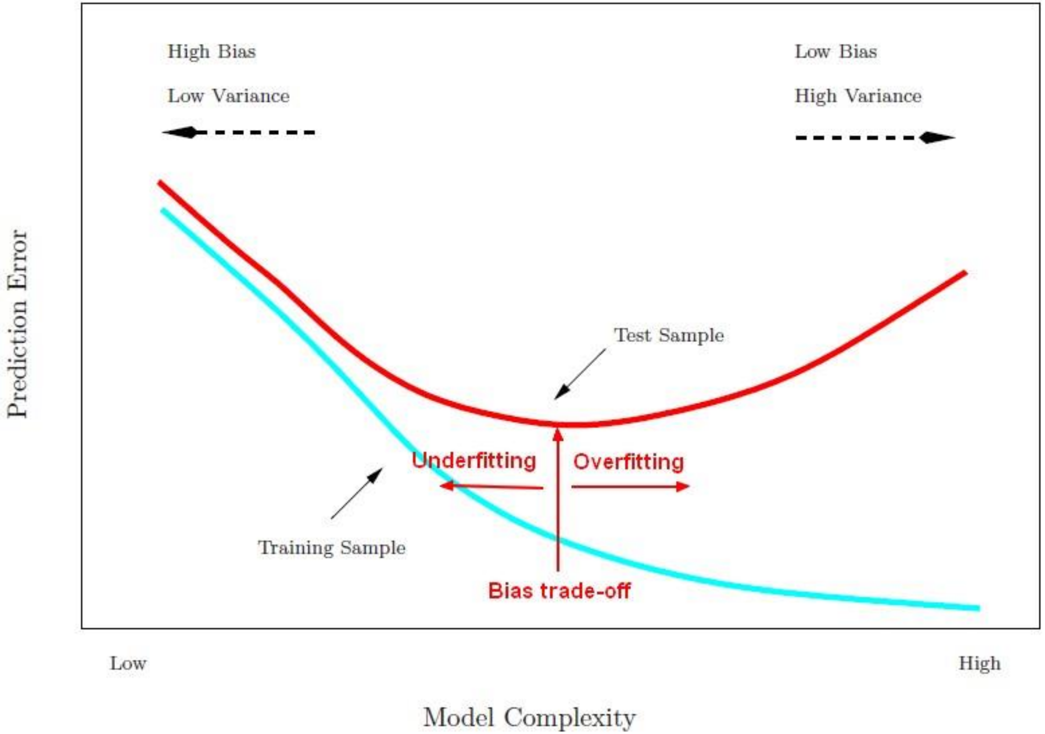
1.Research Background

1.Research Background

- Problems with traditional neural networks
  - overlearning
  - model uncertainty
- Bayesian Neural Network (BNN)

overlearning

"Overlearning" refers to a a phenomenon where a machine learning model becomes too specialized on on the training data, to the the point where it performs performs poorly on new, unseen data.

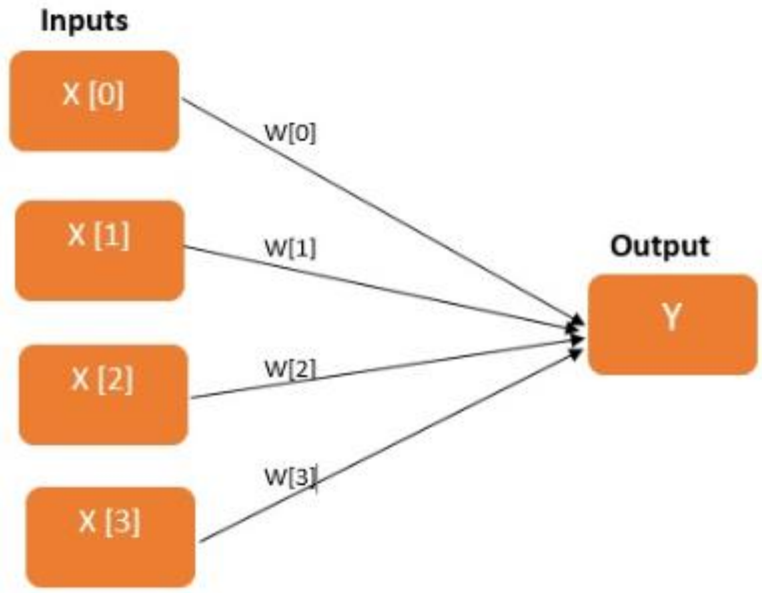


model uncertainty

The term "model uncertainty" refers to the uncertainty of a model and can also be referred to as "model uncertainty estimation".

Bayesian Neural Network (BNN)

Bayesian Neural Network is a neural network model that uses Bayesian statistics to infer the weights and biases of the model, allowing the model to quantify the uncertainty of the prediction.



02

2.Related Research

2.Related Research

BNNs combine neural networks and Bayesian methods.

Many studies have explored the impact of prior weights and learning methods on BNN performance.

Gaussian prior weights work well on the MNIST dataset, while Laplace prior weights work better on the CIFAR-10 dataset.

Variational inference methods and Markov chain Monte Carlo methods are two learning methods that affect BNN performance. In my project, I test the performance of different prior weights and learning methods on various datasets based on existing studies.

03

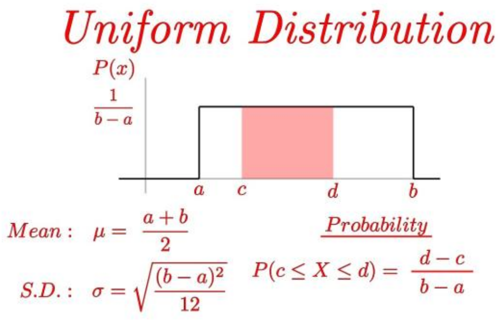
3.Model Design

3.Model Design

- Convolutional neural network-based model architecture
- Bayesian weights
- Weight design with different prior distributions
- Model implementation using the PyTorch framework

Weight design with different prior distributions

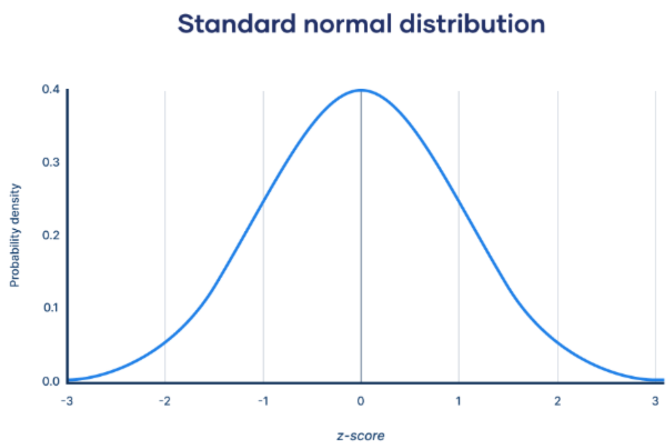
- uniform distribution
- Gaussian distribution
- Laplace distribution



“

"Uniform distribution" is a probability distribution where all values specified have an equal probability. Hence, this distribution is commonly used for random selection where each value has an equal chance of being selected.

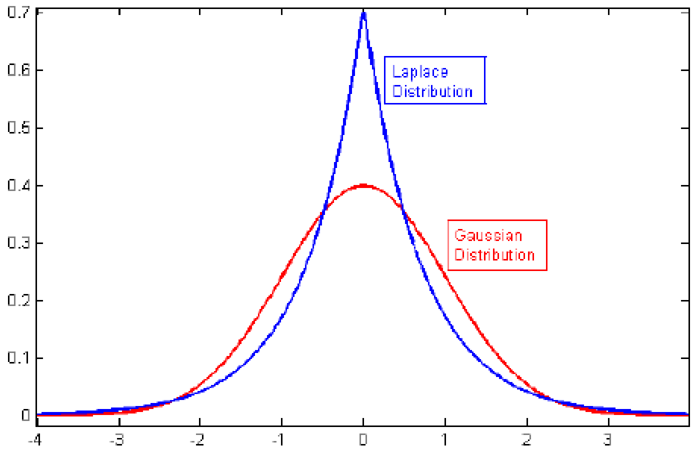
”



“

Gaussian distribution, also known as normal distribution, is a commonly used continuous probability distribution in probability theory and statistics. Its graph has a bell shape, so it is also called a bell curve.

”



> “As to Laplace distribution, also known as double exponential distribution, is a commonly used probability distribution. It is the symmetric version of the exponential distribution, and its probability density function has a sharp peak and a steep shape.”

04

4.Dataset and Pre-processing

4.Dataset and Pre-processing

Dataset description

Data pre-processing methods

Dataset description

MNIST

Omniglot

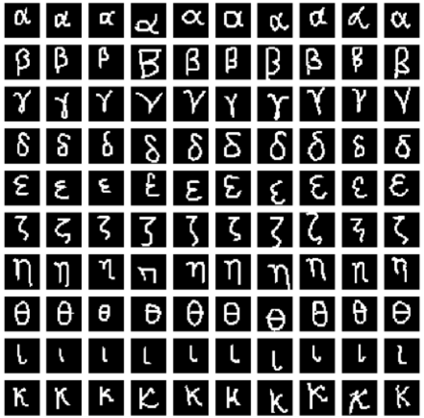
CIFAR-10

MNIST

- MNIST stands for "Modified National Institute of Standards and Technology" and refers to a dataset of images of handwritten digits. It is a popular benchmark dataset for image classification tasks in machine learning and computer vision.

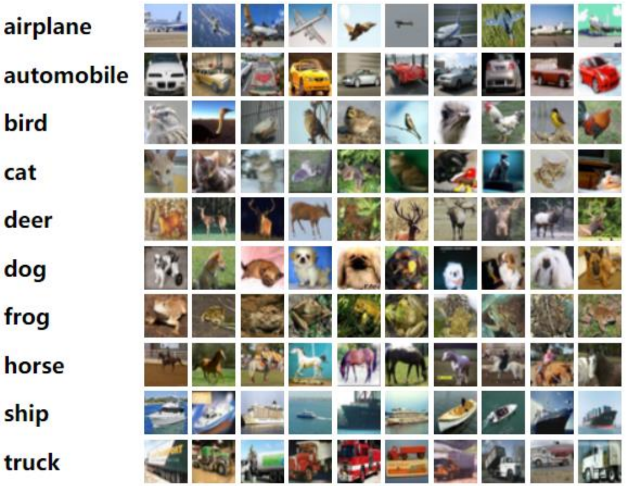
Omniglot

- Omniglot is a dataset of handwritten characters from 50 languages. Each character has only 20 examples, with 15 for training and 5 for testing. Its small size and few samples per character make it a popular benchmark for testing models' generalization ability under few-shot learning settings.



CIFAR-10

- The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.



Data pre-processing methods

- Image scaling
- Normalization
- Data augmentation

05

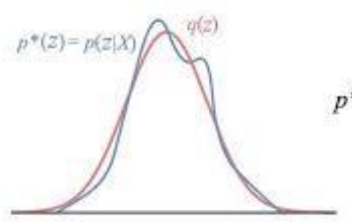
5.Training Methods

5.Training Methods

Variational Inference (VI) method ⟷ Markov chain Monte Carlo (MCMC) method

Variational Inference (VI) method

- Variational inference is a method for approximating the posterior probability distribution. It uses a simple distribution with parameters (e.g. Gaussian) to represent the posterior probability and minimizes the difference between this and the true posterior probability (KL divergence).

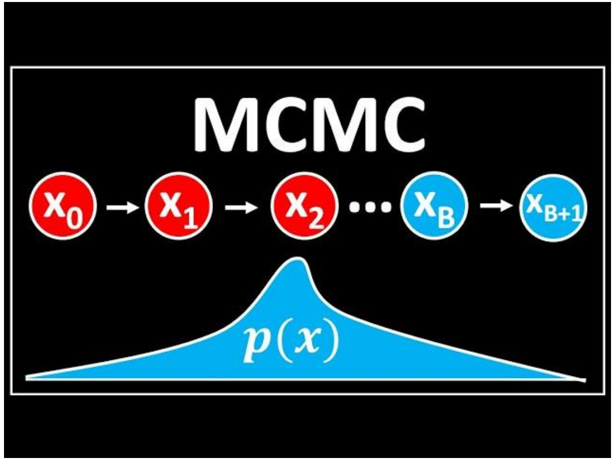


$$p^*(z) = p(z|X) = \frac{p(X,z)}{p(X)}$$

“

The idea is to create a Markov chain with a smooth distribution to be sampled, and use the properties of the chain to eventually converge to the required smooth distribution. This is commonly used in Bayesian statistics for posterior probability distributions and model parameters.

”



06

6.Experiment Results

6.Experiment Results

Comparison of performance using different prior weights and learning methods

Advantages of Gaussian prior weights on simpler datasets

Advantages of Laplace prior weights on complex datasets

Applicability of MCMC and VI methods under different circumstances

Comparison of performance using different prior weights and learning methods

Accuracy of experimental results table
Variable Inference

datasets	Gaussian	Laplace	Equalisation
MNIST	47.8%	47%	47.2%
Omniglot	40.7%	42.1%	45.6%
Cifar-10	43.6%	43.2%	42.3%

- Results showed that different learning methods have advantages depending on the dataset and pre-weights used.

Comparison of performance using different prior weights and learning methods

Accuracy of experimental results table MCMC

datasets	Gaussian	Laplace	Equalisation
MNIST	9.4%	9.1%	10.1%
Omniglot	10.27%	9.8%	9.1%
Cifar-10	10.0%	9.4%	10.4%

- Results showed that different learning methods have advantages depending on the dataset and pre-weights used.

Advantages of Gaussian prior weights on simpler datasets

The Gaussian distribution is smoother and therefore better suited to describe the **uncertainty** in the weights. This works better on **small datasets** where the number of data points is limited and therefore prior knowledge of the weights can provide more information.

It can be optimised by maximum a posteriori probability estimation, a regularisation method that **avoids overfitting**. This is very effective for simple datasets that are prone to overfitting.

Gaussian prior weights provide an intuitive way to **control the complexity of the model**. Using smaller variances reduces the degrees of freedom of the model and thus the risk of overfitting. This is very useful on simple datasets where overly complex models can lead to over-fitting.

Advantages of Laplace prior weights on complex datasets

The Laplace prior allows the neural network parameters to be more **sparse** and **less overfitted** than a Gaussian prior through its "spikes". In particular, on high-dimensional datasets, the Laplace prior can better **control the complexity of the model**.

The Laplace prior can better **handle outliers** by generating sharper losses, allowing the model to focus more on samples with large losses and reducing the impact of outliers.

The tils of the Laplace distribution are **longer** than those of the Gaussian distribution, which makes the effect of the Laplace prior on robustness more pronounced.

Applicability of MCMC and VI methods under different circumstances

- MCMC：
  - Markov chain Monte Carlo methods comprise a class of algorithms for sampling from a probability distribution. By constructing a Markov chain that has the desired distribution as its equilibrium distribution, one can obtain a sample of the desired distribution by recording states from the chain.
- VI：
  - Variational Bayesian methods are a family of techniques for approximating intractable integrals arising in Bayesian inference and machine learning.

07

7.Conclusion

7.Conclusion

- High performance and uncertainty estimation of Bayesian neural networks
- Advantages of training methods

Advantages of training methods

Different training methods have different advantages with different different datasets and prior weights weights

We can choose the appropriate prior weight assignment and training method according to the specific characteristics of the problem

THE END

THANKS