CS6462

Probabilistic and Explainable AI

Lesson 21

Bayesian Neural Networks

*

Posterior Variational Inference

by Emil Vassev

March 10, 2025

BNN Probabilistic Models

Bayesian Neural Network:

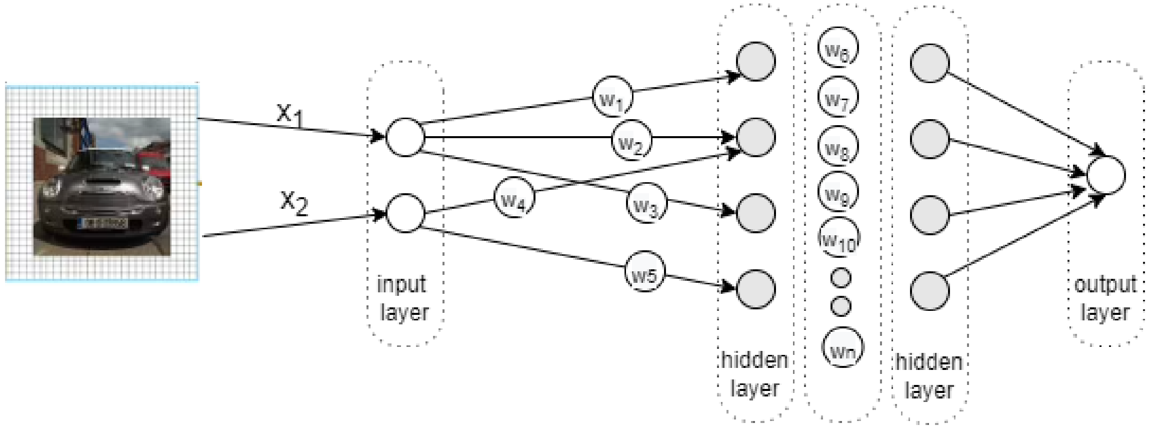
- based on Bayes' Theorem
- probabilistic model of BNN

  $p(y|x,w)$

  - inputs $x = \{x_1, x_2, \dots, x_n\}$
  - weight factors $w = \{w_1, w_2, \dots, w_n\}$
  - for classification:
    - y - set of classes
    - $p(y|x,w)$ - Categorical distribution
  - for regression:
  - y - continuous random variable
    - $p(y|x,w)$ - Gaussian distribution

$$Posterior = \frac{Likelyhood * Prior}{Evidence} \propto Likelyhood * Prior$$

$$p(w|D) = \frac{p(D|w) * p(w)}{p(D)} \propto p(D|w) * p(w)$$



BNN Probabilistic Models (cont.)

BNN Probabilistic model: $p(y|x, w)$

Bayesian Neural Network Training:

- training dataset

$D = \{x^{(i)}, y^{(i)}\}$

- $x^{(i)}$ - the input vector of the i-th training example
- $y^{(i)}$ - class label for the i-th training example

Maximum Likelihood Estimate (MLE):

- likelihood distribution:

  $p(D|w) = \prod_i p\left(y^{(i)}\middle|x^{(i)}, w\right)$ - function of the weight factors w

  maximizing $p(D|w)$ - Maximum Likelihood Estimate (MLE) of w

- optimization objective – negative log likelihood:
  - Categorical distribution: Cross Entropy Error function
  - Gaussian distribution: proportional to the sum of Square Error function
  - MLE can lead to heavy overfitting

BNN Probabilistic Models (cont.)

Maximum a Posteriori (MAP) Estimate:

BNN Probabilistic model: $p(y|x,w)$

- $p(w|D)$ - posterior distribution
- product $p(D|w) * p(w)$ - proportional ($\propto$) to $p(w|D)$

$$p(w|D) \propto p(D|w) * p(w)$$

- maximizing $p(D|w) * p(w)$ - Maximum a Posteriori (MAP) estimate of w
- computing MAP - can prevent overfitting
- optimization objective:
  - negative log likelihood
  - regularization term with log prior

Posterior Predictive Distribution:

$$p(y|x,D) = \int p(y|x,w) * p(w|D) * dw$$

- full posterior distribution over parameters - predictions with weight uncertainty into account
- parameters w are marginalized

Variational Inference

$$p(y|x, D) = \int p(y|x, w) * p(w|D) * dw$$

Posterior with a variational distribution:

- $p(w|D)$ - posterior distribution
- difficult analytical solution to $p(w|D)$ in BNN
- solution: approximate the true posterior with a variational distribution
- $q(w|\theta)$ - variational distribution;
- θ – set of parameters we want to estimate

- new posterior with variational distribution:
  - cost function $\boldsymbol{F}(\boldsymbol{D}, \boldsymbol{\theta})$
  - minimizes the *Kullback–Leibler divergence** between $q(w|\theta)$ and $p(w|D)$

*Kullback–Leibler divergence on Wikipedia

Variational Inference(cont)

Cost function (variational free energy):

*Kullback–Leibler divergence**

- corresponding optimization objective or cost function (*variational free energy*):

$$\boldsymbol{F(D,\theta) = KL\big(q(w|\theta) \parallel p(w)\big) - \mathbb{E}_{q(w|\theta)}[\log(p(D|w))]}$$

$$F(D,\theta) = \textit{complexity cost} - \textit{likelyhood cost}$$

$KL\big(q(w|\theta) \parallel p(w)\big)$ *- measures the statistical distance how* $q(w|\theta)$ *is different from* $p(w)$

$\mathbb{E}_{q(w|\theta)}[\log(p(D|w))]$ - likelihood cost: the expected value of $p(D|w)$ with respect to $q(w|\theta)$

$\mathbb{E}_{q(w|\theta)}$ - energy expectation function

*Kullback–Leibler divergence on Wikipedia

Variational Inference (cont.)

Cost function:

$$F(D,\theta) = KL\big(q(w|\theta) \parallel p(w)\big) - \mathbb{E}_{q(w|\theta)}[\log(p(D|w))]$$

- rearranging the complexity cost component:

$$\boldsymbol{F(D,\theta) = KL\big(q(w|\theta) \parallel p(w)\big) - \mathbb{E}_{q(w|\theta)}[\log(p(D|w))] =}$$

$$\boldsymbol{F(D,\theta) = \mathbb{E}_{q(w|\theta)}[\log(q(w|\theta))] - \mathbb{E}_{q(w|\theta)}[\log(p(w))] - \mathbb{E}_{q(w|\theta)}[\log(p(D|w))]}$$

- all three terms are energy expectations with respect to variational distribution $q(w|\theta)$

Variational Inference (cont.)

Cost function (cont.):

$$\boldsymbol{F(D,\theta) = \mathbb{E}_{q(w|\theta)}[\log(q(w|\theta))] - \mathbb{E}_{q(w|\theta)}[\log(p(w))] - \mathbb{E}_{q(w|\theta)}[\log(p(D|w))]}$$

- cost function $F(D,\theta)$ can be approximated by drawing samples $w^{(i)}$ from $q(w|\theta)$

$$\boldsymbol{F(D,\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\left[log\left(q\left(w^{(i)}\middle|\theta\right)\right) - log\left(p\left(w^{(i)}\right)\right) - log\left(p\left(D\middle|w^{(i)}\right)\right)\right]}$$

Example:

- Gaussian distribution for the variational posterior $q(w|\theta)$, parameterized by $\boldsymbol{\theta=\{\mu,\sigma\}}$
- $\boldsymbol{\mu}$ - mean vector of the distribution
- $\boldsymbol{\sigma}$ - standard deviation vector

We do not parameterize BNN on weights directly but on $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ – we double the number of parameters compared to a plain BNN.

Summary

Bayesian Neural Networks – *Posterior Variational Inference*

BNN Probabilistic Model: $p(y|x, w)$

Bayesian Neural Network Training

- *Training Dataset:* $D = \{x^{(i)}, y^{(i)}\}$
- *Maximum Likelihood Estimate (MLE)*
- *Maximum a Posteriori Estimate (MAP)*

Posterior Variational Inference

- *Variational Distribution:* $q(w|\theta)$
- *Cost Function:* $F(D, \theta) = KL\big(q(w|\theta) \parallel p(w)\big) - \mathbb{E}_{q(w|\theta)}[log(p(D|w))]$

Next Lesson:

- Causal Inference

Thank You!

Questions?