Bayesian Neural Network

Abdo NASSER ALI

Multitel

June 2023

Outline

Motivation

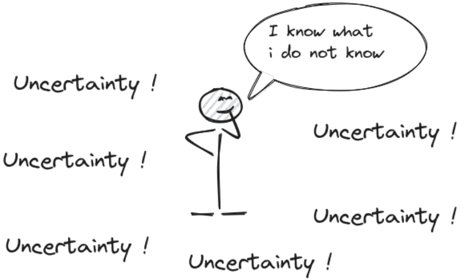
- Fast development of research and engineering in ML.
- AI applications have emerged in various sectors in the real world, including:
  - **Medical**: AI-powered medical imaging analysis for faster decision-making, etc.
  - **Transportation**: Self-driving cars, traffic management systems, etc.
  - **Finance**: Fraud detection systems, algorithmic trading systems, etc.

Motivation

- Fast development of research and engineering in ML.
- AI applications have emerged in various sectors in the real world, including:
  - **Medical**: AI-powered medical imaging analysis for faster decision-making, etc.
  - **Transportation**: Self-driving cars, traffic management systems, etc.
  - **Finance**: Fraud detection systems, algorithmic trading systems, etc.
- However, all of these advancements can be quite dangerous without proper supervision.

Motivation

- Fast development of research and engineering in ML.
- AI applications have emerged in various sectors in the real world, including:
  - **Medical**: AI-powered medical imaging analysis for faster decision-making, etc.
  - **Transportation**: Self-driving cars, traffic management systems, etc.
  - **Finance**: Fraud detection systems, algorithmic trading systems, etc.
- However, all of these advancements can be quite dangerous without proper supervision.



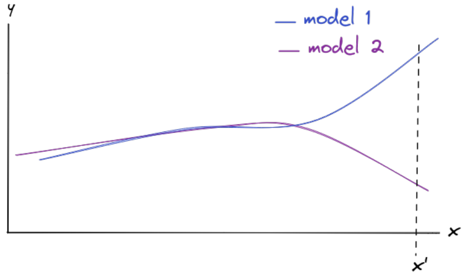
Example: Medical Diagnosis

- **cancAIr**: a new startup.
  - Specializes in detecting melanoma.
  - Achieves 98% accuracy on their 1000 train and test samples.
  - Trains two different models to detect and predict melanoma given input images.

Example: Medical Diagnosis

- **cancAIr**: a new startup.
  - Specializes in detecting melanoma.
  - Achieves 98% accuracy on their 1000 train and test samples.
  - Trains two different models to detect and predict melanoma given input images.

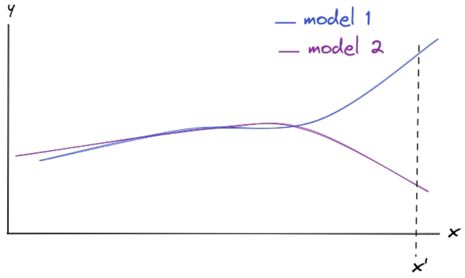
The new patient has been diagnosed by both models.



Example: Medical Diagnosis

- **cancAIr**: a new startup.
  - Specializes in detecting melanoma.
  - Achieves 98% accuracy on their 1000 train and test samples.
  - Trains two different models to detect and predict melanoma given input images.

  The new patient has been diagnosed by both models.

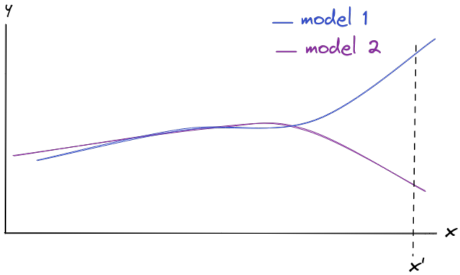


- Which model (Model 1 or 2) would you want the startup to choose for the patient's diagnosis?

Example: Medical Diagnosis

- **cancAIr**: a new startup.
  - Specializes in detecting melanoma.
  - Achieves 98% accuracy on their 1000 train and test samples.
  - Trains two different models to detect and predict melanoma given input images.

The new patient has been diagnosed by both models.



- Which model (Model 1 or 2) would you want the startup to choose for the patient's diagnosis? **None of these of course !**

Example: Self-driving cars

- In 2016, Tesla's assisted driving failed to distinguish the white side of the tractor-trailer against a brightly lit sky.

A Tragic Loss

The Tesla Team, June 30, 2016

We learned yesterday evening that NHTSA is opening a preliminary evaluation into the performance of Autopilot during a recent fatal crash that occurred in a Model S. This is the first known fatality in just over 130 million miles where Autopilot was activated. Among all vehicles in the US, there is a fatality every 94 million miles. Worldwide, there is a fatality approximately every 60 million miles. It is important to emphasize that the NHTSA action is simply a preliminary evaluation to determine whether the system worked according to expectations.

Following our standard practice, Tesla informed NHTSA about the incident immediately after it occurred. What we know is that the vehicle was on a divided highway with Autopilot engaged when a tractor trailer drove across the highway perpendicular to the Model S. Neither Autopilot nor the driver noticed the white side of the tractor trailer against a brightly lit sky, so the brake was not applied. The high ride height of the trailer combined with its positioning across the road and the extremely rare circumstances of the impact caused the Model S to pass under the trailer, with the bottom of the trailer impacting the windshield of the Model S. Had the Model S impacted the front or rear of the trailer, even at high speed, its advanced crash safety system would likely have prevented serious injury as it has in numerous other similar incidents.

Tesla Confirms Automated Driving Systems Were Engaged During Fatal Crash

The February 2023 crash left the driver dead and four firefighters injured.

By **Andy Kalmowitz** Published April 18, 2023 | Comments (59)

Photo: Contra Costa County Fire Protection District (AP)

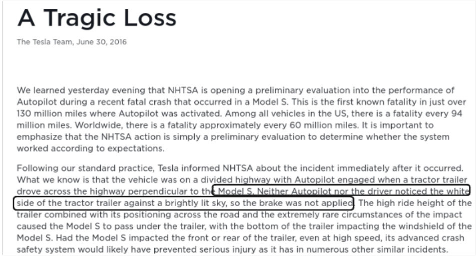
Fun

EVs and Hybrids

Make

Example: Self-driving cars

- In 2016, Tesla's assisted driving failed to distinguish the white side of the tractor-trailer against a brightly lit sky.



A Tragic Loss

The Tesla Team, June 30, 2016

We learned yesterday evening that NHTSA is opening a preliminary evaluation into the performance of Autopilot during a recent fatal crash that occurred in a Model S. This is the first known fatality in just over 130 million miles where Autopilot was activated. Among all vehicles in the US, there is a fatality every 94 million miles. Worldwide, there is a fatality approximately every 60 million miles. It is important to emphasize that the NHTSA action is simply a preliminary evaluation to determine whether the system worked according to expectations.

Following our standard practice, Tesla informed NHTSA about the incident immediately after it occurred. What we know is that the vehicle was on a divided highway with Autopilot engaged when a tractor trailer drove across the highway perpendicular to the Model S. Neither Autopilot nor the driver noticed the white side of the tractor trailer against a brightly lit sky, so the brake was not applied. The high ride height of the trailer combined with its positioning across the road and the extremely rare circumstances of the impact caused the Model S to pass under the trailer, with the bottom of the trailer impacting the windshield of the Model S. Had the Model S impacted the front or rear of the trailer, even at high speed, its advanced crash safety system would likely have prevented serious injury as it has in numerous other similar incidents.

Tesla Confirms Automated Driving Systems Were Engaged During Fatal Crash

The February 2023 crash left the driver dead and four firefighters injured.

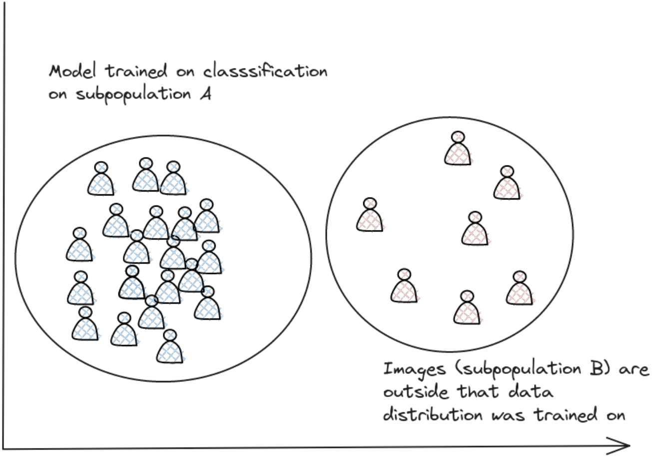
By Andy Kalmowitz Published April 18, 2023 | Comments (59)

Photo: Contra Costa County Fire Protection District (AP)

- If the system had identified its own **uncertainty**? it could have alerted the user to take control.

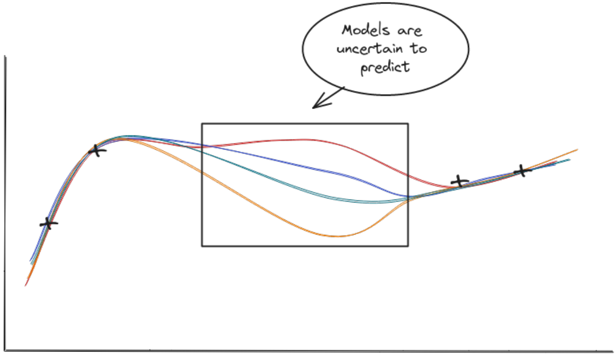
Sources of uncertainty

- **Test data** may be very dissimilar from the training dataset.



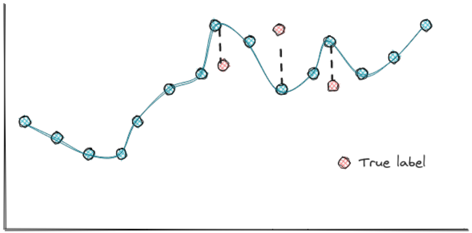
Sources of uncertainty

- **Uncertainty in model parameters**: A large number of models can explain a dataset.



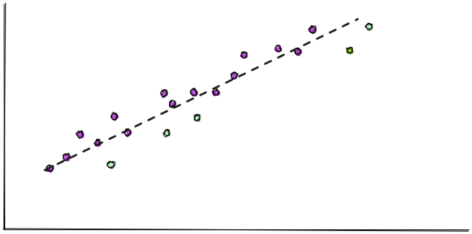
Sources of uncertainty

- **Training labels** are noisy.
  - measurement imprecision (sensor).
  - expert mistakes.



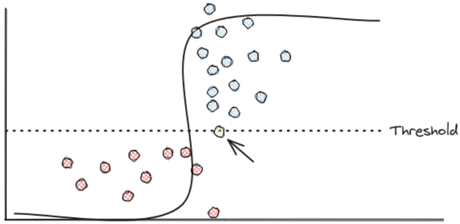
Deep learning models are deterministic

- Classic deep learning does not capture uncertainty
  - Regression models output as single value.



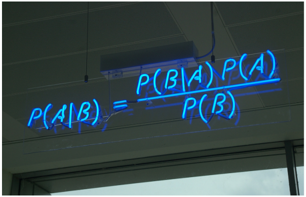
Deep learning models are deterministic

- Classic deep learning does not capture uncertainty.
  - classification models output a probability vector.



Deep learning models are deterministic

- Classic deep learning does not capture uncertainty.
  - Regression models output as single value.
  - classification models output a probability vector
- Combination of **probability theory** and **deep learning** capture uncertainty, known as **Bayesian deep learning**.



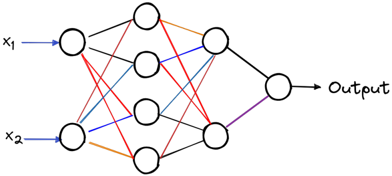
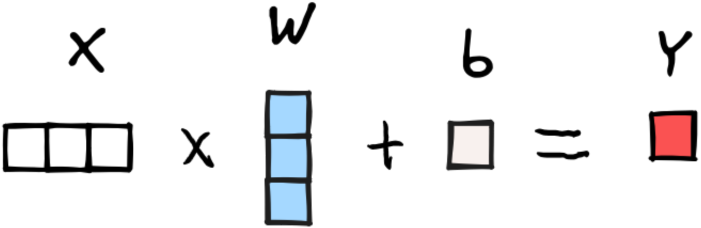


Figure: Bayesian probability [1] + Neural Newtork

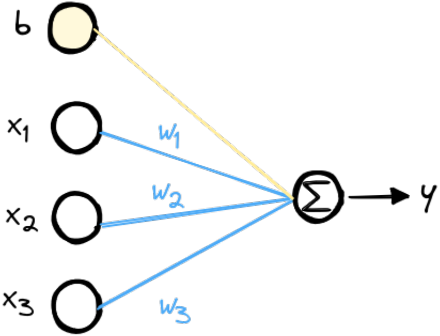
Outline

Linear Regression



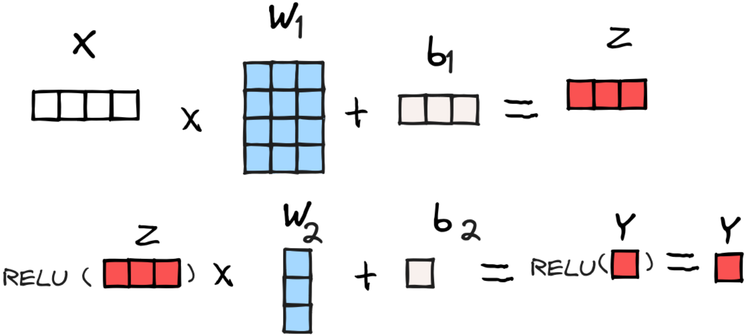
- x : Data
- w : weight
- b : bias
- y : predicted value

Linear Regression



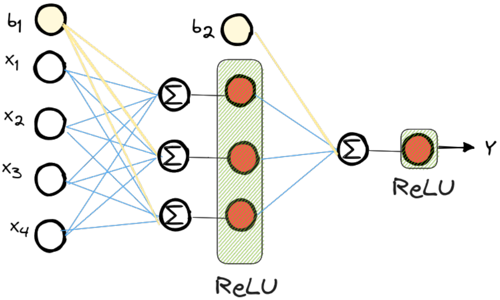
$f(x) = w^T x + b$

Matrices Neural Networks



- x : Data
- w : weight
- b : bias
- y : predicted value

Matrices Neural Networks

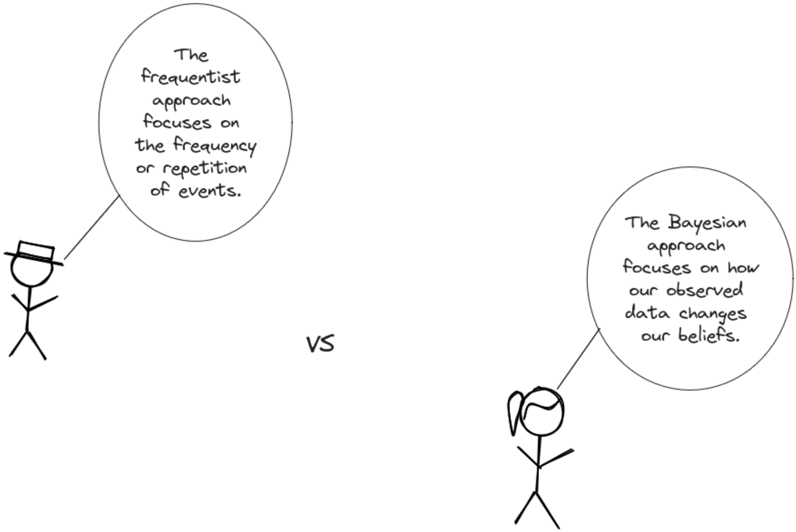


Classic Deep learning

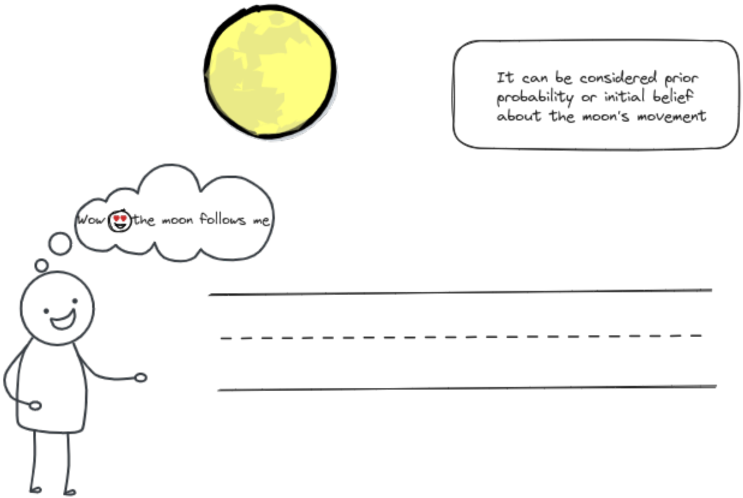
- Given the data-set $\mathbf{D} = (\mathbf{x_i}, \mathbf{y_i})_{\mathbf{i=1}}^{\mathbf{n}}$, model $\mathbf{F}_\theta(\mathbf{x})$ and loss functions $\mathbf{L}(\mathbf{X}, \mathbf{Y}, \theta)$.
- Note that θ takes into account the **weight and bias**.
- Training a model: Minimize the loss function
$$\theta^* = \mathbf{argmin}_\theta \sum_{(\mathbf{x_i}, \mathbf{y_i}) \in \mathbf{D}} \mathbf{L}(\mathbf{F}_\theta(\mathbf{x_i}), \mathbf{y_i}).$$
- Or, Maximize the likelihood
$$\theta^* = \mathbf{argmax}_\theta \sum_{(\mathbf{x_i}, \mathbf{y_i}) \in \mathbf{D}} \mathbf{Log}[\mathbf{p}(\mathbf{y_i} | \mathbf{x_i}), \theta].$$
- Prediction:
$$\hat{\mathbf{y}} = \mathbf{F}_\theta(\hat{\mathbf{x}}) \text{ or } \mathcal{P}(\hat{\mathbf{y}} | \hat{\mathbf{x}}, \theta^*)$$

Outline

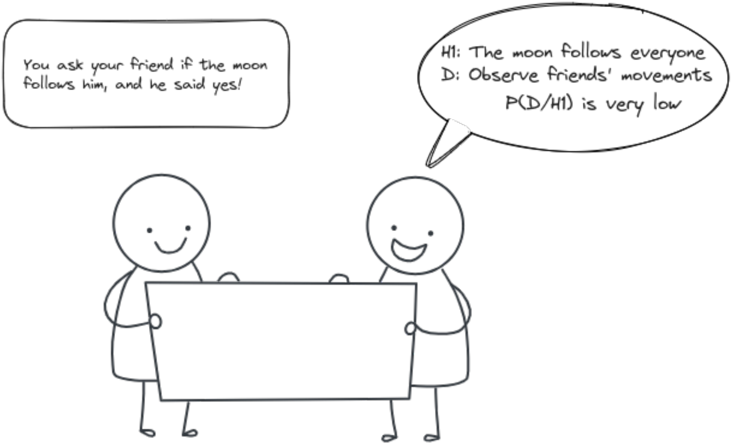
Bayesian Vs Frequentist



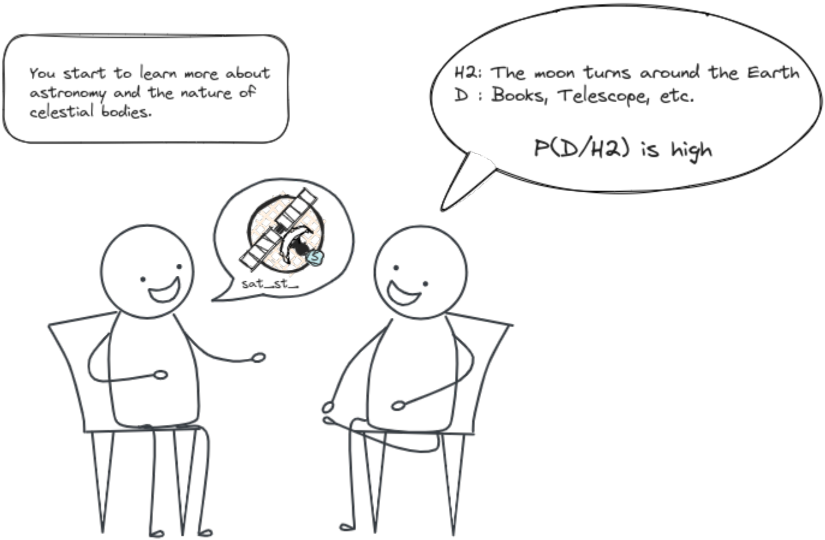
Bayesian Thinking



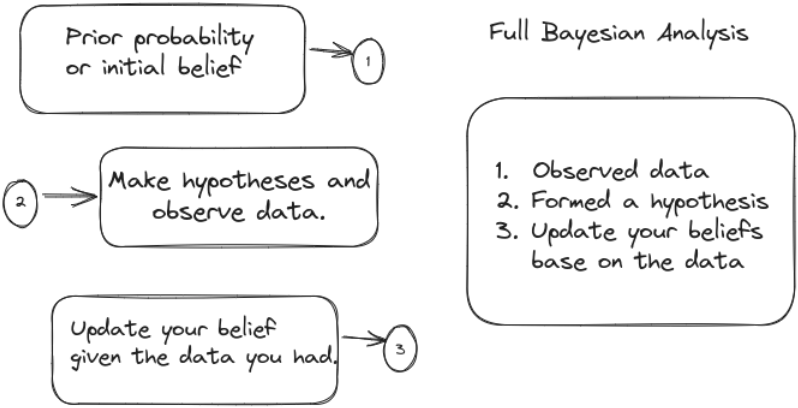
Bayesian Thinking



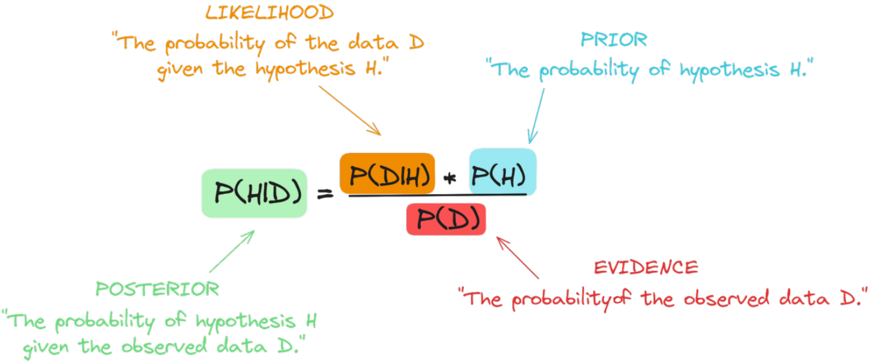
Bayesian Thinking



Bayesian Thinking



Inference


LIKELIHOOD
"The probability of the data D given the hypothesis H."
PRIOR
"The probability of hypothesis H."
P(H|D) = P(D|H) * P(H) / P(D)
POSTERIOR
"The probability of hypothesis H given the observed data D."
EVIDENCE
"The probability of the observed data D."

Outline

Bayesian Neural Networks

- **Bayesian neural networks (BNNs)** combine neural networks with Bayesian inference.

Bayesian Neural Networks

- **Bayesian neural networks (BNNs)** combine neural networks with Bayesian inference.
- Treat **weights** and **outputs** as variables and find marginal distributions that fit the data.

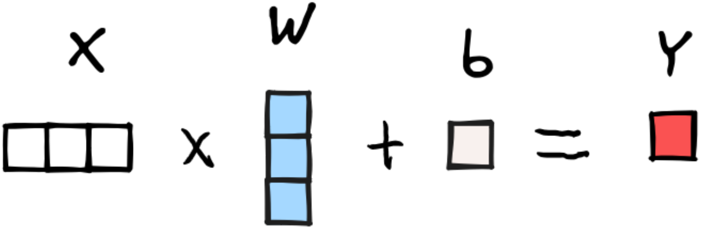
Bayesian Neural Networks

- **Bayesian neural networks (BNNs)** combine neural networks with Bayesian inference.
- Treat **weights** and **outputs** as variables and find marginal distributions that fit the data.
- For the SNN (Standard Neural Network), the parameter estimation is done using **maximum likelihood estimators (MLE) or Loss Functions**, whereas for the BNN (Bayesian Neural Network), the parameter estimation is based on **maximum a posteriori (MAP) or predictive distribution**.

Bayesian Neural Networks

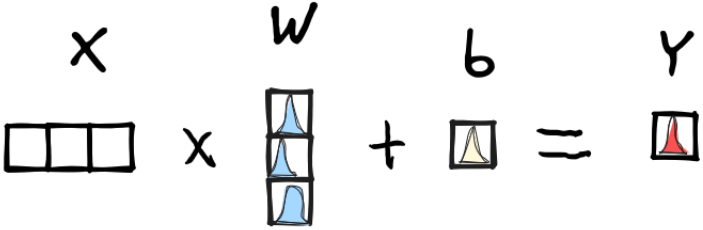
- **Bayesian neural networks (BNNs)** combine neural networks with Bayesian inference.
- Treat **weights** and **outputs** as variables and find marginal distributions that fit the data.
- For the SNN (Standard Neural Network), the parameter estimation is done using **maximum likelihood estimators (MLE) or Loss Functions**, whereas for the BNN (Bayesian Neural Network), the parameter estimation is based on **maximum a posteriori (MAP) or predictive distribution**.
- **Objective**: Quantifying uncertainty in BNN for trustworthy predictions.

Linear Regression



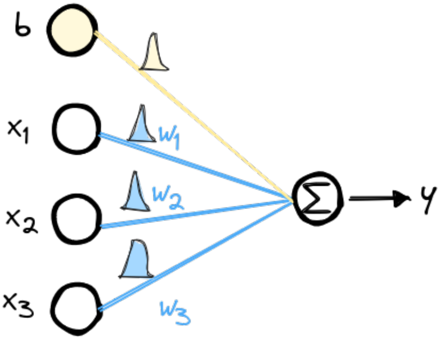
- x : Data
- w : weight
- b : bias
- y : predicted value

Bayesian Linear Regression

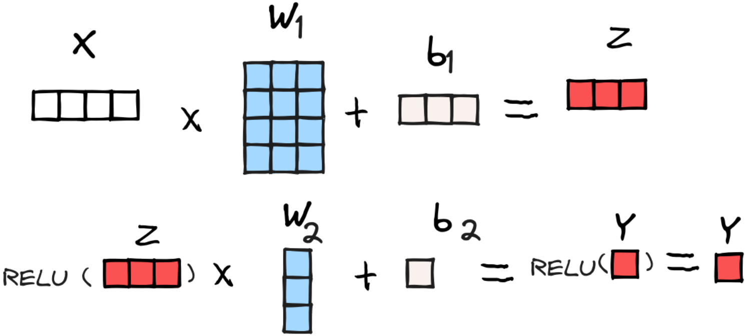


- x : Data
- w : weight
- b : bias
- y : predicted value

Bayesian Linear Regression

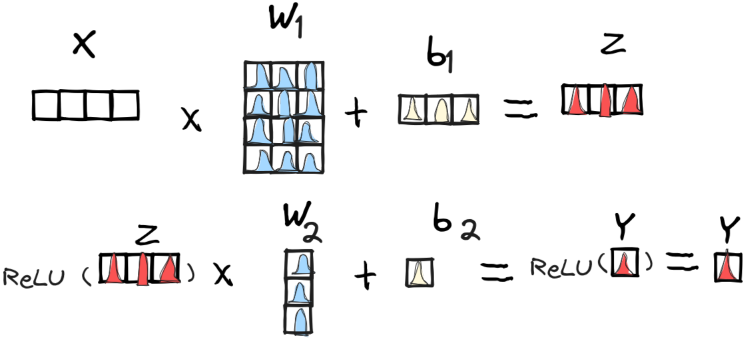


Matrices Neural Networks



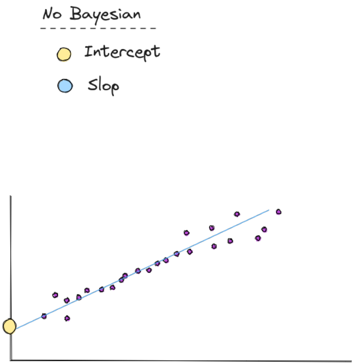
- x : Data
- w : weight
- b : bias
- y : predicted value

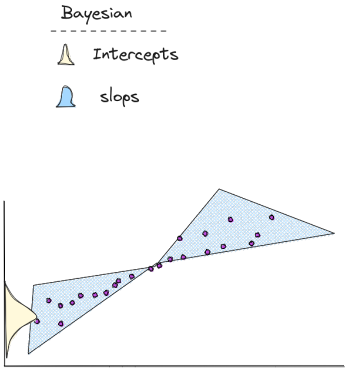
Matrices Bayesian Neural Networks



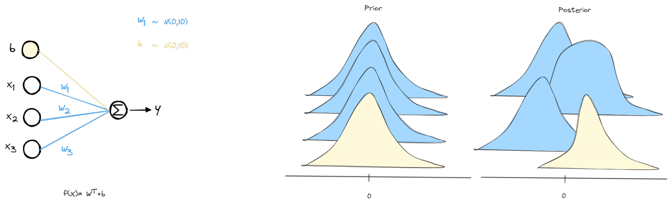
- x : Data
- w : weight
- b : bias
- y : predicted value

Intuition

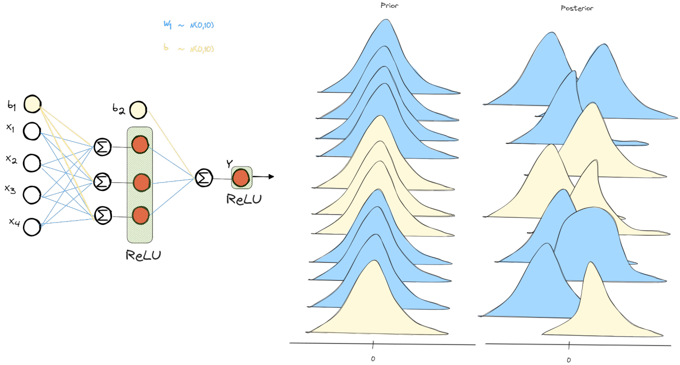




Overview



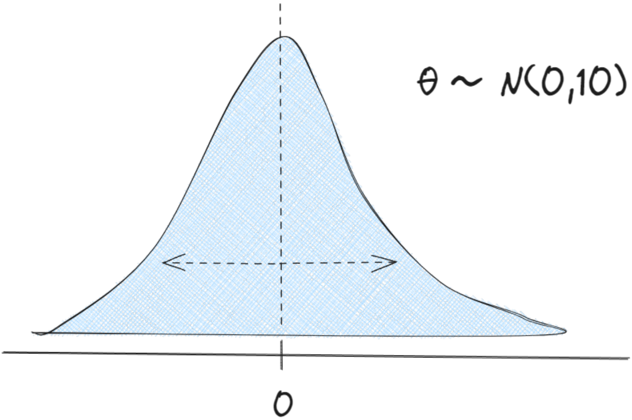
w1 ~ N(0,10)
b ~ N(0,10)
b
x1
x2
x3
w1
w2
w3
Σ
y
f(x)= wT+b
Prior
Posterior
0
0



w1 ~ N(0,10)
b ~ N(0,10)
b1
x1
x2
x3
x4
b2
Σ
ReLU
Y
ReLU
Prior
Posterior
0
0

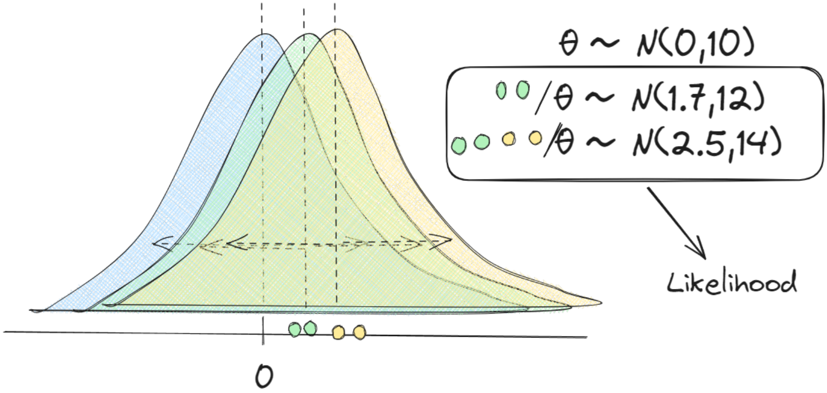
Bayesian Probabilistic model

- **Prior**:[What I believe the parameters look like]



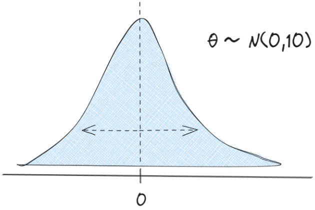
Bayesian Probabilistic model

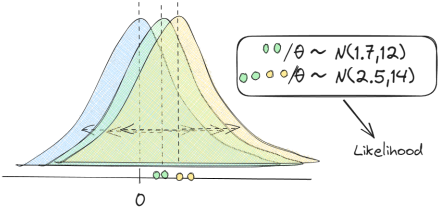
- **Likelihood**:[How I believe data was generated given the parameters]

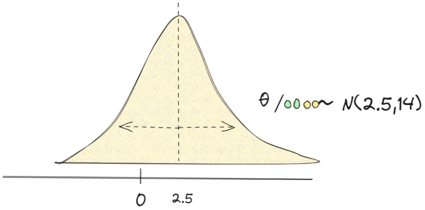


Bayesian Probabilistic model

- Will Update Prior belief on θ conditioned on data you give







Bayesian Probabilistic model

- **Prior**:[What I believe the parameters look like].
$$\theta \sim \mathcal{N}(\mathbf{0}, \mathbf{10})$$
- **Likelihood**:[How I believe data was generated given the parameters]
$$\mathbf{D}|\theta \sim \mathcal{N}(\mu', \sigma'^{\mathbf{2}})$$
- Will Update Prior belief on θ conditioned on data you give
$$\theta|\mathbf{D} \sim \mathcal{N}(\mu'', \sigma''^{\mathbf{2}})$$
- "*With the data you gave me this what i currently think θ could be and i might become more certain if you give me more data.*"

$$\mathbf{P}(\theta|\mathbf{D}) = \frac{\mathbf{P}(\mathbf{D}|\theta) \cdot \mathbf{P}(\theta)}{\mathbf{P}(\mathbf{D})}$$

Training Bayesian neural network

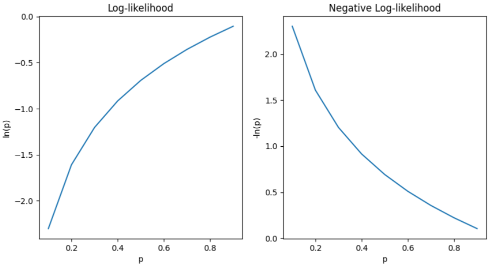
- Loss Function (L):

$$\mathbf{NegativeLogLikelihood}(-\mathbf{L}) = -\sum_{(\mathbf{x_i},\mathbf{y_i})\in\mathbf{D}} \mathbf{ln}(\mathcal{P}(\mathbf{y_i}|\mathbf{x_i},\theta))$$

where :

$$\mathbf{Likelihood}(\mathbf{L}) = -\prod_{(\mathbf{x_i},\mathbf{y_i})\in\mathbf{D}} \mathcal{P}(\mathbf{y_i}|\mathbf{x_i},\theta)$$

- Note that we use the negative log-likelihood because in machine learning, we want to minimize the loss function.



Training Bayesian neural network

- The prediction :

$$\mathcal{P}(\mathbf{y}|\mathbf{x}, \mathbf{D}) = \int \mathcal{P}(\mathbf{y}|\mathbf{x}, \theta)\mathbf{P}(\theta|\mathbf{D})\mathbf{d}\theta$$

Training Bayesian neural network

- The prediction :

$$\mathcal{P}(\mathbf{y}|\mathbf{x}, \mathbf{D}) = \int \mathcal{P}(\mathbf{y}|\mathbf{x}, \theta)\mathbf{P}(\theta|\mathbf{D})\mathbf{d}\theta$$

- However, integrating across all probable values of *theta* is intractable.
- Because, calculating the exact posterior distribution involves integrating over all possible weight configurations.

Training Bayesian neural network

- The prediction :

$$\mathcal{P}(\mathbf{y}|\mathbf{x}, \mathbf{D}) = \int \mathcal{P}(\mathbf{y}|\mathbf{x}, \theta)\mathbf{P}(\theta|\mathbf{D})\mathbf{d}\theta$$

- However, integrating across all probable values of *theta* is intractable.
- Because, calculating the exact posterior distribution involves integrating over all possible weight configurations.
- So, instead of directly solving the integral, we approximate.
- Techniques developed by statisticians **Variational Inference (VI) and Monte Carlo sampling**.
- Focusing on Variational Inference.

Training Bayesian neural network

- **Kullback-Leibler Divergence (KL divergence)**:
  - Can be considered as **regularization**.
  - Quantify how much difference there is from distribution to another distribution.
- Let's say $\mathcal{Q}$ is the predicted distribution, and $\mathcal{P}$ is the true distribution.

$$\mathbf{D_{KL}}[\mathcal{P}||\mathcal{Q}] = \mathbf{H}(\mathcal{P}, \mathcal{Q}) - \mathbf{H}(\mathcal{P})$$

where

$$\mathbf{H}(\mathcal{P}, \mathcal{Q}) = -\mathbf{E}_{(\mathcal{P})} \times \ln \mathcal{P}$$

and

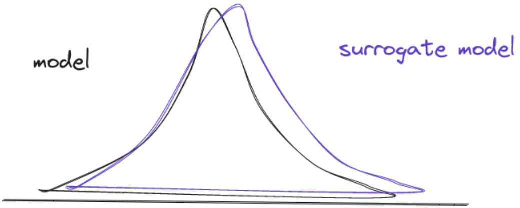
$$\mathbf{H}(\mathcal{P}) = -\mathbf{E}_{(\mathcal{P})} \times \ln \mathcal{P}$$

Variational Inference

- Aim : Approximating the true posterior distribution with a **surrogate** model by minimizing the **evidence lower bound/(ELBO)**. [2]

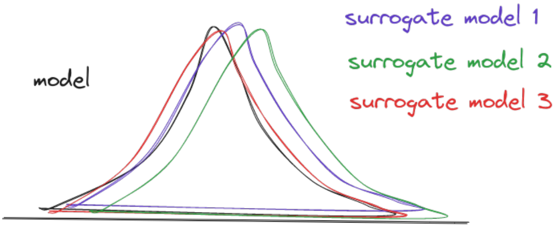
Variational Inference

- Aim : Approximating the true posterior distribution with a **surrogate** model by minimizing the **evidence lower bound/(ELBO)**. [2]
- **A surrogate model** : a model that used to approximate the behavior of the original model.[3]



Variational Inference

- Aim : Approximating the true posterior distribution with a **surrogate** model by minimizing the **evidence lower bound/(ELBO)**. [2]
- **A surrogate model** : a model that used to approximate the behavior of the original model.[3]
- But, there are many surrogate models, how we can ensure that our disturbed prediction is good enough to represent our posterior distribution $P(\theta|D)$?



Variatioanal Inference

- Aim : Approximating the true posterior distribution with a **surrogate** model by minimizing the **evidence lower bound/(ELBO)**. [2]
- **A surrogate model** : a model that used to approximate the behavior of the original model.[3]
- But, there are many surrogate models, how we can ensure that our disturbed prediction is good enough to represent our posterior distribution $P(\theta|D)$?
- By simply trying to find the optimal $\mathcal{Q}^*$ among the surrogate models M.

$$\mathbf{q}^* = argmin_{q \in M} D_{KL}[q(\theta)||\mathcal{P}(\theta|\mathcal{D})] \text{ [4]}$$

Evidence Lower bound (ELBO)

- Rewrite and apply the KL-divergence formula:

$$\mathbf{D_{KL}}[\mathbf{q}(\theta)||\mathcal{P}(\theta|\mathcal{D})] = \mathbf{E_{(q)}}[\mathbf{Ln}(\mathbf{q}(\theta)) - \mathbf{Ln}(\mathbf{P}(\theta|\mathbf{D})$$

$$\mathbf{D_{KL}}[\mathbf{q}(\theta)||\mathcal{P}(\theta|\mathcal{D})] = -\mathbf{E_{(q)}}[\mathbf{Ln}(\frac{\mathbf{P}(\theta,\mathbf{D})}{\mathbf{q}(\theta)})] + \mathbf{Ln}(\mathbf{P}(\mathbf{D}))$$

where $q(\theta)$ is the predicted value.

Evidence Lower bound (ELBO)

- Rewrite and apply the KL-divergence formula :

$$\mathbf{D_{KL}}[\mathbf{q}(\theta)||\mathcal{P}(\theta|\mathcal{D})] = \mathbf{E_{(q)}}[\mathbf{Ln}(\mathbf{q}(\theta)) - \mathbf{Ln}(\mathbf{P}(\theta|\mathbf{D}) \ [4]$$

$$\mathbf{D_{KL}}[\mathbf{q}(\theta)||\mathcal{P}(\theta|\mathcal{D})] = -\mathbf{E_{(q)}}[\mathbf{Ln}(\frac{\mathbf{P}(\theta,\mathbf{D})}{\mathbf{q}(\theta)})] + \mathbf{Ln}(\mathbf{P}(\mathbf{D}))$$

  where $q(\theta)$ is the predicted value.

- $\mathbf{L}(\mathbf{q}) = \mathbf{E_{(q)}}[\mathbf{Ln}(\frac{\mathbf{P}(\theta,\mathbf{D})}{\mathbf{q}(\theta)})]$ [4], the lower bound of evidence (ELBO).

Evidence Lower bound (ELBO)

- Rewrite and apply the KL-divergence formula :

$$\mathbf{D_{KL}}[\mathbf{q}(\theta)||\mathcal{P}(\theta|\mathcal{D})] = \mathbf{E_{(q)}}[\mathbf{Ln}(\mathbf{q}(\theta)) - \mathbf{Ln}(\mathbf{P}(\theta|\mathbf{D})$$

$$\mathbf{D_{KL}}[\mathbf{q}(\theta)||\mathcal{P}(\theta|\mathcal{D})] = -\mathbf{E_{(q)}}[\mathbf{Ln}(\frac{\mathbf{P}(\theta,\mathbf{D})}{\mathbf{q}(\theta)})] + \mathbf{Ln}(\mathbf{P}(\mathbf{D}))$$

  where $q(\theta)$ is the predicted value.

- $\mathbf{L}(\mathbf{q}) = \mathbf{E_{(q)}}[\mathbf{Ln}(\frac{\mathbf{P}(\theta,\mathbf{D})}{\mathbf{q}(\theta)})]$, the lower bound of evidence (ELBO).
- Final formula

$$\mathbf{D_{KL}}[\mathbf{q}(\theta)||\mathcal{P}(\theta|\mathcal{D})] = -\mathbf{L}(\mathbf{q}) + \mathbf{Ln}(\mathbf{P}(\theta|\mathbf{D})\ [4]$$

Evidence Lower bound (ELBO)

- Final formula :

$$\mathbf{D_{KL}}[\mathbf{q}(\theta)||\mathcal{P}(\theta|\mathcal{D})] = -\mathbf{L}(\mathbf{q}) + \mathbf{Ln}(\mathbf{P}(\theta|\mathbf{D})$$

- Given that KL-Divergence $\in \mathbb{R}^{+}$ and $0 < \mathbf{P}(\theta|\mathbf{D}) \leq 1$
- $L(q)$, called the **Evidence Lower Bound**
- The optimal value :

$$\mathbf{q}^{*} = argmin_{q \in M}(-L(q)) \; [4]$$

Overview

Instead of directly solving the integral :

$$\mathcal{P}(\mathbf{y}|\mathbf{x}, \mathbf{D}) = \int \mathcal{P}(\mathbf{y}|\mathbf{x}), \theta)\mathbf{P}(\theta|\mathbf{D})\mathbf{d}\theta$$

we approximate the integral and compute

- The expectation $\mathbb{E}[\mathcal{P}(\mathbf{y}|\mathbf{x}, \mathbf{D})]$ [5]
- The variance $\mathbb{V}[\mathcal{P}(\mathbf{y}|\mathbf{x}, \mathbf{D})]$ [5]

using

- Monte Carlo Sampling
- Variational inference (VI)

Overview

Predicted distribution of $\mathcal{P}(\mathbf{y}|\mathbf{x}, \mathbf{D})$ can be visualized as

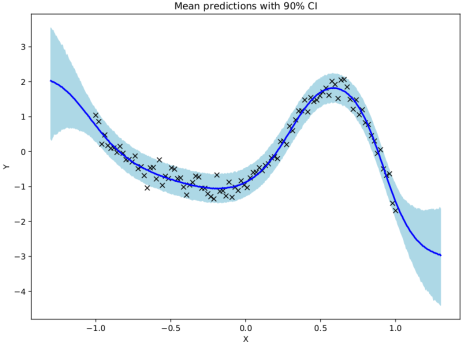


Figure: Bayesian Neural Network with 2 hidden. [6]

- Cyan region is the **uncertainty** from the $\mathbb{V}[\mathcal{P}(\mathbf{y}|\mathbf{x}, \mathbf{D})]$
- Blue is the mean of the prediction $\mathbb{E}[\mathcal{P}(\mathbf{y}|\mathbf{x}, \mathbf{D})]$.

Bayesian Neural Networks

	Standard Neural Network	Bayesian Neural Network
Goal	*Optimization*	*Marginalization*
Weight	*A Single Set*	*Probabilistic Distribution*
Method	*Differentiation* *(Gradient Descent)*	*Markov Chain Monte Carlos* *Variational Inference* *Normalizing Flows*
Estimate	*Maximum Likelihood Estimators*	*Maximum A Posteriori* *Full / Approximate Predictive Distribution*

Figure: Difference between Standard NN and Bayesian NN. [4]

Outline

Example:Uncertainty in self-driving car

Figure: An example of Autonomous car. [7]

Example:Uncertainty in self-driving car

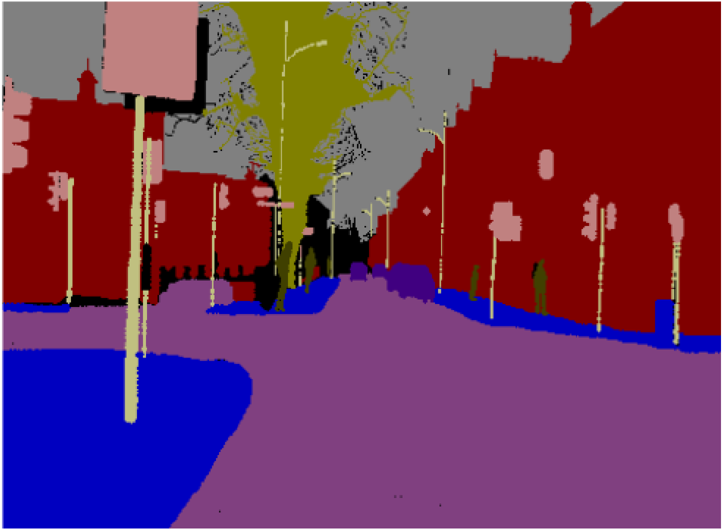


Figure: An example of Autonomous car. [7]

Example:Uncertainty in self-driving car

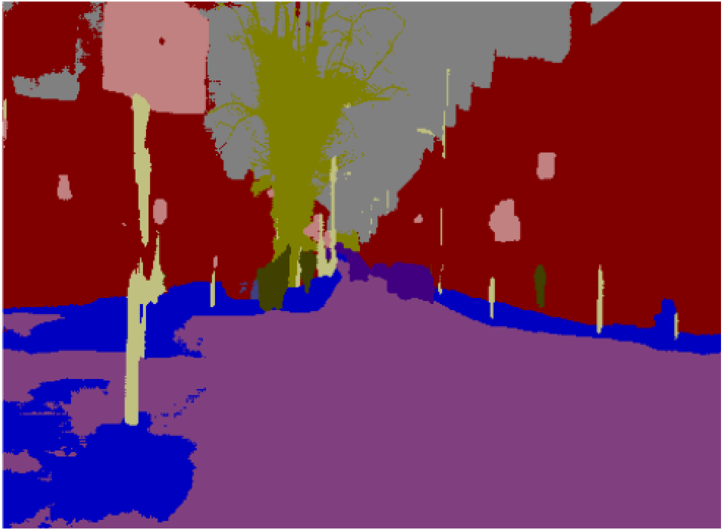


Figure: An example of Autonomous car. [7]

Example:Uncertainty in self-driving car

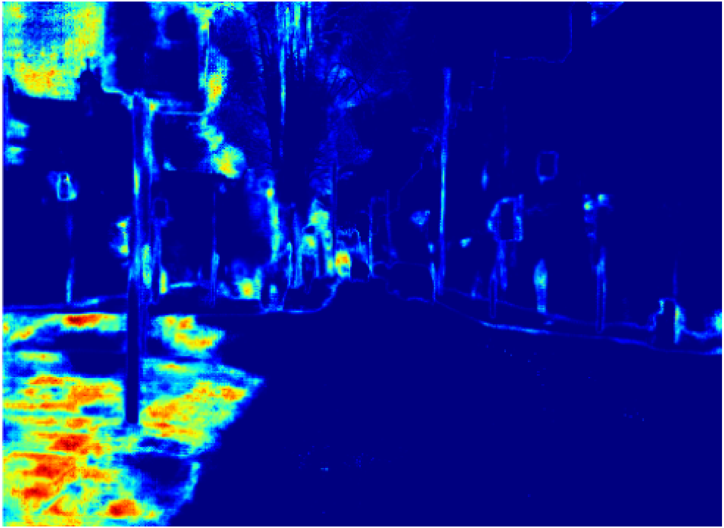


Figure: An example of Autonomous car. [7]

Example : Skin lesions detection

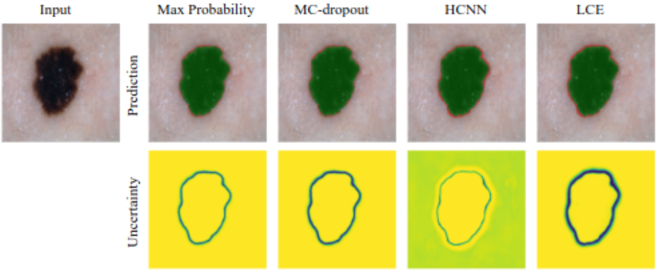


Figure: Different uncertainty estimation methods. In the segmentation predictions, green = true positive, red = false positive, and blue = false negative. [8]

Example : Skin lesions detection

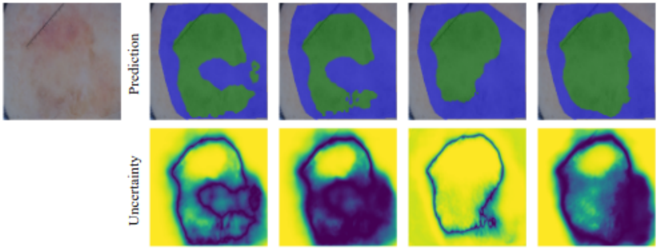


Figure: Different uncertainty estimation methods. In the segmentation predictions, green = true positive, red = false positive, and blue = false negative. [8]

Example : Skin lesions detection

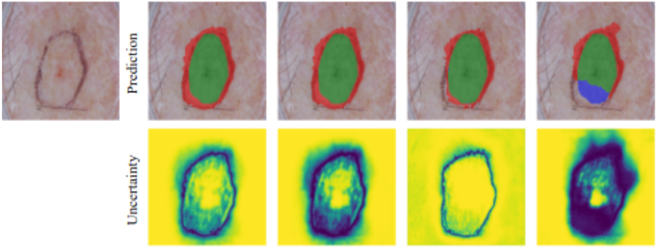


Figure: Different uncertainty estimation methods. In the segmentation predictions, green = true positive, red = false positive, and blue = false negative. [8]

Example : Skin lesions detection

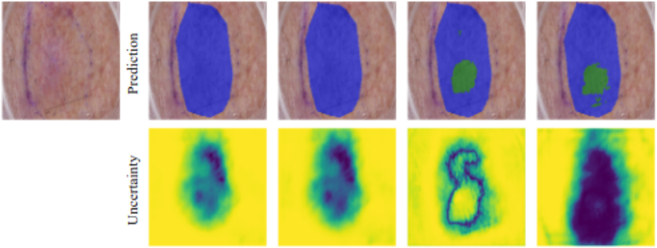


Figure: Different uncertainty estimation methods. In the segmentation predictions, green = true positive, red = false positive, and blue = false negative. [8]

Outline

Advantages and limitations

- **Advantages**:
  - Trustworthiness of predictions explained through uncertainty quantification.
  - Handling variability and ambiguity in data.
  - Informed decision-making based on the level of uncertainty.
- **Limitations**:
  - Demanding mathematics and statistics knowledge.
  - Longer training epochs to converge.

Outline

References I

- "Bayes rules wikipédia," May 2023.
- "Variational bayesian methods," May 2023.
- wikipedia, "Surrogate model," Jun 2023.
- Y. WONG, "8 terms you should know about bayesian neural network," Oct 2021.
- "Bayesian deep learning."
- "Bayesian neural network - numpyro documentation."
- A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," *Advances in neural information processing systems*, vol. 30, 2017.

References II

- T. DeVries and G. W. Taylor, "Leveraging uncertainty estimates for predicting segmentation quality," *arXiv preprint arXiv:1807.00502*, 2018.
- L. Roque, "Primer on bayesian deep learning," Feb 2023.
- "Kullback–leibler divergence wikipedia," May 2021.