RDFIA Practical work 4

Giulia Prosio and Alexander Hölzl

DATE

4-a: BAYESIAN LINEAR REGRESSION

Summary

The aspect we will focus on now is the necessity for the image classificators to be robust - specifically to have a reliable uncertainty estimation in the decision (classification) process.
The goal of this project work and the next ones is to focus and deepen our knowledge regarding the robustness analysis in Deep Learning.

With practical work 4a we use and analyze to use of the Bayesian models in uncertainty estimation - both linear regression and non-linear models, such as the Polynomial and Gaussian.

Part I - Linear Basis Function Model

We start with a linear dataset where we will analyze the behavior of linear basis functions in the framework of Bayesian Linear Regression.

The first step to implement the Bayesian model is to compute the basis function, here being: $\phi : x \longrightarrow (1, x)$

Question 1

Recall closed form of the posterior distribution in linear case. Then, code and visualize posterior sampling. What can you observe?

In a Bayesian linear regression model with a normal prior on the parameters and normally distributed noise, the closed form of the posterior distribution is given by the updating of a conjugate prior.

The closed form solution for the posterior distribution can be derived by updating the prior distribution with the information from the data, which is done by multiplying the prior by the likelihood and normalizing the result. The resulting posterior distribution is given by:

$$p(\boldsymbol{w}|\mathbf{X}, \mathbf{Y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

where

$$\boldsymbol{\mu} = \beta \boldsymbol{\Sigma} \boldsymbol{\Phi}^T \mathbf{Y} \tag{2}$$

$$\boldsymbol{\Sigma} = (\beta \boldsymbol{\Phi}^T \boldsymbol{\Phi} + \boldsymbol{\alpha} I)^{-1} \tag{3}$$

These updated mean and covariance values for the posterior distribution incorporate both the prior beliefs and the information from the data.

As we can observe from the Figure 1 , the bigger the number of points (data from the training set) given to the model, the more accurate the prediction is going to be.
Furthermore, it reduces posterior uncertainty.

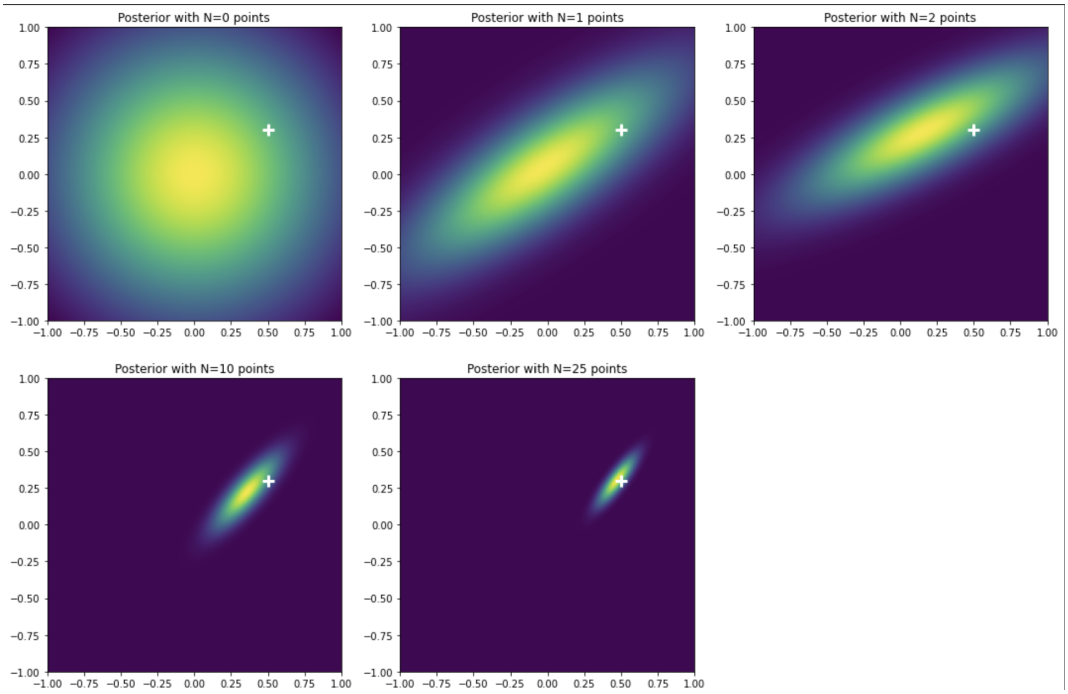

Figure 1: Posterior sampling with different nbr of points N

Question 2

Recall and code closed form of the predictive distribution in linear case.

The closed-form expression for the predictive distribution in linear regression can be written as:

$$p(y \mid x^*, \alpha, \beta) = N(y; \mu^T \mathbf{\Phi}(x^*), (1/\beta) + \mathbf{\Phi}(x^*)^T \mathbf{\Sigma} \mathbf{\Phi}(x^*)) \tag{4}$$

where:

- σ^2 is the variance of the error term and is given by $(1/\beta) + \mathbf{\Phi}(x^*)^T \mathbf{\Sigma} \mathbf{\Phi}(x^*)$
- β is the aleatoric noise representation
- $\mathbf{\Phi}(x^*)^T \mathbf{\Sigma} \mathbf{\Phi}(x^*)$ is the epistemic uncertainty over the parameters w.
  The weights w can be estimated using maximum a posteriori estimation (MAP), given the training data

Question 3

Predict on the test dataset. Analyse these results. Why predictive variance increases far from training distribution? Prove it analytically in the case where α =0 and β =1

Based on the prediction function implemented, we predicted the distribution of the points of the test data set.
The results are shown on Figure 2.

As we can observe from the result, the predictive variance increases far from the training distribution. This is because the predictive variance captures the uncertainty of the target variable given a new input.

When the new input is similar to the training inputs, we have a good understanding of the relationship between the inputs and the target variable, and the uncertainty is low.

However, as the new input becomes more different from the training inputs, our understanding of the relationship becomes less certain, and the predictive variance increases.

α and β control the shape of the prior distribution on the weights.

When α=0 and β=1, the prior distribution on the weights becomes a flat non-informative prior, and the observation noise is considered to have unit variance.
In this case, the posterior distribution of the weights is dominated by the likelihood term, which is proportional to the observed data.
The predictive variance will then reflect only the uncertainty in the observed data and the observation noise, and will be relatively small for inputs close to the training inputs, and increase largely as the input moves away from the training inputs.

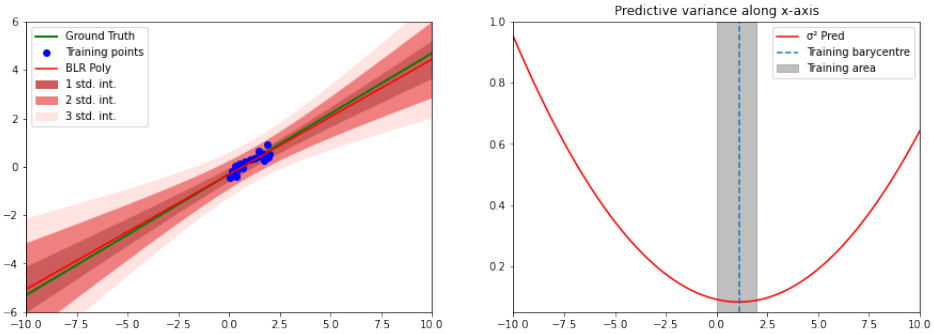

Figure 2: Prediction on test dataset

Part II - Non Linear models

For the second part of the project work we then consider a more complex dataset, an increasing sinusoidal curve, that we analyze with a Polynomial and a Gaussian Basis functions.

Question 4

Code and visualize results on sinusoidal dataset using polynomial basis functions. What can you say about the predictive variance?

For this analysis we used a polynomial basis functions up to a 10th-degree, so the model had the capacity to fit a wide range of functions.
However, the use of higher-degree polynomials can result in overfitting, particularly if the number of training data points is limited.

As we can see from Figure 3, the predictive variance is high for inputs that are far from the training inputs, since the model may be extrapolating beyond the range of the training data and making predictions based on noisy or unreliable features.
The predictive variance is also higher for inputs near the training inputs if the model has overfit the data and is sensitive to the noise in the data.

To address this issue, regularization techniques, such as ridge regression or Bayesian model averaging, can be used to constrain the magnitude of the weights and reduce overfitting. The choice of regularization technique and the value of the regularization hyperparameters will affect the predictive variance.

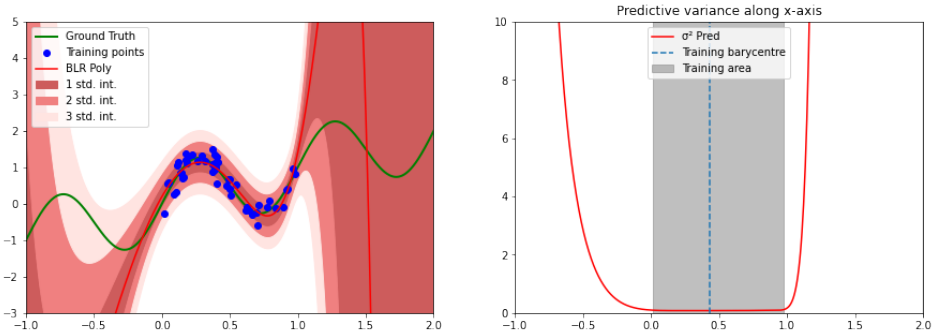

Figure 3: Model with Polynomial Basis Function

Question 5

Code and visualize results on sinusoidal dataset using Gaussian basis functions. What can you say this time about the predictive variance?

The Gaussian basis functions allow for non-linear relationships between the inputs and the target to be captured.
As we can observe from the results obtained on the data in Figure 4, the model is designed to capture important features in the data, such as local patterns or smooth transitions.
This way the obtained variance is low.
On the other hand, if the number of basis functions is too large, the model may overfit the training data, leading to a low training error but a high test error and high predictive variance.
Also in this case, to address this issue, regularization techniques, such as ridge regression or Bayesian model averaging, can be used to constrain the magnitude of the weights and reduce overfitting.

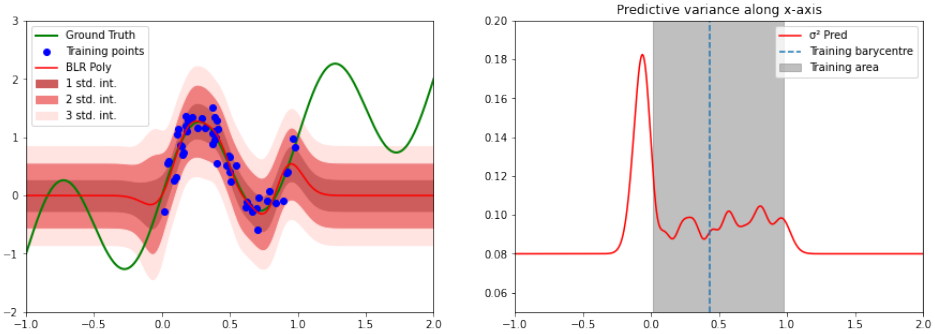

Figure 4: Model with Gaussian Basis Function

Question 6

Explain why in regions far from training distribution, the predictive variance converges to this value when using localized basis functions such as Gaussians

The phenomenon we are observing is epistemic uncertainty.
In Bayesian regression using Gaussian basis functions, epistemic uncertainty refers to the uncertainty in the model parameters caused by the lack of information or knowledge about the true relationship between the inputs and the target.

4-b: APPROXIMATE INFERENCE IN CLASSIFICATION

Summary

This second project work makes a step further the Bayesian Regression Model as the posterior and predictive distributions are too simple to be applied to Neural Networks.

The first part of the project focuses on Bayesian Logistic Regression (BLR), the second Bayesian Neural Networks.

Part I: Bayesian Logistic Regression

Question 1

Analyze the results provided by previous plot. Looking at $p(y = 1|x, w_{MAP})$, what can you say about points far from train distribution?

Points which are father away from train distribution might get classified with very high probabilities, only if points lie on the decision boundary, or close to it the confidence will drop.

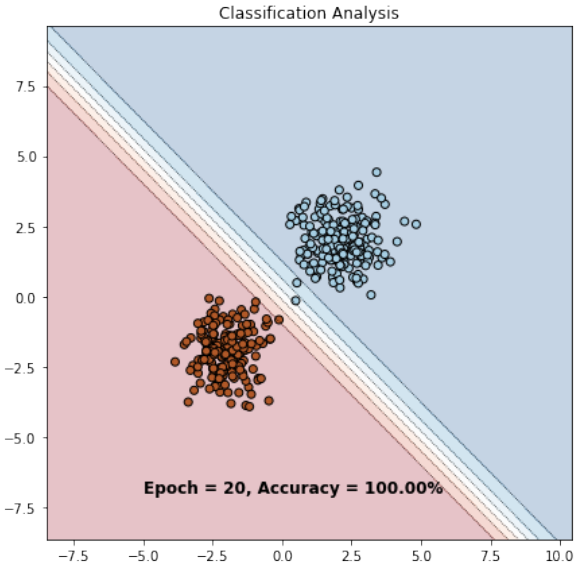

Figure 5: Plot for question 1.1

Question 2

Analyze the results provided by previous plot. Compared to previous MAP estimate, how does the predictive distribution behave?

This predictive distribution is able to to handle points far from train distribution better as the decision boundary is not a 1-dimensional line, but an expanding area. That means it is able to assign lower probabilities to points far from the training distribution.
In addition to that there are also "noisy" areas.

Question 3

Analyze the results provided by previous plot. Compared to previous MAP estimate, how does the predictive distribution behave?

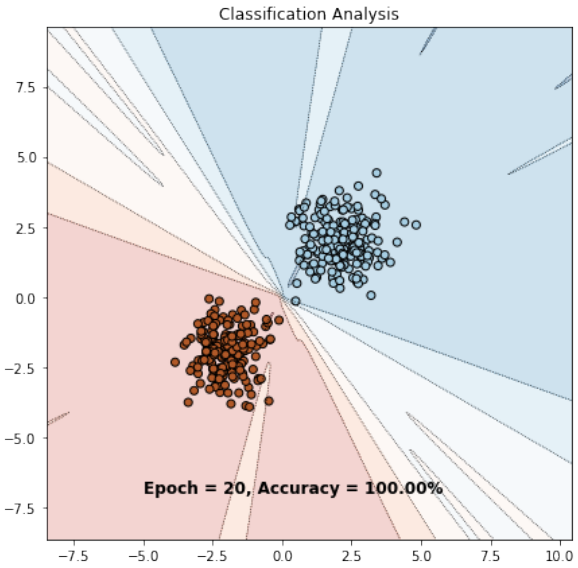

Figure 6: Plot for question 1.2

As the previous distribution, it has a non-linear decision boundary but does not suffer from the noisy areas that the previous approach exhibited.

Part II: Bayesian Neural Networks

Question 1

Again, analyze the results showed on plot. What is the benefit of MC Dropout variational inference over Bayesian Logistic Regression with variational inference?

As can be seen in 8 the decision boundary is a bit noisy. An advantage of this approach is, that it is much less computationally intensive and requires less iterations to converge.

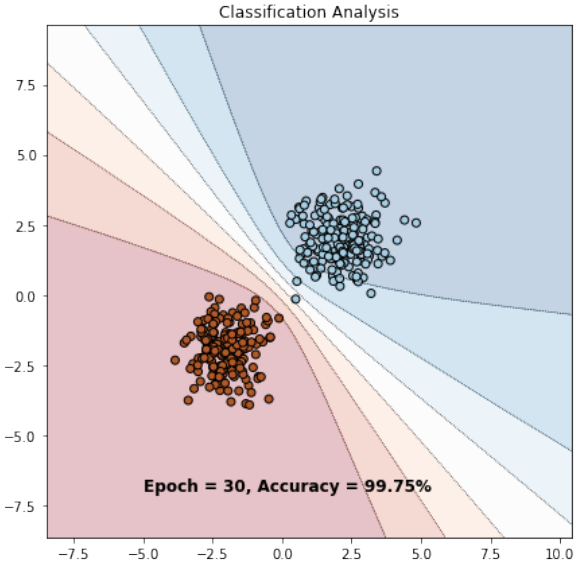

Figure 7: Plot for question 1.3

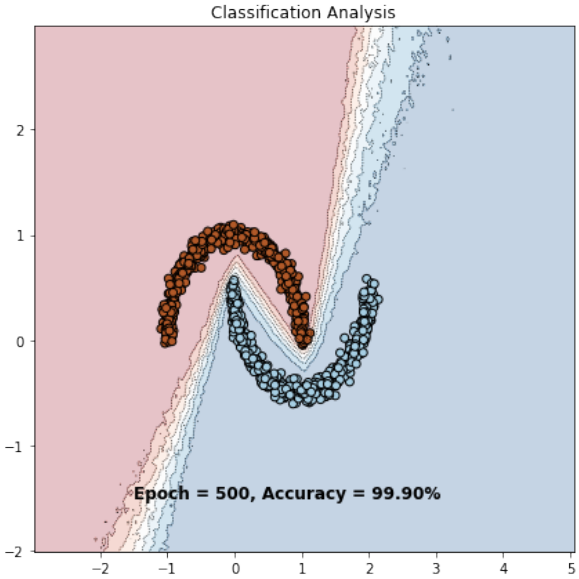

Figure 8: Plot for question 2.1

4-c: UNCERTAINTY APPLICATIONS

Summary

The purpose of this lab was estimating uncertainty for predictions our models do. In the first part we tried to qualitatively evaluate the most uncertain images in our test set. To do that we built a leNet-5 style network with dropout layers in order to implement Monte-Carlo dropout variational inference. We used three different mesures to compute the uncertanity of samples, those are variation-ratios, entropy and mutual information.

The second part deals with failure prediction, which means predicting if the network is doing a misclassification on the current sample. In order to do that, a ConfidNet is used which consists of two parts, the normal classification part and a second network, the confidence net which is trained seperatly. Its job is to predict if the classification network is failing to predict a correct class.

The last part of this course dealt with detecting if an input sample is outside the distribution the network was trained on.

Monte-Carlo Dropout on MNIST

Question 1

What can you say about the images themselves. How do the histograms along them helps to explain failure cases? Finally, how do probabilities distribution of random images compare to the previous top uncertain images?

In the context of the MNIST dataset, Monte Carlo Dropout can be used to obtain a measure of uncertainty for the predictions made by a deep learning model trained on the MNIST dataset.

The basic idea behind Monte Carlo Dropout is to use Dropout during the forward pass to sample multiple dropout masks and obtain multiple predictions for the same input.
The variance of these predictions can be used as a measure of the model's uncertainty. To obtain a prediction with Monte Carlo Dropout, multiple forward passes are made with the Dropout layer active, and the predictions are averaged to obtain the final prediction. The uncertainty can be estimated from the variance of these predictions.

As we can observe from the shown output in Figure 9, this approach gives us multiple predictions, and the final prediction will be the one with the maximal frequency, given by the histogram of the passes.

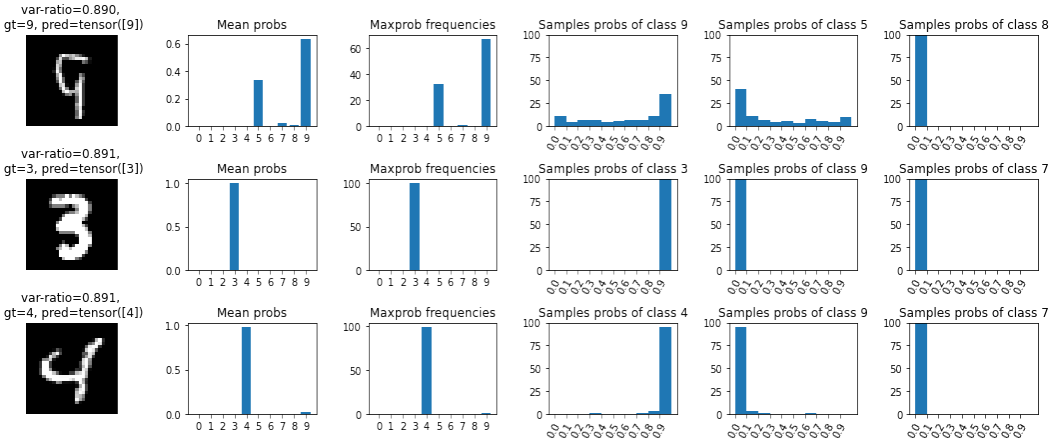

Figure 9: Uncertain images

Part II: Failure prediction

Question 2

Compare the precision-recall curves of each method along with their AUPR values. Why did we use AUPR metric instead of standard AUROC?

The Area Under the ROC Curve (AUC-ROC) is a popular performance metric for binary classification problems.
It measures the ability of the classifier to distinguish between positive and negative examples and is insensitive to the class distribution and the imbalance between positive and negative examples.

On the other hand, the Area Under the Precision-Recall Curve (AUPR) is a metric that focuses on the positive examples and is sensitive to the imbalance between positive and negative examples.

In applications where the positive class is rare or there is a high cost associated with false positive predictions, it may be more important to optimize precision, which is the fraction of true positive predictions out of all positive predictions. In such cases, the AUPR metric may be more appropriate than the AUC-ROC, as it provides a more complete picture of the model's performance.

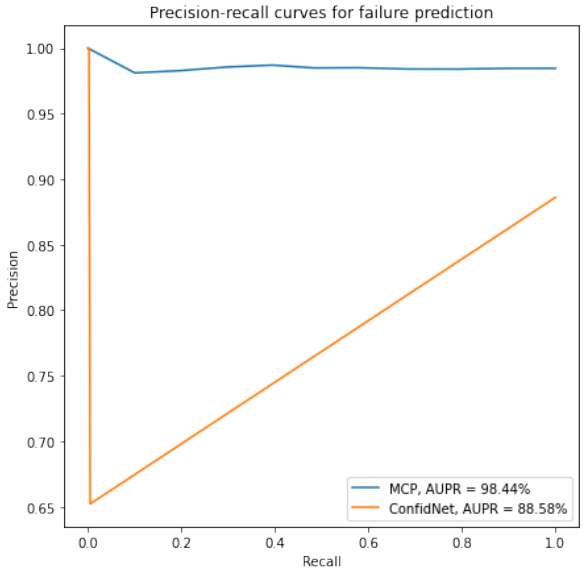

Figure 10: Precision-recall curves for different methods

Part III: Out-of-distribution detection

Question 3

Compare the precision-recall curves of each OOD method along with their AUPR values. Which method perform best and why?

Precision-recall curves and the corresponding AUPR values are used to evaluate the performance of OOD detection methods.
In this context, precision represents the proportion of OOD examples that are correctly identified as such, while recall represents the proportion of OOD examples that are correctly identified among all OOD examples.

The precision-recall curve plots precision against recall for different threshold values, and the AUPR value provides a summary of the area under this curve.

In comparing OOD detection methods, one would look at the shape of the precision-recall curves, as well as their corresponding AUPR values.
A method with a higher AUPR value would generally indicate a better overall performance in detecting OOD examples, while the shape of the precision-recall curve can provide insights into the trade-off between precision and recall for a given method.

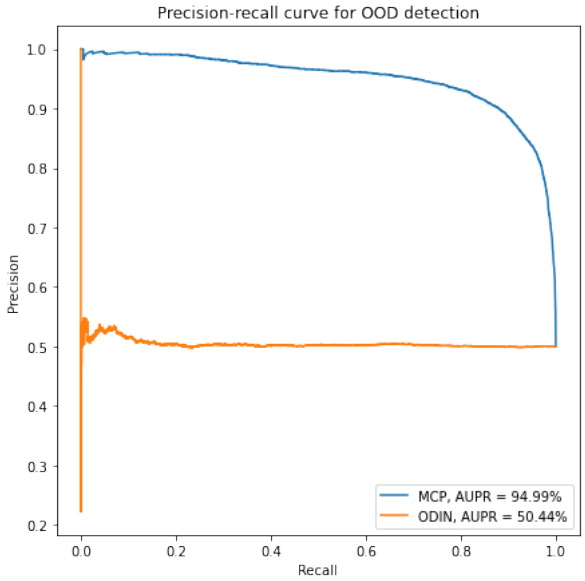

Figure 11: Precision-recall curves for different OOD methods