RDFIA
Bayesian Models and Deep Learning

Johan Pardo - Shubhamkumar Patel

February 13, 2022

Practical 1

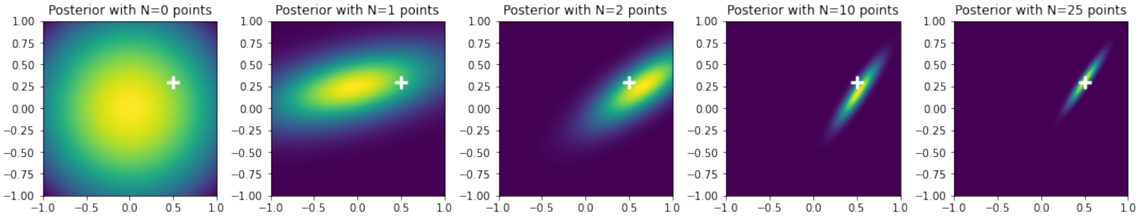
Part 1 : Linear Basis function model

Question 1.1 Recall closed form of the posterior distribution in linear case

The closed form of the posterior distribution in linear case is:

$$p(w/X, Y) = N(w|\mu, \Sigma)$$
$$\Sigma^{-1} = \alpha I + \beta \phi^T \phi$$
$$\mu = \beta \Sigma \phi^T Y$$

Question 1.2: Looking at the visualization of the posterior above, what can you say?



In this part we are experimenting with the posterior sampling. We start with the case of $N = 0$, using 0 points, in this case the posterior is exactly equal to the prior as there is no data likelihood. Next we keep adding more and more points ($N \geq 1$), as we do so we increase the data likelihood in the model. We can see that at the end the posterior gets closer to the white cross which represents our target. This means that adding more points effectively means reducing posterior (epistemic) uncertainty.

Question 1.3: Recall the closed form of the predictive distribution in linear case.

The closed form of the predictive distribution in linear case is:

$$p(y^*|x^*, D, \alpha, \beta) = N(y^*; \mu^T \phi(x^*), \frac{1}{\beta} + \phi(x^*)^T \Sigma \phi(x^*))$$

Question 1.4: Analyse these results. Describe the behavior of the predictive variance for points far from training distribution. Prove it analytically in the case where α=0 and β=1.

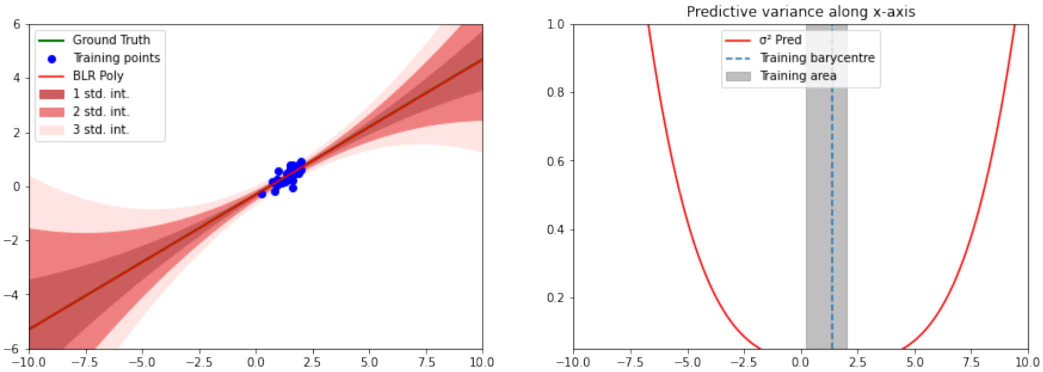


Figure 1: Predictions using a linear ϕ.

As we can see in the Figure 1 the further we move away from the dataset defined by the cluster of training points, the predictive variance along the x-axis gets higher. Here we tried to use a linear basis function for our bayesian linear regression.

In the case where α=0 and β=1, we have $\Sigma^{-1} = \phi^T\phi$ and $\mu = \Sigma\phi^TY$, which bring the problem down to the Maximum Likelihood case.

Question Bonus: What happens when applying Bayesian Linear Regression on the following dataset?

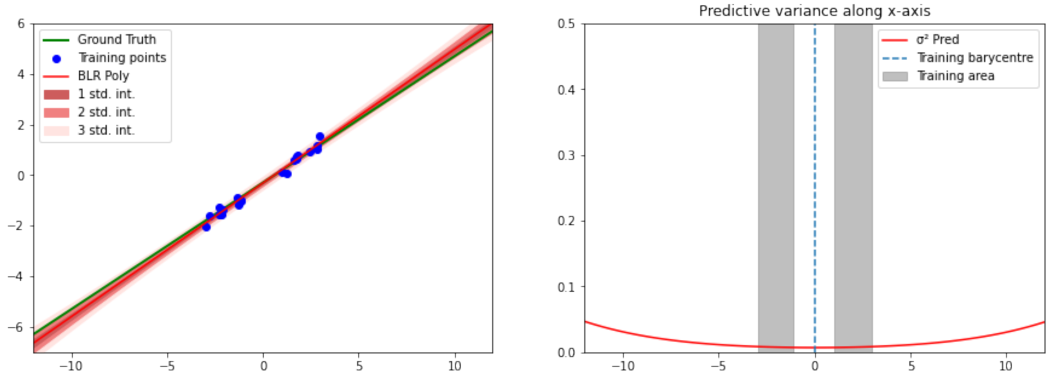


Figure 2: Dataset with hole and associated Ground truth

When applying bayesian linear regression on the following dataset we have better results due to the presence of 2 groups of points which helps our linear regression to converge towards the ground truth more efficiently. The two groups of points here can be considered as two observations which helps us a lot during the training as we can see in the Figure 2. The predictive variance along the x-axis stays closer to 0, this shows us that we are able get very low variance even when we are far away from the training area this means we have a obtained a better regression than in the previous case with a single cluster of training points.

Part 2 : Non Linear models

Question 2.1: What can you say about the predictive variance?

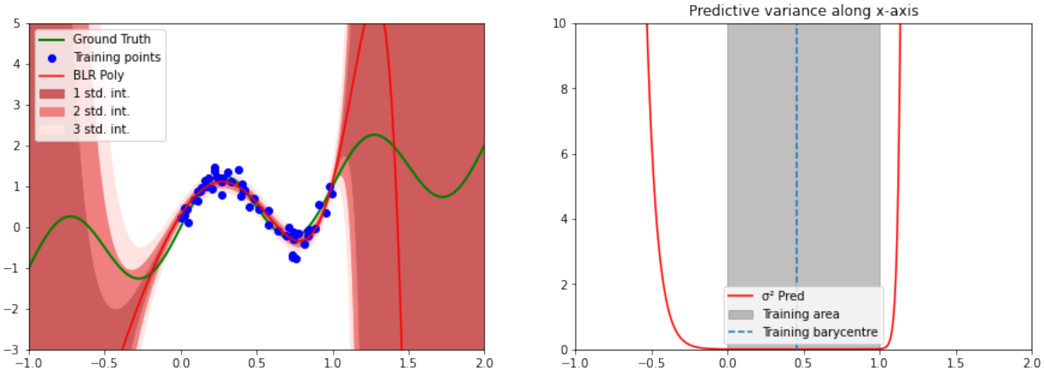
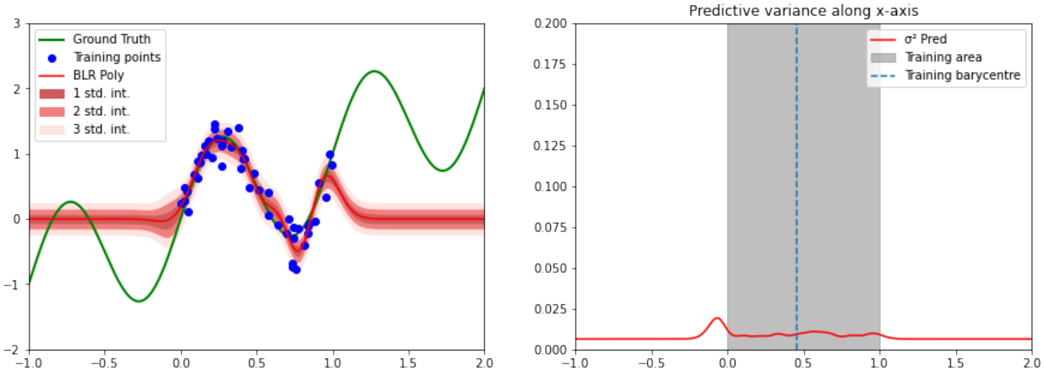


Figure 3: Sine Dataset and associated Ground truth

In this case the chosen basis function is polynomial. This allows us to change the behaviour of the predictive variance along the x-axis to better adapt it to the training points from the input dataset. We can say that the predictive variance doesn't behave well when we try to predict values that are far from the dataset, but we are able to fit around the ground truth in parts of the graph where there are a lot of training points. In the training area we achieved extremely low predictive variance but the limitation of this method lies outside the training area where we get diverging predictive variance.

Question 2.2: What can you say this time about the predictive variance? What can you conclude?



Compared to the previous method where we had divergence of the predictive variance outside the training area we are now getting significantly low variance when we go far from the training points, the predictive variance using Gaussian Basis Function gives good results in the training area but outside we don't get the diverging problems as the values converge towards 0.

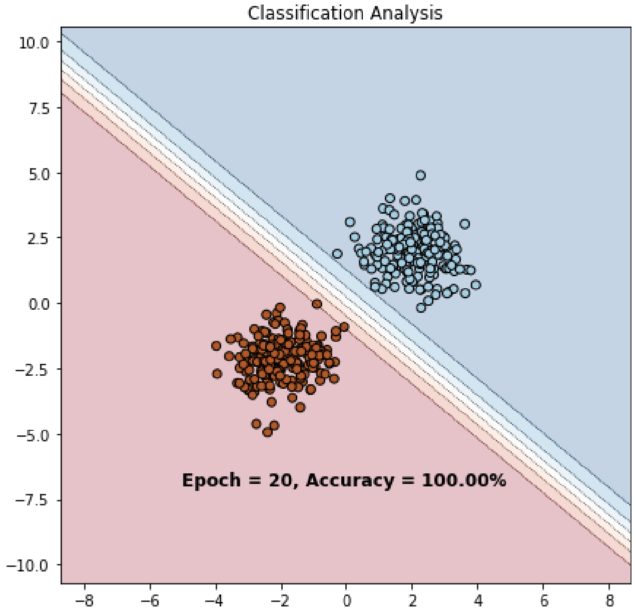
Question 2.3: Explain why in regions far from training distribution, the predictive variance converges to this particular value when using localized basis functions such as Gaussians.

The function ϕ which is a gaussian will converge to zero when we are far from the training points. So we get constant behaviour when we are farther away from the dataset and some constant value when we closer to the dataset. This means the predictive variance also become constant and converges towards the value of $\frac{1}{\beta}$ and with $\beta = \frac{1}{2\sigma^2}$, in our case the predictive variance converges towards the value of **0.08**.

Practical 2

Part 1: Bayesian Logistic Regression

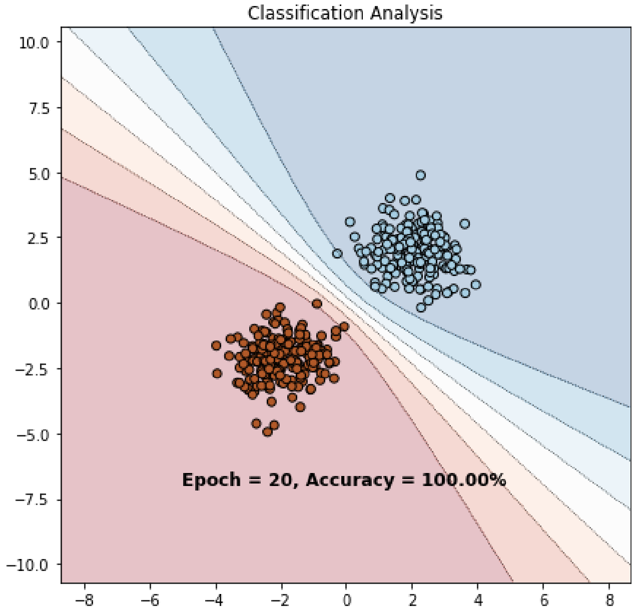
Question 1.1: Analyze the results provided by previous plot.
Looking at $p(y = 1|xx, wwMAP)$, what can you say about points far from train distribution?



We can say that the points that are far from the train distribution will have the same classification based on the hyperplan we will have obtained. We can see that the uncertainty does not increase far from training data, therefore we need for more accurate approximations instead of the current one that is a very coarse approximation of the train distribution.

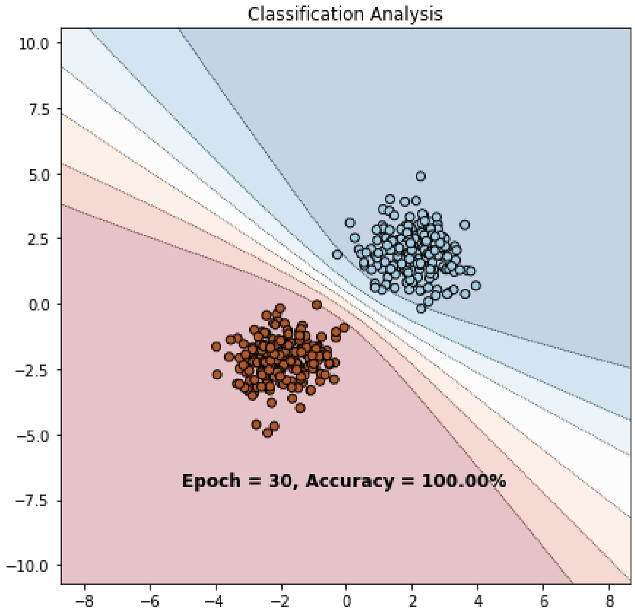
Question 1.2: Analyze the results provided by previous plot. Compared to previous MAP estimate, how does the predictive distribution behave?

Compared to the previous plot the MAP estimate, the current Laplace Approximation method tends to have a more non linear approach and therefore might be more suitable



for current data distribution (compared to the previous one). We can see that as we go further away from the dataset we observe an increasing level of uncertainty which means value that are far from some datasets we might have high variance.

Question 1.3: Analyze the results provided by previous plot. Compared to previous MAP estimate, how does the predictive distribution behave?



Compared to the previous plot for the Laplacian Approximation, the current Variational inference approach tends to give closer results to the distribution of the dataset. But overall we can see that as we go further away from the dataset we observe the similar increasing level of uncertainty as we go away from the data distribution, just as in the previous case.

Part 2: Bayesian Neural Networks

Question: 2.1: Again, analyze the results showed on plot. What is the benefit of MC Dropout variational inference over Bayesian Logistic Regression with variational inference?

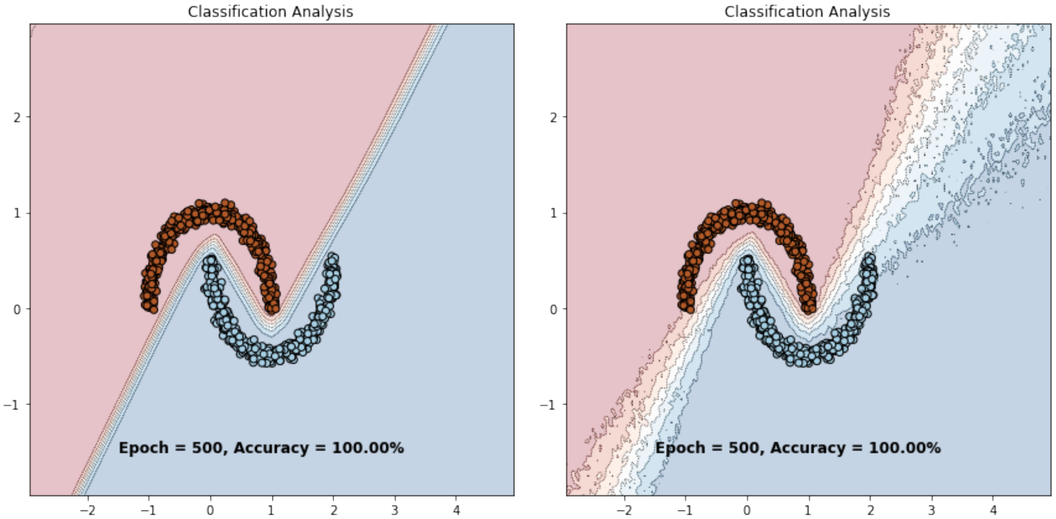


Figure 4: Left : MLP model and Right : MC Dropout

In this case we are using a dropout as a Bayesian approximation which helps us in getting a classification result that is more relevant for the current data distribution. As we can see in the Right image of the Figure 4 we get much better results for the classification, with classification that is adapted according to the area around the distribution. Whereas before on the Left image of the Figure 4 we get straight lines that maybe let us classify the data distribution accurately but visually we can see that the one on the right is better representative of the underlying data distribution.

Practical 3

Part 1 : Monte-Carlo Dropout on MNIST

Question 1.1: 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?

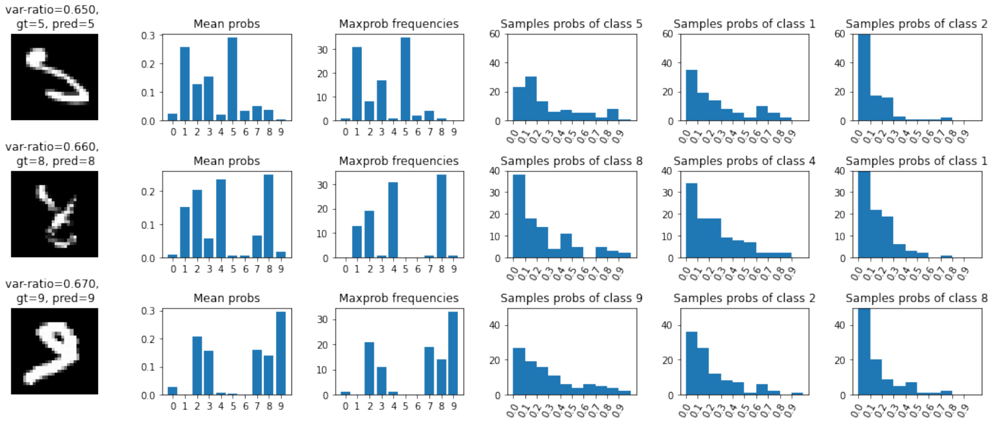


Figure 5: Top-3 most uncertain images along with their var-ratios value

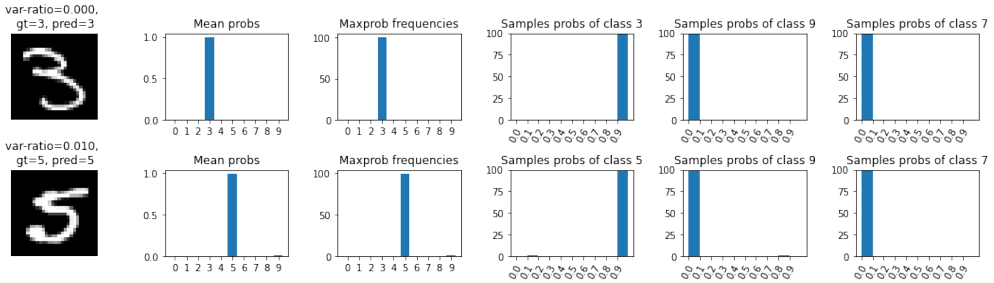


Figure 6: Plots for random images with their var-ratios value

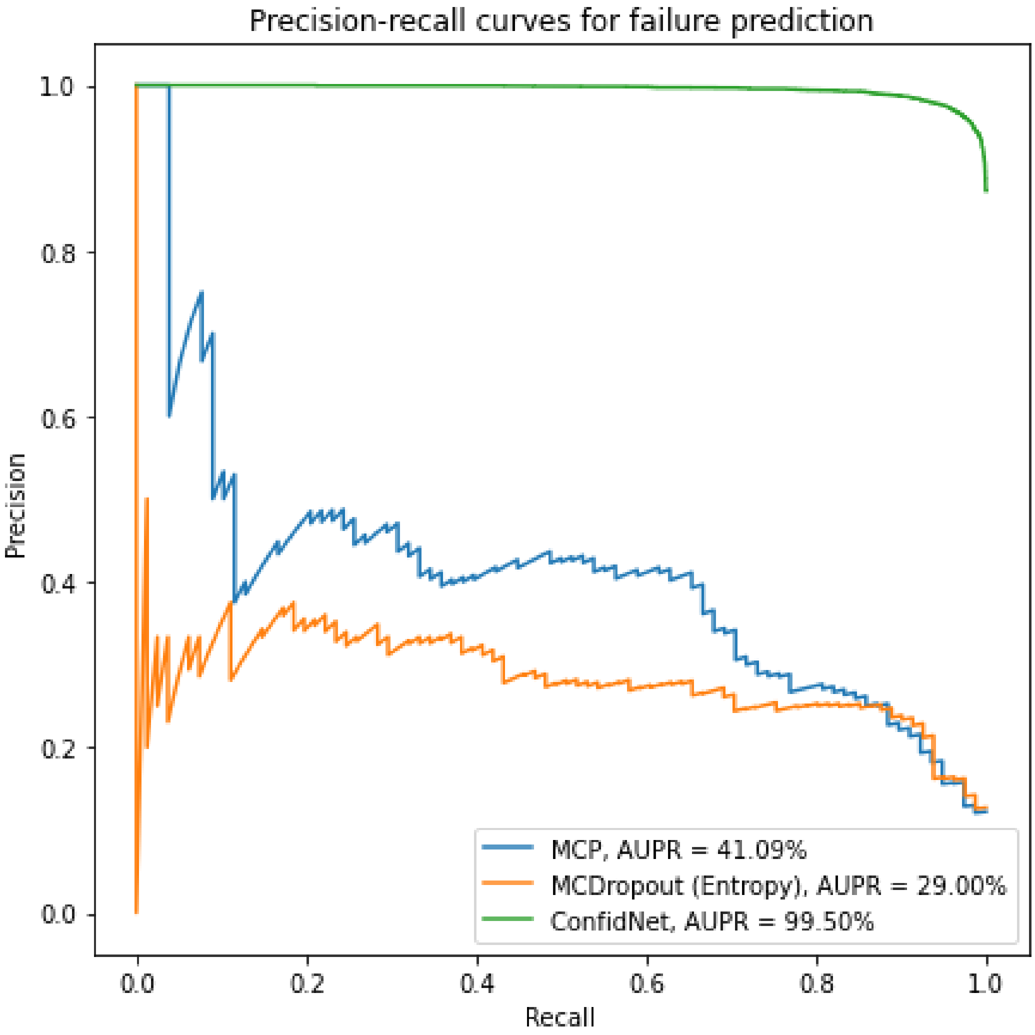
Comparing the top-3 most uncertain images with the random images we can say that for the random images the model seems to be quite certain for the choice of the classification therefore the distribution of probability is located exactly at the value the model choose to classify the image.

Whereas in the top-3 uncertain images we get a distribution of probability that spans across the different classes that the model seems to look at as suitable values for the classification. This also explain why the var-ratio for the top-3 uncertain images is much higher compared to the random images.

To resume, in the probability distribution of the uncertain images the model classifies them into multiple classes whereas for the previous random we didn't have the uncertainty in the classification results as we only had a unique class as a result.

Part 2 : Failure Prediction

Question 2.1: Compare the precision-recall curves of each method along with their AUPR values. Why did we use AUPR metric instead of standard AUROC?

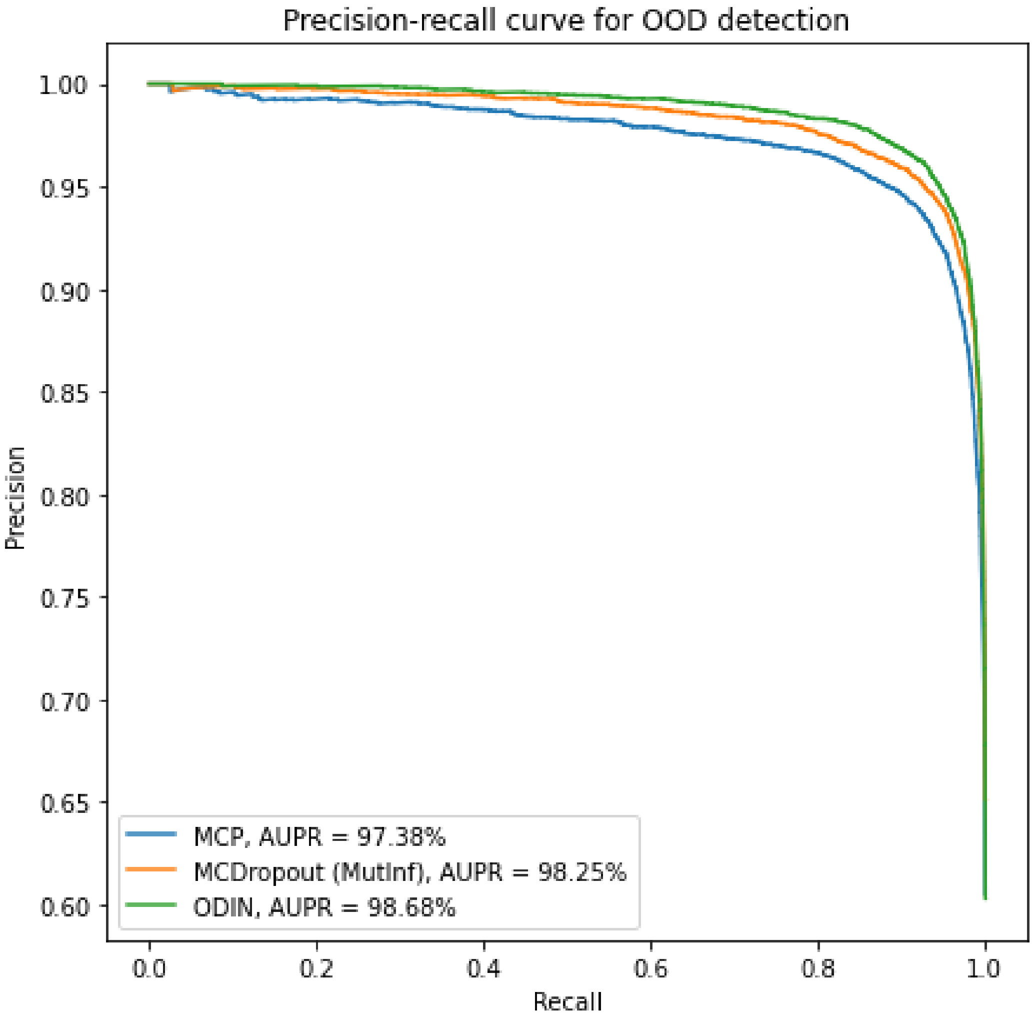


The AUPR values allow us to evaluate a confidence level for the predictions of the model. It allow us to distinguish the correct from the incorrect predictions of the model using the precision and recall metrics. This is helpful because depending on the problematic we encounter we need to choose between precision and recall.

In our case we tested three diffrent method (MCP, MCDropout, ConfidNet) and we can see that both the MCP (AUPR of 41%) and MCDropout(AUPR of 29%) gave us average results (with precision around 0.4) but the ConfidNet give us much better result as we get very high precision overall with an AUPR of 99.50%.

Part 3 : Out-of-distribution detection

Question 3.1: Compare the precision-recall curves of each OOD method along with their AUPR values. Which method perform best and why?



For all the methods (Mcdropout, MCP, ODIN) We can see the area under each curves get very close to 1. Although ODIN seems to be the most efficient of them as it gets an AUPR of 98.25%. ODIN is able to do error prediction by detecting the Out Of Distribution cases, this may be the cause of such a high AUPR value.