RDFIA - TME 4

Lucrezia Tosato & Marie Diez

Contents

1 Bayesian Linear Regression

1.1 Linear Basis function model

1. **Recall closed form of the posterior distribution in linear case.**

$$p(w|D,\alpha,\beta) = N(w,\mu,\Sigma)$$
$$\cdot\Sigma^{-1} = \alpha\boldsymbol{I} + \beta\Phi^T\Phi$$
$$\cdot\boldsymbol{\mu} = \beta\Sigma\Phi^T\mathrm{Y}$$

2. **Looking at the visualization of the posterior above, what can you say?**
As we gain more access to the dataset, we can see that our model is learning the proper parameters that bring its predictions closer to the reality, as demonstrated by the posterior distribution with figure 1 on the parameters' space.
As the number of examined data-points increases, the distribution of our model's parameters W becomes more precise (lower variance) and accurate (with a mean that gets closer and closer to the white target point) - getting closer to the right values in the parameter space.

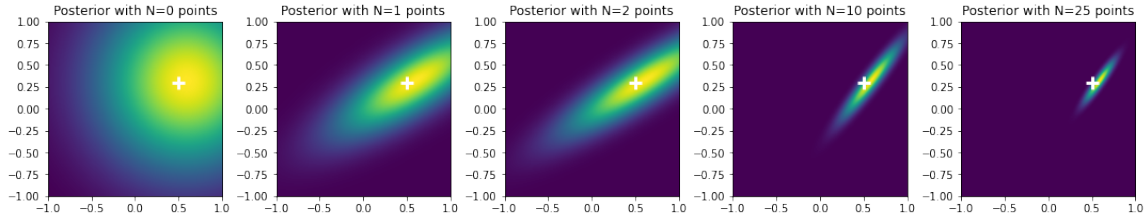

Figure 1: Posterior distribution - convergence as the number of available datapoints grow

3. **Week 1. 1 - Recall the closed form of the predictive distribution in linear case.**

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

4. **Week 1. 2/3 - Analyse these results. Describe the behavior of the predictive variance for points far from training distribution. Prove it analytically in the case where $\alpha = 0$ and $\beta = 1$.**
Figure 2 shows that for a location closer to the dataset, the predictions are both more accurate and precise, correctly predicting the values y for each x. As we go further away from it, our predicted variance grows, as does our error when comparing our prediction red line to the ground truth green line. This occurs because our model becomes less confident as we move away from the dataset in which it was trained, leading the variance to increase.

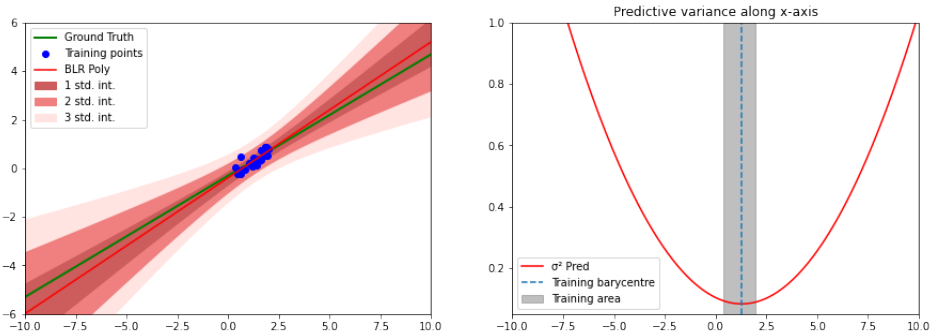

Figure 2: Predictions on the test dataset using a linear ϕ

Analytically, we have that our predicted variance is provided by: with $\alpha = 0$ and $\beta = 1$:

$$\Sigma^{-1} = \alpha \boldsymbol{I} + \beta \Phi^T \Phi = \Phi^T \Phi \ = \begin{pmatrix} N & \sum_i xi \\ \sum_i xi & \sum_i xi^2 \end{pmatrix}$$

So we have,

$$\Sigma = \frac{1}{det(\Sigma^{-1})} \begin{pmatrix} \sum_i xi^2 & -\sum_i xi \\ \sum_i xi & N \end{pmatrix}$$

The variance (epistemic uncertainty) is then :

$$\Phi\left(\mathrm{x}^*\right)^T \Sigma \Phi\left(\mathrm{x}^*\right) = \frac{1}{det(\Sigma^{-1})} (1 \ x^*) * \begin{pmatrix} \sum_i xi^2 & -\sum_i xi \\ \sum_i xi & N \end{pmatrix} * (1 \ x^*).T$$

$$= \frac{1}{det(\Sigma^{-1})} * (\sum_i xi^2 - x^* \sum_i xi - \sum_i + N x^*) * (1 \ x^*).T = \frac{1}{det(\Sigma^{-1})} * (\sum_i xi^2 - 2x \sum_i xi + N(x^*)^2)$$

$$= \frac{1}{det(\Sigma^{-1})} * (\sum_i (xi - x^*)^2)$$

We can see that the far away xi is from the barycenter x^* the more the variance increase.

5. **Week 1. 4 - Bonus: What happens when applying Bayesian Linear Regression on the following dataset?**

We can see that the model does not work well on this type of data. The predicted uncertainty is very bad, it should be strong when we move away from the points, which is not the case here. This is due to the mean that is in this case in between the clusters.

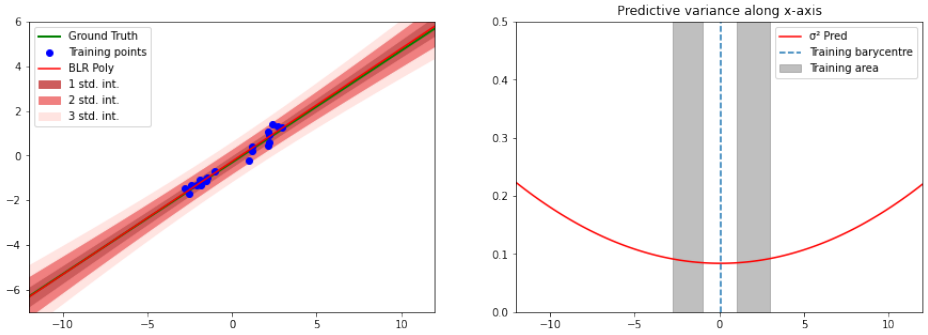

Figure 3: Predictions on the "hole dataset" using a linear ϕ

1.2 Non Linear models

1.2.1 Polynomial basis functions

6. **What can you say about the predictive variance?**
Figure 4 shows that the predictive variance increases as we move away from the points in our dataset, indicating that the model is losing confidence in its predictions. The graph on the right clearly depicts this behavior.

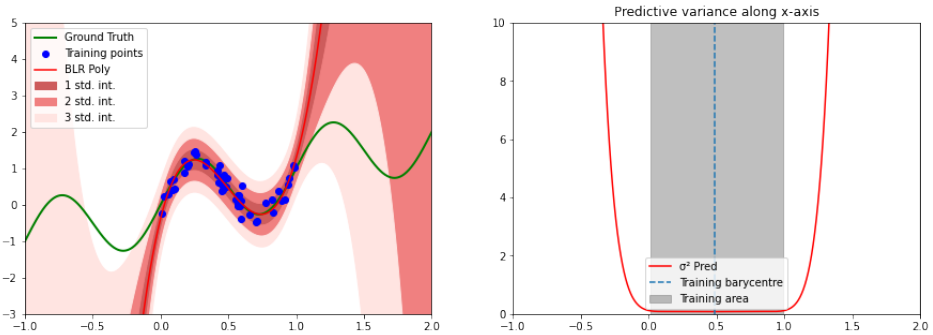

Figure 4: Predictions on the sinusoidal dataset using a polynomial ϕ with D = 5

1.2.2 Gaussian basis functions

7. **What can you say this time about the predictive variance? What can you conclude?**
With Figure 5 we can see that the predictive variance fluctuates a little along the region of our training points but eventually converges to a fixed value after going out of this region, as well represented on the graph of the right. Also, we can see once again with the graph of the left how good the prediction is when we are closer to our training points, and how worse it becomes as we go farther from them. We can conclude that the chosen basis function has an importance weight on the predictive variance behavior.

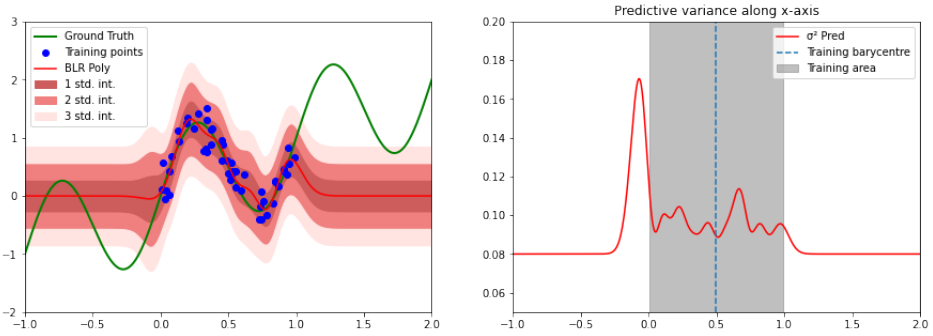

Figure 5
Predictions on the sinusoidal dataset using a gaussian ϕ

8. **Explain why in regions far from training distribution, the predictive variance converges to this particular value when using localized basis functions such as Gaussians.**
When using localized basis functions, such as Gaussians, the problem we have is that the epistemic uncertainty $\phi(x^*)^T \cdot \Sigma \cdot \phi(x^*)$ goes to zero as we are farther from our training points, which should not happen. This way only the constant factor $\frac{1}{\beta}$ remains on the computation of the predictive variance with its complete formulation on the equation below.

$$\sigma^2_{pred}(x^*) = \frac{1}{\beta} + \phi(x^*)^T \cdot \Sigma \cdot \phi(x^*)$$

This happens because localized basis functions tend to degenerate as we go farther from the training data. On the case of our Gaussians, for example, as we go farther from the μ_j values that we are using

on our function ϕ the function output degenerates to zero, which explains this phenomenon expressed in the question 7.

We can check by printing $1/dataset\_linear['BETA']$ and we obtain 0.0800 as seen on Figure 5

2 Approximate Inference

2.1 Bayesian Logistic Regression

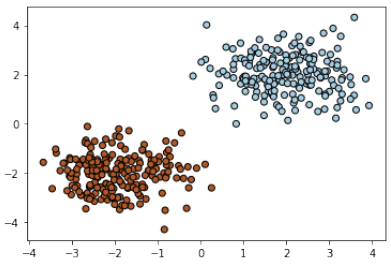

Figure 6: Linear dataset

2.1.1 Maximum A Posteriori Estimate

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? We can see in Figure 7 that the uncertainty does not increase when we are far away from the data and that's because the MAP estimate doesn't provide an estimation of the likelihood $p(w|X, Y)$ only for 1 point, the map estimate w_{map}, so we have $p(w|X, Y) \approx \delta(w - w_{map})$

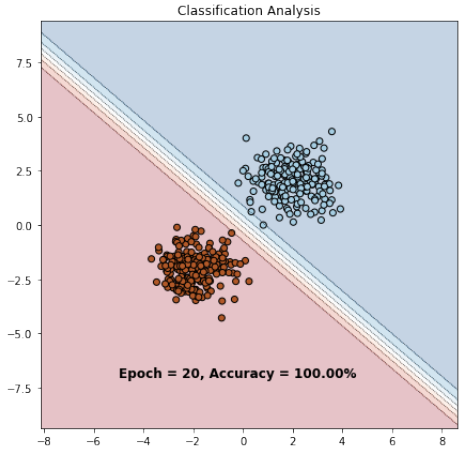

Figure 7: Logistic Regression model with stochastic gradient descent for 20 epochs.

2.1.2 Laplace Approximation

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

With classical logistic regression (Fig. 7), the model doesn't associate a good estimation of the uncertainty. Using Laplace approximation to estimate the intractable posterior (Fig. 8), the certainty decreases with distance. The computation of the hessian matrix is however super computationally heavy, also we will have problem if we assume quadratic posteriors.

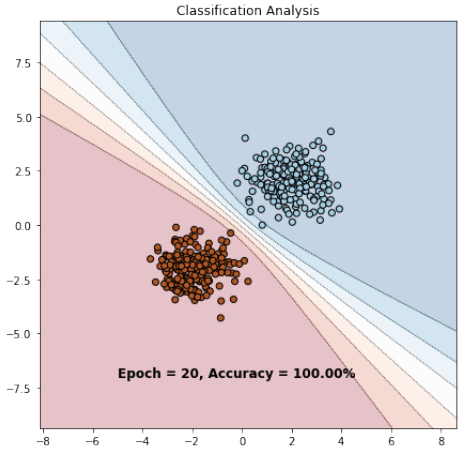

Figure 8: Using Laplace approximation to estimate the intractable posterior

2.1.3 Variational Inference

Week 2. 1 - Part I.3 "Variational inference" : comment the class LinearVariational

LinearVariational class enable to compute an approximation of the distribution $q_\theta(w)$ of the unknown true posterior $p(w|\mathcal{D})$ thanks to the minimization of the KL divergence. For that we define q_θ as a Gaussian distribution, we use the reparametrization trick to enable that the gradient flow through the model. We can finally define the predictive distribution thanks to the monte carlo sampling of w_s where $w_s \sim q_\theta^*$ are samples from the optimum variational distribution.
So in the class we define :

- The logistic regression model $f(x) = \sigma(w^T x + b)$ with σ the sigmoid function
- the sampling function sample $w_{l,s}$ with the monte carlo sampling from the optimal distribution $w_{l,s} \sim \mathcal{N}(\mu_l, \Sigma_l^2)$ where $\varepsilon_s \sim \mathcal{N}(0, \boldsymbol{I}_l)$, with $I_l \in \mathbb{R}^l$ the identity vector of size l.
- Then we can apply the forward pass with the samples w and same for the biais b.

3. Analyze the results provided by previous plot. Compared to previous MAP estimate, how does the predictive distribution behave? With variational inference by minimizing the KL divergence between the likelihood $p(w|X, Y)$ and a parametrize function $q_\theta(w)$ we can approximate the likelihood distribution and have a good results as seen below. Indeed the uncertainty increase when we are far away from the data.

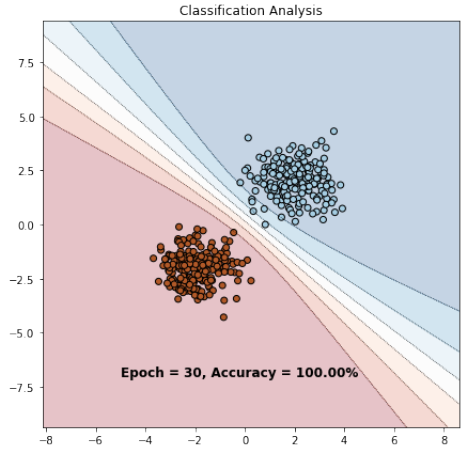

Figure 9: Doing a re implementation of variational inference

2.2 Bayesian Neural Networks

Week 2. 2 - Results and analysis of VI on the non-linear dataset (moons)

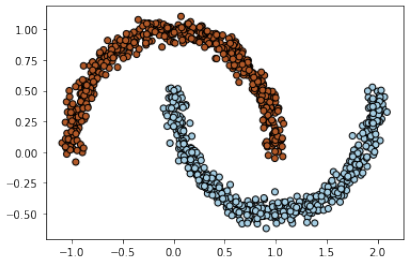

Figure 10: Two moons dataset

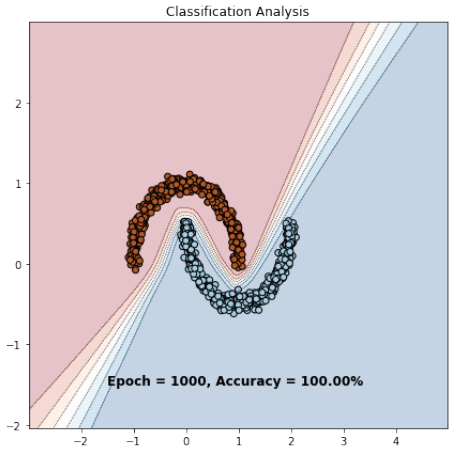

Figure 11: BNN with variational inference layers

Here with MLP with variational inference layers, we can now classify non linear dataset and get an uncertainty when we are far away from the data.

2.2.1 MonteCarlo dropout

1. Week 2. 3 - Again, analyze the results showed on plot. What is the benefit of MC Dropout variational inference over Bayesian Logistic Regression with variational inference?
With linear layer and without dropout we are not able to get an interesting uncertainty estimation. But the dropout enable to fix that by randomly dropping some activations, we can actually be seen as an approximate variational inference method.

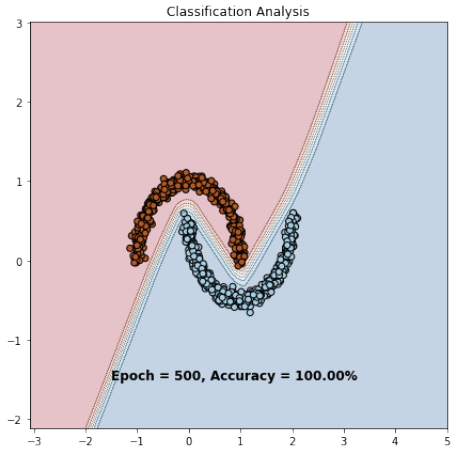

Figure 12: Before Monte Carlo Dropout

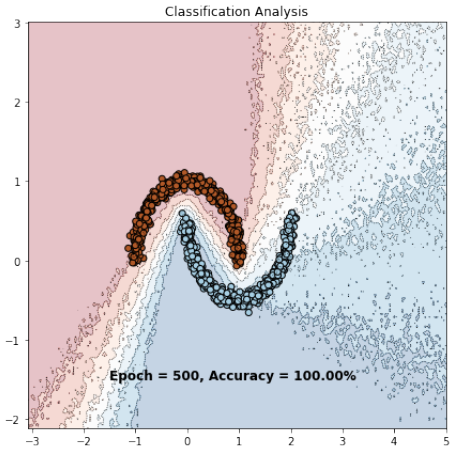

Figure 13: After Monte Carlo Dropout

If we train our model for more epochs (e.g. 1000 epochs), the model will still have an accuracy of 100% however the variance will not be as wide as in the figure, it will be a bit smaller.
The benefits of using drop out are:

- Possibility to change the dropout rate, also at the code level it is easier to implement
- Get results without increasing the complexity

3 Uncertainty Applications

3.1 Monte-Carlo Dropout on MNIST

3.1.1 LeNet-5 network with dropout layers

3.1.2 Investigating most uncertain samples

1. **Week 3. 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?**

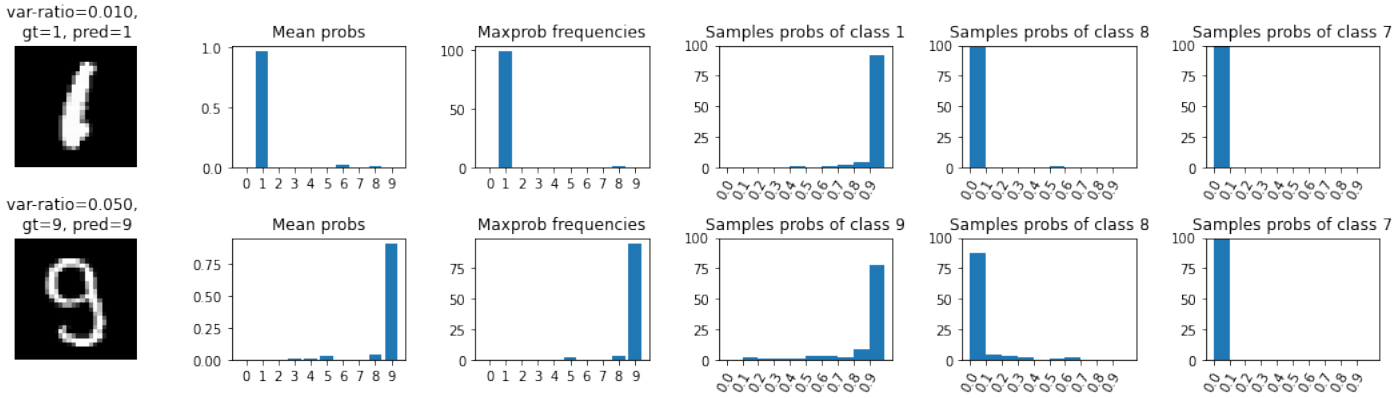

Figure 14: Random sample

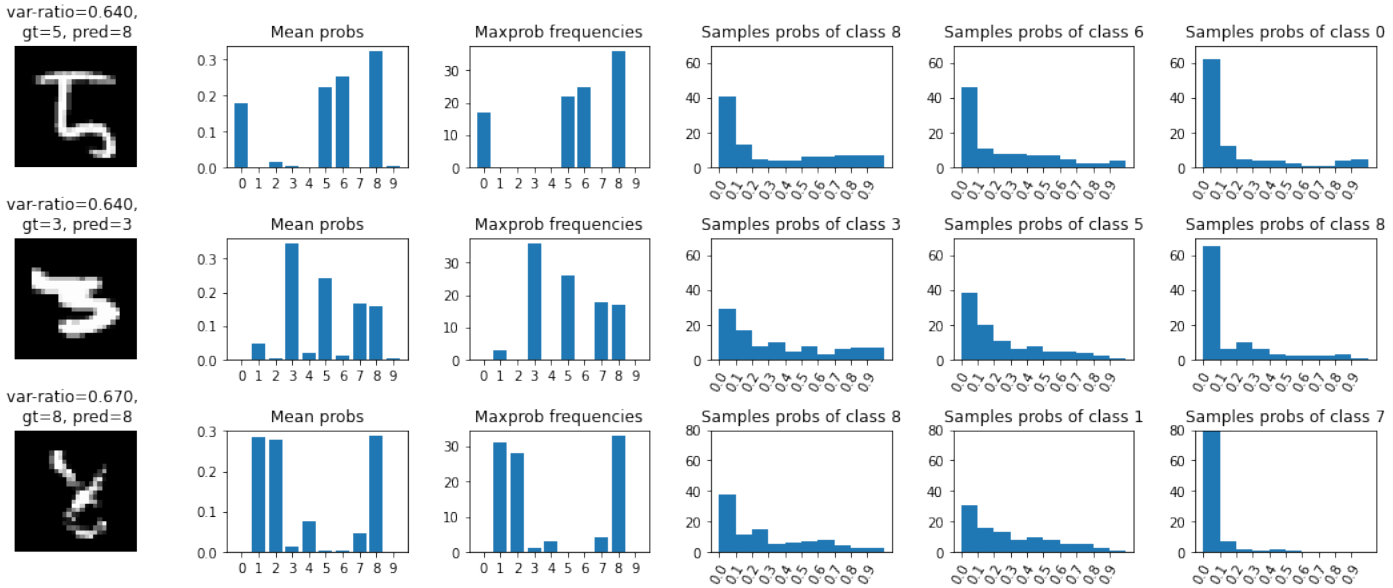

Figure 15: Most uncertain images according to the var-ratio

About the images themselves we can notice that the most uncertain image are indeed not so clear to identify.

- The histogram of the mean and Maxprob probability clarify to us why a certain class was chosen, especially, we can see the probability that was given to each class and of how much a certain class won with respect to the others. We can see that for random image we have very high probability for the the good number, unlike the 3 most uncertain results where the histogram is more homogeneous.
- In the Sample prob we can see the probability to be the sample number. For the random image 1

we can see a high probability of class 1 and very low for class 8 that make sense. It's the same for the second image. However for the 3 most uncertain images we can clearly see that the histograms are not as easy than for the random images.

Indeed as the images are uncertain the histogram reflect this uncertainty.

3.2 Failure prediction

Week 3. 2 - Goal of failure prediction

It's important to have an idea about the failure prediction to have an idea about the uncertainty of the network.

3.2.1 Evaluate failure prediction performances

1. **Week 3. 3 - Compare the precision-recall curves of each method along with their AUPR values. Why did we use AUPR metric instead of standard AUROC?**

- ROC AUC is the area under the curve where x is false positive rate (FPR) and y is true positive rate (TPR)
- PR AUC is the area under the curve where x is recall and y is precision.
  - Precision : Is the proportion of relevant items among all proposed items / How many true positives out of all that have been predicted as positives.
  - Recall : Is the proportion of relevant items proposed among all relevant items / How many positive cases have been classified correctly out of all positive cases in the data.

The main difference between the AUROC and AUPR lies in its tractability for unbalanced classes. When we have an "unbalanced class problem", it's more interesting to work with the precision and recall metrics rather than FPR and TPR (Unbalanced with low error rate).

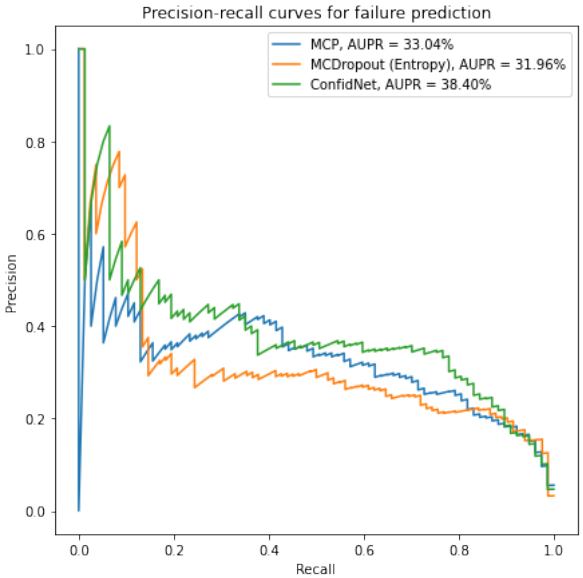

Figure 16: AUPR

Week 3. 4 - Comments on MCP, MCDropout and ConfidNet performances
It can be observed that for the three methods used, the network fails to predict its errors (33% for the MCP, 32% for the MCDropout entropy and 38% for Confidnet). The ideal would be to have curves that tend more towards the upper right-hand side of the graph.
The graph present some irregularities due to computation, we have an improvement but not as much as we actually expected. We can see that CondidNet has a higher AUPR values than MCDropout or MCP thanks to the consideration of the True Class Probability as a suitable uncertainty criterion.

3.3 Out-of-distribution detection

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

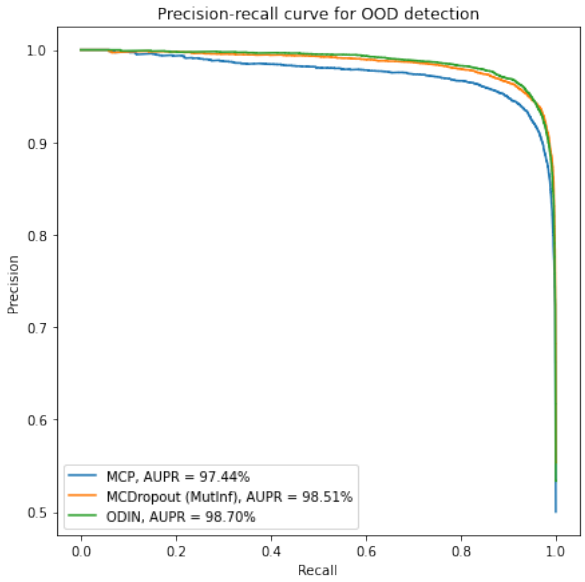

Figure 17: OOD

For out-of-detection samples, the different methods show very good results, the curves stretch almost perfectly towards the upper right side of the graph, for a recall of 80%, we have an accuracy of over 97%. The best performing method is ODIN, as it has the highest curve with respect to the other 2 methods, we can indeed see a higher air under the curve for ODIN.

4 Conclusion

The notion of robustness and uncertainty are important for models in deep learning.

We have seen that the variance is a form of uncertainty and decreases when we are close to the learning points.

We have estimated the posterior distribution by several methods. We have seen that the dropout allows the uncertainty prediction and can be seen as an approximate variational inference method.

The uncertainty can be used for the detection of "Out-of-distribution" samples.