

ROB 313: Assignment 4

Ali Seifeldin 1003894431

In this assignment,

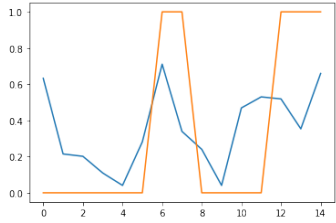
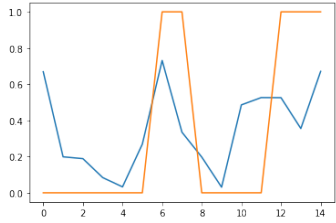
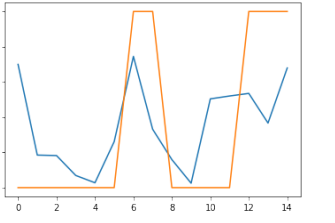
1)

a)

We use the gradient of the logs to update the weights of the model, and we use the Hessian of the model to compute the marginal log likelihood. The process is similar to what was done in assignment 3. Below is a summary of the results

We train our models in order until the max gradient is below 000001 for the first 2 models, and for the 3rd model. 0.0001. We use a learning rate of 0.005.

We get the following result

Variance	0.5	1	2
Plot			
Accuracy	0.73333	0.73333	0.666666666
Marginal log Likelihood	-69.09987109	-65.31653277	-62.15839875
Iteration	1089	1840	1881

Model with variance of 0.5 variance has the highest complexity, as it has the lowest Marginal log Likelihood

Based on marginal Likelihood, we would pick the first model. As it has the largest value.

b)

Attempted but could do nothing sadly.

2)

Engineering Safety in Machine Learning:

The paper discusses safety in Machine learning, specially techniques that can be used to achieve safety. The paper breaks down safety applications into type A and B, where type A applications are those in which safety is important, while type B is where risk minimization is sufficient. A good way to define safety is 'Safety is the reduction or minimization of risk and uncertainty of harmful events.' In order to model safety, we use the different information given in order to come up with a response. The response has an uncertainty measurement associated with it, therefore being able to reduce it increases the level of safety, as opposed to minimizing the loss function.

There are 4 approaches to safety in machine learning:

- Inherent design: Models are able to detect similarities and relationships between data efficiently, sometimes too efficiently. Cleaning up the data and eliminating relationships that are not there in real life can increase safety

- Safety Reserve: By allowing a certain margin of uncertainty in order to get a better approximation of the uncertainty

- Safe fail: Build a system in which if the model is not able to make a prediction due to high uncertainty, it returns nothing, and the system is able to deal with that accordingly

- Procedural Safeguards: Come from testing the model extensively in different scenarios and with different uncertainties to better the behavior

I think the paper did a good job at describing different techniques that can be used to increase safety in machine learning models, as well as provide a direction to the reader as to where they might continue to implement these ideas. I felt the paper could have been clearer and structured in a better format, as well as give more examples of the implementations of the 4 approaches