Alvin Pane [1004281118]

11 April 2020

ROB313: Assignment 4

Objectives

The objective of this assignment is to study Bayesian implementations of machine learning. Bayesian inferencing will be performed for a logistic regression model on the iris dataset with a Bernoulli likelihood and Gaussian prior. In addition, the complexity of 3 different prior variances will be compared, and the most probable predictive posterior on elements of the test set were estimated with a prior variance of 1. Finally, the paper INSERT NAME HERE will be studied, with a brief summary and comments on strength and weaknesses of INSERT HERE given.

Code Structure

The code was designed in a way to optimize the experience for the person running the code. The main strategy was to make the code modular. This was achieved by defining different functions, each responsible for handling a small task. Where applicable, general functions were written such that they could be called multiple times with different inputs. This was in the interest of space efficiency. The code uses print statements frequently to present data to the user in a clear manner. The main section calls functions depending on what question is to be answered. There are 2 variables: Q1a and Q1b, all initialized to False. To run a question, simply set the variable equal to True.

- Q1a uses 1 function, called marginal_LL(..), which computes the log marginal likelihood, given an input variance. It makes use of 3 helper functions, sigmoid(..), LL(..), and Hess(..). These compute the sigmoid function, the log likelihood, and the Hessian, respectively
- Q1b makes use the function sampling(..) with a slew of helper functions, which, for a proposal distribution, computes the test and validation negative log-likelihood, and the test and validation accuracy ratios.

Q1 ***Bayesian Inferencing***

a) Using a Laplace approximation, the log marginal likelihood was computed for the merged training and validation sets for the iris dataset. First, gradient descent was applied (with a learning rate = 0.001) as in Assignment 3, to compute the MAP solution. Inferences were made using the sigmoid function:

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

This sigmoid function returns a number on the range 0 to 1, which represents the conditional probability of a data point being Class 1. The log marginal likelihood was then computed using

the following equation at 3 different prior variances, $\sigma^2 = 0.5$, $\sigma^2 = 1$, $\sigma^2 = 2$. The results are summarized in Table 1.

Table 1: Log Marginal Likelihoods at Different Variance Values

Prior Variance	Log Marginal Likelihood
0.5	-75.014376
1	-74.50592642
2	-75.035156

The complexity of each model can be deduced by analyzing this log marginal likelihood. For a dataset, the more complex a model is, the smaller the log marginal likelihood becomes. This means that the most complex model here will be the one with the most negative log marginal likelihood. By this analysis, the most complex model is the one with a prior variance of 2, and the least complex model is the one with a prior variance of 1. It is important to note that the difference in log marginal likelihoods between the 3 models is quite small. This indicates that they are relatively similar in complexity, especially the models with prior variances of 0.5 and 2. When two models have equivalent fits to a dataset, the model with lower complexity will give less error on any future data. This being said, we expect the model with a prior variance of 1 to thus have less error and perhaps to be the best selection among the three.

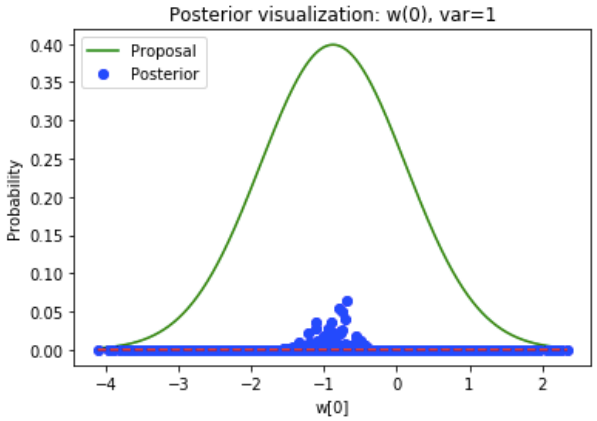
b) With an importance sampling approach, the predictive posterior class on each element of the Iris test set was computed. A multivariate Gaussian with variance $\sigma^2_{proposal}$ centered at the MAP solution (with prior variance of 1) was used for the proposal distribution. The means, $\mu_{proposal}$ are computed from the solution of Q1a. The optimal proposal was selected by looking at the negative log-likelihood on the validation set, and choosing the parameters which minimized the validation negative log-likelihood. Sample sizes in the set [10, 25, 50, 100, 200, 300] and proposal variances in the set [1, 2, 2.5, 5, 10] were considered. Selecting the optimal parameters for sample size and proposal variance, the predictive posterior was computed on the test set, giving the results summarized in Table 2 below.

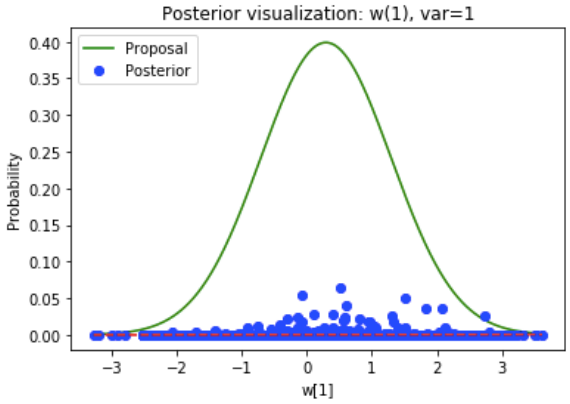
Table 2: Summary of Test Predictive Posterior Results

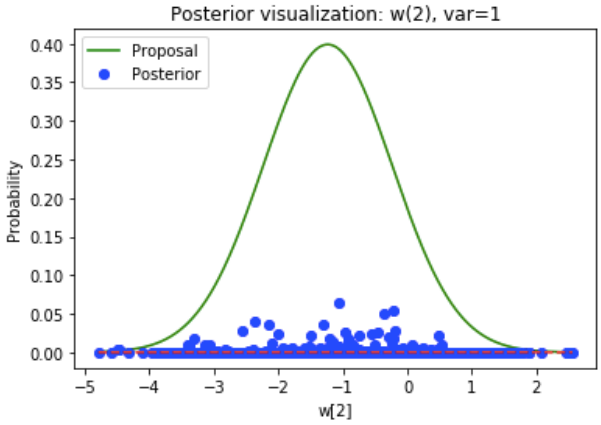
$\sigma^2_{proposal}$ (Proposal Variance)	Sample Size	Test Negative Log-Likelihood	Test Accuracy Ratio	Validation Negative Log-Likelihood	Validation Accuracy Ratio
1	50	7.003	0.800	14.392	0.774

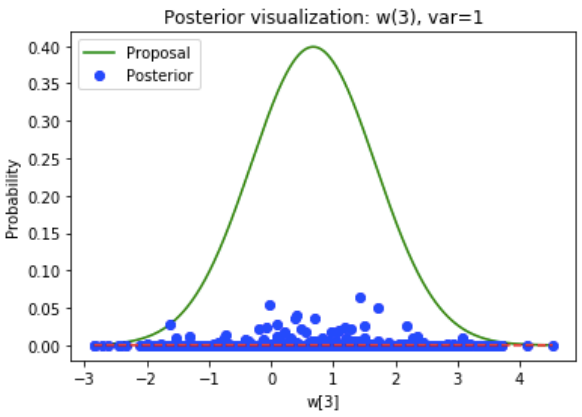
As can be seen, this proposal yields a relatively high accuracy on both sets, at a value of roughly 80% for both. Another thing of note is that the validation log likelihood is about double the test log likelihood. This is likely due to the fact that there is an unequal amount of data in each set.

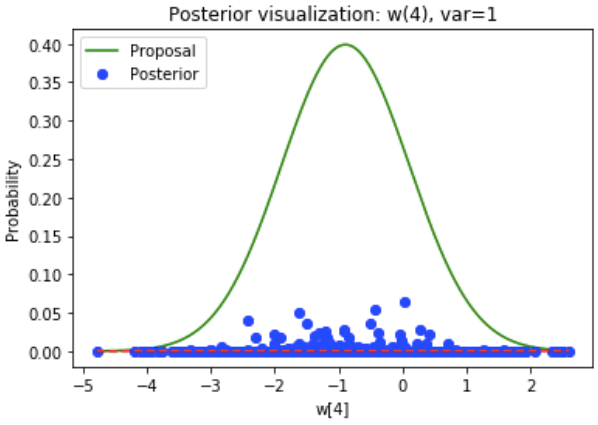
The validation set has 31 data points, roughly twice more than the 15 in the test set, which may account for this. To further justify this selected proposal, we can visualize the posterior plotted with the proposal distribution plotted for each component of the weight vector, w, in Figures 1-5, below.











Figures 1-5: Plots of Proposal and Posterior Distributions, for each component of w

As shown by these figures, there is significant overlap between the posterior and proposal distributions, especially in areas of high probability mass. Thus, the accuracy of the proposal and its validity as being the optimal selection is justified.

Q2 ***Literature Review***

Paper Chosen: *A Few Useful Things to Know about Machine Learning,* Pedro Domingos, University of Washington

The premise of this paper is to give some key tips for machine learning practitioners to write classification algorithms. He breaks things down into 3 components:

- Representation (Model Type)
- Evaluation (Evaluation Function)
- Optimization (Techniques)

The goal of the paper is to design ML models that are best capable of generalizing to new data, and have a low variance and bias. The paper also touches upon common challenges in designing a good classifier, with techniques to bypass them.

A successful ML model needs to embody knowledge and assumptions beyond just the data. This is important for effectively generalizing. A model needs to reason from inductance, generating large output knowledge, from little input. The writer makes it clear that his belief is that knowledge and data are two separate things. I believe that this is a weakness in the paper. This concept of "embodiment of knowledge" fails to make it clear that any knowledge in the model is first transferred from the data, and without data there can be no knowledge. Having good, clean data is equally important to creating a good model. The author then goes on to discuss the Curse of Dimensionality. He writes that the issue with high dimensions comes from the fact that we live in a 3-dimensional world. I believe the author makes a strong argument here as lack of intuition was one of my own struggles when trying to understand higher dimensional problems. The last few sections covers the Blessing of Non-Uniformity, feature engineering, more data in contrast to algorithm complexity, and warns about some incorrect implications that ML designers make. The author does well here to support all his arguments with supporting examples.