Assignment 4: Bayesian Classification

Zhengmin Yang
University of Toronto

April 11, 2020

1 Introduction

Code Structure The code is again very modularized. For user convenience, importance sampling and log marginal likelihood are separated into two functions, `importance_sampling` and `lml`. In addition, there is a sanity check in the main function in case the user wants to check that the fundamental algorithm for calculating the MAP estimate works. This can be useful for debugging in case the user changes anything in the program and causes any errors in the code. They can check whether the error occurred during the MAP estimate or whether it occurred in other parts such as importance sampling or log marginal likelihood.
The main function has three boolean values that can be changed to produce the desired results by the user. `sanity_check` can be triggered to check if the gradient descent algorithm works, `log_marginal_likelihood` can be set to True to check if the log marginal likelihood function works, and `im_sampling` can be checked if the user wants to see the results of importance sampling. In addition, the number of samples drawn can be changed by changing `num_samples`. This method ensures easy readability and modifiability without the user having to look through the source code directly.

Objectives In this report we will explore using Bayesian techniques to model classification problems in machine learning. We will only consider binary classification. For binary classification, the prediction function is a sigmoid function (1) and the negative log likelihood, which is also the loss function, is given (2).

$$sigmoid(x) = \frac{1}{1+e^{x}} \tag{1}$$

$$\log Pr(y|w,X) = \sum_{i=1}^{N} y^{(i)} \log(\hat{f}(x^{(i)};w)) + (1-y^{(i)}) \log(1-\hat{f}(x^{(i)};w)) \tag{2}$$

The Bayesian approach calculates a posterior distribution of the weights **w** rather than a specific weight, as shown:

$$Pr(w|y,X) = \frac{Pr(y|w,X)Pr(w)}{Pr(y|X)} \tag{3}$$

Where the denominator, $Pr(y|X)$ is known as the marginal likelihood and is a high dimensional integral that is difficult to compute under ordinary circumstances. However, using Laplace's approximation we can directly approximate the posterior as a normal distribution without having to calculate the marginal likelihood:

$$\mathcal{N}(w^{(*)}, -H^{(-1)}) \tag{4}$$

Where $w^{(*)}$ is the MAP solution to the optimization problem, found back in Assignment 3 using Gradient Descent or Stochastic Gradient Descent. H is the Hessian matrix given by:

$$\begin{aligned} H &= \nabla^2 \log Pr(y|w,X)Pr(w) \\ &= \nabla^2 \log Pr(y|w,X) + \nabla^2 \log Pr(w) \\ &= \sum_{i=1}^{N} \hat{f}(x^{(i)};w)[\hat{f}(x^{(i)};w)-1]\bar{x}^{(i)} - \frac{1}{\sigma^2} I \end{aligned} \tag{5}$$

For our current problem I will use the Gradient Descent MAP solution as the gradient descent solution converges in the fewest number of iterations and produces almost constant average loss values after convergence unlike SGD. Note that the weights for Gradient Descent (GD) are updated as following:

$$w^{(k+1)} = w^{(k)} - \eta \nabla \log Pr(y|w^{(k)}, X) \tag{6}$$

We will also assume that the prior distribution, $Pr(w)$ is a multivariate Normal distribution centred at $\mathbf{0}$ with variance σ^2, i.e.

$$Pr(w) = \mathcal{N}(w|0, \sigma^2 I) \tag{7}$$

2 Log Marginal Likelihood

The log marginal likelihood is calculated by rearranging (3) and taking the log of both sides of the equation. Plugging in $w = w^{(*)}$ and (4) and simplifying, we get that the log marginal likelihood is:

$$\log Pr(y|X) = \log Pr(y|w^{(*)}, X) + \log Pr(w^{(*)}) + \frac{D+1}{2}\log(2\pi) - \frac{1}{2}\log|-H| \tag{8}$$

We will find the log marginal likelihood with three different variances on the prior, 0.5, 1, and 2. The highest variance, 2, gives a model with the highest complexity since the marginal likelihood would be more ”spread out” and would have nonzero probability values for a wider range of y’s. The log marginal likelihood values are shown below:

Variance	Log Marginal Likelihood
0.5	-81.2
1	-77.0
2	-75.7

Table 1: Log Marginal Likelihood for Different Variance Values

As we can see, the most complex model also gives the highest log marginal likelihood and therefore the highest probability that randomly sampled weights from $Pr(w)$ would give the most accurate predictions y.

3 Importance Sampling

Importance sampling uses Monte Carlo sampling to find the marginal likelihood by sampling weights from a proposal distribution $q(w)$ and approximating the marginal likelihood as a sum evaluated at the weights.

$$Pr(y|X) \approx \frac{1}{s}\sum_{i=1}^{s} r(w^{(i)}), w^{(i)} \; q(w) \tag{9}$$

where

$$r(w) = \frac{Pr(y|w, X)Pr(w)}{q(w)} \tag{10}$$

It can be combined with the formula for calculating the predictive posterior, leading to a Monte Carlo approximation of the predictive posterior.

$$Pr(y^{(*)}|y, X, x^{(*)}) \approx \sum_{i=1}^{s} Pr(y^{(*)}|w^{(i)}, x^{(*)}) \frac{r(w^{(i)})}{\sum_{j=1}^{s} r(w^{(j)})} \tag{11}$$

Our goal is to find the probability that the test points map to a single class. In this case we choose $y^{(*)} = 1$. We also choose the proposal distribution q(w) to be a Gaussian with a mean of w_{MAP} and the required variance $1I$. This is because the MAP estimate is representative of all of the $w^{(*)}$’s that give the most accurate results; by the law of large numbers all posterior solutions will converge to w_{MAP}. Therefore choosing a proposal with the maximum probability at the w_{MAP} will ensure that the two distributions overlap the best and therefore have a decent testing accuracy of 0.73 for 200 samples. Note that this matches the testing accuracy when using only the MAP estimate in Assignment 3. The large number of samples $w^{(*)}$, 200, ensured that the accuracy converged according to the law of large numbers.

4 Literature Review

We will analyse a paper on the aspect of safety within machine learning, starting with an overview of the paper and then delving into its strengths and weaknesses.

4.1 Overview

In Engineering Safety in Machine Learning, Varney discusses safety in machine learning, the present risks, and strategies to increase safety in machine learning. He defines safety in terms of the algorithm being able to minimize risk and epistemic uncertainty; risk is defined as the expected cost of harm and epistemic uncertainty is harm that results from insufficient knowledge about a problem. The four solutions to the problem that the author presented are having inherently safe design, safety reserves, safe fail, and procedural safeguards. Having inherently safe design means that we try to exclude parts of the machine learning procedure that could cause damage. Safety reserves means putting in buffers to the machine learning system, such as limiting the maximum error or making the risk of harm roughly the same across all demographic groups. Varney describes safe fail whereby measures are put into place such that the system does not cause harm even in the case when it fails, such as rejecting unreliable predictions. Lastly procedural safeguards are measures that involve aspects outside the machine learning process, such as creating open source code that allows for public auditing and creating user-friendly UX's that guide users in setting up the program safely.

4.2 Analysis

The author was able to define the concept of safety very rigorously, thus providing a solid foundation for his arguments. In addition, in the first 3 solutions his examples of these solutions applied to other common engineering applications made it very clear to the reader the nature of the solution. In addition, he linked the examples of the effects of each solution back to his definitions of safety, further strengthening his argument. However Varney's fourth argument on Procedural Safeguards lacked this. In addition, arguments for how procedural safeguards increase safety are vague and do not use his previous definitions of safety in terms of risk minimization and epistemic uncertainty minimization. Instead, arguments were provided holistically without quantitative evidence. Either a stronger link back to his definitions on safety or quantitative evidence on the effectiveness of procedural safeguards would make pursing solutions in that category more enticing.