FIT5201 Assignment - 1 Report

Semester 1 2020

Jaimon Thyparambil Thomas
Student ID: 29566428
Email : jthy0001@monash.student.edu
Monash University

May 09, 2020

Question 4 [Bayes Rule, 20 Marks]

Recall the simple example from Appendix A of Module 1. Suppose we have one red, one blue, and one yellow box. In the red box we have 3 apples and 5 oranges, in the blue box we have 4 apples and 4 orange, and in the yellow box we have 3 apples and 1 orange. Now suppose we randomly selected one of the boxes and picked a fruit. If the picked fruit is an apple, what is the probability that it was picked from the yellow box?

Note that the chances of picking the red, blue, and yellow boxes are 50%, 30%, and 20% respectively and the selection chance for any of the pieces from a box is equal for all the pieces in that box. Please show your work in your PDF report.

Hint: You can formulize this problem following the denotations in "Random Variable" paragraph in Appendix A of Module 1.

Let x denote the colour of the box and y denote the fruit

It is Given that

Probability of selecting the red box = $P(x = red) = 0.5$

Probability of selecting the blue box = $P(x = blue) = 0.3$

Probability of selecting the yellow box = $P(x = yellow) = 0.2$

Probability of selecting an apple from yellow box =

$P(x = yellow, y = apple) = \frac{3}{20}$

Probability of selecting an apple = $P(y = apple) = \frac{10}{20}$

Probability of selecting an apple given that we are selecting it from yellow box = $P(x = yellow \mid y = apple) = \frac{P(x=yellow,y=apple)}{P(y=apple)} = \frac{3}{20} * \frac{20}{10} = \frac{3}{10} = 0.3$

Therefore, the probability of fruit being picked from yellow box given that the picked fruit is apple is 0.3

Question 5 [Ridge Regression, 25 Marks]

I. Given the gradient descent algorithms for linear regression (discussed in Chapter 2 of Module 2), derive weight update steps of stochastic gradient descent (SGD) for linear regression with L2 regularisation norm. Show your work with enough explanation in your PDF report; you should provide the steps of SGD.

Hint: Recall that for linear regression we defined the error function E. For this assignment, you only need to add an L2 regularization term to the error function (error term plus the regularization term). This question is similar to Activity 1 of Module 2.

We Know that the Error function without regularisation constant for Stochastic gradient descent(SGD) is

$$E(w^{\tau}) \coloneqq \frac{1}{2}(t_{\tau} - w^{\tau}.\phi(x_{\tau}))^2$$

On Applying L2 regularisation on the error function. The error function now becomes

$$E(w^{\tau}) \coloneqq \frac{1}{2}(t_{\tau} - w^{\tau}.\phi(x_{\tau}))^2 \; + \frac{\lambda}{2}\sum_{j=0}^{M-1}(w_j^{\tau})^2$$

Equation 1

On applying derivative of E(w) with respect to w we get

$$\nabla E(w^{\tau}) \coloneqq -(t_{\tau} - w^{\tau}.\phi(x_{\tau})) \; + \lambda\sum_{j=0}^{M-1} w_j^{\tau}$$

Equation 2

We also know that in stochastic gradient descent the weights have to be updated using the formula

$$w^{(t)} = w^{(t-1)} - \; \eta'\nabla E(w^{(t-1)})$$

$$w^{(\tau)} = w^{(\tau-1)} + \; \eta'((t_{\tau} - w^{\tau}.\phi(x_{\tau})) - \lambda\sum_{j=0}^{M-1} w_j^{\tau})$$

Algorithm for applying stochastic gradient descent using L2 regularisation is

Algo 1 - this is the algorithm used in the Sgd function in code

```
• Initialise the parameters to $w^{(0)}$, $\eta'$ ,$\tau$=1,
• $\tau_{max} = 20 * length(train\ data)$
• while a stopping condition is not met do:
    o while $E(w) > \epsilon\ or\ \tau < \tau_{max}$ do
        ▪ Shuffle the data and labels
        ▪ i=1
        ▪ while $E(w^{(\tau)}) > \epsilon\ or\ \tau < \tau_{max}\ or\ i < length(train\ data)$ do
            ▪ $w^{(\tau)} = w^{(\tau-1)} + \eta'((t_{\tau} - w^{\tau}.\phi(x_{\tau})) - \lambda\sum_{j=0}^{M-1} w_j^{\tau})$

            ▪ $\tau = \tau + 1$
```

Algo 2 - This is the algorithm not used as it is taking too much time for executing

Note that this Algo portion is commented in the code

```
• Initialise the parameters to $w^{(0)}$, $\eta, \tau = 1$ ,
  $\tau_{max} = 20 * length(train\ data)$
• while a stopping condition is not met do:
    o while $E(w) > \epsilon\ or\ \tau < \tau_{max}$ do
        ▪ Shuffle the data and labels
        ▪ i=1
        ▪ while $E\left(w^{(\tau)}\right) > \epsilon\ or\ \tau < \tau_{max}\ or\ i < length(train\ data)$ do
            ▪ $\eta' = \eta$
            ▪ $terminate$ = true
            ▪ While($terminate$)
                ▪ $w^{(\tau)} = w^{(\tau-1)} + \eta'((t_{\tau} - w^{\tau}.\phi(x_{\tau})) - \lambda \sum_{j=0}^{M-1} w_j^{\tau})$
                ▪ $terminate = E\left(w^{(\tau)}\right) > E(w^{(\tau-1)})$
                ▪ $\eta' = \frac{\eta'}{2}$
            ▪ $\tau = \tau + 1$
```

Also Note that E used in both the algorithms is the same values that we derived earlier in equation 1.