FIT 5201 Assignment 1 Report

Name: Akshay Sapra

ID: 29858186

Section C. Probabilistic Machine Learning

Bayes Rule (Question 4)

Recall the simple example from Appendix A of Module 1. Suppose we have one red and one blue box. In the red box we have 2 apples and 6 oranges, whilst in the blue 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 orange, what is the probability that it was picked from the blue box?

Note that the chance of picking the red box is 60% and the selection chance for any of the pieces from a box is equal for all the pieces in that box.

=>

- the event of the box to be chosen is denoted by **B**
- There can be two possibilities for this event:
  - **r** which means box chosen is red
  - **b** which means chosen box is blue

- Similarly, the identity of the fruit will be denoted by **F**.
- There can be two possibilities for this event:
  - '**a**' for apples
  - '**o**' for oranges

P(B = r)= 6/10 & P (B = b) = 4/10
Given, picked fruit is orange, we have to find probability of picking it from blue box
= P(B = b| F = o)

From **Conditional Probability**, it is known that

$$P(X|Y) = \frac{P(X \text{ and } Y)}{P(Y)} = \frac{P(Y|X) * P(X)}{P(Y)}$$

⇨ $P(B = b| F = o) = \frac{P(F = o| B = b) * P(B = b)}{P(F = o)} = \frac{P(F = o| B = b) * P(B = b)}{P(F = o| B = b) * P(B = b) + P(F = o| B = r) * P(B = r)}$

⇨ $\frac{1}{4} * \frac{4}{10} / (\frac{1}{4} * \frac{4}{10} + \frac{6}{8} * \frac{6}{10}) = 0.10/0.10+0.45$

⇨ 0.1818 or 18.18%

Section D. Ridge Regression

Ridge Regression (Question 5)

Task 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) as well as batch gradient descent (BGD) for linear regression with L2 regularisation norm. Show your work with enough explanation in your PDF report; you should provide the steps of SGD and BGD, separately.

<u>Note: Solution on next page</u>

Section D - Ridge Regression

Question 5)

Task 1)

$T \rightarrow$ denotes iteration number.

$\eta^{(t)} \rightarrow$ denotes learning rate parameter

According to linear basis functions

$$y(x, w) := w_0 + w_1 x_1 + w_2 x_2 + \dots + w_d x_d$$

or

$$y(x, w) := w \cdot \phi(x)$$

Generic Algorithm for gradient descent

- Initialise the parameter to $w^{(0)}$ & $t = 1$
- While stopping is not met
  - $\rightarrow \eta' = \eta$
  - $\rightarrow$ while $\eta' > \epsilon$ do
    - $w := w^{(t-1)} - \eta' \nabla E(w^{(t-1)})$
    - if $E(w) < E(w^{(t-1)})$ then break
    - $\eta' = \eta' / 2$
  - $\rightarrow w^{(t)} := w$
  - $\rightarrow t = t + 1$

In case of liner regression, Error function to minimise is:

$$E(w) := \frac{1}{2} \sum_{n=1}^{N} [t_n - w \cdot \phi(x_n)]^2$$

where $D := \{(x_n, t_n)\}_{n=1}^{N}$ is the training data

Gradient of Training objective is:

$$\nabla E(w) = -\sum_{n=1}^{N} [t_n - w \cdot \phi(x_n)] \phi(x_n)$$

Plugging it into Ridge regression, Cost function will be

$$C(w) = \frac{1}{2} \sum_{n=1}^{N} (t_n - w \cdot \phi(x_n))^2 + \frac{\lambda}{2} \sum_{j=0}^{M-1} w_j^2$$

$$C(w) = \frac{1}{2} \sum_{n=1}^{N} \left[(t_n - w \cdot \phi(x_n))^2 + \frac{\lambda}{N} w^T \cdot w \right]$$

$\lambda \equiv$ regularisation parameter

$$C_n(w) = \frac{1}{2} (t_n - w \phi(x_n))^2 + \frac{\lambda}{2N} w^T \cdot w$$

Weight update step for SGD:

$$w^{(T)} = w^{(T-1)} - \eta^{(T-1)} \nabla C_n(w^{(T-1)}) \quad \text{—(i)}$$

where $\nabla C_n(w) = -\phi(x_n)(t_n - w \cdot \phi(x_n)) + \frac{\lambda}{N} w$ —(ii)

using (i) & (ii)

$$w^{(T)} = w^{(T-1)} + \eta^{(T-1)} \left[\phi(x_n)(t - w^{(T-1)} \cdot \phi(x_n)) - \frac{\lambda}{N} w^{(T-1)} \right]$$

---

for SGD, weight update step:

$$\Rightarrow w^{(T)} = w^{(T-1)} + \eta^{(T-1)} \left[\phi(x_n)\left(t_n - w^{(T-1)} \cdot \phi(x_n)\right) - \frac{\lambda}{N} w^{(T-1)} \right]$$

Similarly for BGD

using,

$$w^{(T)} = w^{(T-1)} - \eta^{(T-1)} \nabla C\left(w^{(T-1)}\right) \quad \text{—(iii)}$$

&

$$\nabla C(w) = \sum_{n=1}^{N} -\phi(x_n)\left[t_n - w \cdot \phi(x_n)\right] + \lambda w \quad \text{—(iv)}$$

using (iii) & (iv), we will get weight update step for BGD

$$\Rightarrow \quad w^{(T)} := w^{(T-1)} + \eta^{(T-1)}\left[\sum_{n=1}^{N} \phi(x_n)\left[t_n - w^{(T-1)} \cdot \phi(x_n)\right] - \lambda w^{(T-1)}\right]$$