Section C. Probabilistic Machine Learning

Question 4 [Bayes Rule, 20 Marks]

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 apple, what is the probability that it was picked from the blue box?

Note that the chance of picking the red box is 40% 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.

P (Box = blue box) = $\frac{3}{5}$, P (Box = red box) = $\frac{2}{5}$

P (Fruit = apple) = $\frac{3+2}{4+12} = \frac{5}{12}$, P (Fruit = orange) = $\frac{1+7}{4+12} = \frac{7}{12}$

P (Fruit = apple | Box = red box) = $\frac{2}{2+6} = \frac{1}{4}$, P (Fruit = apple | Box = blue box) = $\frac{3}{4}$

P (Box = blue box | Fruit = apple) =

$$\frac{\text{P (Fruit = apple | Box = blue box)}*\text{P (Box = blue box)}}{\text{P (Fruit = apple | Box = blue box)}*\text{ P (Box = blue box)}+\text{ P (Fruit = apple | Box = red box)}*\text{P (Box = red box)}}$$

$$= \frac{3}{4} * \frac{3}{5} / \left(\frac{3}{4} * \frac{3}{5} + \frac{1}{4} * \frac{2}{5}\right) = \frac{9}{11}$$

Question 5 [Maximum Likelihood, 20 Marks]

As opposed to a coin which has two faces, a dice has 6 faces. Suppose we are given a dataset which contains the outcomes of 10 independent tosses of a dice: D:={1,4,5,3,1,2,6,5,6,6}. We are asked to build a model for this dice, i.e. a model which tells what is the probability of each face of the dice if we toss it. Using the maximum likelihood principle, please determine the best value for our model parameters. Please show your work in your PDF report.

Step 1: Take a simple probabilistic model for our dice, in which the probability of coming face is denoted by the *parameter* w where $0 \leq w \leq 1$.

P (x = 1) =W_1 for face 1, P (x ≠ 1) = 1 – W_1

P (x = 2) =W_2 for face 1, P (x ≠ 2) = 1 – W_2

P (x = 3) =W_3 for face 1, P (x ≠ 3) = 1 – W_3

$P(x = 4) = W_4$ for face 1, $P(x \neq 4) = 1 - W_4$

$P(x = 5) = W_5$ for face 1, $P(x \neq 5) = 1 - W_5$

$P(x = 6) = W_6$ for face 1, $P(x \neq 6) = 1 - W_6$

Step 2: Calculate the probability of having the dataset generated from the dice.

$D = W_1 * W_4 * W_5 * W_3 * W_1 * W_2 * W_6 * W_5 * W_6 * W_6 = W_1^2 * W_4 * W_5^2 * W_3 * W_2 * W_6^3$

Step 3: Compute the probability of the dataset based on our parametric model for the dice

$P(D|\text{dice model}) = \Pi_{i=1}^{10} P(x_i|\text{dice model}) = W_1^2 * (1-W_1)^8$

$\frac{dL(w)}{dw} = 0 \quad \Rightarrow \quad \frac{d[2logw+8\log(1-w)}{dw} = 0 \quad \Rightarrow \frac{2}{w} - \frac{8}{1-w} = 0 \quad \Rightarrow \quad W_1 = \frac{2}{2+8} = \frac{1}{5}$

$P(D|\text{dice model}) = \Pi_{i=1}^{10} P(x_i|\text{dice model}) = W_2 * (1-W_2)^9$

$\frac{dL(w)}{dw} = 0 \quad \Rightarrow \quad \frac{d[logw+9\log(1-w)}{dw} = 0 \quad \Rightarrow \frac{1}{w} - \frac{9}{1-w} = 0 \quad \Rightarrow \quad W_2 = \frac{1}{1+9} = \frac{1}{10}$

$P(D|\text{dice model}) = \Pi_{i=1}^{10} P(x_i|\text{dice model}) = W_3 * (1-W_3)^9$

$\frac{dL(w)}{dw} = 0 \quad \Rightarrow \quad \frac{d[logw+9\log(1-w)}{dw} = 0 \quad \Rightarrow \frac{1}{w} - \frac{9}{1-w} = 0 \quad \Rightarrow \quad W_3 = \frac{1}{1+9} = \frac{1}{10}$

$P(D|\text{dice model}) = \Pi_{i=1}^{10} P(x_i|\text{dice model}) = W_4 * (1-W_4)^9$

$\frac{dL(w)}{dw} = 0 \quad \Rightarrow \quad \frac{d[logw+9\log(1-w)}{dw} = 0 \quad \Rightarrow \frac{1}{w} - \frac{9}{1-w} = 0 \quad \Rightarrow \quad W_4 = \frac{1}{1+9} = \frac{1}{10}$

$P(D|\text{dice model}) = \Pi_{i=1}^{10} P(x_i|\text{dice model}) = W_5^2 * (1-W_5)^8$

$\frac{dL(w)}{dw} = 0 \quad \Rightarrow \quad \frac{d[2logw+8\log(1-w)}{dw} = 0 \quad \Rightarrow \frac{2}{w} - \frac{8}{1-w} = 0 \quad \Rightarrow \quad W_5 = \frac{2}{2+8} = \frac{1}{5}$

$P(D|\text{dice model}) = \Pi_{i=1}^{10} P(x_i|\text{dice model}) = W_6^3 * (1-W_6)^7$

$\frac{dL(w)}{dw} = 0 \quad \Rightarrow \quad \frac{d[3logw+7\log(1-w)}{dw} = 0 \quad \Rightarrow \frac{3}{w} - \frac{7}{1-w} = 0 \quad \Rightarrow \quad W_6 = \frac{3}{3+7} = \frac{3}{10}$

Step 4: The best value for our model parameters

$W_1 = W_5 = 0.2, \quad W_2 = W_3 = W_4 = 0.1, \quad W_6 = 0.3$

Section D. Ridge Regression

Question 6 [Ridge Regression, 25 Marks]

1.Given the gradient descent algorithms for linear regression (discussed in Chapter 2 of Module2), 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.

SGD

The standard gradient descent algorithm updates the parameters θ of the objective $J(\theta)$ as, $\theta = \theta - \alpha\nabla_\theta E[J(\theta)]$, where the expectation in the above equation is approximated by evaluating the cost and gradient over the full training set.

Update θ, $\theta = \theta - \alpha\nabla_\theta J(\theta;\, x^{(i),} y^{(i)})$ with a pair $(x^{(i),} y^{(i)})$ from the training set.

Use one example in each iteration.

```
a) Random shuffle dataset)
b) Repeat {
           For i = 1… {
                $\theta'_j$ = $\theta_j$+ ($y^i$ - $h_\theta$($x^i$ ))$X_j^i$
             (for j = 0,… )
           }

   }
```

BGD

Use all example in each iteration.

```
Repeat {

     $\theta'_j$ = $\theta_j$ + $\frac{1}{m}\sum_{i=1}^{m}$( $y^i$- $h_\theta$($x^i$ ))$X_j^i$

     (for every j = 0,…)

}
```