School of Computer Science and Applied Mathematics
UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

COMS4030 Assignment 2 - report

Wesley Earl Stander - 1056114 - Coms Hons

June 20, 2020

1 Question 1

1.1 Question 1.1

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 + \theta_6 x_6 + \theta_7 x_7 + \theta_8 x_8 \tag{1}$$

$$\begin{aligned} \text{where } x_2 &= x_1^2 \\ x_3 &= x_1^3 \\ x_4 &= x_1^4 \\ x_5 &= x_1^5 \\ x_6 &= x_1^6 \\ x_7 &= x_1^7 \\ x_8 &= x_1^8 \end{aligned} \tag{2}$$

Using the model in equation 1,θ is acquired experimentally from gradient descent as:

$$\theta = \begin{bmatrix} 1 \\ 0.00202016 \\ -0.50843811 \\ 0.0151267 \\ 0.02739159 \\ 0.0077916 \\ -0.00392878 \\ 0.00047956 \\ -0.00001908 \end{bmatrix}^T \tag{3}$$

Hence to validate that the model is accurate the following values accuracy are calculated using the model and their absolute error.

$$\begin{gathered}
\text{Let } x = 0 \\
\therefore h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 + \theta_6 x_6 + \theta_7 x_7 + \theta_8 x_8 \\
h_\theta(0) = (1) + (0.00202016)(0) + (-0.50843811)(0)^2 + (0.0151267)(0)^3 + (0.02739159)(0)^4 + \\
(0.0077916)(0)^5 + (-0.00392878)(0)^6 + (0.00047956)(0)^7 + (-0.00001908)(0)^8 = 1 \\
cos(0) = 1 \\
|h_\theta(0) - cos(0)| = |1-1| = 0 \leq 10^{-3} \therefore \text{accurate for this record}
\end{gathered} \tag{4}$$

$$\begin{gathered}
\text{Let } x = \frac{\pi}{4} \\
\therefore h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 + \theta_6 x_6 + \theta_7 x_7 + \theta_8 x_8 \\
h_\theta(\frac{\pi}{4}) = (1) + (0.00202016)(\frac{\pi}{4}) + (-0.50843811)(\frac{\pi}{4})^2 + (0.0151267)(\frac{\pi}{4})^3 + (0.02739159)(\frac{\pi}{4})^4 + \\
(0.0077916)(\frac{\pi}{4})^5 + (-0.00392878)(\frac{\pi}{4})^6 + (0.00047956)(\frac{\pi}{4})^7 + (-0.00001908)(\frac{\pi}{4})^8 = 0.707199 \\
cos(\frac{\pi}{4}) = 0.707106 \\
|h_\theta(\frac{\pi}{4}) - cos(\frac{\pi}{4})| = |0.707199 - 0.707106| = 0.000093 \leq 10^{-3} \therefore \text{accurate for this record}
\end{gathered} \tag{5}$$

$$\begin{gathered}
\text{Let } x = \frac{\pi}{2} \\
\therefore h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 + \theta_6 x_6 + \theta_7 x_7 + \theta_8 x_8 \\
h_\theta(\frac{\pi}{2}) = (1) + (0.00202016)(\frac{\pi}{2}) + (-0.50843811)(\frac{\pi}{2})^2 + (0.0151267)(\frac{\pi}{2})^3 + (0.02739159)(\frac{\pi}{2})^4 + \\
(0.0077916)(\frac{\pi}{2})^5 + (-0.00392878)(\frac{\pi}{2})^6 + (0.00047956)(\frac{\pi}{2})^7 + (-0.00001908)(\frac{\pi}{2})^8 = 0.000145 \\
cos(\frac{\pi}{2}) = 0 \\
|h_\theta(\frac{\pi}{2}) - cos(\frac{\pi}{2})| = |0.000145 - 0| = 0.000145 \leq 10^{-3} \therefore \text{accurate for this record}
\end{gathered} \tag{6}$$

1.2 Question 1.2

The same polynomial could be used to model the logistic (Sigmoid) function $f(x) = 1/(1+e^{-x})$. The accuracy can be validated by performing similar calculations as above. Using the model in equation 1, θ is acquired experimentally from gradient descent as:

$$\theta = \begin{bmatrix} 0.5 \\ 0.24220219 \\ 0.00002787 \\ -0.01488692 \\ -0.00000652 \\ 0.00061516 \\ 0.00000049 \\ -0.00001007 \\ -0.00000001 \end{bmatrix}^T \tag{7}$$

$$\begin{gathered} \text{Let } x = 0 \\ \therefore h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 + \theta_6 x_6 + \theta_7 x_7 + \theta_8 x_8 \\ h_\theta(0) = (0.5) + (0.24220219)(0) + (0.00002787)(0)^2 + (-0.01488692)(0)^3 + (-0.00000652)(0)^4 + \\ (0.00061516)(0)^5 + (0.00000049)(0)^6 + (-0.00001007)(0)^7 + (-0.00000001)(0)^8 = 0.5 \\ 1/(1+e^{-0}) = 0.5 \\ |h_\theta(0) - 1/(1+e^{-0})| = |0.5 - 0.5| = 0 \leq 10^{-3} \therefore \text{accurate for this record} \end{gathered} \tag{8}$$

$$\begin{gathered} \text{Let } x = 2 \\ \therefore h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 + \theta_6 x_6 + \theta_7 x_7 + \theta_8 x_8 \\ h_\theta(2) = (0.5) + (0.24220219)(2) + (0.00002787)(2)^2 + (-0.01488692)(2)^3 + (-0.00000652)(2)^4 + \\ (0.00061516)(2)^5 + (0.00000049)(2)^6 + (-0.00001007)(2)^7 + (-0.00000001)(2)^8 = 0.88374 \\ 1/(1+e^{-2}) = 0.8808 \\ |h_\theta(2) - 1/(1+e^{-2})| = |0.88374 - 0.8808| = 0.00294 \nleq 10^{-3} \therefore \text{not accurate for this record} \end{gathered} \tag{9}$$

It can be seen from $x = 2$ that the same polynomial does not have the accuracy required for the model to be justified according to the accuracy heuristic. It can be seen however that the model is very close and if the accuracy heuristic was 3 times larger, the model would be accurate to model the logistic (Sigmoid) function.

2 Question 2

2.1 Question 2.1

$$\begin{aligned}
s_1 &= x_1\theta_1 + x_2\theta_3 \\
s_2 &= x_1\theta_2 + x_2\theta_4 \\
s_3 &= g(s_1)\theta_5 + g(s_2)\theta_7 = g(x_1\theta_1 + x_2\theta_3)\theta_5 + g(x_1\theta_2 + x_2\theta_4)\theta_7 \\
s_4 &= g(s_1)\theta_6 + g(s_2)\theta_8 = g(x_1\theta_1 + x_2\theta_3)\theta_6 + g(x_1\theta_2 + x_2\theta_4)\theta_8 \\
u_1 &= g(s_1) = g(x_1\theta_1 + x_2\theta_3) \\
u_2 &= g(s_2) = g(x_1\theta_2 + x_2\theta_4) \\
u_3 &= g(s_3) = g(g(x_1\theta_1 + x_2\theta_3)\theta_5 + g(x_1\theta_2 + x_2\theta_4)\theta_7) \\
u_4 &= g(s_4) = g(g(x_1\theta_1 + x_2\theta_3)\theta_6 + g(x_1\theta_2 + x_2\theta_4)\theta_8)
\end{aligned} \tag{10}$$

2.2 Question 2.2

$$g(z) = z \text{ and } \frac{\delta}{\delta z}(g(z)) = \frac{\delta(z)}{\delta z} \tag{11}$$

$$\begin{gathered}
\frac{\delta J(\theta)}{\delta\theta_k} = \frac{\delta}{\delta\theta_k}[\frac{1}{2N}\sum_{n=1}^{N}[(\hat{y}_1^{(n)} - y_1^{(n)})^2 + (\hat{y}_2^{(n)} - y_2^{(n)})^2 + \frac{\lambda}{2}\sum_{i=1}^{8}\theta_i^2]] \\
\text{where } \hat{y}_1^{(n)} = u_3^{(n)} = g(s_3)^{(n)} \text{ and } \hat{y}_2^{(n)} = u_4^{(n)} = g(s_4)^{(n)} \\
\text{and } g(s_1)^{(n)} = s_1^{(n)} = x_1^{(n)}\theta_1 + x_2^{(n)}\theta_3 \\
\text{and } g(s_2)^{(n)} = s_2^{(n)} = x_1^{(n)}\theta_2 + x_2^{(n)}\theta_4 \\
\text{and } g(s_3)^{(n)} = s_3^{(n)} = g(x_1^{(n)}\theta_1 + x_2^{(n)}\theta_3)\theta_5 + g(x_1^{(n)}\theta_2 + x_2^{(n)}\theta_4)\theta_7 = g(s_1)^{(n)}\theta_5 + g(s_2)^{(n)}\theta_7 \\
\text{and } g(s_4)^{(n)} = s_4^{(n)} = g(x_1^{(n)}\theta_1 + x_2^{(n)}\theta_3)\theta_6 + g(x_1^{(n)}\theta_2 + x_2^{(n)}\theta_4)\theta_8 = g(s_1)^{(n)}\theta_6 + g(s_2)^{(n)}\theta_8
\end{gathered} \tag{12}$$

For $k = 8$

$$\frac{\delta J(\theta)}{\delta\theta_8} = \frac{\delta}{\delta\theta_8}[\frac{1}{2N}\sum_{n=1}^{N}[(\hat{y}_1^{(n)} - y_1^{(n)})^2 + (\hat{y}_2^{(n)} - y_2^{(n)})^2 + \frac{\lambda}{2}\sum_{i=1}^{8}\theta_i^2]]$$
$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_4)^{(n)} - y_2^{(n)})\frac{\delta}{\delta\theta_8}(g(s_4)^{(n)} - y_2^{(n)}) + \frac{\lambda\theta_8}{2}]$$
$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_4)^{(n)} - y_2^{(n)})\frac{\delta}{\delta\theta_8}(s_4^{(n)}) + \frac{\lambda\theta_8}{2}]$$
$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_4)^{(n)} - y_2^{(n)})(g(s_2)^{(n)}) + \frac{\lambda\theta_8}{2}]$$
$$\therefore \theta_8 \leftarrow \theta_8 - \alpha[\frac{1}{N}\sum_{n=1}^{N}[(g(s_4)^{(n)} - y_2^{(n)})(g(s_2)^{(n)}) + \frac{\lambda\theta_8}{2}]] \tag{13}$$

For $k = 7$

$$\frac{\delta J(\theta)}{\delta\theta_7} = \frac{\delta}{\delta\theta_7}[\frac{1}{2N}\sum_{n=1}^{N}[(\hat{y}_1^{(n)} - y_1^{(n)})^2 + (\hat{y}_2^{(n)} - y_2^{(n)})^2 + \frac{\lambda}{2}\sum_{i=1}^{8}\theta_i^2]]$$
$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_1^{(n)})\frac{\delta}{\delta\theta_7}(g(s_3)^{(n)} - y_1^{(n)}) + \frac{\lambda\theta_7}{2}]$$
$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_2^{(n)})\frac{\delta}{\delta\theta_7}(s_3^{(n)}) + \frac{\lambda\theta_7}{2}]$$
$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_2^{(n)})(g(s_2)^{(n)}) + \frac{\lambda\theta_7}{2}]$$
$$\therefore \theta_7 \leftarrow \theta_7 - \alpha[\frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_2^{(n)})(g(s_2)^{(n)}) + \frac{\lambda\theta_7}{2}]] \tag{14}$$

For $k = 6$

$$\frac{\delta J(\theta)}{\delta\theta_6} = \frac{\delta}{\delta\theta_6}[\frac{1}{2N}\sum_{n=1}^{N}[(\hat{y}_1^{(n)} - y_1^{(n)})^2 + (\hat{y}_2^{(n)} - y_2^{(n)})^2 + \frac{\lambda}{2}\sum_{i=1}^{8}\theta_i^2]]$$
$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_4)^{(n)} - y_2^{(n)})\frac{\delta}{\delta\theta_6}(g(s_4)^{(n)} - y_2^{(n)}) + \frac{\lambda\theta_6}{2}]$$
$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_4)^{(n)} - y_2^{(n)})\frac{\delta}{\delta\theta_6}(s_4^{(n)}) + \frac{\lambda\theta_6}{2}]$$
$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_4)^{(n)} - y_2^{(n)})(g(s_1)^{(n)}) + \frac{\lambda\theta_6}{2}]$$
$$\therefore \theta_6 \leftarrow \theta_6 - \alpha[\frac{1}{N}\sum_{n=1}^{N}[(g(s_4)^{(n)} - y_2^{(n)})(g(s_1)^{(n)}) + \frac{\lambda\theta_6}{2}]] \tag{15}$$

For $k = 5$

$$\frac{\delta J(\theta)}{\delta \theta_5} = \frac{\delta}{\delta \theta_5}[\frac{1}{2N}\sum_{n=1}^{N}[(\hat{y}_1^{(n)} - y_1^{(n)})^2 + (\hat{y}_2^{(n)} - y_2^{(n)})^2 + \frac{\lambda}{2}\sum_{i=1}^{8}\theta_i^2]]$$

$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_1^{(n)})\frac{\delta}{\delta \theta_5}(g(s_3)^{(n)} - y_1^{(n)}) + \frac{\lambda \theta_5}{2}]$$

$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_2^{(n)})\frac{\delta}{\delta \theta_5}(s_3^{(n)}) + \frac{\lambda \theta_5}{2}]$$

$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_2^{(n)})(g(s_1)^{(n)}) + \frac{\lambda \theta_5}{2}]$$

$$\therefore \theta_5 \leftarrow \theta_5 - \alpha[\frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_2^{(n)})(g(s_1)^{(n)}) + \frac{\lambda \theta_5}{2}]] \tag{16}$$

For $k = 4$

$$\frac{\delta J(\theta)}{\delta \theta_4} = \frac{\delta}{\delta \theta_4}[\frac{1}{2N}\sum_{n=1}^{N}[(\hat{y}_1^{(n)} - y_1^{(n)})^2 + (\hat{y}_2^{(n)} - y_2^{(n)})^2 + \frac{\lambda}{2}\sum_{i=1}^{8}\theta_i^2]]$$

$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_1^{(n)})\frac{\delta}{\delta \theta_4}(g(s_3)^{(n)} - y_1^{(n)}) + (g(s_4)^{(n)} - y_2^{(n)})\frac{\delta}{\delta \theta_4}(g(s_4)^{(n)} - y_2^{(n)}) + \frac{\lambda \theta_4}{2}]$$

$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_2^{(n)})\frac{\delta}{\delta \theta_4}(s_3^{(n)}) + (g(s_4)^{(n)} - y_2^{(n)})\frac{\delta}{\delta \theta_4}(s_4^{(n)}) + \frac{\lambda \theta_4}{2}]$$

$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_2^{(n)})\theta_7\frac{\delta}{\delta \theta_4}(s_2^{(n)}) + (g(s_4)^{(n)} - y_2^{(n)})\theta_8\frac{\delta}{\delta \theta_4}(s_2^{(n)}) + \frac{\lambda \theta_4}{2}]$$

$$= \frac{1}{N}\sum_{n=1}^{N}[x_2^{(n)}[(g(s_3)^{(n)} - y_2^{(n)})\theta_7 + (g(s_4)^{(n)} - y_2^{(n)})\theta_8] + \frac{\lambda \theta_4}{2}]$$

$$\therefore \theta_4 \leftarrow \theta_4 - \alpha[\frac{1}{N}\sum_{n=1}^{N}[x_2^{(n)}[(g(s_3)^{(n)} - y_2^{(n)})\theta_7 + (g(s_4)^{(n)} - y_2^{(n)})\theta_8] + \frac{\lambda \theta_4}{2}]] \tag{17}$$

$$\text{For } k = 3$$

$$\frac{\delta J(\theta)}{\delta\theta_3} = \frac{\delta}{\delta\theta_3}[\frac{1}{2N}\sum_{n=1}^{N}[(\hat{y}_1^{(n)} - y_1^{(n)})^2 + (\hat{y}_2^{(n)} - y_2^{(n)})^2 + \frac{\lambda}{2}\sum_{i=1}^{8}\theta_i^2]]$$

$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_1^{(n)})\frac{\delta}{\delta\theta_3}(g(s_3)^{(n)} - y_1^{(n)}) + (g(s_4)^{(n)} - y_2^{(n)})\frac{\delta}{\delta\theta_3}(g(s_4)^{(n)} - y_2^{(n)}) + \frac{\lambda\theta_3}{2}]$$

$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_2^{(n)})\frac{\delta}{\delta\theta_3}(s_3^{(n)}) + (g(s_4)^{(n)} - y_2^{(n)})\frac{\delta}{\delta\theta_3}(s_4^{(n)}) + \frac{\lambda\theta_3}{2}]$$

$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_2^{(n)})\theta_5\frac{\delta}{\delta\theta_3}(s_1^{(n)}) + (g(s_4)^{(n)} - y_2^{(n)})\theta_6\frac{\delta}{\delta\theta_3}(s_1^{(n)}) + \frac{\lambda\theta_3}{2}]$$

$$= \frac{1}{N}\sum_{n=1}^{N}[x_2^{(n)}[(g(s_3)^{(n)} - y_2^{(n)})\theta_5 + (g(s_4)^{(n)} - y_2^{(n)})\theta_6] + \frac{\lambda\theta_3}{2}]$$

$$\therefore \theta_3 \leftarrow \theta_3 - \alpha[\frac{1}{N}\sum_{n=1}^{N}[x_2^{(n)}[(g(s_3)^{(n)} - y_2^{(n)})\theta_5 + (g(s_4)^{(n)} - y_2^{(n)})\theta_6] + \frac{\lambda\theta_3}{2}]] \tag{18}$$

$$\text{For } k = 2$$

$$\frac{\delta J(\theta)}{\delta\theta_2} = \frac{\delta}{\delta\theta_2}[\frac{1}{2N}\sum_{n=1}^{N}[(\hat{y}_1^{(n)} - y_1^{(n)})^2 + (\hat{y}_2^{(n)} - y_2^{(n)})^2 + \frac{\lambda}{2}\sum_{i=1}^{8}\theta_i^2]]$$

$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_1^{(n)})\frac{\delta}{\delta\theta_2}(g(s_3)^{(n)} - y_1^{(n)}) + (g(s_4)^{(n)} - y_2^{(n)})\frac{\delta}{\delta\theta_2}(g(s_4)^{(n)} - y_2^{(n)}) + \frac{\lambda\theta_2}{2}]$$

$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_2^{(n)})\frac{\delta}{\delta\theta_2}(s_3^{(n)}) + (g(s_4)^{(n)} - y_2^{(n)})\frac{\delta}{\delta\theta_2}(s_4^{(n)}) + \frac{\lambda\theta_2}{2}]$$

$$= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_2^{(n)})\theta_7\frac{\delta}{\delta\theta_2}(s_2^{(n)}) + (g(s_4)^{(n)} - y_2^{(n)})\theta_8\frac{\delta}{\delta\theta_2}(s_2^{(n)}) + \frac{\lambda\theta_2}{2}]$$

$$= \frac{1}{N}\sum_{n=1}^{N}[x_1^{(n)}[(g(s_3)^{(n)} - y_2^{(n)})\theta_7 + (g(s_4)^{(n)} - y_2^{(n)})\theta_8] + \frac{\lambda\theta_2}{2}]$$

$$\therefore \theta_2 \leftarrow \theta_2 - \alpha[\frac{1}{N}\sum_{n=1}^{N}[x_1^{(n)}[(g(s_3)^{(n)} - y_2^{(n)})\theta_7 + (g(s_4)^{(n)} - y_2^{(n)})\theta_8] + \frac{\lambda\theta_2}{2}]] \tag{19}$$

$$\text{For } k = 1$$

$$\begin{aligned}
\frac{\delta J(\theta)}{\delta \theta_1} &= \frac{\delta}{\delta \theta_1}[\frac{1}{2N}\sum_{n=1}^{N}[(\hat{y}_1^{(n)} - y_1^{(n)})^2 + (\hat{y}_2^{(n)} - y_2^{(n)})^2 + \frac{\lambda}{2}\sum_{i=1}^{8}\theta_i^2]] \\
&= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_1^{(n)})\frac{\delta}{\delta \theta_1}(g(s_3)^{(n)} - y_1^{(n)}) + (g(s_4)^{(n)} - y_2^{(n)})\frac{\delta}{\delta \theta_1}(g(s_4)^{(n)} - y_2^{(n)}) + \frac{\lambda \theta_1}{2}] \\
&= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_2^{(n)})\frac{\delta}{\delta \theta_1}(s_3^{(n)}) + (g(s_4)^{(n)} - y_2^{(n)})\frac{\delta}{\delta \theta_1}(s_4^{(n)}) + \frac{\lambda \theta_1}{2}] \\
&= \frac{1}{N}\sum_{n=1}^{N}[(g(s_3)^{(n)} - y_2^{(n)})\theta_5\frac{\delta}{\delta \theta_3}(s_1^{(n)}) + (g(s_4)^{(n)} - y_2^{(n)})\theta_6\frac{\delta}{\delta \theta_3}(s_1^{(n)}) + \frac{\lambda \theta_1}{2}] \\
&= \frac{1}{N}\sum_{n=1}^{N}[x_1^{(n)}[(g(s_3)^{(n)} - y_2^{(n)})\theta_5 + (g(s_4)^{(n)} - y_2^{(n)})\theta_6] + \frac{\lambda \theta_1}{2}] \\
\therefore \theta_1 &\leftarrow \theta_1 - \alpha[\frac{1}{N}\sum_{n=1}^{N}[x_1^{(n)}[(g(s_3)^{(n)} - y_2^{(n)})\theta_5 + (g(s_4)^{(n)} - y_2^{(n)})\theta_6] + \frac{\lambda \theta_1}{2}]]
\end{aligned} \tag{20}$$

3 Question 3

3.1 Question 3.1

$$\begin{aligned}
H(S) &= -p_{\oplus}\log_2(p_{\oplus}) - p_{\ominus}\log_2(p_{\ominus}) \\
&= -\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5}) \\
&= 0.971
\end{aligned} \tag{21}$$

$$IG(S,A) = H(S) - \sum_{a \in \text{Values}(A)} \frac{|S_a|}{|S|}H(S_a) \tag{22}$$

$$\begin{aligned}
IG(S,x_1) &= H(S) - \sum_{a \in \{0,1\}} \frac{|S_a|}{|S|}H(S_a) \\
H(S_0) &= -\frac{2}{4}\log_2(\frac{2}{4}) - \frac{2}{4}\log_2(\frac{2}{4}) = 1 \\
H(S_1) &= -1\log_2(1) = 0 \\
IG(S,x_1) &= 0.971 - \frac{4}{5}1 - \frac{1}{5}0 = 0.171
\end{aligned} \tag{23}$$

$$
\begin{aligned}
IG(S,x_2) &= H(S) - \sum_{a\in\{0,1\}} \frac{|S_a|}{|S|} H(S_a) \\
H(S_0) &= -\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3}) = 0.918 \\
H(S_1) &= -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1 \\
IG(S,x_2) &= 0.971 - \frac{3}{5}0.918 - \frac{2}{5}1 = 0.020
\end{aligned} \tag{24}
$$

$$
\begin{aligned}
IG(S,x_3) &= H(S) - \sum_{a\in\{0,1\}} \frac{|S_a|}{|S|} H(S_a) \\
H(S_0) &= -\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3}) = 0.918 \\
H(S_1) &= -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1 \\
IG(S,x_3) &= 0.971 - \frac{3}{5}0.918 - \frac{2}{5}1 = 0.020
\end{aligned} \tag{25}
$$

Hence attribute x_1 has the highest information gain.

3.2 Question 3.2

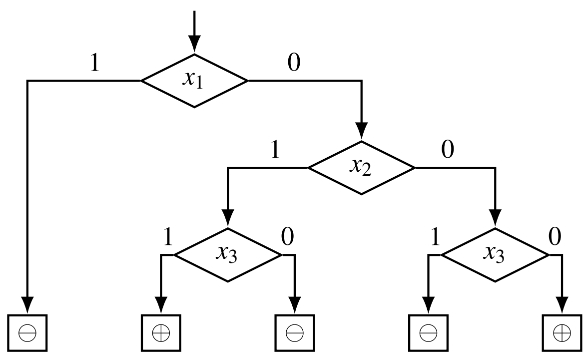
Attribute x_1 has the highest information gain and hence will be the root node for the tree. The one branch will only have 1 value whereas the other branch will have the remaining 4 values where x_1 is 0. The entropy of this sub-set is required now to continue working out information gain.

$$
\begin{aligned}
H(S) &= -p_{\oplus}\log_2(p_{\oplus}) - p_{\ominus}\log_2(p_{\ominus}) \\
&= -\frac{2}{4}\log_2(\frac{2}{4}) - \frac{2}{4}\log_2(\frac{2}{4}) \\
&= 1
\end{aligned} \tag{26}
$$

$$
\begin{aligned}
IG(S,x_2) &= H(S) - \sum_{a\in\{0,1\}} \frac{|S_a|}{|S|} H(S_a) \\
H(S_0) &= -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1 \\
H(S_1) &= -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1 \\
IG(S,x_2) &= 1 - \frac{2}{4}1 - \frac{2}{4}1 = 0
\end{aligned} \tag{27}
$$

$$
\begin{aligned}
IG(S,x_3) &= H(S) - \sum_{a\in\{0,1\}} \frac{|S_a|}{|S|} H(S_a) \\
H(S_0) &= -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1 \\
H(S_1) &= -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1 \\
IG(S,x_3) &= 1 - \frac{2}{4}1 - \frac{2}{4}1 = 0
\end{aligned} \tag{28}
$$

The information gain for both x_2 and x_3 is 0 so the feature which is chosen to split on is arbitrary and hence x_2 is chosen at random. Then x_3 is split on both branches of x_2 to produce all predictions.



4 Question 4

4.1 Question 4.1

$$
\mu = \begin{bmatrix} 2.43 & 2.25 \end{bmatrix} \tag{29}
$$

$$
\Sigma = \begin{bmatrix} 0.79 & 0.68 \\ 0.68 & 0.68 \end{bmatrix} \tag{30}
$$

$$
P = \begin{bmatrix} -0.73 & 0.68 \\ -0.68 & -0.73 \end{bmatrix} \tag{31}
$$

$$
\begin{aligned}
w^{(1)} &= (\begin{bmatrix} 2.2 & 2.2 \end{bmatrix} - \begin{bmatrix} 2.43 & 2.25 \end{bmatrix}) \cdot \begin{bmatrix} -0.73 & 0.68 \\ -0.68 & -0.73 \end{bmatrix} \\
&= \begin{bmatrix} -0.23 & -0.05 \end{bmatrix} \cdot \begin{bmatrix} -0.73 & 0.68 \\ -0.68 & -0.73 \end{bmatrix} \\
&= \begin{bmatrix} 0.2 & -0.12 \end{bmatrix} \therefore 0.2 \geq 0 \implies w^{(1)} \in \text{class 0 (normal)}
\end{aligned} \tag{32}
$$

$$\begin{aligned}w^{(2)} &= (\begin{bmatrix}4.23 & 3.64\end{bmatrix} - \begin{bmatrix}2.43 & 2.25\end{bmatrix}) \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}1.8 & 1.39\end{bmatrix} \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}-2.26 & 0.21\end{bmatrix} \therefore -2.26 \leq 0 \implies w^{(2)} \in \text{class 1 (anomaly)}\end{aligned} \tag{33}$$

$$\begin{aligned}w^{(3)} &= (\begin{bmatrix}1.91 & 2.07\end{bmatrix} - \begin{bmatrix}2.43 & 2.25\end{bmatrix}) \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}-0.52 & -0.18\end{bmatrix} \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}0.5 & -0.22\end{bmatrix} \therefore 0.5 \geq 0 \implies w^{(3)} \in \text{class 0 (normal)}\end{aligned} \tag{34}$$

$$\begin{aligned}w^{(4)} &= (\begin{bmatrix}2.19 & 2.12\end{bmatrix} - \begin{bmatrix}2.43 & 2.25\end{bmatrix}) \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}-0.24 & -0.13\end{bmatrix} \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}0.26 & -0.07\end{bmatrix} \therefore 0.26 \geq 0 \implies w^{(4)} \in \text{class 0 (normal)}\end{aligned} \tag{35}$$

$$\begin{aligned}w^{(5)} &= (\begin{bmatrix}1.92 & 1.82\end{bmatrix} - \begin{bmatrix}2.43 & 2.25\end{bmatrix}) \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}-0.51 & -0.43\end{bmatrix} \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}0.66 & -0.03\end{bmatrix} \therefore 0.66 \geq 0 \implies w^{(5)} \in \text{class 0 (normal)}\end{aligned} \tag{36}$$

$$\begin{aligned}w^{(6)} &= (\begin{bmatrix}1.67 & 1.49\end{bmatrix} - \begin{bmatrix}2.43 & 2.25\end{bmatrix}) \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}-0.76 & -0.76\end{bmatrix} \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}1.07 & 0.04\end{bmatrix} \therefore 1.07 \geq 0 \implies w^{(6)} \in \text{class 0 (normal)}\end{aligned} \tag{37}$$

$$\begin{aligned}w^{(7)} &= (\begin{bmatrix}3.84 & 3.89\end{bmatrix} - \begin{bmatrix}2.43 & 2.25\end{bmatrix}) \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}1.41 & 1.64\end{bmatrix} \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}-2.14 & -0.24\end{bmatrix} \therefore -2.14 \leq 0 \implies w^{(7)} \in \text{class 1 (anomaly)}\end{aligned} \tag{38}$$

$$\begin{aligned} w^{(8)} &= (\begin{bmatrix}2.34 & 1.66\end{bmatrix} - \begin{bmatrix}2.43 & 2.25\end{bmatrix}) \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}-0.09 & -0.59\end{bmatrix} \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}0.47 & 0.37\end{bmatrix} \therefore 0.47 \geq 0 \implies w^{(8)} \in \text{class 0 (normal)} \end{aligned} \tag{39}$$

$$\begin{aligned} w^{(9)} &= (\begin{bmatrix}2.34 & 1.78\end{bmatrix} - \begin{bmatrix}2.43 & 2.25\end{bmatrix}) \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}-0.09 & -0.47\end{bmatrix} \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}0.39 & 0.28\end{bmatrix} \therefore 0.39 \geq 0 \implies w^{(9)} \in \text{class 0 (normal)} \end{aligned} \tag{40}$$

$$\begin{aligned} w^{(10)} &= (\begin{bmatrix}2.34 & 1.66\end{bmatrix} - \begin{bmatrix}2.43 & 2.25\end{bmatrix}) \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}-0.8 & -0.39\end{bmatrix} \cdot \begin{bmatrix}-0.73 & 0.68 \\ -0.68 & -0.73\end{bmatrix} \\ &= \begin{bmatrix}0.85 & -0.26\end{bmatrix} \therefore 0.85178865 \geq 0 \implies w^{(10)} \in \text{class 0 (normal)} \end{aligned} \tag{41}$$

4.2 Question 4.2

$$\mu = \frac{1}{8}\sum_{n=1}^{8} x^n = \begin{bmatrix}2.03 & 1.88\end{bmatrix} \tag{42}$$

$$\Sigma = \begin{bmatrix}0.08 & 0.02 \\ 0.02 & 0.06\end{bmatrix} \tag{43}$$

$$\begin{aligned} p(x^{(t)} &= \frac{1}{2\pi^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x^{(t)} - \mu)\Sigma^{-1}(x^{(t)} - \mu)^T\right) \\ |\Sigma|^{1/2} &= 0.07 \\ \frac{1}{2\pi^{2/2}(0.07)} &= 2.27 \\ \Sigma^{-1} &= \begin{bmatrix}13.35 & -4.09 \\ -4.09 & 18.31\end{bmatrix} \end{aligned} \tag{44}$$

$$\begin{aligned} p(x^{(1)}) &= 2.27\exp\left(-\frac{1}{2}(\begin{bmatrix}0.18 & 0.33\end{bmatrix})\Sigma^{-1}(\begin{bmatrix}0.18 & 0.33\end{bmatrix})^T\right) \\ &= 2.27\exp(-0.97) = 0.86 > 0.5 \implies x^{(1)} \in \text{class 0 (normal)} \end{aligned} \tag{45}$$

$$p(x^{(2)}) = 2.27\exp\left(-\frac{1}{2}(\begin{bmatrix}2.21 & 1.77\end{bmatrix})\Sigma^{-1}(\begin{bmatrix}2.21 & 1.77\end{bmatrix})^T\right)$$
$$= 2.27\exp(-45.28) = 0.0 < 0.5 \implies x^{(2)} \in \text{class 1 (anomaly)} \tag{46}$$

$$p(x^{(3)}) = 2.27\exp\left(-\frac{1}{2}(\begin{bmatrix}-0.12 & 0.2\end{bmatrix})\Sigma^{-1}(\begin{bmatrix}-0.12 & 0.2\end{bmatrix})^T\right)$$
$$= 2.27\exp(-0.56) = 1.3 > 0.5 \implies x^{(3)} \in \text{class 0 (normal)} \tag{47}$$

$$p(x^{(4)}) = 2.27\exp\left(-\frac{1}{2}(\begin{bmatrix}0.17 & 0.25\end{bmatrix})\Sigma^{-1}(\begin{bmatrix}0.17 & 0.25\end{bmatrix})^T\right)$$
$$= 2.27\exp(-0.59) = 1.26 > 0.5 \implies x^{(4)} \in \text{class 0 (normal)} \tag{48}$$

$$p(x^{(5)}) = 2.27\exp\left(-\frac{1}{2}(\begin{bmatrix}-0.1 & -0.05\end{bmatrix})\Sigma^{-1}(\begin{bmatrix}-0.1 & -0.05\end{bmatrix})^T\right)$$
$$= 2.27\exp(-0.07) = 2.12 > 0.5 \implies x^{(5)} \in \text{class 0 (normal)} \tag{49}$$

$$p(x^{(6)}) = 2.27\exp\left(-\frac{1}{2}(\begin{bmatrix}-0.36 & -0.38\end{bmatrix})\Sigma^{-1}(\begin{bmatrix}-0.36 & -0.38\end{bmatrix})^T\right)$$
$$= 2.27\exp(-1.63) = 0.45 < 0.5 \implies x^{(6)} \in \text{class 1 (anomaly)} \tag{50}$$

$$p(x^{(7)}) = 2.27\exp\left(-\frac{1}{2}(\begin{bmatrix}1.82 & 2.02\end{bmatrix})\Sigma^{-1}(\begin{bmatrix}1.82 & 2.02\end{bmatrix})^T\right)$$
$$= 2.27\exp(-44.43) = 0.0 < 0.5 \implies x^{(7)} \in \text{class 1 (anomaly)} \tag{51}$$

$$p(x^{(8)}) = 2.27\exp\left(-\frac{1}{2}(\begin{bmatrix}0.31 & -0.21\end{bmatrix})\Sigma^{-1}(\begin{bmatrix}0.31 & -0.21\end{bmatrix})^T\right)$$
$$= 2.27\exp(-1.31) = 0.61 > 0.5 \implies x^{(8)} \in \text{class 0 (normal)} \tag{52}$$

$$p(x^{(9)}) = 2.27\exp\left(-\frac{1}{2}(\begin{bmatrix}0.31 & -0.09\end{bmatrix})\Sigma^{-1}(\begin{bmatrix}0.31 & -0.09\end{bmatrix})^T\right)$$
$$= 2.27\exp(-0.83) = 0.99 > 0.5 \implies x^{(9)} \in \text{class 0 (normal)} \tag{53}$$

$$p(x^{(10)}) = 2.27\exp\left(-\frac{1}{2}(\begin{bmatrix}-0.4 & -0.01\end{bmatrix})\Sigma^{-1}(\begin{bmatrix}-0.4 & -0.01\end{bmatrix})^T\right)$$
$$= 2.27\exp(-1.05) = 0.79 > 0.5 \implies x^{(10)} \in \text{class 0 (normal)} \tag{54}$$

5 Question 5

5.1 Question 5.1

$$
\begin{gathered}
\hat{Q}(s,a) \leftarrow r + \gamma \max_{a'} Q(s,a') \\
\hat{Q}(s_{10}, a_{right}) = 200 \\
\hat{Q}(s_7, a_{up}) = 100 \\
\hat{Q}(s_9, a_{right}) = 0 + (0.2)\max\{200, 0\} = 40 \\
\hat{Q}(s_4, a_{up}) = 0 + (0.2)\max\{100, 0\} = 20 \\
\hat{Q}(s_{12}, a_{down}) = 0 + (0.2)\max\{40, 0, 0, 0\} = 8 \\
\hat{Q}(s_8, a_{right}) = 0 + (0.2)\max\{40, 0, 0, 0\} = 8 \\
\hat{Q}(s_6, a_{up}) = 0 + (0.2)\max\{40, 0, 0, 0\} = 8 \\
\hat{Q}(s_5, a_{left}) = 0 + (0.2)\max\{20, 0, 0\} = 4 \\
\hat{Q}(s_3, a_{right}) = 0 + (0.2)\max\{20, 0, 0\} = 4 \\
\hat{Q}(s_2, a_{right}) = 0 + (0.2)\max\{4, 0\} = 0.8 \\
\hat{Q}(s_2, a_{up}) = 0 + (0.2)\max\{8, 0\} = 1.6 \\
\hat{Q}(s_1, a_{right}) = 0 + (0.2)\max\{1.6, 0.8, 0\} = 0.32 \\
\hat{Q}(s_9, a_{up}) = 0 + (0.2)\max\{8\} = 1.6 \\
\hat{Q}(s_9, a_{left}) = 0 + (0.2)\max\{8\} = 1.6 \\
\hat{Q}(s_2, a_{left}) = 0 + (0.2)\max\{0.32\} = 0.064 \\
\hat{Q}(s_4, a_{right}) = 0 + (0.2)\max\{4\} = 0.8 \\
\hat{Q}(s_{10}, a_{left}) = 0 + (0.2)\max\{40, 1.6, 1.6, 0\} = 8 \\
\hat{Q}(s_7, a_{down}) = 0 + (0.2)\max\{20, 0.8, 0\} = 4 \\
\hat{Q}(s_9, a_{down}) = 0 + (0.2)\max\{8, 0\} = 1.6 \\
\hat{Q}(s_6, a_{down}) = 0 + (0.2)\max\{0.064, 1.6, 0.8\} = 0.32 \\
\hat{Q}(s_7, a_{down}) = 0 + (0.2)\max\{20, 0.8\} = 4 \\
\hat{Q}(s_3, a_{left}) = 0 + (0.2)\max\{1.6, 0.8, 0.32\} = 0.32 \\
\hat{Q}(s_4, a_{left}) = 0 + (0.2)\max\{4, 0.32\} = 0.8
\end{gathered}
\tag{55}
$$

	Actions			
State	$\rightarrow$	$\leftarrow$	$\uparrow$	$\downarrow$
s_1	0.32	-	-	-
s_2	0.8	0.064	1.6	-
s_3	4	0.32	-	-
s_4	0.8	0.8	20	-
s_5	-	4	-	-
s_6	-	-	8	0.32
s_7	-	-	100	4
s_8	8	-	-	-
s_9	40	1.6	1.6	1.6
s_{10}	200	8	-	-
s_{11}	-	-	-	-
s_{12}	-	-	-	8

5.2 Question 5.2

$$\begin{gathered}
V^*(s) = \max_{a'} Q(s,a') \\
V^*(s_1) = 0.32 \\
V^*(s_2) = 1.6 \\
V^*(s_3) = 4 \\
V^*(s_4) = 20 \\
V^*(s_5) = 4 \\
V^*(s_6) = 8 \\
V^*(s_7) = 100 \\
V^*(s_8) = 8 \\
V^*(s_9) = 40 \\
V^*(s_{10}) = 200 \\
V^*(s_{11}) = 0 \\
V^*(s_{12}) = 8
\end{gathered} \tag{56}$$

5.3 Question 5.3

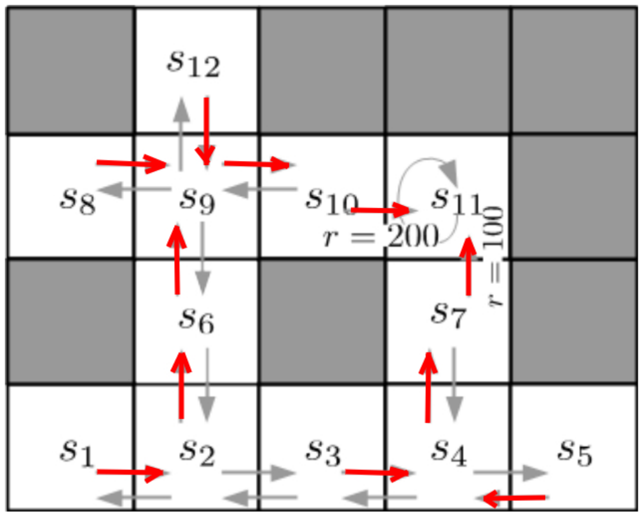


Figure 1: Optimal policy

5.4 Question 5.4

If γ is changed so that $\gamma = 0.8$ then the only optimal policy that changes is $\pi^*(s_3)$ which changes to a_{left} as $V^*(s_3) = 81.92$ where $Q(s_3, a_{left}) = 81.92$ and $Q(s_3, a_{right}) = 64$.

5.5 Question 5.5

$$
\begin{gathered}
V^*(s) = \max_{a'} Q(s,a') \text{ and } Q(s,a) = E[r(s,a)] + \gamma \sum_{s'} P(s'|s,a) V^*(s') \\
V^*(s_{10}) = (0.5)(100) + (0.2)(0.5)(V^*(s_{10})) \therefore V^*(s_{10}) = 111.11 \\
V^*(s_7) = (0.5)(100) + (0.2)(0.5)(V^*(s_7)) \therefore V^*(s_7) = 55.56 \\
V^*(s_1) = 0.18 \\
V^*(s_2) = 0.89 \\
V^*(s_3) = 2.22 \\
V^*(s_4) = 11.11 \\
V^*(s_5) = 2.22 \\
V^*(s_6) = 4.44 \\
V^*(s_8) = 4.44 \\
V^*(s_9) = 22.22 \\
V^*(s_{11}) = 0 \\
V^*(s_{12}) = 4.44
\end{gathered} \tag{57}
$$

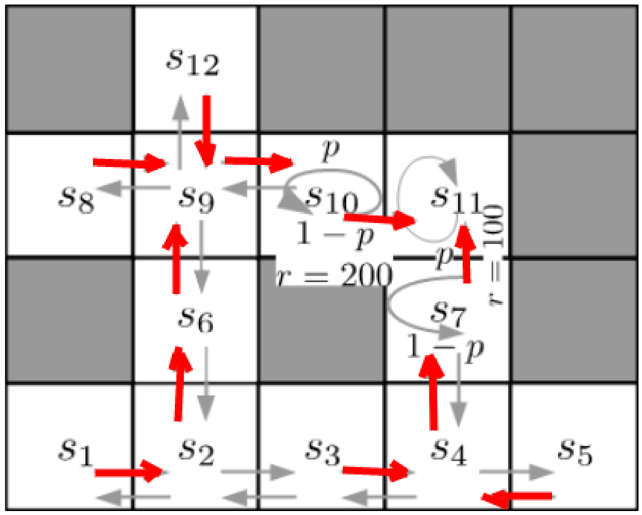


Figure 2: Optimal Policy

6 Question 6

6.1 Question 6.1

An example of a machine learning problem is to predict the winner of a match of the game *League of Legends*. The task is to predict the winner of the match using the experience of many games. The performance is measured by the overall accuracy of all predictions.

6.2 Question 6.2

The test data-set is required to check the accuracy of the model that was trained on the training data-set. The test data-set is a set of data that the model has never seen and is used to determine the accuracy of the model on data that it has never seen. This is essential as if the model can only be accurate on data that it was trained on then the model is not general enough to be used on new data and hence is useless to predict new data entries which is the point of supervised learning.

6.3 Question 6.3

Supervised learning and reinforcement learning are similar as they both use statistical measures to generate a model and they both require data or samples to generate the model.

6.4 Question 6.4

An over-fitting model can be determined by acquiring the accuracy of the model on a set of data that has never been seen by the model and if the accuracy is low compared to the training data accuracy then the model is most likely over-fitting.

6.5 Question 6.5

I don’t agree with the statement as the centroids will converge at the location which is central to it’s group of data points.