Q1-1.

a) $7/8 = 0.875$
$0.875 \times 2 = 1.750$ **1**
$0.750 \times 2 = 1.500$ **1**
$0.500 \times 2 = 1.000$ **1**
$7/8$ is (0.111) in binary notation.

b)

$0.500 \times 2 = 1.0$ **1**
$0.5 = 0.1$ in binary
$10 = 8 + 2$
$10 = 1010$ in binary
$10.5 = 1010.1$ in binary

c)

$0.8 \times 2 = 1.6$ **1**
$0.6 \times 2 = 1.2$ **1**
$0.2 \times 2 = 0.4$ **0**
$0.4 \times 2 = 0.8$ **0**
$0.8 \times 2 = 1.6$ **1**
and then it repeats as it goes.
$0.8 = 0.\overline{1100}$
$12 = 8 + 4$
$12 = 1100$
$12.8 = 1100.\overline{1100}$

d)

$2/3 = 0.\overline{6}$
$1/3 = 1/4 + 1/16 + 1/64$ ..
$1/3 = 0.0101\overline{01}$
$2/3 = 1/3 + 1/3 = 0.1010\overline{10}$
$2/3 = 0.\overline{10}$

e)

$0.2 \times 2 = 0.4$ **0**
$0.4 \times 2 = 0.8$ **0**
$0.8 \times 2 = 1.6$ **1**
$0.6 \times 2 = 1.2$ **1**
$0.2 \times 2 = 0.4$ **0**
$0.2 = 0.\overline{0011}$
$3 = (11)$
$3.2 = 11.\overline{0011}$

Q1-2.

a)

1/3 in binary = $0.\overline{01}$ 1.0101... x 2^{-2}
sign bit is 0
exponent is 01111111101 = $2^{1023-1025} = 2^{-2}$
fraction is 01010101....01 (52 bits)
We truncate the repetitive part since 53rd bit is 0
$0.\overline{01}$ x 2^{-2} x 2^{-52} = -(1/3 x 2^{-54}) = error between fl(1/3) - 1/3

b)

3 = (11)
0.3 x 2 = 0.6 **0**
0.6 x 2 = 1.2 **1**
0.2 x 2 = 0.4 **0**
0.4 x 2 = 0.8 **0**
0.8 x 2 = 1.6 **1**
0.6 x 2 = 1.2 **1**

$0.3 = 0.0\overline{1001}$
$3.3 = 11.0\overline{1001}$
$3.3 = 1.10\overline{1001}$ x 2^{-1}
sign = 0
exponent = 01111111110 = $2^{1023-1024} = 2^{-1}$
fraction = 1010011001....100110 (52 bits)
We truncate the repetitive part since 53rd bit is 0
$0.\overline{0110}$ x 2^{-1} x 2^{-52}
x = $0.\overline{0110}$
16x = $110.\overline{0110}$
15x = 110
x = 6/15 = 0.375
$0.\overline{0110}$ x 2^{-1} x 2^{-52} = -(0.375 * 2^{-53}) = fl(3.3)-3.3

c)

9/7 = 1 + 2/7
2/7 = 1/7 + 1/7
1/7 = 1/8 + 1/64 + 1/512 ...
$1/7 = 0.001001\overline{001}$
$2/7 = 0.0100100\overline{100}$
$9/7 = 1.0\overline{100}$
sign = 0
exp = 01111111111
We round up since 53rd bit is 1, and get rid of the repetitive part.

fraction = 0100100...101 (52 bits)
$0.\overline{100}$ x 2^{-52} is the truncated part.
x = $0.\overline{100}$
8x = $100.\overline{100}$
7x = 100 , x = 4/7
Error = 1x2^{-52} - 4/7 x 2^{-52} = 3/7 x 2^{-52} = fl(9/7) - 9/7

Q1-3.

4.3 = $100.0\overline{1001}$
$100.0\overline{1001}$ = $1.000\overline{1001}$ x 2^{-2}
sign = 0
exponent = 01111111101 = $2^{1023-1025}$ = 2^{-2}
fraction = 00010011001....10011 (52 bits)
We round down since 53rd bit is 0
0011 x 2^{-52} x 2^{-2} is truncated.
x = $0.\overline{0011}$
16x = $11.\overline{0011}$ 15x = 11 , x = 1/5 = 0.2
fl(4.3) = 4.3 - (0.2 x 2^{-54})

3.3 = $11.0\overline{1001}$ = $1.10\overline{1001}$ x 2 x 2^{-1}
sign = 0
exponent = 01111111110 = $2^{1023-1024}$ = 2^{-1}
fraction = 1010011001...100110 (52 bits)
We round down since 53rd bit is 0
.0110 x 2^{-52} x 2^{-1}
x = $0.\overline{0110}$
16x = $110.\overline{0110}$
15x = 110 , x = 6/15 = 0.4
0.4 x 2^{-52} x 2^{-1} = 0.4 x 2^{-53}
fl(3.3) = 3.3 - (0.4 x 2^{-53})

fl(4.3) - fl(3.3) = (4.3 - (0.1 x 2^{-53})) - (3.3 - (0.4 x 2^{-53})) =
fl(4.3) - fl(3.3) = 1 + 0.3 x 2^{-53}

Q2-1.

x^3-9 = 0

k	ak	f(ak)	ck	f(ck)	bk	f(bk)
0	2.000000	-	2.250000	+	2.500000	+
1	2.000000	-	2.125000	+	2.250000	+
2	2.062500	-	2.062500	-	2.125000	+
3	2.062500	-	2.093750	+	2.125000	+
4	2.078125	-	2.078125	-	2.093750	+
5	2.078125	-	2.085937	+	2.093750	+
6	2.078125	-	2.082031	+	2.085937	+
7	2.080078	-	2.080078	-	2.082031	+
8	2.080078	-	2.081054	+	2.082031	+
9	2.080078	-	2.080566	+	2.081054	+
10	2.080078	-	2.080322	+	2.080566	+
11	2.080078	-	2.080200	+	2.080322	+
12	2.080078	-	2.080139	+	2.080200	+
13	2.080078	-	2.080108	+	2.080139	+
14	2.080078	-	2.080093	+	2.080108	+
15	2.080078	-	2.080085	+	2.080093	+
16	2.080078	-	2.080081	-	2.080085	+
17	2.080081	-	2.080083	-	2.080085	+
18	2.000083	-	2.080084	+	2.080085	+
19	2.000083	-	2.080083	-	2.080084	+

The root to six correct decimal places is 2.080083

Q2-2.

cos(x)-sin(x) = 0

k	ak	f(ak)	ck	f(ck)	bk	f(bk)
0	0.000000	+	0.500000	+	1.000000	-
1	0.500000	+	0.750000	+	1.000000	-
2	0.750000	+	0.875000	-	1.000000	-
3	0.750000	+	0.812500	-	0.875000	-
4	0.750000	+	0.781250	+	0.812500	-
5	0.781250	+	0.796875	-	0.812500	-
6	0.781250	+	0.789062	-	0.796875	-
7	0.781250	+	0.785156	-	0.789062	-
8	0.785156	+	0.787109	-	0.789062	-
9	0.785156	+	0.786132	+	0.787109	-
10	0.785156	+	0.785644	+	0.786132	-
11	0.785156	+	0.785400	+	0.785644	-
12	0.785156	+	0.785278	+	0.785400	-
13	0.785278	+	0.785339	+	0.785400	-
14	0.785339	+	0.785369	+	0.785400	-
15	0.785369	+	0.785384	+	0.785400	-
16	0.785384	+	0.785392	+	0.785400	-
17	0.785392	+	0.785396	+	0.785400	-
18	0.785396	+	0.785398	+	0.785400	-
19	0.785398	+	0.785399	-	0.785400	-
20	0.785398	+	0.785398	+	0.785399	-

The root to six correct decimal places is 0.785398.

Q2-3.

g(x) = x^2 + x/2 -1/2
Fixed points are 1 and -0.5
Function is locally convergent to -0.5 as shown below
g(-0.25) = -0.5625
g(-0.5625) = -0.4648
g(-0.4648) = -0.5163
g(-0.5163) = -0.4915
g(-0.4915) = -0.5041
g(-0.5041) = -0.4979
We get closer to -0.5 every iteration , which implies locally convergence.

Function is not locally convergent to 1 as shown below
g(1.2) = 1.54
g(1.54) = 2.6416
g(2.6416) = 7.7988
g(7.7988) = 64.2206

Reason for this divergence is that the absolute value of the g(x) functions derivative is bigger than 1.

Q2-5.

Forward error = $| r - x_a |$ => $|0.75 - 0.74| = 0.01$
$Backwarderror = |f(x_a)|$=>|((4 x 0.74)-3)^2|= 0.0016

Q2-6.

$\frac{df(x)}{dx} = 3x^2 + 2x$

$x_1 = x_0$ - $((x_0)^3 + (x_0)^2 - 1)/(3x_0^2 + 2x_0)$
= 1 - (1+1+1)/(3+2) = 4/5 = x_1

$x_2 = x_1$ - $((x_1)^3 + (x_1)^2 - 1)/(3x_1^2 + 2x_1)$
= 4/5 - (64/125 - 16/25 - 1)/(3 x 16/25 + 2 x 4/5) = 0.8 - (1.128/3.52) = 1.12