Name:

USC ID:

Notes:

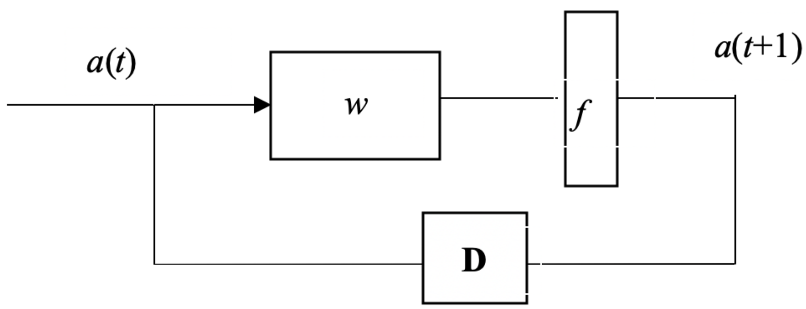
- Write your name and ID number in the spaces above.
- No books, cell phones or other notes are permitted. Only two letter size cheat sheet (back and front) and a calculator are allowed.
- Problems are not sorted in terms of difficulty. Please avoid guess work and long and irrelevant answers.
- Show all your work and your final answer. Simplify your answer as much as you can.
- Open your exam only when you are instructed to do so.
- Make sure you submit ALL pages of your answers. Answers submitted after the exam is adjourned WILL NOT BE ACCEPTED.

Problem	Score	Earned
1	25	
2	20	
3	20	
4	25	
5	25	
Total	115	

1. Consider the RNN shown below with one layer and one neuron in the layer. The activation function of the neural network is $f(n) = \frac{1}{1+e^{-n}}$. The initial weight is $w = -1$. There are no bias terms. Assume that the initial condition of the network is $a(0) = 1$. We would like the output of the network after two time steps to be $a(2) = y = 1/2$.

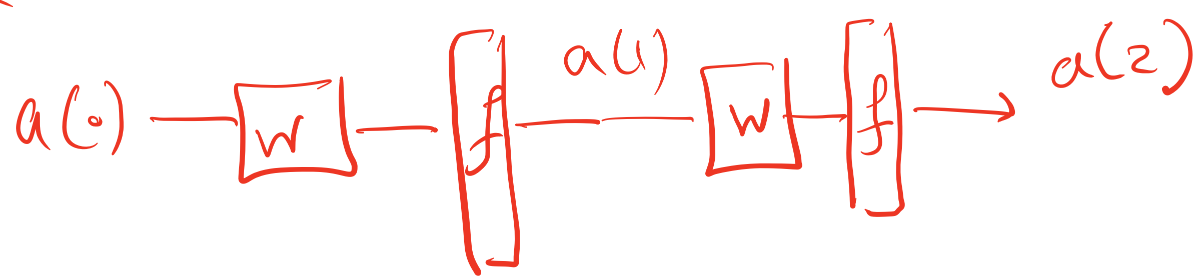
   (a) Unfold the RNN through time and show its block diagram as a feedforward neural network through time, whose input is $a(0)$ and whose output is $a(2)$.

   (b) Perform one step of the Stochastic Gradient Descent algorithm by applying the backpropagation algorithm to the unfolded network, assuming the learning rate $\alpha = 0.1$. Use the objective function $J = (y - a(2))^2$. This means that you should calculate the updated weight.



(a)

5pts



(b) Forward path

5 pts

$n(1) = w\,a(0) = -1 \times 1 = -1$

$a(1) = f(n(1)) = \frac{1}{1+e} = .269$

$$n(2) = w\,a(1) = -1 \times .269 = -.269$$

$$a(2) = f(n(2)) = \frac{1}{1+e^{-.269}} = 0.567$$

Backward path

$$F'^{(2)}(n(2)) = \left.\frac{df}{dn}\right|_{n(2)} = \left.\frac{e^{-n}}{(1+e^{-n})^2}\right|_{n=-.269}$$

$$= \frac{e^{.269}}{(1+e^{-.269})^2} = .420$$

5 pts

$$F'^{(1)}(n(1)) = \left.\frac{e^{-n}}{(1+e^{-n})^2}\right|_{n=-1} = .080$$

$$S(2) = -2F'^{(2)}(n^{(2)})(y-a) = (-2)(.420)(-1-.567)$$

$$= 1.316$$

$$S(1) = F'^{(1)}(n(1))\, w\, S(2) = (.080)(-1)(1.316)$$

$$= -.105$$

5 pts

$$w^{new} = -1 - \alpha\, a(1)\, S(2) = -1 - (.1)(.269)(1.316)$$

$$= -1.035$$

$$w^{new} = -1.035 - \alpha\, a(0)\, S(1) = -1.035 - .1(1)(-.105)$$

$$= -1.025$$

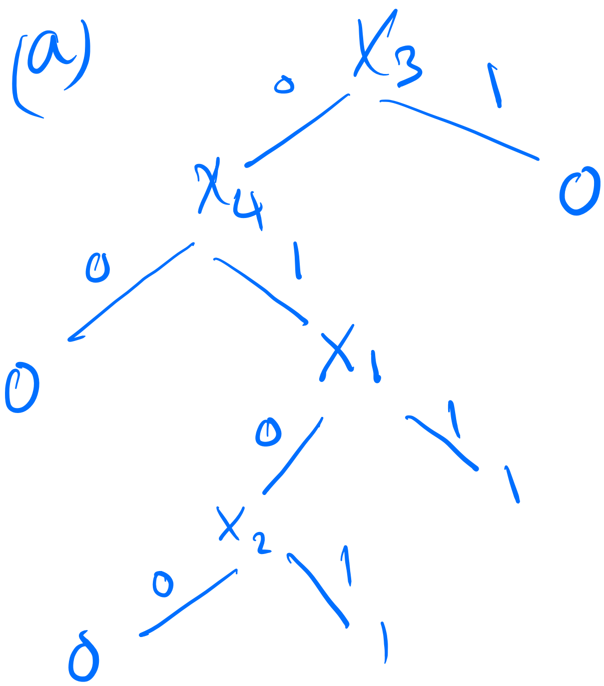
5 pts

2. Draw the decision trees that calculate the following Boolean functions:

   (a) $Y = (X_1 \vee X_2) \wedge (\neg X_3 \wedge X_4)$, $X_i \in \{0, 1\}$. The root of the decision tree must be X_3.

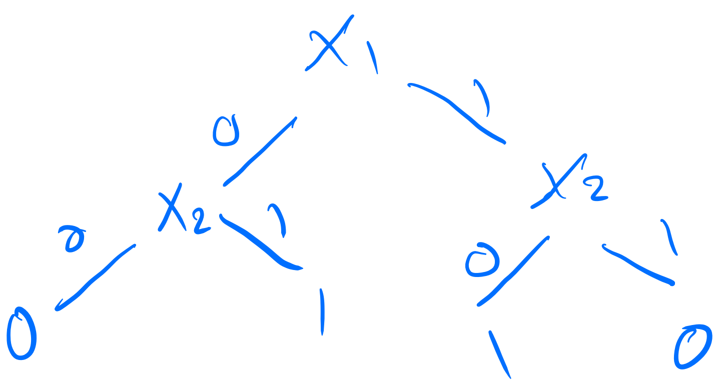
   (b) $Y = X_1 \otimes X_2$, $X_i \in \{0, 1\}$, where $\otimes$ is the exclusive or (XOR) function.

(a)



10 pts

(b)



10 pts

3. We are trying to estimate the temperature of consecutive years based on *observations* on tree ring sizes. Possible ring sizes are Very Small = VS, Small = S, Medium = M, Large = L, and Very Large = VL. Years can be Cold = C or Hot = H. Assume that we observed VS, VL tree ring sizes in two consecutive years. Also, Assume that π = [0.3 0.7] shows the initial distribution of C and H, respectively. Which of the following HMMs is more likely to have given rise to the observation O= {VS,VL} and why? First rows of A_1, B_1, A_2, B_2 represent C and second rows represent H.

(a)

$$A_1 = \begin{array}{c} \begin{array}{cc} C & H \end{array} \\ \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix} \end{array} \quad B_1 = \begin{array}{c} \begin{array}{ccccc} VS & S & M & L & VL \end{array} \\ \begin{bmatrix} 0.1 & 0.2 & 0.2 & 0.2 & 0.3 \\ 0.2 & 0.1 & 0.4 & 0.1 & 0.2 \end{bmatrix} \end{array}$$

(b)

$$A_2 = \begin{array}{c} \begin{array}{cc} C & H \end{array} \\ \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} \end{array} \quad B_2 = \begin{array}{c} \begin{array}{ccccc} VS & S & M & L & VL \end{array} \\ \begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.6 \\ 0.5 & 0.1 & 0.1 & 0.1 & 0.2 \end{bmatrix} \end{array}$$

(a)

State	P(O, State)
CC	$(.3)(.1)(.7)(.3) = 63 \times 10^{-4}$
CH	$(.3)(.1)(.3)(.2) = 18 \times 10^{-4}$
HC	$(.7)(.2)(.3)(.3) = 126 \times 10^{-4}$
HH	$(.7)(.2)(.7)(.2) = 196 \times 10^{-4}$

9pts

$P(O) = 403 \times 10^{-4}$

Solution:

(b)

State	$P(O, \text{State})$
CC	$(.3)(.1)(.5)(.6) = 90 \times 10^{-4}$
CH	$(.3)(.1)(.5)(.2) = 30 \times 10^{-4}$
HC	$(.7)(.5)(.5)(.6) = 1050 \times 10^{-4}$
HH	$(.7)(.5)(.5)(.2) = 350$

9pts

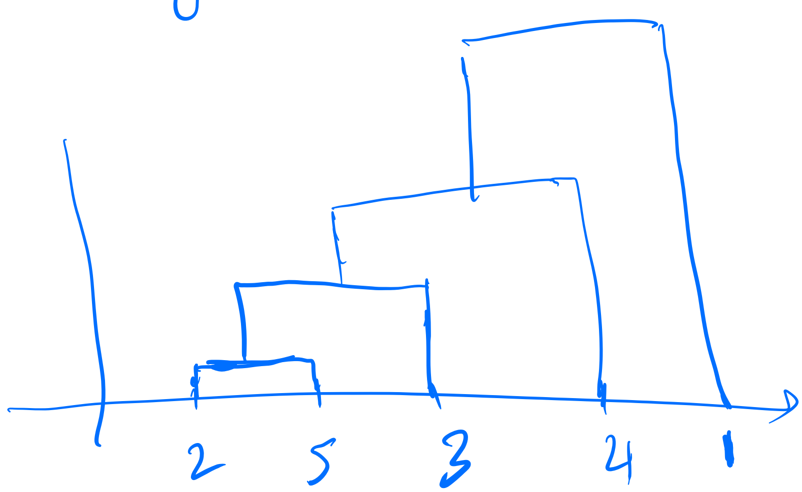
$P(O) = 1620 \times 10^{-4}$

(b) is more likely 2pts

4. Use the similarity matrix in the following Table to perform single and complete linkage hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged. Note that p_i's represent data points.

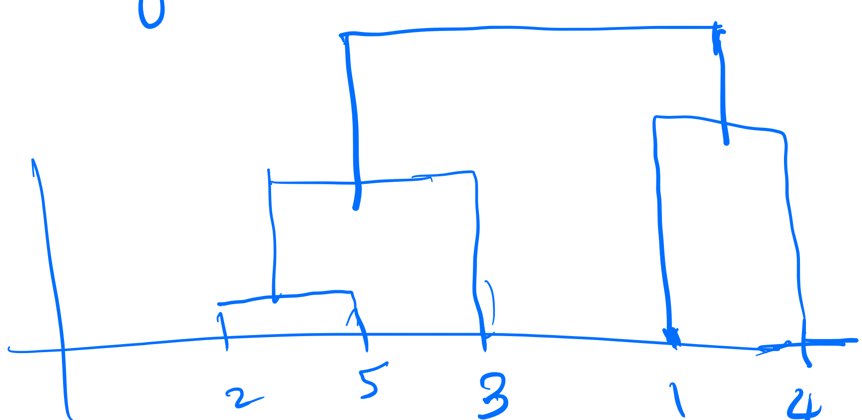
	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

Dendrogram for Single link



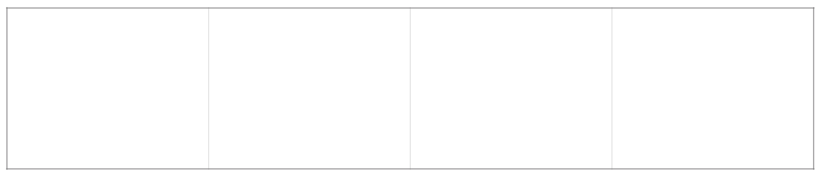
12 pts

Dendrogram for complete link



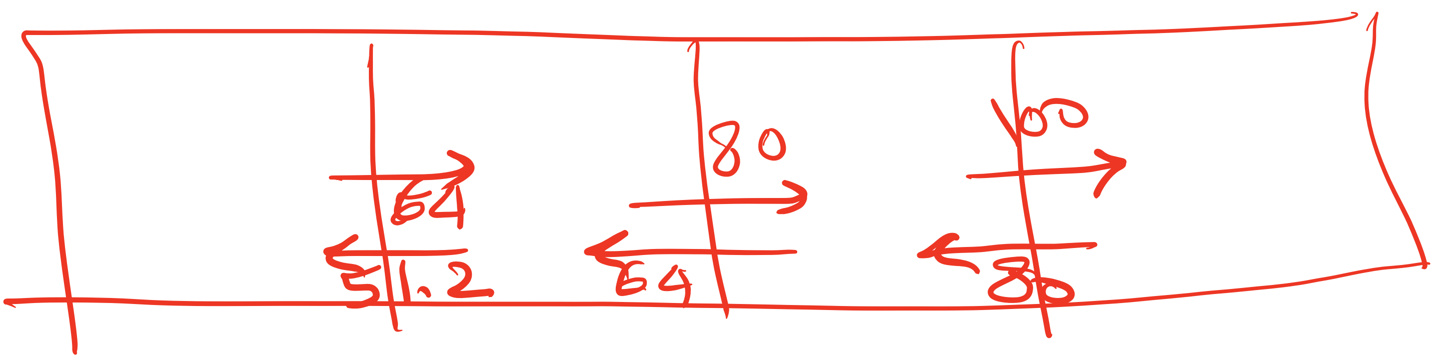
13 pts

5. Assume the deterministic 4-state world below (each cell is a state) where the immediate reward is 0 for entering all states, except the rightmost state, for which the reward is 10, and which is an absorbing state (when in this state, the agent stays in it). The only actions are move right and move left (only one of which is available when in the leftmost border cell). Assuming a discount factor of $\gamma = 0.8$, give the final optimal Q values for each action in each state and describe an optimal policy.

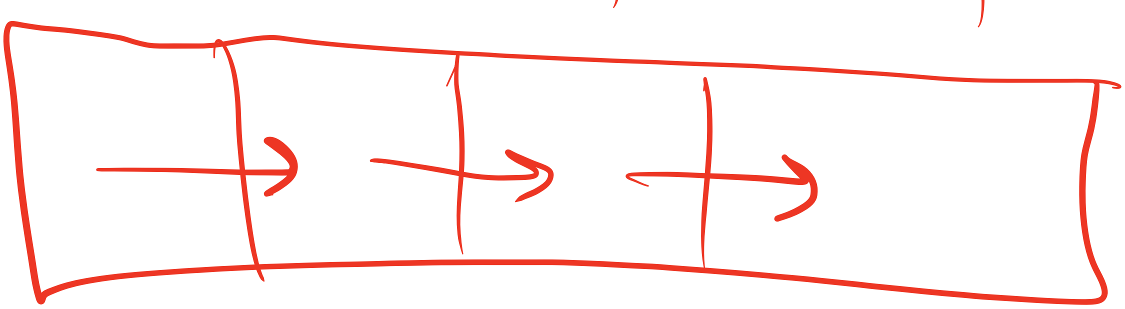


$Q(s,a)$

20pts



optimal policy



5pts

Scratch paper

Name:

USC ID:

Scratch paper

Name:

USC ID: