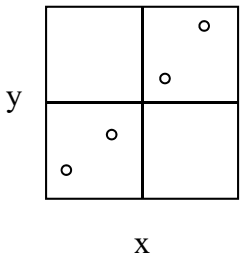
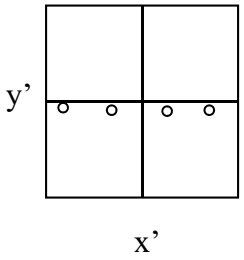
COMP5331 Knowledge Discovery in Databases (Fall Semester 2019)
Homework 2 Solution

Q1.(a) No.

In the original coordinate system, we have the following 4 data points. The threshold is set to 4. In the following, we could not find any clusters in any subspace. There is no grid unit involved in the original coordinate system.



However, in the new coordinate system, we will obtain the following.



We could find one grid unit on the 1-dimensional axis, y', with the same threshold of 4.
Thus, the total number of grid units involved in all clusters based on the new coordinate system is greater than that based on the original coordinate system

(b) When the size of the subspace is larger, the no. of grid units is larger. The probability that a grid unit contains a sufficient number of points is smaller.

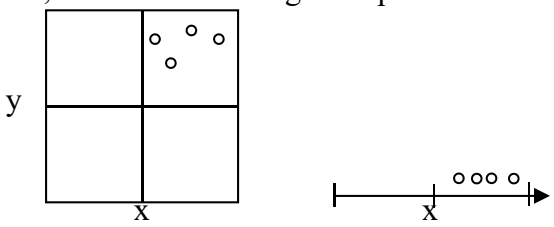
(c) (i) No.

Apriori-like algorithm is valid when the following property holds. If there exists a dense unit *w.r.t* a subspace S, then there exists a dense unit *w.r.t* a subspace S' where $S' \subseteq S$.

Now, we give a counter-example that the above property does not hold under Condition 1. Consider $T_2 = 4$ and $T_1 = 5$ (in count instead of density)

$\because T_2 < T_1 \quad \therefore$ Condition 1 holds.

However, from the following example



we find that there exists a dense unit in subspace{x, y}but not in subspace {x}.

(ii) No.

Let us use the example from Part (c)(i).

In this example, c is equal to 2.

Suppose that $T_2 = 4$. In this case, according to Condition 1, we have $T_1 = 2\ T_2 = 2 \cdot 4 = 8$.

Q2.

(a)

$$\text{mean vector} = \begin{pmatrix} \frac{6+8+5+9}{4} \\ \frac{6+8+9+5}{4} \end{pmatrix} = \begin{pmatrix} 7 \\ 7 \end{pmatrix} \begin{pmatrix} \frac{7+9+6+10}{4} \\ \frac{7+9+10+6}{4} \end{pmatrix} = \begin{pmatrix} 8 \\ 8 \end{pmatrix}$$

For data (7, 7), difference from mean vector $= \begin{pmatrix} 6-7 \\ 6-7 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$

For data (9, 9), difference from mean vector $= \begin{pmatrix} 8-7 \\ 8-7 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

For data (6, 10), difference from mean vector $= \begin{pmatrix} 5-7 \\ 9-7 \end{pmatrix} = \begin{pmatrix} -2 \\ 2 \end{pmatrix}$

For data (10, 6), difference from mean vector $= \begin{pmatrix} 9-7 \\ 5-7 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$

$$Y = \begin{pmatrix} -1 & 1 & -2 & 2 \\ -1 & 1 & 2 & -2 \end{pmatrix}$$

$$\Sigma = \frac{1}{4} YY^T = \frac{1}{4} \begin{pmatrix} -1 & 1 & -2 & 2 \\ -1 & 1 & 2 & -2 \end{pmatrix} \begin{pmatrix} -1 & -1 \\ 1 & 1 \\ -2 & 2 \\ 2 & -2 \end{pmatrix}$$

$$= \frac{1}{4} \begin{pmatrix} 10 & -6 \\ -6 & 10 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{5}{2} & -\frac{3}{2} \\ -\frac{3}{2} & \frac{5}{2} \end{pmatrix}$$

$$\begin{vmatrix} \frac{5}{2}-\lambda & -\frac{3}{2} \\ -\frac{3}{2} & \frac{5}{2}-\lambda \end{vmatrix} = 0 \quad \Longrightarrow \quad (\frac{5}{2}-\lambda)^2 - (-\frac{3}{2})^2 = 0 \quad \Longrightarrow \quad \lambda = 4 \quad or \quad \lambda = 1$$

when $\lambda = 4$,

$$\begin{pmatrix} \frac{5}{2}-4 & -\frac{3}{2} \\ -\frac{3}{2} & \frac{5}{2}-4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} -\frac{3}{2} & -\frac{3}{2} \\ -\frac{3}{2} & -\frac{3}{2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow x_1 + x_2 = 0$$

We choose the eigenvector of unit length: $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix}$.

When $\lambda = 1$,

$$\begin{pmatrix} \frac{5}{2}-1 & -\frac{3}{2} \\ -\frac{3}{2} & \frac{5}{2}-1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \end{pmatrix}\Rightarrow\begin{pmatrix} \frac{3}{2} & -\frac{3}{2} \\ -\frac{3}{2} & \frac{3}{2} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \end{pmatrix}\Rightarrow\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \end{pmatrix}\Rightarrow x_1 - x_2 = 0$$

We choose the eigenvector of unit length: $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}=\begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$.

Thus, $\Phi=\begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}$, $Y=\Phi^T X=\begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}X$.

For data (6, 6), Y = $\begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}\begin{pmatrix} 6 \\ 6 \end{pmatrix}=\begin{pmatrix} 0 \\ 8.49 \end{pmatrix}$

For data (8, 8), Y = $\begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}\begin{pmatrix} 8 \\ 8 \end{pmatrix}=\begin{pmatrix} 0 \\ 11.31 \end{pmatrix}$

For data (5, 9), Y = $\begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}\begin{pmatrix} 5 \\ 9 \end{pmatrix}=\begin{pmatrix} -2.83 \\ 9.90 \end{pmatrix}$

For data (9, 5), Y = $\begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}\begin{pmatrix} 9 \\ 5 \end{pmatrix}=\begin{pmatrix} 2.83 \\ 9.90 \end{pmatrix}$

Thus, (6, 6) is reduced to (8.49);
(8, 8) is reduced to (11.31);
(5, 9) is reduced to (9.90);
(9, 5) is reduced to (9.90).

(b)(i)

$\alpha = 2$
$\beta = -7$

(ii)

Yes. It is possible.
We know that we could express

$x_q = 2 \cdot x_p - 7$ ……(*)
$y_q = 2 \cdot y_p - 7$ ……(**)

According to the steps in Part (a), when we substitute equations (*) and (**) into the derivation in Part (a), it is easy to verify that we could obtain the same matrix Φ.

$$\Phi = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}$$

Thus, we have $Y = \Phi^T X = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} X$.

For data (5, 5), Y = $\begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} 5 \\ 5 \end{pmatrix} = \begin{pmatrix} 0 \\ 7.07 \end{pmatrix}$

For data (9, 9), Y = $\begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} 9 \\ 9 \end{pmatrix} = \begin{pmatrix} 0 \\ 12.73 \end{pmatrix}$

For data (3, 11), Y = $\begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} 3 \\ 11 \end{pmatrix} = \begin{pmatrix} -5.66 \\ 9.90 \end{pmatrix}$

For data (11, 3), Y = $\begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} 11 \\ 3 \end{pmatrix} = \begin{pmatrix} 5.66 \\ 9.90 \end{pmatrix}$

Thus, (5, 5) is reduced to (7.07);
(9, 9) is reduced to (12.73);
(3, 11) is reduced to (9.90);
(11, 3) is reduced to (9.90).

(c)
Yes. It is possible.
According to the steps in Part (a), we could calculate the same matrix ∑ based on these 8 data points. The remaining steps in Part (a) are the same.

The answer of this part is as follows.
(6, 6) (two occurrences) is reduced to (8.49);
(8, 8) (two occurrences) is reduced to (11.31);
(5, 9) (two occurrences) is reduced to (9.90);
(9, 5) (two occurrences) is reduced to (9.90).

(d)(i) $P(X=1|Y=1)=0 \quad P(X=1|Y=2)=1 \quad P(X=1|Y=3)=0.5$

$P(X=2|Y=1)=0.8 \quad P(X=2|Y=2)=0 \quad P(X=2|Y=3)=0.5$

$P(X=3|Y=1)=0.2 \quad P(X=3|Y=2)=0 \quad P(X=3|Y=3)=0$

$H(X|Y=1)=-0.8\log 0.8-0.2\log 0.2=0.7219,\ H(X|Y=1)=0,\ H(X|Y=3)=1$

$$\therefore H(X|Y)=\frac{5}{8}H(X|Y=1)+\frac{1}{8}H(X|Y=2)+\frac{1}{4}H(X|Y=3)$$

$$=\frac{5}{8}\times 0.7219+\frac{1}{8}\times 0+\frac{1}{4}\times 1$$

$$=0.7012$$

(ii)

$$H(X|Y)=-\frac{1}{2}\log 0.8-\frac{1}{8}\log 0.2-\frac{1}{8}\log 1-\frac{1}{8}\log\frac{1}{2}-\frac{1}{8}\log\frac{1}{2}$$

$$=0.7012$$

Q3

(a)(i)

Info(T) = $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

For attribute NoOfPapers,

Info(T_{enough})= $-1\log 1-0\log 0=0$

Info(T_{many})= $-\frac{3}{4}\log\frac{3}{4}-\frac{1}{4}\log\frac{1}{4}=0.8113$

Info(T_{few})= $-0\log 0-1\log 1=0$

Info(NoOfPapers)= 1/8 Info(T_{enough}) + ½ Info (T_{many}) + 3/8 Info (T_{few}) = 0.405

SplitInfo(NoOfPapers)= $-\frac{1}{8}\log\frac{1}{8}-\frac{1}{2}\log\frac{1}{2}-\frac{3}{8}\log\frac{3}{8}=1.4056$

Gain(NoOfPapers, T)= $\frac{1-0.405}{1.4056}=0.4233$

For attribute Age,

Info(T_{young})= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info(T_{old})= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info(Age, T)= $\frac{1}{2}Info(T_{young})+\frac{1}{2}Info(T_{old})=1$

SplitInfo(Age)= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Gain(Age, T)= $\frac{1-1}{1}=0$

For attribute Major,

Info($T_{ComputerScience}$)= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info($T_{CivilEngineering}$)= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

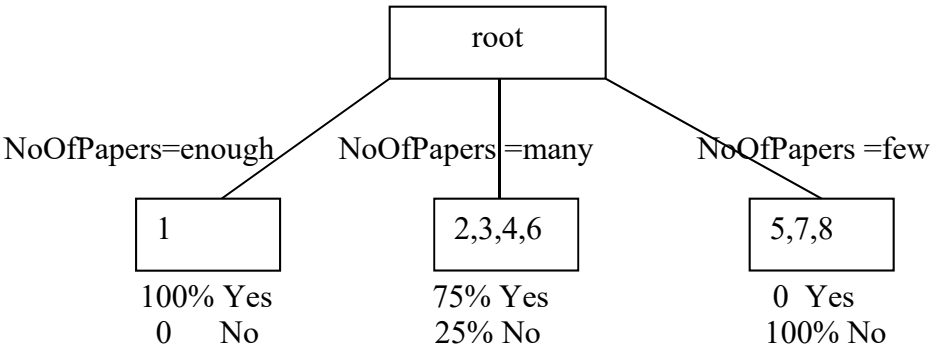
Info($T_{DataScience}$)= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info(Major, T)=1/2 Info($T_{ComputerScience}$) + ¼ Info($T_{CivilEngineering}$) + ¼ Info ($T_{DataScience}$) = 1

SplitInfo(Major)= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{4}\log\frac{1}{4}-\frac{1}{4}\log\frac{1}{4}=1.5$

Gain(Major, T)= $\frac{1-1}{1.5}=0$

We choose attribute **NoOfPapers** for Splitting:



Consider the node for "NoOfPapers=many"

Info(T)= $-\frac{3}{4}\log\frac{3}{4}-\frac{1}{4}\log\frac{1}{4}=0.8113$

For attribute Age,

Info(T_{young})= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info(T_{old})= $-1\log 1-0\log 0=0$

Info(Age, T)= $\frac{1}{2}Info(T_{young})+\frac{1}{2}Info(T_{old})=0.5$

SplitInfo(Age)= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Gain(Age, T)= $\frac{0.8113-0.5}{1}=0.3113$

For attribute Major,

Info($T_{ComputerScience}$)= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info($T_{CivilEngineering}$)= $-1\log 1-0\log 0=0$

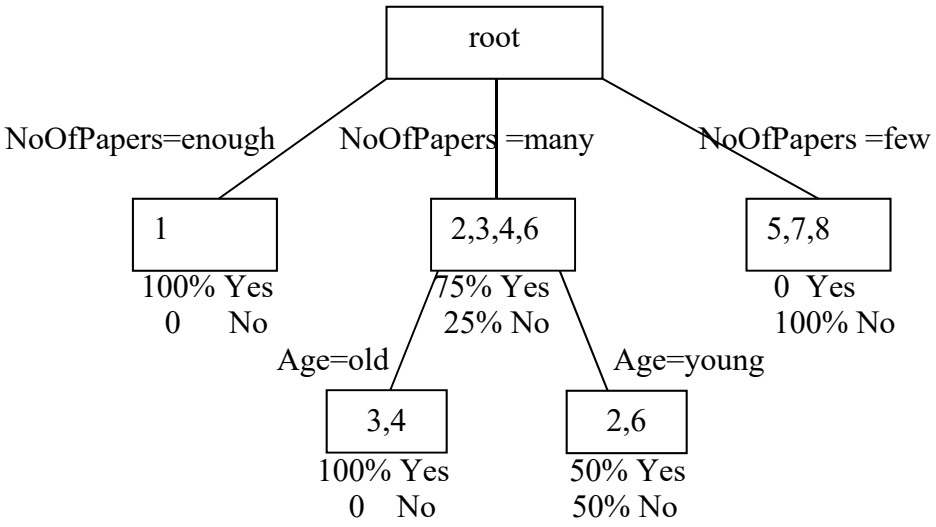
Info($T_{DataScience}$)= $-1\log 1-0\log 0=0$

Info(Major)=1/2 Info($T_{ComputerScience}$) + ¼ Info($T_{CivilEngineering}$) + ¼ Info($T_{DataScience}$) = 0.5

SplitInfo(Major)= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{4}\log\frac{1}{4}-\frac{1}{4}\log\frac{1}{4}=1.5$

Gain(Major, T)= $\frac{0.8113-0.5}{1.5}=0.2075$

We choose attribute **Age** for Splitting:



(ii) It is 50% probability that this PhD student will become a professor.

(b)(i) Person A said this because his claim is based on this "instance" of the "training set" only and based on some specific parameter values for this decision tree.

(ii) Person B said this because he thought that the result is just based on this "instance" of the training set. It is possible that other attributes could be found in X based on other instances of the training set.

(iii) Person A is more unreasonable since the "instance" of the training set is just a "small" snapshot of the whole population.

(c)
Difference:
The definition of the gain used in C4.5 is different from that used in ID3.
The gain used in C4.5 is equal to the gain used in ID3 divided by SplitInfo.

The reason why there is a difference is described as follows.
In ID3, there is a higher tendency to choose an attribute containing more values (e.g., attribute identifier and attribute HKID). Thus, splitInfo in C4.5 is used to penalize an attribute containing more values. If this value is larger, the penalty is larger.

Q4

(a) P(SIR = Yes) = $\sum_{x\in\{Yes,No\}}\sum_{y\in\{Yes,No\}} P(SIR = Yes|AP = x, P = y)P(AP = x, P = y)$

$= 0.7\times0.3\times0.6+0.45\times0.3\times0.4+0.55\times0.7\times0.6+0.2\times0.7\times0.4$

$= 0.467$

P(SIR = Yes | AP = Yes, P = Yes, WBC = Low)

$= \frac{P(WBC=Low|AP=Yes,P=Yes,SIR=Yes)}{P(WBC=Low|AP=Yes,P=Yes)} P(SIR = Yes|AP = Yes, P = Yes)$

$= \frac{P(WBC=Low|SIR=Yes)P(SIR=Yes|AP=Yes,P=Yes)}{\sum_{x\in\{Yes,No\}} P(WBC=Low|SIR=x)P(SIR=x|AP=Yes,P=Yes)}$

$= \frac{0.4*0.7}{0.4*0.7+0.7*0.3}$

$= 0.5714$

P(SIR = No | AP = Yes, P = Yes, WBC = Low) = 1- 0.5714 =0.4286

Since P(SIR = Yes | AP = Yes, P = Yes, WBC = Low) > P(SIR = No | AP = Yes, P = Yes, WBC = Low), it is more likely that the person has systemic inflammation reaction.

(a) Disadvantages:
The Bayesian Belief network classifier requires a predefined knowledge about the network.
The Bayesian Belief Network classifier cannot work directly when the network contains cycles.

Q5

(a)(i)

Consider t = 1.	Consider t = 2.
$f_1 = \sigma(W_f [x_1, y_0] + b_f)$ $= \sigma(\begin{pmatrix}0.8\\0.4\\0.1\end{pmatrix}\begin{pmatrix}0.3\\0.6\\0\end{pmatrix} + 0.2)$ $= \sigma(0.8 \cdot 0.3 + 0.4 \cdot 0.6 + 0.1 \cdot 0 + 0.2)$ $= \sigma(0.68)$ $= 0.6637$	$f_2 = \sigma(W_f [x_2, y_1] + b_f)$ $= \sigma(\begin{pmatrix}0.8\\0.4\\0.1\end{pmatrix}\begin{pmatrix}0.1\\1.0\\0.2376\end{pmatrix} + 0.2)$ $= \sigma(0.8 \cdot 0.1 + 0.4 \cdot 1.0 + 0.1 \cdot 0.2376 + 0.2)$ $= \sigma(0.7038)$ $= 0.6690$
$i_1 = \sigma(W_i [x_1, y_0] + b_i)$ $= \sigma(\begin{pmatrix}0.9\\0.8\\0.7\end{pmatrix}\begin{pmatrix}0.3\\0.6\\0\end{pmatrix} + 0.5)$ $= \sigma(0.9 \cdot 0.3 + 0.8 \cdot 0.6 + 0.7 \cdot 0 + 0.5)$ $= \sigma(1.25)$ $= 0.7773$	$i_2 = \sigma(W_i [x_2, y_1] + b_i)$ $= \sigma(\begin{pmatrix}0.9\\0.8\\0.7\end{pmatrix}\begin{pmatrix}0.1\\1.0\\0.2376\end{pmatrix} + 0.5)$ $= \sigma(0.9 \cdot 0.1 + 0.8 \cdot 1.0 + 0.7 \cdot 0.2376 + 0.5)$ $= \sigma(1.5564)$ $= 0.8258$
$a_1 = \tanh(W_a [x_1, y_0] + b_a)$ $= tanh(\begin{pmatrix}0.4\\0.2\\0.1\end{pmatrix}\begin{pmatrix}0.3\\0.6\\0\end{pmatrix} + 0.3)$ $= \tanh(0.4 \cdot 0.3 + 0.2 \cdot 0.6 + 0.1 \cdot 0 + 0.3)$ $= \tanh(0.54)$ $= 0.4930$	$a_2 = \tanh(W_a [x_2, y_1] + b_a)$ $= tanh(\begin{pmatrix}0.4\\0.2\\0.1\end{pmatrix}\begin{pmatrix}0.1\\1.0\\0.2376\end{pmatrix} + 0.3)$ $= \tanh(0.4 \cdot 0.1 + 0.2 \cdot 1.0 + 0.1 \cdot 0.2376 + 0.3)$ $= \tanh(0.5638)$ $= 0.5108$
$s_1 = f_1 \cdot s_0 + i_1 \cdot a_1$ $= 0.6637 \cdot 0 + 0.7773 \cdot 0.4930$ $= 0.3832$	$s_2 = f_2 \cdot s_1 + i_2 \cdot a_2$ $= 0.6690 \cdot 0.3832 + 0.8258 \cdot 0.5108$ $= 0.6782$
$o_1 = \sigma(W_o [x_1, y_0] + b_o)$ $= \sigma(\begin{pmatrix}0.6\\0.4\\0.1\end{pmatrix}\begin{pmatrix}0.3\\0.6\\0\end{pmatrix} + 0.2)$ $= \sigma(0.6 \cdot 0.3 + 0.4 \cdot 0.6 + 0.1 \cdot 0 + 0.2)$ $= \sigma(0.62)$ $= 0.6502$	$o_2 = \sigma(W_o [x_2, y_1] + b_o)$ $= \sigma(\begin{pmatrix}0.6\\0.4\\0.1\end{pmatrix}\begin{pmatrix}0.1\\1.0\\0.2376\end{pmatrix} + 0.2)$ $= \sigma(0.6 \cdot 0.1 + 0.4 \cdot 1.0 + 0.1 \cdot 0.2376 + 0.2)$ $= \sigma(0.6838)$ $= 0.6646$
$y_3 = o_1 \cdot \tanh(s_1)$ $= 0.6502 \cdot \tanh(0.3832)$ $= 0.2376$	$y_2 = o_2 \cdot \tanh(s_2)$ $= 0.6646 \cdot \tanh(0.6782)$ $= 0.3923$

(ii)

When t = 1,	When t = 2,
error = $y_1 - y$ $= 0.2376 - 0.2$ $= 0.0376$	error = $y_2 - y$ $= 0.3923 - 0.4$ $= -0.0077$

(b) (i)

Consider t = 1.	Consider t = 2.
$r_1 = \sigma(W_r [x_1, y_0] + b_r)$	$r_2 = \sigma(W_r [x_2, y_1] + b_r)$

$= \sigma\left(\begin{pmatrix}0.3\\0.2\\0.1\end{pmatrix}\begin{pmatrix}0.3\\0.6\\0\end{pmatrix} + 0.5\right)$ $= \sigma(0.3 \cdot 0.3 + 0.2 \cdot 0.6 + 0.1 \cdot 0 + 0.5)$ $= \sigma(\ 0.71\ )$ $= 0.6704$ $a_1 = \tanh(W_a\ [x_1, r_1 \cdot y_0] + b_a)$ $= tanh\left(\begin{pmatrix}0.4\\0.3\\0.1\end{pmatrix}\begin{pmatrix}0.3\\0.6\\0.6704 \cdot 0\end{pmatrix} + 0.1\right)$ $= \tanh(0.4 \cdot 0.3 + 0.3 \cdot 0.6 + 0.1 \cdot 0 + 0.1)$ $= \tanh(\ 0.4\ )$ $= 0.3799$ $u_1 = \sigma(W_u\ [x_1, y_0] + b_u)$ $= \sigma\left(\begin{pmatrix}0.4\\0.2\\0.1\end{pmatrix}\begin{pmatrix}0.3\\0.6\\0\end{pmatrix} + 0.1\right)$ $= \sigma(0.4 \cdot 0.3 + 0.2 \cdot 0.6 + 0.1 \cdot 0 + 0.1)$ $= \sigma(\ 0.34\ )$ $= 0.5842$ $y_1 = (1 - u_1) \cdot y_0 + u_1 \cdot a_1$ $= (1 - 0.5842) \cdot 0 + 0.5842 \cdot 0.3799$ $= 0.2220$	$= \sigma\left(\begin{pmatrix}0.3\\0.2\\0.1\end{pmatrix}\begin{pmatrix}0.1\\1.0\\0.2220\end{pmatrix} + 0.5\right)$ $= \sigma(0.3 \cdot 0.1 + 0.2 \cdot 1.0 + 0.1 \cdot 0.2220 + 0.5)$ $= \sigma(\ 0.7522\ )$ $= 0.6797$ $a_2 = \tanh(W_a\ [x_2, r_2 \cdot y_1] + b_a)$ $= tanh\left(\begin{pmatrix}0.4\\0.3\\0.1\end{pmatrix}\begin{pmatrix}0.1\\1.0\\0.6797 \cdot 0.2220\end{pmatrix} + 0.1\right)$ $= \tanh(0.4 \cdot 0.1 + 0.3 \cdot 1.0 + 0.1 \cdot 0.1509 + 0.1)$ $= \tanh(\ 0.4551\ )$ $= 0.4261$ $u_2 = \sigma(W_u\ [x_2, y_1] + b_u)$ $= \sigma\left(\begin{pmatrix}0.4\\0.2\\0.1\end{pmatrix}\begin{pmatrix}0.1\\1.0\\0.2220\end{pmatrix} + 0.1\right)$ $= \sigma(0.4 \cdot 0.1 + 0.2 \cdot 1.0 + 0.1 \cdot 0.2220 + 0.1)$ $= \sigma(\ 0.3622\ )$ $= 0.5896$ $y_2 = (1 - u_2) \cdot y_1 + u_2 \cdot a_2$ $= (1 - 0.5896) \cdot 0.2220 + 0.5896 \cdot 0.4261$ $= 0.3423$

(ii)

When t = 1, error = $y_1 - y$ $= 0.2220 - 0.2$ $= 0.0220$	When t = 2, error = $y_2 - y$ $= 0.3423 - 0.4$ $= -0.0577$

(c) The neural network has an assumption that records in the training set are "independent".
In some cases, records in the training set are related/correlated to (or dependent on) other records in the training set. Thus, the neural network could not capture this "dependent" scenario well if the training set has dependent records.