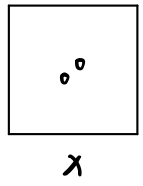
Q1

(a) This is because the number of grid units is larger when the subspace is larger, hence the probability of a dense subspace is smaller.

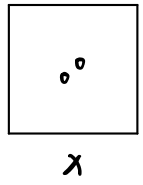
(b)

(i) NO. the Apriori-like algorithm is useless is the following example Consider $T_1 = 3$, $T_2 = 2$, the condition 1 satisfied, however in this case there is an example



There is a dense unit in subspace $\{x, y\}$, however there is no dense unit in $\{x\}$ or $\{y\}$

(ii) I think in this new form of Condition 1, we still cannot adopt the Apriori-like algorithm, α is at least 1 and if $\alpha = 1$, it is the same as Condition 2: for any i and j, $T_i = T_j$, however when $\alpha = 2$ or above



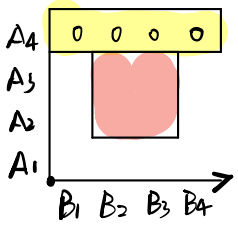
We can use the same example, when $T_2 = 2$, $T_1 = 4$
There is no dense unit in $\{x\}$ or $\{y\}$

So the greatest possible value is 1, but there is no difference between condition 1 and 2.

(c) We can adopt the Apriori-like algorithm in this case

The Apriori-like Algorithm is like this:

① $k = 1$. look for 1-dimension dense unit, store them in C_1



② Loop

③ $K = k+1$, $j = 1$

④ Loop (Join step)

⑤ $j = 2^{n-1} / i$

⑥ Look for $j \cdot i$ (for exam, in k-dimesion, j, i means width and length respectively) sized subspace in the permutation of C_{k-1}, restore them in L_k ($\{A_4\} * \{B_1, B_2, B_3, B_4\}$ in the figure)

⑦

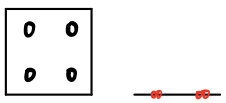
⑧

```
    i = i * 2
    If (i = 2 * 2^n!) break
Loop
Check the subspaces in L_i is dense, if true restore them in C_k (prune step)
if k==n, break
Loop
for (i=1; i<=n; i++)
    retue Ci
```

Q2

(a) False, the size of subspace doesn't have impart on whether the subspace has a good clustering.

For example:



In the case:
the data of 1-demension seems to have a better clustering that the two-dimension.

(b) (i) mean vector $= \begin{pmatrix} \frac{7+C+9+C+6+C+10+C}{4} \\ \frac{7+C+9+C+10+C+6+C}{4} \end{pmatrix} = \begin{pmatrix} 8+C \\ 8+C \end{pmatrix}$

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

② For data $(9+C, 9+C)$, difference from mean vector $\begin{pmatrix} 9+C-8-C \\ 9+C-8-C \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

③ For data $(6+C, 10+C)$, difference from mean vector $\begin{pmatrix} 6+C-8-C \\ 10+C-8-C \end{pmatrix} = \begin{pmatrix} -2 \\ 2 \end{pmatrix}$

④ For data $(10+C, 6+C)$, difference from mean vector $\begin{pmatrix} 10+C-8-C \\ 6+C-8-C \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$

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

$$\Sigma = \frac{1}{4} Y Y^T = \frac{1}{4} \begin{pmatrix} -1 & 1 & -2 & 2 \\ -1 & 1 & 2 & -2 \end{pmatrix} \begin{pmatrix} -1 & -1 \\ 1 & 1 \\ -2 & 2 \\ 2 & -2 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 10 & -6 \\ -6 & 10 \end{pmatrix} = \begin{pmatrix} \frac{5}{2} & -\frac{3}{2} \\ -\frac{3}{2} & \frac{5}{2} \end{pmatrix}$$

$$\begin{vmatrix} \frac{5}{2}-\lambda & -\frac{3}{2} \\ -\frac{3}{2} & \frac{5}{2}-\lambda \end{vmatrix} = 0 \Rightarrow \left(\frac{5}{2}-\lambda\right)^2 - \left(-\frac{3}{2}\right)^2 = 0 \Rightarrow \lambda = 4 \text{ or } \lambda = 1$$

when $\lambda = 4$

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

so 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 $\lambda = 1$

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

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

$$\Phi = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}, \quad Y = \Phi^T X = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} X = \begin{pmatrix} \frac{\sqrt{2}}{2}(y-x) \\ \frac{\sqrt{2}}{2}(x+y) \end{pmatrix}$$

For data $(7+C, 7+C)$, $Y = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} 7+C \\ 7+C \end{pmatrix} = \begin{pmatrix} 0 \\ 7\sqrt{2}+\sqrt{2}C \end{pmatrix} = \begin{pmatrix} 0 \\ 9.90+\sqrt{2}C \end{pmatrix}$

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

For data $(6+C, 10+C)$, $Y = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} 6+C \\ 10+C \end{pmatrix} = \begin{pmatrix} -2\sqrt{2} \\ 8\sqrt{2}+\sqrt{2}C \end{pmatrix} = \begin{pmatrix} -2.83 \\ 11.31+\sqrt{2}C \end{pmatrix}$

For data $(10+C, 6+C)$, $Y = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} 10+C \\ 6+C \end{pmatrix} = \begin{pmatrix} 2\sqrt{2} \\ 8\sqrt{2}+\sqrt{2}C \end{pmatrix} = \begin{pmatrix} 2.83 \\ 11.31+\sqrt{2}C \end{pmatrix}$

Therefore: $(7+C, 7+C) \longrightarrow 9.90+\sqrt{2}C$

$(9+C, 9+C) \longrightarrow 12.73+\sqrt{2}C$

$(6+C, 10+C) \longrightarrow 11.31+\sqrt{2}C$

$(10+C, 6+C) \longrightarrow 11.31+\sqrt{2}C$

(II) Yes.

We can do this in this way

for a $(7+C, 7+C)$ in dataset 1

$a'(7-d, 7-C) = (2-C, 2-C)$ in dataset 2

We can express a' in dataset 2 in a linear form involving a in dataset 1

$$\begin{cases} x_2 = -1 \cdot x_1 + 5 & ① \\ y_2 = -1 \cdot y_1 + 5 & ② \end{cases}$$

According to the steps in (i), we can easily obtain the same matrix

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

$$Y = \Phi^T X_2 = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} X_2 = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} -x_1+5 \\ -y_1+5 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2}(y_1-x_1) \\ -\frac{\sqrt{2}}{2}(x_1+y_1)+5\sqrt{2} \end{pmatrix}$$

$$X_1 = \begin{matrix} (7+C, 7+C) \\ (9+C, 9+C) \\ (6+C, 10+C) \\ (10+C, 6+C) \end{matrix} \Longrightarrow \begin{matrix} (0, -7\sqrt{2}C-2\sqrt{2}) & \longrightarrow (-2\sqrt{2}C-2\sqrt{2}) \\ (0, -7\sqrt{2}C-4\sqrt{2}) & \longrightarrow (-\sqrt{2}C-4\sqrt{2}) \\ (2\sqrt{2}, -\sqrt{2}C-3\sqrt{2}) & \longrightarrow (-\sqrt{2}C-3\sqrt{2}) \\ (-2\sqrt{2}, -\sqrt{2}C-3\sqrt{2}) & \longrightarrow (-\sqrt{2}C-3\sqrt{2}) \end{matrix}$$

(iii) Yes. If we want to make use of the answers, we need to find the linear mapping relationship in the two coordinates systems, and we can get a linear mapping relationship like this:

$$\begin{cases} x_1 = a+c \\ y_1 = b+c \end{cases} \Rightarrow \begin{cases} a = x_1 - c \\ b = y_1 - c \end{cases}$$

$$\begin{cases} x_2 = ac \\ y_2 = bc \end{cases} \Rightarrow \begin{cases} x_2 = (x_1-c)c = x_1c - c^2 \\ y_2 = (y_1-c)c = y_1c - c^2 \end{cases}$$

According to the steps in (i), we can easily obtain the same matrix

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

$$Y = \Phi^T X_2 = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} cx_1 - c^2 \\ cy_1 - c^2 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2}(y_1 - x_1) \\ \frac{\sqrt{2}c}{2}(y_1 + x_1) - \sqrt{2}c^2 \end{pmatrix}$$

$$X_1 = \begin{matrix} (7+c, 7+c) \\ (9+c, 9+c) \\ (6+c, 10+c) \\ (10+c, 6+c) \end{matrix} \Longrightarrow \begin{matrix} (0, 7\sqrt{2}c) \to (7\sqrt{2}c) \\ (0, 9\sqrt{2}c) \to (9\sqrt{2}c) \\ (2\sqrt{2}, 8\sqrt{2}c) \to (8\sqrt{2}c) \\ (-2\sqrt{2}, 8\sqrt{2}c) \to (8\sqrt{2}c) \end{matrix}$$

Because c is a positive real number but we do not know the exact value we can get the answer of the first two data, however, we need to know the specific value of c so that we get to know the last two number.

So the conclusion is that we can make use of the answers in (b)(i), but we need to know the exact number of c.

Q3

(a) (i) $I(p) = 1 - \sum_j p_j^2$

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

For attribute Income:

$Info(T_{high}) = 1 - 1^2 - 0^2 = 0$

$Info(T_{medium}) = 1 - (\frac{1}{5})^2 - (\frac{4}{5})^2 = \frac{8}{25} = 0.32$

$Info(Income, T) = \frac{3}{8} \times Info(T_{high}) + \frac{5}{8} \times Info(T_{medium})$

$= \frac{3}{8} \times 0 + \frac{5}{8} \times \frac{8}{25} = 0.2$

$Gain(Income, T) = Info(T) - Info(Income, T)$

$= \frac{1}{2} - 0.2 = 0.3$

For attribute Have_iPhone

$Info(T_{yes}) = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = \frac{1}{2}$

$Info(T_{no}) = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = \frac{1}{2}$

$Info(Have\_iphone, T) = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = 0.5$

$Gain(Have, iphone, T) = \frac{1}{2} - \frac{1}{2} = 0$

For attribute Have_ipad

$Info(T_{yes}) = 1 - 1^2 - 0^2 = 0$

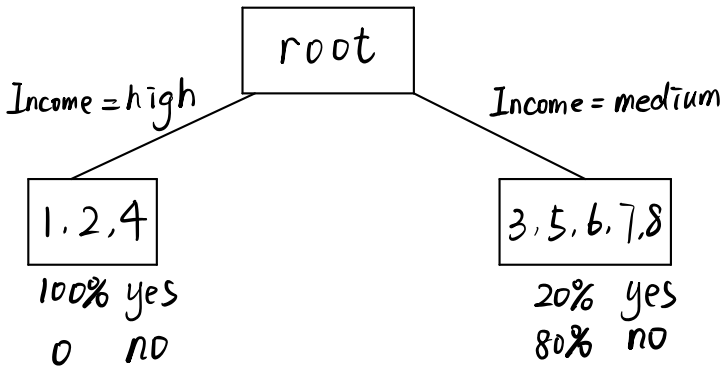
$Info(T_{no}) = 1 - (\frac{2}{6})^2 - (\frac{4}{6})^2 = \frac{4}{9} = 0.44$

$Info(Have\_ipad, T) = \frac{2}{8} \times 0 + \frac{6}{8} \times \frac{4}{9} = \frac{1}{3} = 0.33$

$Gain(Have\_ipad, T) = \frac{1}{2} - 0.33 = 0.17$

$max\ Gain = Gain(Income, T) = 0.3$

We can arribute Income for splitting



Consider the node for "Income = Medium"

$Info(T) = 1 - (\frac{1}{5})^2 - (\frac{4}{5})^2 = \frac{8}{25}$

For attribute Have_iphone

$Info(T_{yes}) = 1 - (\frac{1}{3})^2 - (\frac{2}{3})^2 = \frac{4}{9} = 0.44$

$Info(T_{no}) = 1 - 1^2 - 0^2 = 0$

$Info(Have\_iphone, T) = \frac{3}{5} \times \frac{4}{9} + \frac{2}{5} \times 0 = \frac{4}{15} = 0.27$

$Gain(Have\_iphone, T) = 0.5 - 0.27 = 0.23$

$maxGain = Gain(Have\_iphone, T) = 0.23$

For arribute Have_ipad

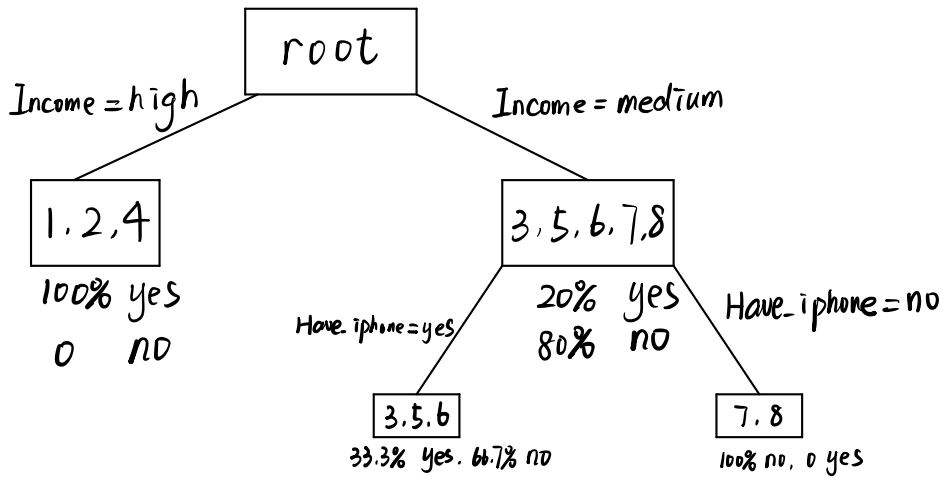
$Info(T_{yes}) = 1 - 1^2 - 0^2 = 0$

$Info(T_{no}) = 1 - (\frac{1}{5})^2 - (\frac{4}{5})^2 = \frac{8}{25}$

$Info(Have\_ipad, T) = 1 \times \frac{8}{25} = 0.32$

$Gain(Have\_ipad, T) = 0.18$

so we choose "Have_iphone" attribute for spitting



(ii) the costumer 66.7% not likely to buy the ipad-mini, 33.3% will buy the ipad.

(b) Difference:

The definition of Gain used in C4.5 is different from that used in ID3

$$Gain(A,T)_{C4.5} = Gain(A,T) / SplitInfo(A)$$

In ID3, there is a higher tendency to choose an attribute containing more values but C4.5 is used to penalize an attribute containing more values.

Q4

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

$= 0.7\times0.3\times0.6 + 0.45\times0.3\times0.6 + 0.55\times0.7\times0.6 + 0.2\times0.7\times0.4 = 0.467$

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

$$= \frac{P(WBC=Low|SIR=Yes)P(SIR=Yes|AP=Yes, P=Yes)}{\Sigma_{x\in\{Yes,No\}} P(WBC=Low|SIR=x)\cdot P(SIR=x|AP=Yes, P=Yes)} = \frac{0.4\times0.7}{0.4\times0.7+0.7\times0.3} = 0.57$$

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

$$= 1 - 0.57 = 0.43 < 0.57$$

So the person is more likely to have SIR.

(b)

① The Bayesian Belief Network Classifier requires a predefined knowledge, can only exploit causal influences that are recognized by programmers.

② There is no universally acknowledged method for constructing networks.

③ It fails to define cycle relationships

④ It is expensive to build.

⑤ It performs poorly on high dimensional data.

⑥ It is tough to interpret and require copula function to separate effects and causes.

Q5 (a)

(i) when $t=1$

$f_1 = \sigma(W_f[x_1, y_0] + b_f)$

$= \sigma\left(\begin{pmatrix}0.8\\0.4\\0.1\end{pmatrix}\begin{pmatrix}0.3\\0.6\\0\end{pmatrix} + 0.2\right)$

$= \sigma(0.48 + 0.2) = \sigma(0.68) = 0.6637$

$i_1 = \sigma(W_i[x_1, y_0] + b_i)$

$= \sigma\left(\begin{pmatrix}0.9\\0.8\\0.7\end{pmatrix}\begin{pmatrix}0.3\\0.6\\0\end{pmatrix} + 0.5\right)$

$= \sigma(0.75 + 0.5) = \sigma(1.25) = 0.7773$

$a_1 = \tanh(W_a[x_1, y_0] + b_a)$

$= \tanh\left(\begin{pmatrix}0.4\\0.2\\0.1\end{pmatrix}\begin{pmatrix}0.3\\0.6\\0\end{pmatrix} + 0.3\right)$

$= \tanh(0.24 + 0.3) = \tanh(0.54)$

$= 0.4930$

$S_1 = f_1 \cdot S_0 + i_1 \cdot a_1 = 0.6637 \times 0 + 0.7773 \times 0.4930 = 0.3832$

$O_1 = \sigma(W_o[x_1, y_0] + b_o) = \sigma\left(\begin{pmatrix}0.6 & 0.3\\0.4 & 0.6\\0.1 & 0\end{pmatrix} + 0.2\right) = \sigma(0.42 + 0.2) = \sigma(0.62)$

$= 0.6502$

$y_3 = O_1 \cdot \tanh(S_1) = 0.6502 \times \tanh(0.3832) = 0.2376$

when $t=2$

$f_2 = \sigma(W_f[x_2, y_1] + b_f) = \sigma\left(\begin{pmatrix}0.8\\0.4\\0.1\end{pmatrix}\begin{pmatrix}0.1\\1.0\\0.2376\end{pmatrix} + 0.2\right) = \sigma(0.5038 + 0.2) = \sigma(0.7038) = 0.6690$

$i_2 = \sigma(W_i[x_2, y_1] + b_i) = \sigma\left(\begin{pmatrix}0.9\\0.8\\0.7\end{pmatrix}\begin{pmatrix}0.1\\1.0\\0.2376\end{pmatrix} + 0.5\right) = \sigma(1.0564 + 0.5) = \sigma(1.5564) = 0.8258$

$a_2 = \tanh(W_a[x_2, y_1] + b_a) = \tanh\left(\begin{pmatrix}0.4\\0.2\\0.1\end{pmatrix}\begin{pmatrix}0.1\\1.0\\0.2376\end{pmatrix} + 0.3\right) = \tanh(0.2638 + 0.3) = \tanh(0.5638) = 0.5108$

$S_2 = f_2 \cdot S_1 + i_2 \cdot a_1 = 0.6690 \times 0.3832 + 0.8258 \times 0.5108 = 0.6708$

$O_2 = \sigma(W_o[x_2, y_1] + b_o) = \sigma\left(\begin{pmatrix}0.6\\0.4\\0.1\end{pmatrix}\begin{pmatrix}0.1\\1.0\\0.2376\end{pmatrix} + 0.2\right) = \sigma(0.4838 + 0.2) = \sigma(0.6838) = 0.6646$

$y_2 = O_2 \cdot \tanh(S_2) = 0.6646 \times \tanh(0.6782) = 0.3923$

(ii) when $t=1$: $\text{error} = y_1 - y = 0.2376 - 0.2 = 0.0376$

when $t=2$:

$\text{error} = y_2 - y = 0.3923 - 0.4 = -0.077$

(b)

(i)

when $t=1$

$r_1 = \sigma(W_r[x_1, y_0] + b_r) = \sigma\left(\begin{pmatrix}0.3\\0.2\\0.1\end{pmatrix}\begin{pmatrix}0.3\\0.6\\0\end{pmatrix} + 0.5\right) = \sigma(0.21 + 0.5) = \sigma(0.71) = 0.6704$

$a_1 = \tanh(W_a[x_1, r_1 \cdot y_0] + b_a) = \tanh\left(\begin{pmatrix}0.4\\0.3\\0.1\end{pmatrix}\begin{pmatrix}0.3\\0.6\\0.6704 \times 0\end{pmatrix} + 0.1\right) = \tanh(0.3 + 0.1) = \tanh(0.4) = 0.3799$

$u_1 = \sigma(W_u[x_1, y_0] + b_u) = \sigma\left(\begin{pmatrix}0.4\\0.2\\0.1\end{pmatrix}\begin{pmatrix}0.3\\0.6\\0\end{pmatrix} + 0.1\right) = \sigma(0.24 + 0.1) = \sigma(0.34) = 0.5842$

$y_1 = (1 - u_1) \cdot y_0 + u_1 \cdot a_1 = (1 - 0.5842) \times 0 + 0.5842 \times 0.3799 = 0.2220$

When $t=2$

$r_2 = \sigma(W_r[x_2, y_1] + b_r) = \sigma\left(\begin{pmatrix}0.3\\0.2\\0.1\end{pmatrix}\begin{pmatrix}0.1\\1\\0.2220\end{pmatrix} + 0.5\right) = \sigma(0.2522 + 0.5) = \sigma(0.7522) = 0.6797$

$a_2 = \tanh(W_a[x_2, r_2 \cdot y_1] + b_a) = \tanh\left(\begin{pmatrix}0.4\\0.3\\0.1\end{pmatrix}\begin{pmatrix}0.1\\1\\0.6797 \times 0.2220\end{pmatrix} + 0.1\right) = \tanh(0.3551 + 0.1) = \tanh(0.4551) = 0.4261$

$u_2 = \sigma(W_u[x_2, y_1] + b_u] = \sigma\left(\begin{pmatrix}0.4\\0.2\\0.1\end{pmatrix}\begin{pmatrix}0.1\\1\\0.2220\end{pmatrix} + 0.1\right) = \sigma(0.2622 + 0.1) = \sigma(0.3622) = 0.5896$

$y_2 = (1 - u_2) y_1 + u_2 \cdot a_2 = (1 - 0.5896) \times 0.2220 + 0.5896 \times 0.4261 = 0.3423$

(ii) When $t=1$:

$$\text{error} = y_1 - y = 0.2220 - 0.2 = 0.0220$$

When $t=2$:

$$\text{error} = y_2 - y = 0.3423 - 0.4 = -0.0577$$

(c) The traditional neural network has an assumption that records in the table are independent. In some cases, the current record is "related" to the "previous" records in the table, the traditional neural networks could not capture these dependent features in the model.