COMP1942 Exploring and Visualizing Data (Spring Semester 2016)
Final Examination (Answer Sheet)
Date: 26 May, 2016 (Thu)
Time: 16:30-19:30
Duration: 3 hours

Student ID:__________________ Student Name:__________________________________

Seat No. :__________________

Instructions:
(1) Please answer **all** questions in **Part A** in this paper.
(2) You can **optionally** answer the bonus question in **Part B** in this paper. You can obtain additional marks for the bonus question if you answer it correctly.
(3) The total marks in Part A are 200.
(4) The total marks in Part B are 20.
(5) The total marks you can obtain in this exam are 200 only.
If you answer the bonus question in Part B correctly, you can obtain additional marks.
But, if the total marks you obtain from Part A and Part B are over 200, your marks will be truncated to 100 only.
(6) You can use a calculator.

Answer Sheet

Part	Question	Full Mark	Mark
A	Q1	20	
	Q2	20	
	Q3	20	
	Q4	20	
	Q5	20	
	Q6	20	
	Q7	20	
	Q8	20	
	Q9	20	
	Q10	20	
Total (Part A)		200	
B	Q11 (OPTIONAL)	20	
Total (Parts A and B)		200	

Part A (Compulsory Short Questions)

Q1 (20 Marks)
(a)

a
a, b
a, b
a, b
a, c
a, b, c
a, b, c
b
b
b
b, c

(b) (i)

Yes. This is because L_2 is exactly equal to the set of itemsets in C_2 which frequency is at least a given support threshold.

(ii)

No. Suppose that L_1 = {{A}, {B}}. Then, C_2 = {{A, B}}. In this case, the number of itemsets in C_2 is smaller than the number of itemsets in L_1.

Q2 (20 Marks)

$$
\begin{array}{c|c}
 & \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} &
\begin{pmatrix}
0 & & & & & & \\
10 & 0 & & & & & \\
7 & 7 & 0 & & & & \\
30 & 23 & 21 & 0 & & & \\
29 & 25 & 22 & 7 & 0 & & \\
38 & 34 & 31 & 10 & 11 & 0 & \\
42 & 36 & 36 & 13 & 17 & 9 & 0
\end{pmatrix}
\end{array}
$$

A = {1 }

B = {2, 3, 4, 5, 6, 7}

D(1, *) = 26.0

D(2, *) = 22.5

D(3, *) = 20.7

D(4, *) = 17.3

D(5, *) = 18.5

D(6, *) = 22.2

D(7, *) = 25.5

Q2 (Continued)

$$
\begin{array}{c|ccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
1 & 0 & & & & & & \\
2 & 10 & 0 & & & & & \\
3 & 7 & 7 & 0 & & & & \\
4 & 30 & 23 & 21 & 0 & & & \\
5 & 29 & 25 & 22 & 7 & 0 & & \\
6 & 38 & 34 & 31 & 10 & 11 & 0 & \\
7 & 42 & 36 & 36 & 13 & 17 & 9 & 0
\end{array}
$$

D(2, A) = 10	D(2, B) = 25.0	Δ_2 = 15.0
D(3, A) = 7	D(3, B) = 23.4	Δ_3 = 16.4
D(4, A) = 30	D(4, B) = 14.8	Δ_4 = -15.2
D(5, A) = 29	D(5, B) = 16.4	Δ_5 = -12.6
D(6, A) = 38	D(6, B) = 19.0	Δ_6 = -19.0
D(7, A) = 42	D(7, B) = 22.2	Δ_7 = -19.8

A = {1, 3 }

B = {2, 4, 5, 6, 7}

$$
\begin{array}{c|ccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
1 & 0 & & & & & & \\
2 & 10 & 0 & & & & & \\
3 & 7 & 7 & 0 & & & & \\
4 & 30 & 23 & 21 & 0 & & & \\
5 & 29 & 25 & 22 & 7 & 0 & & \\
6 & 38 & 34 & 31 & 10 & 11 & 0 & \\
7 & 42 & 36 & 36 & 13 & 17 & 9 & 0
\end{array}
$$

D(2, A) = 8.5	D(2, B) = 29.5	Δ_2 = 21.0
D(4, A) = 25.5	D(4, B) = 13.2	Δ_4 = -12.3
D(5, A) = 25.5	D(5, B) = 15.0	Δ_5 = -10.5
D(6, A) = 34.5	D(6, B) = 16.0	Δ_6 = -18.5
D(7, A) = 39.0	D(7, B) = 18.6	Δ_7 = -20.4

A = {1, 3, 2}

B = { 4, 5, 6, 7}

Q2 (Continued)

$$
\begin{array}{c}
\\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7
\end{array}
\begin{array}{c}
\begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\
\begin{pmatrix}
0 & & & & & & \\
10 & 0 & & & & & \\
7 & 7 & 0 & & & & \\
30 & 23 & 21 & 0 & & & \\
29 & 25 & 22 & 7 & 0 & & \\
38 & 34 & 31 & 10 & 11 & 0 & \\
42 & 36 & 36 & 13 & 17 & 9 & 0
\end{pmatrix}
\end{array}
$$

$\boxed{D(4, A) = 24.7}$ $\boxed{D(4, B) = 10.0}$ $\boxed{\Delta_4 = -14.7}$

$\boxed{D(5, A) = 25.3}$ $\boxed{D(5, B) = 11.7}$ $\boxed{\Delta_5 = -13.6}$

$\boxed{D(6, A) = 34.3}$ $\boxed{D(6, B) = 10.0}$ $\boxed{\Delta_6 = -24.3}$

$\boxed{D(7, A) = 38.0}$ $\boxed{D(7, B) = 13.0}$ $\boxed{\Lambda_7 = -25.0}$

$\boxed{A = \{1, 3, 2\}}$

$\boxed{B = \{4, 5, 6, 7\}}$

Q2 (Continued)

Q3 (20 Marks)
(a)

Make initial guesses of the means $m_1, m_2, \ldots, m_k$
Set the counts $n_1, n_2, \ldots, n_k$ to zero
Until interrupted
  Acquire the next example x
  If m_i is closest to x
    Increment n_i
    Replace m_i by $m_i + 1/n_i (x - m_i)$

(b)

x_j : the j-th example in cluster i
$m_i(t)$: the mean vector of cluster i containing t examples

Consider that x is the t-th example in cluster i

$$m_i(t-1) = \frac{x_1 + x_2 + \ldots + x_{t-1}}{t-1}$$

$$m_i(t) = \frac{x_1 + x_2 + \ldots + x_{t-1} + x_t}{t}$$

$$= \frac{m_i(t-1) \times (t-1) + x_t}{t}$$

$$= \frac{t \times m_i(t-1) + x_t - m_i(t-1)}{t}$$

$$= m_i(t-1) + \frac{1}{t}(x_t - m_i(t-1))$$

Q4 (20 Marks)

(a) P(LC=Yes) = $\sum_{x\in\{Yes,No\}}\sum_{y\in\{Yes,No\}} P(LC = Yes \mid FH = x, S = y)P(FH = x, S = y)$

$= 0.7\times 0.3\times 0.6 + 0.45\times 0.3\times 0.4 + 0.55\times 0.7\times 0.6 + 0.2\times 0.7\times 0.4$

$= 0.467$

$$
\begin{aligned}
&P(LC = Yes \mid FH = Yes, Smoker = No, PR = Yes) \\
&= \frac{P(PR = Yes \mid FH = Yes, Smoker = No, LC = Yes)}{P(PR = Yes \mid FH = Yes, Smoker = No)} P(LC = Yes \mid FH = Yes, Smoker = No) \\
&= \frac{P(PR = Yes \mid LC = Yes)\times P(LC = Yes \mid FH = Yes, Smoker = No)}{\sum_{x\in\{Yes,No\}} P(PR = Yes \mid LC = x)P(LC = x \mid FH = Yes, Smoker = No)} \\
&= \frac{0.85\times 0.45}{0.85\times 0.45 + 0.45\times 0.55} \\
&= 0.607143
\end{aligned}
$$

$P(LC = No \mid FH = Yes, Smoker = No, PR = Yes) = 1 - 0.601743 = 0.392857$

$\because P(LC = Yes \mid FH = Yes, Smoker = No, PR = Yes) > P(LC = No \mid FH = Yes, Smoker = No, PR = Yes)$ $\therefore$ It is more likely that the person is likely to have Lung Cancer.

Q4 (Continued)

(b) 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 (20 Marks)
(a)

Classifier 1: No
Classifier 2: Yes
Classifier 3: Yes

The overall predicted results is “Yes” (since the majority of the results is “Yes”).

(b)

The target attribute of this new record is “Yes”.
The 3 nearest neighbors are 8, 9 and 13.

(c)

No.

P(Insurance = Yes) = 1/2
P(Insurance = No) = 1/2

P(Insurance = Yes | A = 0) = 3/4
P(Insurance = No | A = 0) = 1/4

P(Insurance = Yes | A = 1) = 3/4
P(Insurance = No | A = 1) = 1/4

P(A = 0) = 1/2
P(A = 1) = 1/2

P(Insurance = Yes | B = 0) = 1
P(Insurance = No | B = 0) = 0

P(Insurance = Yes | B = 1) = 1/3
P(Insurance = No | B = 1) = 2/3

P(B = 0) = 1/8
P(B = 1) = 7/8

Q5 (Continued)

Consider ID3.
Info(T) = - 1/2 log 1/2 - 1/2 log 1/2 = 1

Consider attribute A.
Info(T_0) = -3/4 log 3/4 - 1/4 log 1/4 = 0.8113
Info(T_1) = -3/4 log 3/4 - 1/4 log 1/4 = 0.8113
Info(A, T) = 1/2 Info(T_0) + 1/2 Info(T_1) = 0.8113
Gain(A, T) = Info(T) - Info(A, T) = 1 - 0.8113 = 0.1887

Consider attribute B.
Info(T_0) = -1 log 1 - 0 log 0 = 0
Info(T_1) = -1/3 log 1/3 - 2/3 log 2/3 = 0.9183
Info(B, T) = 1/8 Info(T_0) + 7/8 Info(T_1) = 0.8035
Gain(b, T) = Info(T) - Info(B, T) = 1 - 0.8035 = 0.1965

Here, Gain(A, T) < Gain(B, T)

Under ID3, Imp-ID3(A) = Gain(A, T)
and Imp-ID3(B) = Gain(B, T).
Thus, we have "Imp-ID3(A) < Imp-ID3(B)".

Consider CART.
Info(T) = 1 - $(1/2)^2$ - $(1/2)^2$ = 1/2

Consider attribute A.
Info(T_0) = 1 - $(3/4)^2$ - $(1/4)^2$ = 0.375
Info(T_1) = 1 - $(3/4)^2$ - $(1/4)^2$ = 0.375
Info(A, T) = 1/2 Info(T_0) + 1/2 Info(T_1) = 0.375
Gain(A, T) = Info(T) - Info(A, T) = 1/2 - 0.375 = 0.125

Consider attirbute B.
Info(T_0) = 1 - 1^2 - 0^2 = 0
Info(T_1) = 1 - $(1/3)^2$ - $(2/3)^2$ = 0.444
Info(B, T) = 1/8 Info(T_0) + 7/8 Info(T_1) = 0.3885
Gain(B, T) = Info(T) - Info(B, T) = 1/2 - 0.3885 = 0.1115

Here, Gain(A, T) > Gain(B, T).

Under CART, Imp-CART(A) = Gain(A, T)
and Imp-CART(B) = Gain(B, T).
Thus, we have "Imp-CART(A) > Imp-CART(B)".

In conclusion, it is possible that
"Imp-CART(A) > Imp-CART(B)" but "Imp-ID3(A) < Imp-ID3(B)".

Q5 (Continued)

Q6 (20 Marks)
(a)

We can transform the data into a higher dimensional space using a "nonlinear" mapping.
Then, we can use the neural network containing only one neuron in this high-dimensional space for classification.

(b)

Cluster 1: {1, 2, 4, 5, 6}
Cluster 2: {3, 7, 8, 9, 10}

Q7 (20 Marks)

(a)

accuracy = 0.857 (= (7+17)/28)

(b)

recall = 0.875 (=7/8)

(c)

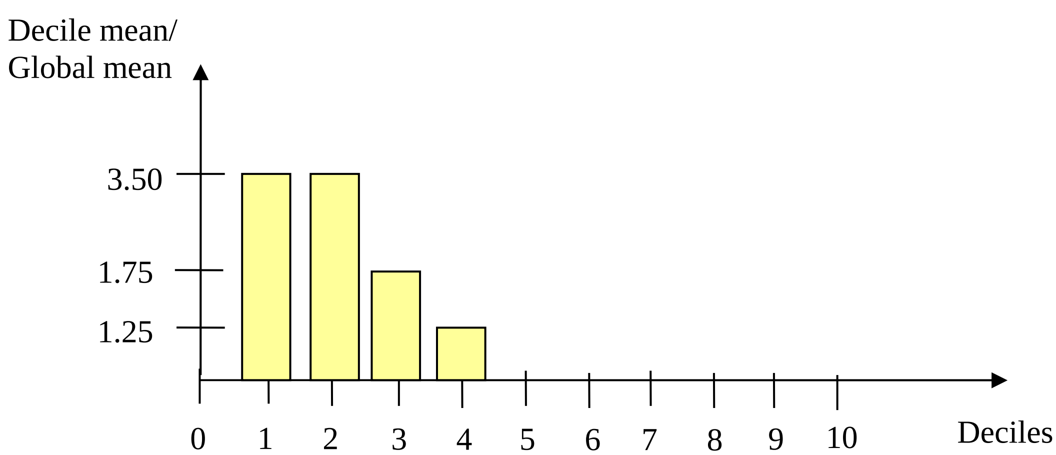
f-measure = 0.778 (= 2 x 0.7 x 0.875 /(0.7 + 0.875))

(d)

no. of false negatives = 1

(e)

Decile-wise Lift Chart:



Q8 (20 Marks)

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

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

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

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

For data (9, 5), 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 $\lambda = 4$,

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

Q8 (Continued)

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

The mean vector of the above transformed data points is $\begin{pmatrix} \frac{0+0+(-2.83)+2.83}{4} \\ \frac{8.49+11.31+9.90+9.90}{4} \end{pmatrix} = \begin{pmatrix} 0 \\ 9.90 \end{pmatrix}$

The final transformed data points are:

Q8 (Continued)

For data (6, 6), final transformed vector = $\begin{pmatrix} 0-0 \\ 8.49-9.90 \end{pmatrix} = \begin{pmatrix} 0 \\ -1.41 \end{pmatrix}$

For data (8, 8), final transformed vector = $\begin{pmatrix} 0-0 \\ 11.31-9.90 \end{pmatrix} = \begin{pmatrix} 0 \\ 1.41 \end{pmatrix}$

For data (5, 9), final transformed vector = $\begin{pmatrix} -2.83-0 \\ 9.90-9.90 \end{pmatrix} = \begin{pmatrix} -2.83 \\ 0 \end{pmatrix}$

For data (9, 5), final transformed vector = $\begin{pmatrix} 2.83-0 \\ 9.90-9.90 \end{pmatrix} = \begin{pmatrix} 2.83 \\ 0 \end{pmatrix}$

Thus, (6, 6) is reduced to (0);
(8, 8) is reduced to (0);
(5, 9) is reduced to (-2.83);
(9, 5) is reduced to (2.83).

(Note: Another possible answer is
(6, 6) is reduced to (0);
(8, 8) is reduced to (0);
(5, 9) is reduced to (2.83);
(9, 5) is reduced to (-2.83).
This is because the eigenvectors used in this case are:

$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$ and $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$.)

Q8 (Continued)

Q8 (Continued)

Q9 (20 Marks)
(a) (i)

Yes.
For each part p, we count the total number of records in the answer of Q3 and insert a record (p, C) into the answer of Q4.

(ii)

No
We need one additional kind of information, the answer of Q3, in addition to the answer of Q5.
The total access cost is 2GB only (the minimum access cost).

For each part p, we do the following.
  we initialize the SUM variable to be 0.
  we initialize the TOTAL variable to be 0.
  For each combination of part and customer c where the part is equal to p,
    we obtain the average price A for (c, p) from the answer of Q5
      and the total number of records in T for (c, p) from the answer of Q3.
    we increment SUM by A x C.
    we increment TOTAL by C.
  we construct a record (p, AVG) where AVG = SUM/TOTAL
  we insert this record into the answer of Q6.

Q9 (Continued)

(b) (i)

We want to transform the objective function from a non-linear form to a quadratic form.
Then, the problem becomes a form of quadratic programming which has many existing efficient techniques for that.

(ii)

200

Q10 (20 Marks)

(a)

Yes.

$$\begin{array}{c c} & \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{array}{c} a \\ b \\ c \\ d \end{array} & \begin{pmatrix} 0 & 0.5 & 0.33 & 0.25 \\ 0.5 & 0.5 & 0 & 0.25 \\ 0 & 0 & 0.33 & 0.25 \\ 0.5 & 0 & 0.33 & 0.25 \end{pmatrix} \end{array}$$

(b) (i)

w_1, w_2, w_6, w_7, w_8

(ii)

w_1, w_2, w_3, w_6, w_7, w_8, w_9

Part B (Bonus Question)

Note: The following bonus question is an **OPTIONAL** question. You can decide whether you will answer it or not.

Q11 (20 Additional Marks)

Since the memory is divided into four equal parts, each part occupies 4096/4 = 1024 bytes.

Consider a single part.
This part stores a summary X which contains two components.
The second component X.p occupies 4 bytes.
Thus, the first component X.E occupies 1024-4=1020 bytes.
Since each entry occupies 12 bytes, the first component X.E can store 1020/12 = 85 entries.
The greatest error in any estimated frequency (in fraction) within this summary X is equal to 1/85 = 0.011765.

Consider 4 parts together.
The greatest error in any estimated frequency (in count) is equal to 1/85 x 4B = 4B/85.
The greatest error in any estimated frequency (in fraction) is equal to 4B/85 x 1/4B = 1/85 = 0.011765.

Q11 (Continued)

End of Answer Sheet