COMP1942 Exploring and Visualizing Data (Spring Semester 2018)
Final Examination (Answer Sheet)
Date: 25 May, 2018 (Fri)
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 200 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)

The reason why we cannot simply output C as the final output is that not all itemsets in C are frequent (i.e., not all itemsets in C can be in the final output).

Let us use the size-2 itemset generation for illustration.

Originally, L_1 = {P, Q, S, T}
After the counting step and the pruning step, we have
C_2 = { PQ, PS, PT, QS, QT, ST }
Not all itemsets in C_2 have frequency at least 2.
E.g., ST is not frequent since its frequency is equal to 1. Thus, ST is not in the output.

(b)

Itemset	Frequency
{a, b, c}	4
{a, b}	7
{a, c}	4
{b, c}	4
a	15
b	7
c	4

Q2 (20 Marks)

(a)(i)

- Make initial guesses for the means m_1, m_2, …, m_k
- Until Interrupted
  - Acquire the next example x
  - If m_i is closest to x,
    - replace m_i by $m_i + a(x - m_i)$

(ii)

$$
\begin{aligned}
m_n &= m_{n-1} + a(x_n - m_{n-1}) \\
&= (1-a)m_{n-1} + ax_n \\
&= (1-a)[(1-a)m_{n-2} + ax_{n-1}] + ax_n \\
&= (1-a)^2 m_{n-2} + (1-a)ax_{n-1} + ax_n \\
&=(1-a)^2[(1-a)m_{n-3} + ax_{n-2}] + (1-a)ax_{n-1} + ax_n \\
&=(1-a)^3 m_{n-3} + (1-a)^2 ax_{n-2} + (1-a)ax_{n-1} + ax_n \\
&= \ldots \\
&= (1-a)^n m_0 + \sum_{p=1}^{n} (1-a)^{n-p} ax_p
\end{aligned}
$$

$X = (1-a)^n$

$Y = (1-a)^{n-p}a$

Q2 (Continued)

(b)

Consider the correlation between A and B.

B\A	1	0
1	2	0
0	1	1

$X_{AB}^2 = 1.33$

Consider the correlation between A and C.

C\A	1	0
1	1	1
0	2	0

$X_{AC}^2 = 1.33$

Consider the correlation between B and C.

C\B	1	0
1	0	2
0	2	0

$X_{BC}^2 = 4$

For attribute A,

$X_{AB}^2 + X_{AC}^2 = 1.33 + 1.33 = 2.66$

For attribute B,

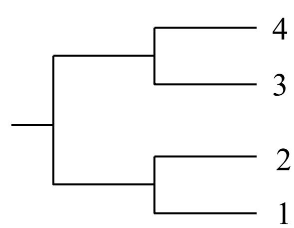
$X_{AB}^2 + X_{BC}^2 = 1.33 + 4 = 5.33$

For attribute C,

$X_{AC}^2 + X_{BC}^2 = 1.33 + 4 = 5.33$

We choose attribute B for splitting since it has the largest value.
We divide the data into two groups, namely {1, 2} and {3, 4}.

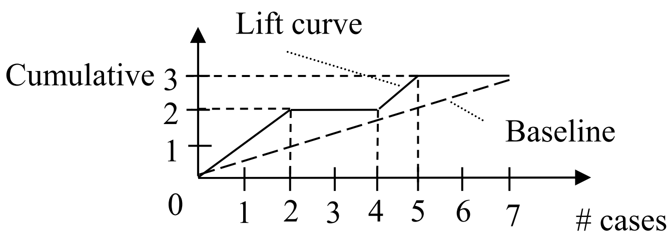
Dendrogram:



Q3 (20 Marks)
(a)(i)

	Predicted Class	
Actual Class	Yes	No
Yes	2	1
No	2	2

(ii)



Q3 (Continued)
(b)

No. This is because we do not know the distance between cluster (a, b) and cluster (c d) and the distance between (a b) and e.

Q4 (20 Marks)

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

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

$= 0.467$

$$P(LC=Yes \mid FH=Yes, Smoker=No, PR=Yes)$$
$$=\frac{P(PR=Yes \mid FH=Yes, Smoker=No, LC=Yes)}{P(PR=Yes \mid FH=Yes, Smoker=No)} P(LC=Yes \mid FH=Yes, Smoker=No)$$
$$=\frac{P(PR=Yes \mid LC=Yes)\times P(LC=Yes \mid FH=Yes, Smoker=No)}{\sum_{x\in\{Yes,No\}} P(PR=Yes \mid LC=x)P(LC=x \mid FH=Yes, Smoker=No)}$$
$$=\frac{0.85\times 0.45}{0.85\times 0.45+0.45\times 0.55}$$
$$=0.607143$$
$$P(LC=No \mid FH=Yes, Smoker=No, PR=Yes)=1-0.601743=0.392857$$

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

(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) Yes.

specificity = 3/4
= 0.75 (or 75%)

(b) Yes.

precision = 3/4
= 0.75 (or 75%)

(c) Yes.

recall = 3/4
= 0.75 (or 75%)

(d) Yes.

f-measure = 2 x Precision x Recall /(Precision + Recall)
= 2 x 0.75 x 0.75 /(0.75 + 0.75)
= 0.75 (or 75%)

Q6 (20 Marks)

(a)

$P(X, Y \mid Z)$

$= \dfrac{P(X,Y,Z)}{P(Z)}$

$= \dfrac{P(X,Y,Z)}{P(Y,Z)} \times \dfrac{P(Y,Z)}{P(Z)}$

$= P(X|Y, Z) \times P(Y|Z)$

$= P(X|Z) \times P(Y|Z)$

(b)

The curse of dimensionality can be described as follows.
When the number of dimensions increases, the distance between any two points is nearly the same.

(c)

Iteration 1:

$(x_1, x_2, y) = (0, 0, 0)$

$net = x_1 w_1 + x_2 w_2 + b$

$= 0 * 0.1 + 0 * 0.1 + 0.1 = 0.1$

b	w_1	w_2
0.1	0.1	0.1

y = 1 $\boxed{\text{Incorrect!}}$

$w_1 = w_1 + \alpha(d - y)x_1$
$= 0.1 + 0.5*(0 - 1) * 0$
$= 0.1$

$w_2 = w_2 + \alpha(d - y)x_2$
$= 0.1 + 0.5*(0 - 1) * 0$
$= 0.1$

$b = b + \alpha(d - y)$
$= 0.1 + 0.5*(0 - 1)$
$= -0.4$

Q6 (Continued)

Iteration 2:

$(x_1, x_2, y) = (0, 1, 0)$

$net = x_1 w_1 + x_2 w_2 + b = 0.3$

y = 0 $\boxed{\text{Correct!}}$

b	w_1	w_2
-0.4	0.1	0.1

$w_1 = w_1 + \alpha(d - y)x_1$
$= 0.1 + 0.5*(0 - 0) * 0$
$= 0.1$
$w_2 = w_2 + \alpha(d - y)x_2$
$= 0.1 + 0.5*(0 - 0) * 1$
$= 0.1$
$b = b + \alpha(d - y)$
$= -0.4 + 0.5*(0 - 0)$
$= -0.4$

Iteration 3:

$(x_1, x_2, y) = (1, 0, 0)$

$net = x_1 w_1 + x_2 w_2 + b = -0.3$

y = 0 $\boxed{\text{Correct!}}$

b	w_1	w_2
-0.4	0.1	0.1

$w_1 = w_1 + \alpha(d - y)x_1$
$= 0.1 + 0.5*(0 - 0) * 1$
$= 0.1$
$w_2 = w_2 + \alpha(d - y)x_2$
$= 0.1 + 0.5*(0 - 0) * 0$
$= 0.1$
$b = b + \alpha(d - y)$
$= -0.4 + 0.5*(0 - 0)$
$= -0.4$

Iteration 4:

$(x_1, x_2, y) = (1, 1, 1)$

$net = x_1 w_1 + x_2 w_2 + b = -0.2$

y = 0 $\boxed{\text{Incorrect!}}$

b	w_1	w_2
-0.4	0.1	0.1

$w_1 = w_1 + \alpha(d - y)x_1$
$= 0.1 + 0.5*(1 - 0) * 1$
$= 0.6$
$w_2 = w_2 + \alpha(d - y)x_2$
$= 0.1 + 0.5*(1 - 0) * 1$
$= 0.6$
$b = b + \alpha(d - y)$
$= -0.4 + 0.5*(1 - 0)$
$= 0.1$

Q6 (Continued)

Iteration 5:

$(x_1, x_2, y) = (0, 0, 0)$

$net = x_1 w_1 + x_2 w_2 + b = 0.1$

y = 1 $\boxed{\text{Incorrect!}}$

b	w_1	w_2
0.1	0.6	0.6

$w_1 = w_1 + \alpha(d - y)x_1$
$= 0.6 + 0.5*(0 - 1) * 0$
$= 0.6$

$w_2 = w_2 + \alpha(d - y)x_2$
$= 0.6 + 0.5*(0 - 1) * 0$
$= 0.6$

$b = b + \alpha(d - y)$
$= 0.1 + 0.5*(0 - 1)$
$= -0.4$

b	w_1	w_2
-0.4	0.6	0.6

Q6 (Continued)

Q7 (20 Marks)
(a)

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

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 = $\begin{pmatrix} 9-7 \\ 5-7 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$

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

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

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

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

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

when $\lambda = 4$,

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

Q7 (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:

Q7 (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:

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

Q7 (Continued)

Q7 (Continued)

Q7 (Continued)

(b)

(5, 5) is reduced to (0);
(7, 7) is reduced to (0);
(4, 8) is reduced to (-2.83);
(8, 4) is reduced to (2.83).

(Note: Another possible answer is
(5, 5) is reduced to (0);
(7, 7) is reduced to (0);
(4, 8) is reduced to (2.83);
(8, 4) is reduced to (-2.83).)

(c)

(18, 18) is reduced to (0);
(24, 24) is reduced to (0);
(15, 27) is reduced to (-8.49);
(27, 15) is reduced to (8.49).

(Note: Another possible answer is
(18, 18) is reduced to (0);
(24, 24) is reduced to (0);
(15, 27) is reduced to (8.49);
(27, 15) is reduced to (-8.49).)

Q8 (20 Marks)

(a)
For a shorter query time (or for performing data analysis).

(b)
The greedy algorithm discussed in class can be modified be changing the heuristics function from the computation of the benefit of a view to the computation of the benefit of a view per “unit space”.
i.e.
Let C(v) be the cost of view v(the number of rows in v)
Algorithm:
$S \leftarrow \{top\ view\}$;
$X \leftarrow X - C(v)$ where v is the top view ;
While there exists a view v not in S $s.t.\ C(v) \leq X$
  Select the view v not in S $s.t.$
    $C(v) \leq X$
    $B(v, S)/C(v)$ is maximized
  $S \leftarrow S \cup \{v\}$
  $X \leftarrow X - C(v)$
output S.

Q9 (20 Marks)
(a)

No.
We know that the whole dataset can be split into two clusters, {a, b} and {c, d, e}.
Consider cluster {c, d, e}.
We do not know the hierarchy for points c, d, and e.
We need two kinds of additional information, D({c}, {e}) and D({d}, {e}) to draw the dendrogram.

(b)

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

Q10 (20 Marks)

(a)

Yes.

$$\begin{array}{cc} & \begin{array}{ccc} x & y & z \end{array} \\ \begin{array}{c} x \\ y \\ z \end{array} & \begin{pmatrix} 0 & 1 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0 \end{pmatrix} \end{array}$$

(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)
(a)

Yes.

We can regard "Gain(V ∪ {top view}, {top view})" as the cost reduction of materializing views in V compared with that of materializing the top view only.

Similarly, we can regard "Gain(V ∪ {top view, view A}, {top view, view A})" as the cost reduction of materializing views in V compared with that of materializing the top view and the view A only.

Since materializing view A reduces the cost of accessing views which could be affected by the views in V, we know that Gain(V ∪ {top view}, {top view}) ≥ Gain(V ∪ {top view, view A}, {top view, view A}).

Q11 (Continued)
(b)

No.

Consider the following example.
Let P = "view b" and C = "view c".

We know that

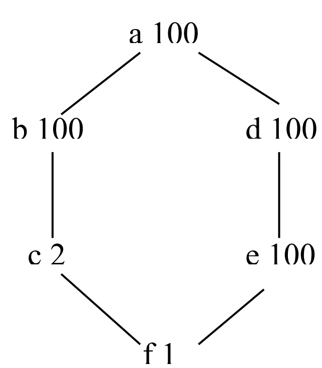
Gain({view P} ∪ {top view}, {top view}) = 0

and

Gain({view C} ∪ {top view}, {top view}) = 92.

Thus,

Gain({view P} ∪ {top view}, {top view}) < Gain({view C} ∪ {top view}, {top view})



Q11 (Continued)
(c)

Yes.

We can regard "Gain({x} ∪ S ∪ {top view}, {top view}) - Gain(S ∪ {top view}, {top view})" as the cost reduction of materializing view x compared with that of materializing the top view and the views in S only.

Similarly, we can regard "Gain({x} ∪ T ∪ {top view}, {top view}) - Gain(T ∪ {top view}, {top view}) " as the cost reduction of materializing view x compared with that of materializing the top view and the views in T only.

Since S ⊆ T, materializing views in S reduces the cost of accessing views which could be affected by the views in T, we know that

Gain({x} ∪ S ∪ {top view}, {top view}) - Gain(S ∪ {top view}, {top view})
≥ Gain({x} ∪ T ∪ {top view}, {top view}) - Gain(T ∪ {top view}, {top view})

End of Answer Sheet