COMP1942 Exploring and Visualizing Data (Spring Semester 2021)
Online Final Examination (Question Paper)
Date: 24 May, 2021 (Monday)
Time: 4:30pm-6:30pm
Duration: 2 hours

Instructions:

(1) Guideline
- (a) Please follow **all** instructions about the exam guideline (e.g., your face video capturing) stated in the Canvas website.
- (b) For the sake of space, we do not write them again.

(2) Question
- (a) Please answer **all** questions. The total scores in this exam are 200.

(3) Answer Sheet
- (a) Please submit your answers in PDF to the Canvas website.
- (b) Please use the cover page stated in the Canvas website as the **first** page of your PDF file. This cover page includes your information and an agreement with your signature.
- (c) Please start to write your answers starting on the **second** page of your PDF file.
- (d) The PDF file should "clearly" show your answers without any blurred images. No marks will be given to any "blurred" parts in the PDF file. Please make sure that the PDF file shows your answers clearly.

(4) Online Exam
- (a) This is an online exam where you could access all online materials.
  However, it is **not** allowed to communicate with other people (except the instructor and the tutors in this course) in any form (including but not limited to orally, electronically and in writing) during the entire exam period together with the pre-15-minute period and the post-15-minute period.

(5) File Submission
- (a) We allow a 15-minute buffer for your PDF file upload. Remember to upload your file at around 6:30pm. We allow your file uploading time at most 15 minutes. Canvas will terminate any file uploading process at 6:45pm if your file is still being uploaded at 6:45pm.

(6) Zero-Score Regulation
- (a) If your face could not be shown in your video for at least 10 seconds in the exam period, your exam score will be set to 0 (even though you submit your PDF file in Canvas).
- (b) If you do not submit the first cover page, your exam score will be set to 0.
- (c) We only mark your latest PDF file uploaded by 6:45pm. Your exam score will be set to 0 if we could not see any PDF file uploaded by 6:45pm (even though you do the question paper or you "could" upload your PDF file after 6:45pm).

Q1 (20 Marks) (Version A)

Consider eight data points.

x_1: (1, 2), x_2: (9, 6), x_3: (2, 3), x_4: (0, 4), x_5: (12, 5), x_6: (4, 2), x_7: (7, 8), x_8: (10, 14)

Note: All numbers in this question should be expressed up to 4 decimal places.

(a) Please write down the pairwise distance matrix based on the above 8 data points (using the Euclidean distance).
(b) Please use the divisive (polythetic) approach to divide these eight points into two groups/clusters by using the group average linkage. You are required to show all steps.

Q1 (20 Marks) (Version B)

Consider eight data points.

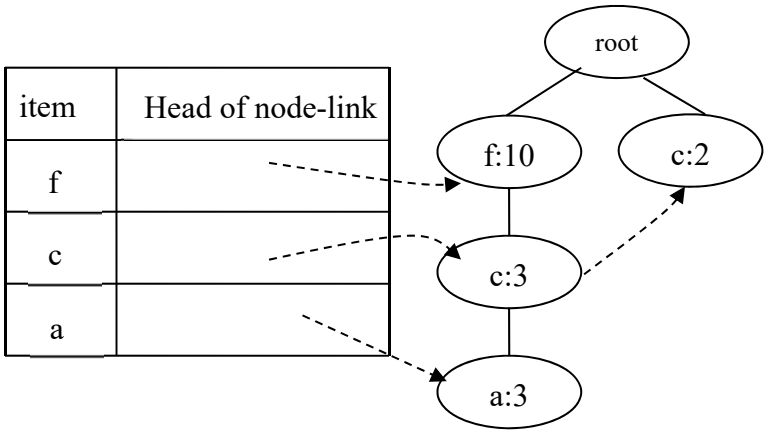
x_1: (22, 15), x_2: (14, 12), x_3: (17, 18), x_4: (20, 24), x_5: (11, 12), x_6: (19, 16), x_7: (12, 13), x_8: (10, 14)

Note: All numbers in this question should be expressed up to 4 decimal places.

(a) Please write down the pairwise distance matrix based on the above 8 data points (using the Euclidean distance).
(b) Please use the divisive (polythetic) approach to divide these eight points into two groups/clusters by using the group average linkage. You are required to show all steps.

Q2 (20 Marks)

(a) Consider a dataset containing items a, c and f. Let the support threshold = 1. We constructed the following FP-tree.



(i) Is it possible to construct back all transactions according to the above FP-tree? If yes, please write down all transactions. In this case, you are allowed to write down each transaction with a number where the number denotes the total number of occurrences of this transaction. Otherwise, please explain it.

(ii) Please write down all frequent itemsets generated according to the FP-growth algorithm. You are not required to show any steps. Note that the support of each itemset should be given too.

(b) Suppose that we are given a dataset with some transactions in binary format, and the support threshold = 2. Finally, we obtain the set X of all frequent itemsets equal to

{ {Q}, {R}, {S}, {T},
{Q, T}, {Q, R}, {Q, S}, {R, T}, {S, T},
{Q, R, T}, {Q, S, T} }

There are many possible datasets which have the same set X as the set of all frequent itemsets.
Please give one possible dataset which has the minimum number of transactions in binary format.
Assume that each transaction in this dataset contains items P, Q, R, S or T.

(c) In the Apriori algorithm, we know how to find some sets L_1, C_2, L_2, ….

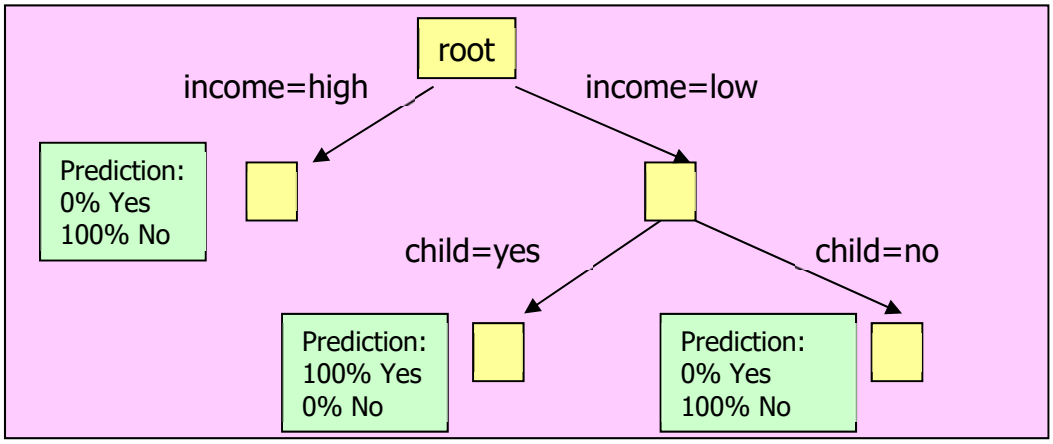
(i) Is it always true that the number of itemsets in L_2 is smaller than or equal to the number of itemsets in C_2? If yes, please explain it. Otherwise, please give a counter example.

(ii) Is it always true that the number of itemsets in C_2 is larger than or equal to the number of itemsets in L_1? If yes, please explain it. Otherwise, please give a counter example.

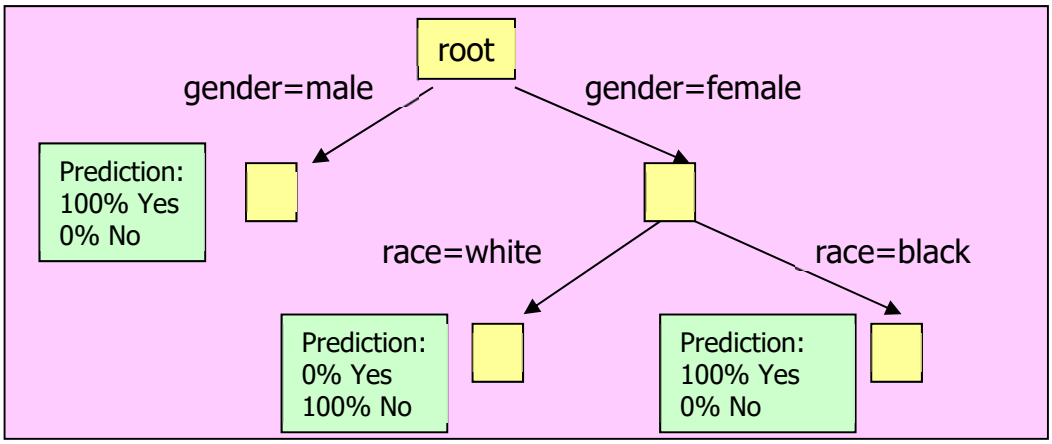
Q3 (20 Marks)

(a) The insurance company is given a table with five input attributes, namely Race, Gender, Married, Income and Child, and one target attribute, namely Insurance. Based on this table, the insurance company constructed three classifiers based on different criteria, namely Classifier 1, Classifier 2 and Classifier 3.

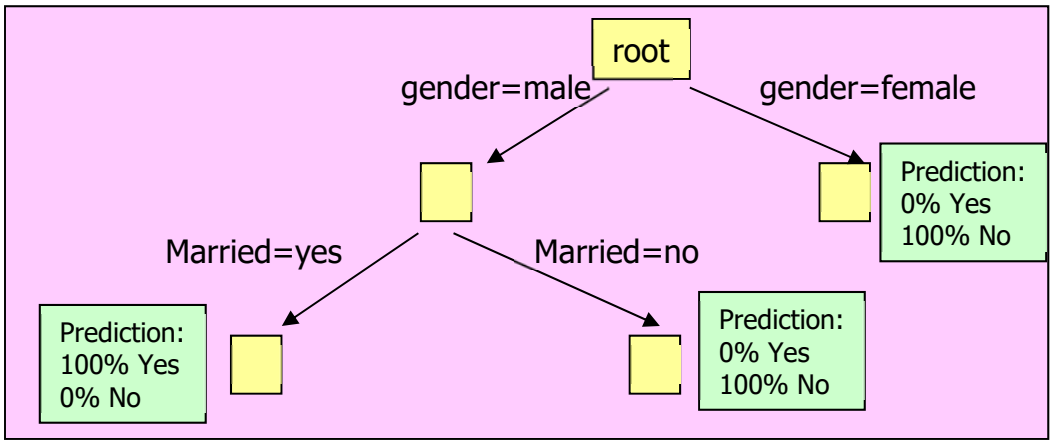
Classifier 1



Classifier 2



Classifier 3



Consider a group of 3 classifiers called an "ensemble" studied in class. Consider a new customer. All input attribute values of this new customer are known to the insurance company. The company uses this "ensemble" to do the prediction and predicts that this new customer will buy an insurance policy. Suppose that we are very "curious" about the input attribute values of this new customer. What we know about the new customer is female and the "predicted" result is that this new customer will buy an insurance policy. We also know the 3 exact classifiers used in the insurance company. Is it possible for us to find the values of some input attribute values of this customer? If yes, please state (1) all these input attribute values and (2) all input attribute(s) that could not be found with their values. Otherwise, please write down the reason why we could not find those values.

(b) We are given the following records.

Record ID	Input Attribute 1	Input Attribute 2	Target Attribute
1	4	3	No
2	3	2	Yes
3	4	9	No
4	9	1	Yes
5	10	4	Yes
6	6	2	No
7	8	3	Yes
8	1	9	Yes
9	2	10	Yes
10	4	8	No
11	6	7	No
12	8	6	Yes
13	7	5	No
14	5	6	No

We want to predict the target attribute of the new record with the input attribute 1 equal to 7 and the input attribute 2 equal to 7. Suppose that we want to use a 3-nearest neighbor classifier and we adopt the Euclidean distance as a distance measurement between two given points. What is the target attribute of this record? Please write down the target attribute of this record and the record IDs of the corresponding 3 nearest neighbors.

(c) We know how to compute the information gain of an attribute A under the ID3 decision tree, denoted by InfoGain-ID3(A). We also know how to compute the information gain of an attribute A under the CART decision tree, denoted by InfoGain-CART(A). Consider two attributes A and B. Is it always true that if InfoGain-CART(A) > InfoGain-CART(B), then InfoGain-ID3(A) > InfoGain-ID3(B)? If yes, please show that it is true. Otherwise, please give a counter example (containing at most 8 records) showing that this is not true and then explain it.

Q4 (20 Marks)

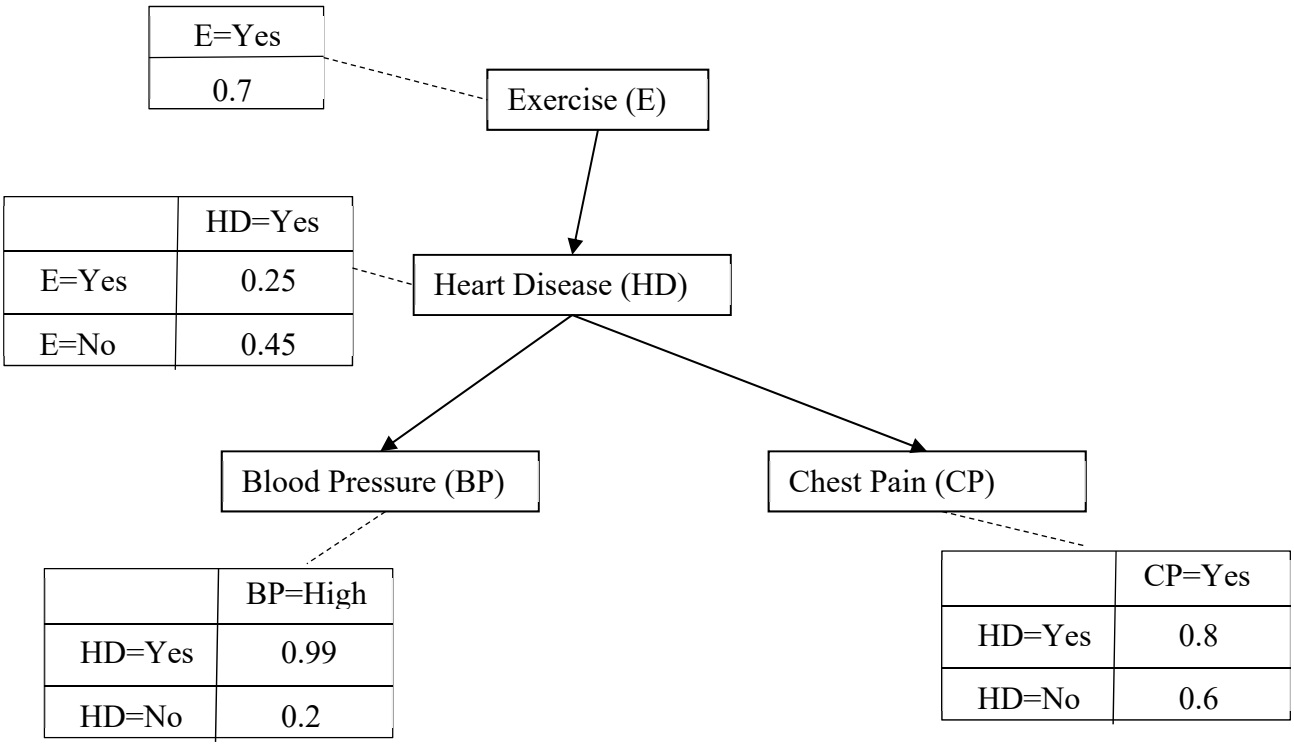
Suppose that we know the following confusion matrix for a classifier.

	Predicted Class	
Actual Class	Yes	No
Yes	7	6
No	3	14

(a) Is it possible to find the number of false positives and the number of false negatives? If yes, please write down the number of false positives and the number of false negatives. Otherwise, please explain it.
(b) Is it possible to find the error report? If yes, please write down the error report. Otherwise, please explain it.
(c) Is it possible to find the lift chart? If yes, please write down the life chart where we regard "Yes" as a success. Otherwise, please explain it.
(d) Is it possible to find the decile-wise lift chart? If yes, please write down the decile-wise lift chart where we regard "Yes" as a success. Otherwise, please explain it.
(e) Is it possible to find the specificity? If yes, please write down the specificity. Otherwise, please explain it.

Q5 (20 Marks)

We have the following Bayesian Belief Network.



Suppose that there is a new person. We know that

(1) he does exercises
(2) he has high blood pressure
(3) he has chest pain

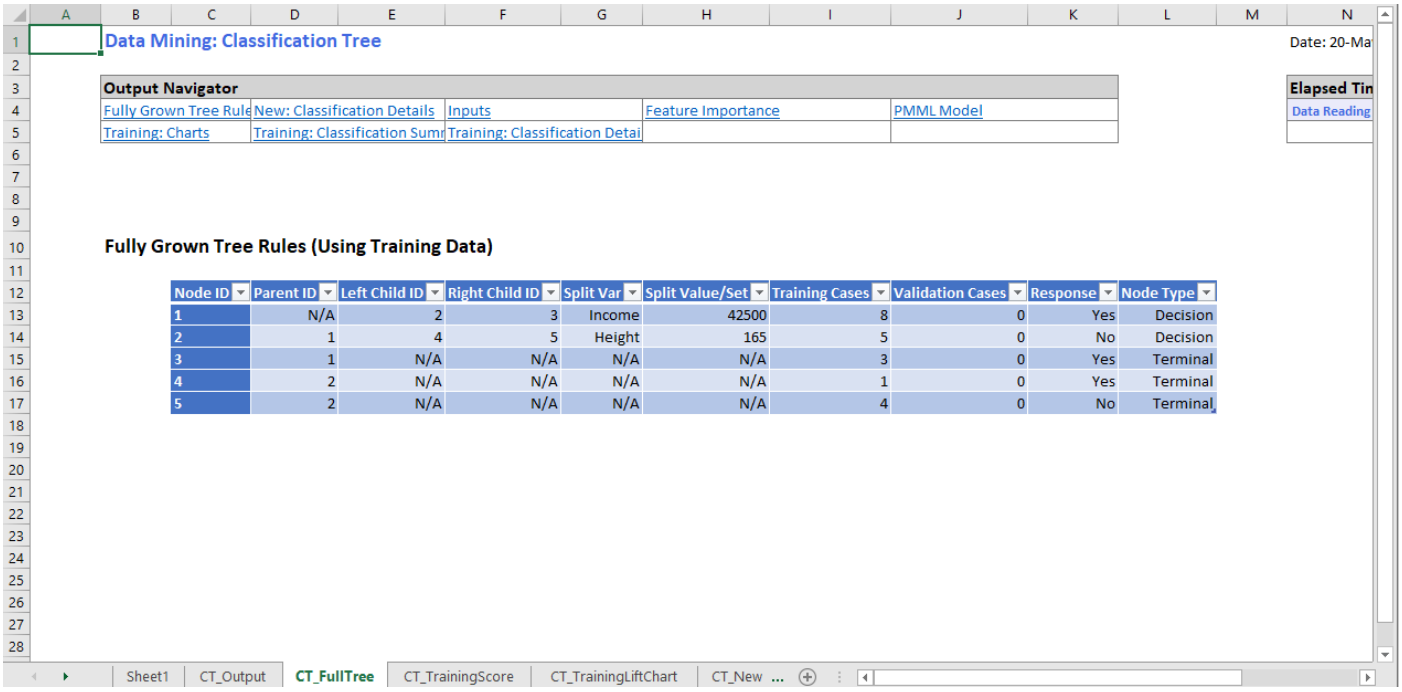
We would like to know whether he is likely to have Heart Disease.

Exercise	Blood Pressure	Chest Pain	Heart Disease
Yes	High	Yes	?

Please use Bayesian classifier with the use of Bayesian Belief Network to predict whether he is likely to have Heart Disease. You are required to show all steps.

Q6 (20 Marks)

(a) We obtained the following output from XLMiner for the decision tree.



Data Mining: Classification Tree

Date: 20-Ma

Output Navigator				
Fully Grown Tree Rule	New: Classification Details	Inputs	Feature Importance	PMML Model
Training: Charts	Training: Classification Sum	Training: Classification Detai		

Elapsed Tin

Data Reading

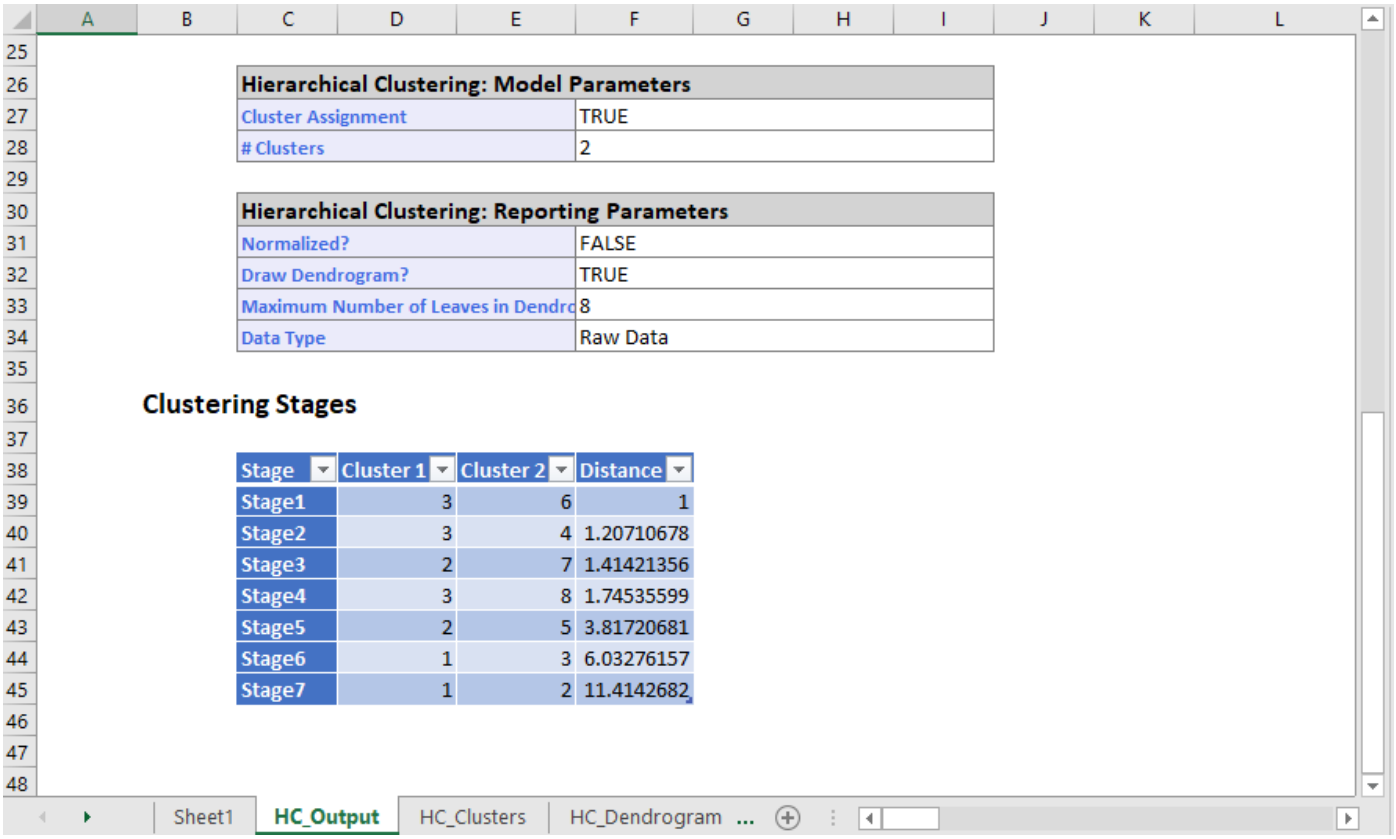
Fully Grown Tree Rules (Using Training Data)

Node ID	Parent ID	Left Child ID	Right Child ID	Split Var	Split Value/Set	Training Cases	Validation Cases	Response	Node Type
1	N/A	2	3	Income	42500	8	0	Yes	Decision
2	1	4	5	Height	165	5	0	No	Decision
3	1	N/A	N/A	N/A	N/A	3	0	Yes	Terminal
4	2	N/A	N/A	N/A	N/A	1	0	Yes	Terminal
5	2	N/A	N/A	N/A	N/A	4	0	No	Terminal

Sheet1 | CT_Output | CT_FullTree | CT_TrainingScore | CT_TrainingLiftChart | CT_New ...

Is it possible that we could draw the corresponding decision tree? If yes, please draw the decision tree which is consistent with the lecture notes format except that for each terminal node, you just need to write down the expected predicted value on the target attribute without writing the percentage of the values in the target attribute. Otherwise, please state all kinds of additional information so that we could draw the corresponding decision tree.

(b) We obtained the following output from XLMiner for clustering.



Hierarchical Clustering: Model Parameters	
Cluster Assignment	TRUE
# Clusters	2

Hierarchical Clustering: Reporting Parameters	
Normalized?	FALSE
Draw Dendrogram?	TRUE
Maximum Number of Leaves in Dendrc	8
Data Type	Raw Data

Clustering Stages

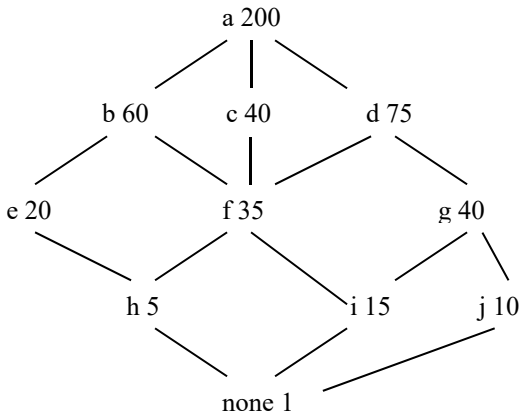
Stage	Cluster 1	Cluster 2	Distance
Stage1	3	6	1
Stage2	3	4	1.20710678
Stage3	2	7	1.41421356
Stage4	3	8	1.74535599
Stage5	2	5	3.81720681
Stage6	1	3	6.03276157
Stage7	1	2	11.4142682

Sheet1 | HC_Output | HC_Clusters | HC_Dendrogram ...

Is it possible that we could draw the corresponding dendrogram? If yes, please draw the dendrogram which is consistent with the lecture notes format (which has the distance axis placed horizontally). In this case, please indicate all (distance) numbers each of which corresponds to the merging operation between two sub-clusters in the dendrogram. Otherwise, please state all kinds of additional information so that we could draw the corresponding dendrogram.

Q7 (20 Marks) (Version A)

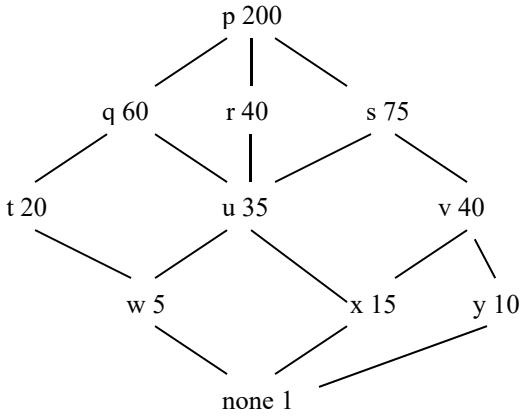
We are given the following lattice containing nodes where each node corresponds to a possible query. The number associated with each query corresponds to the cost of answering this query. Note that each view is placed at an appropriate level to indicate that one view could be derived from another view in an upper level.



Assume that we do not consider "none 1". Suppose 3 views are to be materialized (other than the top view). Apply the greedy algorithm and find the resulting views. You are required to show all steps.

Q7 (20 Marks) (Version B)

We are given the following lattice containing nodes where each node corresponds to a possible query. The number associated with each query corresponds to the cost of answering this query. Note that each view is placed at an appropriate level to indicate that one view could be derived from another view in an upper level.



Assume that we do not consider "none 1". Suppose 3 views are to be materialized (other than the top view). Apply the greedy algorithm and find the resulting views. You are required to show all steps.

Q8 (20 Marks)

The following shows a table where the input attributes are x_1 and x_2 and y is the target attribute. There are two possible values in the target attribute y, namely "Yes" and "No". Attribute ID corresponds to the ID of each tuple.

ID	x_1	x_2	y
p	24	26	No
q	22	24	No
r	26	26	No
s	24	22	No
t	0	0	Yes
u	6	-4	Yes
v	8	2	Yes
w	2	6	Yes

Consider a classification task by SVM.

(a) In order to perform the classification task by SVM, we need to transform the above table T to another table T' such that the target attribute y in the transformed table contains numeric values instead of categorical values. What is this transformed table T'?

(b) Formulate this SVM problem by a quadratic programming. Please list the objective function and all constraints in this quadratic programming.

(c) In the above quadratic programming, the objective function to be minimized is not exactly equal to the margin we want to maximize in our initial plan. Why do we need to write another objective function to be minimized instead of using the original margin?

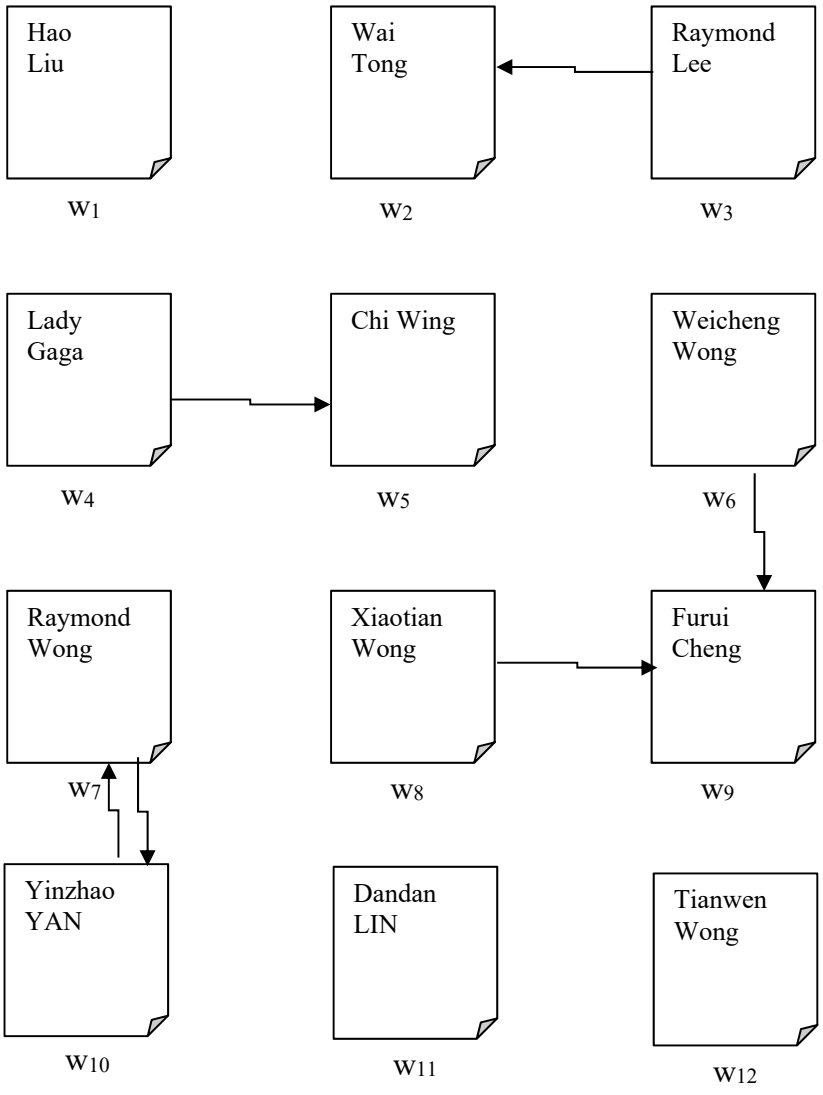
Q9 (20 Marks)

We are given the following adjacency matrix according to five sites, namely a, b, c, d and e.

$$\begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 \end{pmatrix} \end{array}$$

(a) Is it possible to find the corresponding stochastic matrix? If yes, write down the stochastic matrix. Otherwise, please explain it.

(b) We are given the following 12 webpages, namely w_1, w_2, …, w_{12}.



The query terms typed by the user are "Raymond" and "Wong".

(i) What is the root set in this query? Please list the webpages in this set.

(ii) What is the base set in this query? Please list the webpage in this set.

Q10 (20 Marks) (Version A)

Consider the four 2-dimensional data points:

a:(6, 7), b:(8, 9), c:(5, 10) and d:(8, 5)

We can make use of PCA for dimensionality reduction. In dimensionality reduction, given an L-dimensional data point, we want to transform this point to a K-dimensional data point where K < L such that the information loss during the transformation is minimized.
Please illustrate with the above example when L=2 and K=1. Also, shows your results in a diagram. You are required to show all steps. All numbers should be shown up to 4 decimal places.

Q10 (20 Marks) (Version B)

Consider the four 2-dimensional data points:

a:(12, 14), b:(16, 18), c:(10, 20) and d:(16, 10)

We can make use of PCA for dimensionality reduction. In dimensionality reduction, given an L-dimensional data point, we want to transform this point to a K-dimensional data point where K < L such that the information loss during the transformation is minimized.
Please illustrate with the above example when L=2 and K=1. Also, shows your results in a diagram. You are required to show all steps. All numbers should be shown up to 4 decimal places.

End of Paper