COMP5331 Knowledge Discovery in Databases (Fall Semester 2019)
Homework 2
Deadline: 4 Nov, 2019 10:30am
(Please hand in during lecture.)
Full Mark: 100 Marks

Coupon Instructions:

1. You can use a coupon to waive any question you want and obtain full marks for this question.
2. You can waive at most one question in each assignment.
3. You can also answer the question you will waive. We will also mark it but will give full marks to this question.
4. The coupon is non-transferrable. That is, the coupon with a unique ID can be used only by the student who obtained it in class.
5. Please staple the coupon to the submitted assignment.
6. Please write down the question no. you want to waive on the coupon.

Q1 [20 Marks]

Consider the density-based subspace clustering. The size of a subspace is defined to be the total number of dimensions for this subspace. For example, subspace {A, B} is of size 2. For each single dimension, the number of grid units is fixed to a constant c where c is a positive integer greater than 1.

(a) In class, we learnt that the major idea in the KL-transform is to transform the original coordinate system to a new coordinate system such that we could find clusters in subspaces from the new coordinate system. Suppose that we use the KL-transform to transform all data points from the original coordinate system to a new coordinate system (without using Step 6 of the KL-transform (i.e., choosing a subset of attribute values)). Then, all points are now represented in the new coordinate system. Based on the new coordinate system, we adopt the density-based subspace clustering to find clusters in some subspaces. Is it always true that the total number of grid units involved in all clusters based on the new coordinate system is smaller than that based on the original coordinate system? If yes, please give some justifications without any formal proof. If no, similarly, please give some counter examples for illustration.

(b) When the size of the subspace is larger, it is less likely that a grid unit with respect to the subspace is dense. Please explain it.

(c) In order to overcome the weakness described in (b), instead of setting a fixed density threshold for the subspace of any size, we use a smaller density threshold for the subspace of larger size. Specifically, let T_i be the density threshold for the subspace of size i. If $i < j$, then $T_i > T_j$. Let Condition 1 be "$T_i > T_j$ for any $i < j$".
Let Condition 2 be "for any i and j, $T_i = T_j$". We know that if Condition 2 is satisfied, then the original Apriori-like algorithm studied in class can find all subspaces containing dense units.
   (i) Under Condition 1, is it always true that we can still adopt the Apriori-like algorithm? If yes, please describe how to adopt the algorithm. Otherwise, please give reasons why it cannot be adopted.
   (ii) Suppose that we modify Condition 1 to the following form. Let Condition 1 be "$T_i = cT_{i+1}$ for each positive integer i". Assume that we adopt this new form of Condition 1. Under this new form of Condition 1, is it always true that we can still adopt the Apriori-like algorithm? If yes, please describe how to adopt the algorithm. Otherwise, please give reasons why it cannot be adopted.

Q2 [20 Marks]

(a) Consider a set P containing the following four 2-dimensional data points.

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

We can make use of the KL-Transform to find a transformed subspace containing a cluster. Let L be the total number of dimensions in the original space and K be the total number of dimensions in the projected subspace. Please illustrate the KL-transform technique with the above example when L=2 and K=1.

(b) Consider a set Q containing the following four 2-dimensional data points.

e:(5, 5), f:(9, 9), g:(3, 11), h:(11, 3)

(i) Let $p = (x_p, y_p)$ be a point in P and $q = (x_q, y_q)$ be a point in Q. In fact, we could express x_q in a linear form involving x_p such that $x_q = \alpha \cdot x_p + \beta$ where α and β are 2 real numbers. Similarly, we could express y_q in the same linear form involving y_p. Please write down the values of α and β.

(ii) Similar to Part (a), we want to make use of the KL-Transform to find a transformed subspace containing a cluster for the set Q where L = 2 and K = 1. One "straightforward" or "naïve" method is to use the same method in Part (a) to obtain the answer. Is it possible to make use of the result in Part (a) and the result in Part (b)(i) to obtain the answer very quickly? If yes, please explain briefly and give the answer. There is no need to give a formal proof. A brief description it accepted. If no, please give an explanation briefly. In this case, derive the answer by using the method in Part (a).

(c) Consider Part (a). It is independent of Part (b). In Part (a), we know that there are 4 points. Suppose that we have 4 additional points which are identical to the original 4 points. That is, we have the following 4 additional points. Totally, we have 8 data points.

(6, 6), (8, 8), (5, 9), (9, 5)

One "straightforward" or "naïve" method is to use the same method in Part (a) to obtain the answer. Is it possible to make use of the result in Part (a) to obtain the answer very quickly? If yes, please explain briefly and give the answer. There is no need to give a formal proof. A brief description it accepted. If no, please give an explanation briefly. In this case, derive the answer by using the method in Part (a).

(d) Consider two random variables X and Y with the following probabilistic table.

X \ Y	1	2	3
1	0	1/8	1/8
2	1/2	0	1/8
3	1/8	0	0

(i) Calculate the conditional entropy of H(X|Y) by using the original definition of the conditional entropy.

(ii) Calculate H(X|Y) as

$$- \sum_{x \in A} \sum_{y \in B} p(x, y) \log p(x|y)$$

where A = {1, 2, 3} and B = {1, 2, 3}.

Q3 [20 Marks]

The following shows a history of PhD students with their numbers of published papers, their ages and their majors. We also indicate whether they become professors or not after their PhD graduation in the last column. Note that the first column "No." is for us to refer the record number only.

No.	NoOfPapers	Age	Major	Become_Professor
1	enough	young	ComputerScience	yes
2	many	young	ComputerScience	yes
3	many	old	CivilEngineering	yes
4	many	old	DataScience	yes
5	few	young	CivilEngineering	no
6	many	young	ComputerScience	no
7	few	old	DataScience	no
8	few	old	ComputerScience	no

(a) We want to train a C4.5 decision tree classifier to predict whether a PhD student will become a professor or not. We define the value of attribute Become_Professor to be the *label* of a record.
   (i) Please find a C4.5 decision tree according to the above example. In the decision tree, whenever we process (1) a node containing at least 80% records with the same label or (2) a node containing at most 2 records, we stop to process this node for splitting.
   (ii) Consider a young PhD student majoring in computer science who published many papers. Please estimate the probability that this PhD student will become a professor.

(b) Let X be the set of attributes involved in the decision tree found in Part (a). Person A said that we just need to consider all attributes in X *only* to determine whether a student will become a professor. Person B said that we should also consider attributes outside X (in addition to attributes in X) to determine whether a student will become a professor.
   (i) Please give a possible reason why Person A said in this way.
   (ii) Please give a possible reason why Person B said in this way.
   (iii) Which Person (Person A or Person B) is more unreasonable in general?

(c) What is the difference between the C4.5 decision tree and the ID3 decision tree? Why is there a difference?

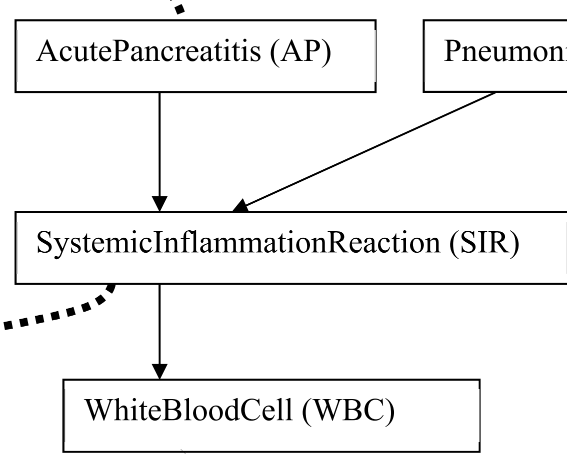
Q4 [20 Marks]

We have the following Bayesian Belief Network.

AP = Yes
0.3

P = Yes
0.6

	SIR = Yes
AP = Yes P = Yes	0.7
AP = Yes P = No	0.45
AP = No P = Yes	0.55
AP = No P = No	0.2



	WBC = High
SIR = Yes	0.6
SIR = No	0.3

Suppose that there is a new patient. We know that

(1) he has acute pancreatitis
(2) he has pneumonia
(3) his result of white blood cell is low

We would like to know whether he is likely to have systemic inflammation reaction.

Acute Pancreatitis	Pneumonia	White Blood Cell	Systemic Inflammation Reaction
Yes	Yes	Low	?

(a) Please use Bayesian Belief Network classifier with the use of Bayesian Belief Network to predict whether he is likely to have systemic inflammation reaction.

(b) Although Bayesian Belief Network classifier does not have an independent assumption among all attributes (compared with the naïve Bayesian classifier), what are the disadvantages of this classifier?

Q5 [20 Marks]

We are given two data points with 2 different timestamps.
At the timestamp t = 1, we have a data point (x_1, x_2, y) where (x_1, x_2) = (0.3, 0.6) and y = 0.2.
At the timestamp t = 2, we have a data point (x_1, x_2, y) where (x_1, x_2) = (0.1, 1.0) and y = 0.4.
Here, x_1 and x_2 are 2 input variables. y is the output variable.

(a) Consider the traditional LSTM model. Initially, we have the following internal weight vectors and bias variables as follows.

$$W_f = \begin{pmatrix} 0.8 \\ 0.4 \\ 0.1 \end{pmatrix} \quad b_f = 0.2$$

$$W_i = \begin{pmatrix} 0.9 \\ 0.8 \\ 0.7 \end{pmatrix} \quad b_i = 0.5$$

$$W_a = \begin{pmatrix} 0.4 \\ 0.2 \\ 0.1 \end{pmatrix} \quad b_a = 0.3$$

$$W_o = \begin{pmatrix} 0.6 \\ 0.4 \\ 0.1 \end{pmatrix} \quad b_o = 0.2$$

In the model, we have the following status variables. For each t = 1, 2, ….

1. forget gate variable f_t
2. input gate variable i_t
3. input activation variable a_t
4. internal state variable s_t
5. output gate variable o_t
6. final output variable y_t

Suppose that $y_0 = 0$ and $s_0 = 0$.

Consider the input forward propagation step only.

(i) What are the values of the above status variables when t = 1 and when t = 2? Please show each answer up to 4 decimal places.

(ii) What are the errors of the final output variables when t = 1 and when t = 2? Please show each answer up to 4 decimal places.

(b) Consider the GRU model. Initially, we have the following internal weight vectors and bias variables as follows.

$$W_r = \begin{pmatrix} 0.3 \\ 0.2 \\ 0.1 \end{pmatrix} \qquad b_r = 0.5$$

$$W_a = \begin{pmatrix} 0.4 \\ 0.3 \\ 0.1 \end{pmatrix} \qquad b_a = 0.1$$

$$W_u = \begin{pmatrix} 0.4 \\ 0.2 \\ 0.1 \end{pmatrix} \qquad b_u = 0.1$$

In the model, we have the following status variables. For each t = 1, 2, ….

1. reset gate variable r_t
2. input activation variable a_t
3. update gate variable u_t
4. final output variable y_t

Suppose that $y_0 = 0$.

Consider the input forward propagation step only.

(i) What are the values of the above status variables when t = 1 and when t = 2? Please show each answer up to 4 decimal places.

(ii) What are the errors of the final output variables when t = 1 and when t = 2? Please show each answer up to 4 decimal places.

(c) What is the major disadvantage of the traditional neural network model compared with the recurrent neural network model?