CSIT5210 Data Mining and Knowledge Discovery (Fall Semester 2021)
Homework 2
Deadline: 2 Nov, 2021 3pm
(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.

(a) 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.

(b) In order to overcome the weakness described in (a), 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, can we 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 "$T_i = \alpha T_{i+1}$ for each positive integer i" where α is a positive real number at least 1 and is given by users. Assume that we adopt this new form of Condition 1. If we set α to some values, we cannot adopt the Apriori-like algorithm. If we set α to other values, we can adopt the Apriori-like algorithm. What is the greatest possible value of α such that we can adopt the Apriori-like algorithm? Please explain it.

(c) In order to overcome the weakness described in (a), we use the same number of grid units for any subspace, says N, and thus the density threshold is set to a fixed value T for any subspace. For example, if N = 4, the total number of grid units with respect to a subspace of size 2 is 4 (Figure 1a) and the total number of grid units with respect to a subspace of size 1 is also 4 (Figure 1b). Let n be the total number of dimensions. Suppose we set $N = 2^{n!}$. Same as before, we want to find all subspaces containing dense units under Condition 2. Can we still adopt the Apriori-like algorithm? If yes, please describe how to adopt the algorithm. Otherwise, please give reasons why it cannot be adopted.

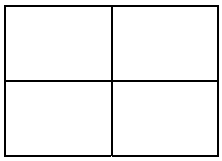


Figure 1a

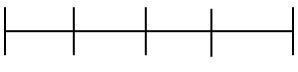


Figure 1b

Q2 [20 Marks]

Consider the entropy-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.

(a) Is the following statement true? If yes, please give a formal proof. If no, please give a counter example.

"When the size of the subspace is larger,
it is less likely or equally likely that the subspace has a good clustering."

(b) Suppose that c is a positive real number where we do not know the exact value.
Similarly, d is also another positive real number where d is equal to c+5.

(i) Consider the four 2-dimensional data points:

a:(7 + c, 7 + c), b:(9 + c, 9 + c), c:(6 + c, 10 + c), d:(10 + c, 6 + c)

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. Suppose that L = 2 and K = 1. Please illustrate with the above example.

(ii) Consider the four 2-dimensional data points:

a:(7 – d, 7 – d), b:(9 – d, 9 – d), c:(6 – d, 10 – d), d:(10 – d, 6 – d)

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. Suppose that L = 2 and K = 1.
Can we make use of the answers in part (b)(i) to perform the KL-Transform? If yes, please write down each transformed data point. If no, please write down the reasons why we cannot make use of the answers of part (b)(i).

(iii) Consider the four 2-dimensional data points:

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

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. Suppose that L = 2 and K = 1.
Can we make use of the answers in part (b)(i) to perform the KL-Transform? If yes, please write down each transformed data point. If no, please write down the reasons why we cannot make use of the answers of part (b)(i).

Q3 [20 Marks]

The following shows a history of customers with their incomes, an attribute called "Have_iPhone" and another attribute called "Have_iPad" . "Have_iPhone" denotes whether a customer has an iPhone and "Have_iPad" denotes whether a customer has an iPad. We also denote whether they will buy an iPadMini or not in the last column.

No.	Income	Have_iPhone	Have_iPad	Buy_iPadMini
1	high	yes	yes	yes
2	high	no	yes	yes
3	medium	yes	no	yes
4	high	no	no	yes
5	medium	yes	no	no
6	medium	yes	no	no
7	medium	no	no	no
8	medium	no	no	no

(a) We want to train a CART decision tree classifier to predict whether a new customer will buy an iPadMini or not. We define the value of attribute Buy_iPadMini is the *label* of a record.

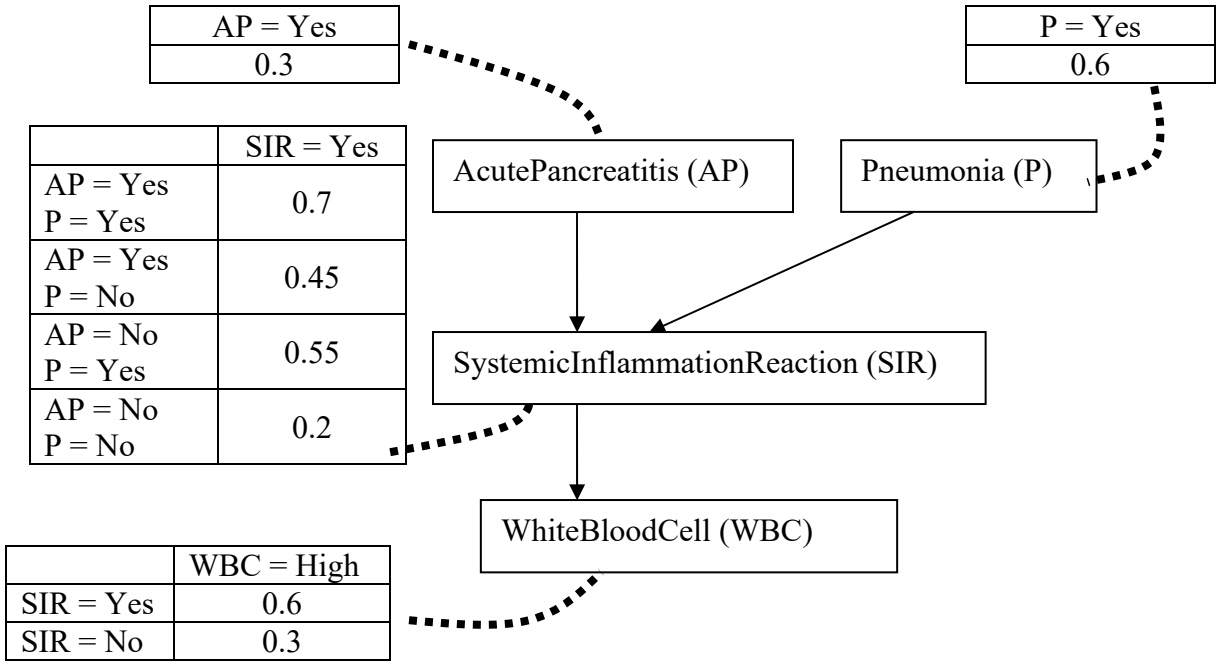
  (i) Please find a CART decision tree according to the above example. In the decision tree, whenever we process a node containing at most 3 records, we stop to process this node for splitting.

  (ii) Consider a new customer whose income is medium and he has both an iPhone and iPad. Please predict whether this new customer will buy an iPadMini or not.

(b) 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.



	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.

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

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

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

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?