UIUC-CS412 “An Introduction to Data Warehousing and Data Mining” (Spring 2017)

Final Exam

- There are 13 problems total worth 100 points as shown in each question.
- You must not communicate with other students during this test.
- No books, notes allowed.
- No other electronic device except calculators are allowed. You cannot use your mobile as calculators.
- This is a 180 minute exam.
- Do not turn this page until instructed to.
- If you can’t finish an answer within the given space, please use the blank pages at the end of the exam. Please put the corresponding question number while using blank pages.

1. Fill in your information:

Full Name: ______________________________

NetID: ______________________________

1/1. (5 points)

The following example shows a candidate generation step in SPADE algorithm. Given in the tables below is the sequence-id and element-id of candidate sequences of length 2(<ab>, <ba>). Complete the table for candidate sequences, <bab> and <aba> given the following information.

	ab			ba	
SID	EID (a)	EID(b)	SID	EID (b)	EID(a)
1	1	2	1	2	3
2	1	3	2	3	4
3	2	5			
4	3	5			

Solution.

SID	EID(a)	EID(b)	EID(a)
1	1	2	3
2	1	3	4

SID	EID(b)	EID(a)	EID(b)

2/1. (5 points) Below is the contingency table that summarizes a supermarket transaction data of milk and cereal sell.

	$Cereal$	$\overline{Cereal}$	$\sum_{col}$
$Milk$	700	120	820
$\overline{Milk}$	170	10	180
$\sum_{row}$	870	130	1000

(a) (3 points) Calculate two measures of correlation: *Lift* and *Kulczynski*.
$Hint: Lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}$, $Kulc(A, B) = \frac{1}{2}(P(A|B) + P(B|A)$.

(b) (1 point) What kind of correlation exists between milk and cereal sell? Is there a positive, negative or no correlation?

(c) (1 point) What is the main difference between the *Lift* and *Kulczynski* measures?

Solution.

(a)

$$Lift = (700/1000)/((820/1000) \times (870/1000)) = 0.9812,$$
$$Kulczynski = 0.5 \times ((700/820) + (700/870)) = 0.8291.$$

(b) Positive.

(c) *Kulczynski* is null-invariant.

3/1. (5 points) An education expert conducts an interview for predicting education level. She uses attributes such as “income-level”, “gender” and “age-group” for predicting the education level of people. However, while training the dataset she observes that some people have not shared their attribute. Give two ways how she could still use the survey dataset for training a decision tree classifier.

Solution. ignore missing cases, most common, most probable, probability, random, etc.

4/1. (10 points)

Sequence-id	Sequence
1	<a(ac)d(cf)a(abc)>
2	<a(bc)(ae)(ab)>
3	<ca(ab)(df)cba(bc)>
4	<(ac)bc(ab)aab>
5	<a(cd)bc(ab)(ab)>

A. Given the above sequence database, identify all sequences that may have a suffix <**ab**>.

B. For the above database, identify all the sequences that may have a prefix <**ac**>.

C. For the above database, what would be the <**a**>-projected database.

Solution.

A. <(ac)bc(ab)aab>.

B. <a(cd)bc(ab)(ab)>.

C.

Sequence-id	Sequence
1	<(ac)d(cf)a(abc)>
2	<(bc)(ae)(ab)>
4	<(-c)bc(ab)aab>
5	<(cd)bc(ab)(ab)>

5/1. (5 points) Below is a transaction dataset and item prices of each transaction. Assume each kind of item has one distinct positive price. Use the constrained FP-growth algorithm to find the FP tree that satisfy the requirements $min\_sup = 2$ and $sum\{S.price\} \geq 10$.

TID	Items (Price)
T1	1, 3, 4
T2	2, 3, 4, 5
T3	1, 3, 5
T4	2, 4, 5

Solution.

TID	Items (Price)
T2	2, 3 (remove), 4, 5
T4	2, 4, 5

FP-tree: single branch of 5-4-2.

6/1. (15 points) Consider the construction of a decision tree from the following training dataset. The training examples comprises predicting whether a person will buy-computer (yes) or not (no) given two categorical attributes - student and education level.

Student	Education	Buy computer
yes	high	yes
yes	low	no
no	high	no
no	high	yes
no	middle	yes
yes	low	no
yes	high	yes
no	low	yes
no	middle	no
yes	high	yes

(a) (5 points) Calculate the Information Gain obtained by splitting the decision tree by age and education respectively. Which of the two attributes is ideal for splitting?
$Hint: Info(D) = -\sum_i p_i log_2(p_i), \quad Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j), \quad Gain(A) = Info(D) - Info_A(D)$.

(b) (5 points) Calculate the Gain Ratio obtained by splitting the decision tree by age and education respectively. Which of the two attributes is ideal for splitting?
$Hint: SplitInfo(A) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} log_2(\frac{|D_j|}{|D|}), \quad GainRatio(A) = Gain(A)/SplitInfo(A)$.

(c) (5 points) Calculate the Gini Index obtained by splitting the decision tree by age and education respectively. Which of the two attributes is ideal for splitting?
$Hint: gini(D) = 1 - \sum_i p_i^2, \quad gini_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} gini(D_j), \quad \Delta gini(A) = gini(D) - gini_A(D)$.

Solution.

(a)

$$\begin{aligned}
Info(D) &= -(0.6 * log_2(0.6) + 0.4 * log_2(0.4)) = 0.97 \\
Info_{Student}(D) &= 0.5 * (-(0.6 * log_2(0.6) + 0.4 * log_2(0.4))) + 0.5 * (-(0.6 * log_2(0.6) + 0.4 * log_2(0.4))) \\
&= 0.5 * 0.97 + 0.5 * 0.97 = 0.97 \\
Info_{Education}(D) &= 0.5 * (-(0.8 * log_2(0.8) + 0.2 * log_2(0.2))) + 0.3 * (-(0.33 * log_2(0.33) + 0.67 * log_2(0.67))) \\
&+ 0.2 * (-(0.5 * log_2(0.5) + 0.5 * log_2(0.5))) \\
&= 0.5 * 0.72 + 0.3 * 0.91 + 0.2 * 1 = 0.83 \\
Gain(Student) &= 0.97 - 0.97 = 0 \\
Gain(Education) &= 0.97 - 0.83 = 0.14
\end{aligned}$$

So we should choose the attribute Education for splitting.

(b)

$$\begin{aligned}
SplitInfo_{Student}(D) &= -(0.5 * log_2(0.5) + 0.5 * log_2(0.5)) = 1 \\
SplitInfo_{Education}(D) &= -(0.5 * log_2(0.5) + 0.3 * log_2(0.3) + 0.2 * log_2(0.2)) = 1.49 \\
GainRatio(Student) &= 0/1 = 0 \\
GainRatio(Education) &= 0.14/1.49 = 0.09
\end{aligned}$$

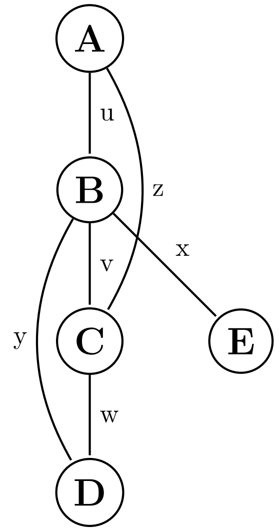
So we should choose the attribute Education for splitting.

(c)

$$\begin{aligned}
gini(D) &= 1 - (0.6^2 + 0.4^2) = 0.48 \\
gini_{Student}(D) &= 0.5 * (1 - (0.6^2 + 0.4^2)) + 0.5 * (1 - (0.6^2 + 0.4^2)) \\
&= 0.24 + 0.24 = 0.48 \\
gini_{Education}(D) &= 0.5 * (1 - (0.8^2 + 0.2^2)) + 0.3 * (1 - (0.33^2 + 0.67^2)) + 0.2 * (1 - (0.5^2 + 0.5^2)) \\
&= 0.16 + 0.13 + 0.1 = 0.39 \\
\Delta gini(Student) &= gini(D) - gini_{Student}(D) \\
&= 0.48 - 0.48 = 0 \\
\Delta gini(Education) &= gini(D) - gini_{Education}(D) \\
&= 0.48 - 0.39 = 0.09
\end{aligned}$$

So we should choose the attribute Education for splitting.

7/1. (5 points) In the graph shown above, there are five nodes (A, B, C, D, E) and six edges (u, v, w, x, y, z). Show five different DFS search trees rooted at node C and their right most paths respectively.



Solution.

- C-D-B-A-E, w-y-x.
- C-D-B-E-A, w-y-u.
- C-B-A-D-E, v-x.
- C-B-A-E-D, v-y.
- C-B-D-A-E, v-x.
- C-B-D-E-A, v-u.
- C-B-E-A-D, v-y.
- C-B-E-D-A, v-u.
- C-A-B-D-E, z-u-x.
- C-A-B-E-D, z-u-y.

8/1. (5 points) Below is the confusion matrix of a cancer prediction model.

Actual class \ Predicted Class	cancer=yes	cancer=no	Total
cancer=yes	90	210	300
cancer=no	140	9560	9700
Total	230	9770	10000

Calculate accuracy, precision and recall of this model. Is this a good model? Justify your answer with a few sentences of explanation.
Hint:

$$\begin{aligned} Accuracy &= \frac{TP + TN}{N} \\ Precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \end{aligned}$$

Solution.

$$\begin{aligned} Accuracy &= \frac{90 + 9560}{10000} = 0.965 \\ Precision &= \frac{90}{90 + 140} = 0.39 \\ Recall &= \frac{90}{90 + 210} = 0.3 \end{aligned}$$

Not good model. Low precision and low recall.

9/1. (15 points)

A. [4] People say that if each classifier is better than random guess, ensemble of multiple such classifiers will lead to a nontrivial increase of classification accuracy. Do you agree with this statement? Give reasoning on it.

B. [5] In what scenario, will you prefer bagging over boosting as an ensemble method?

C. [6] What are the major differences among the three methods for the evaluation of the accuracy of a classifier : (1) **hold-out method**, (2) **cross-validation**, and (3) **boot- strap**?

Solution.

A. Agree. One reasoning could be if the err rate of each classifier r is less than 0.5. For 2K classifiers, the accumulative error rate for k classifier will be r^k, which will be much smaller than a single r.

   Another reasoning could be: Since the classifiers are training using different subsets or weights on the data, they are less likely to be trapped by the same noise or inaccuracy, and thus the majority voting are unlikely to be wrong. Moreover, by giving less weights to the classifiers that tend to make mistakes, the performance can be further improved.

B. We will prefer bagging when there is a chance of boosting overfitting to the noisy tuples.

C. (a) hold-out method: use part of the data (e.g. $\frac{2}{3}$) for training and the remaining for testing.

   (b) cross-validation: partition data into (relatively even) k portions, D1, ..., Dk, use Di for testing and the other k 1 portions for training for any i, and then merge the results.

   (c) bootstrap: works for small data set, it samples the given training data uniformly with replacement, e.g., 0.632 bootstrap.

10/1. (5 points) Suppose we have 4 training points as listed in the following table.

x	y	z
1	0	+1
-1	0	+1
0	1	-1
0	-1	-1

Each point has 2 attributes i.e x and y, and a label z. Apply **Adaboost Algorithm** in this question. Suppose initially the weight for each point is uniform i.e $w_1(j) = \frac{1}{4}$ for $j = 1, .., 4$
In each round, we sample with replacement according to the weights, and get a training dataset $D_i, i = 1, 2$. Suppose, based on D_i, we learn a classifier $M_i, i = 1, 2$ which has the following rule:

$$M_1 : \hat{z} = \begin{cases} +1, & \text{if } x \geq 0.5 \\ -1, & \text{if } x < 0.5 \end{cases}$$

$$M_2 : \hat{z} = \begin{cases} +1, & \text{if } y \geq -0.5 \\ -1, & \text{if } y < -0.5 \end{cases}$$

Please answer:

A. Compute the (weighted) error ϵ_i of model M_i(i = 1,2) and weights for each point after 1st and 2nd round i.e w_2 and w_3

[Hint : Correctly classified tuples are weighted by $\frac{\epsilon_i}{1-\epsilon_i}$]

B. Combine the 2 classifiers based on your calculation in the above step. Give the ensemble classifier generated as a result. [Hint: Classifiers are weighted by log $\frac{1-\epsilon_i}{\epsilon_i}$]

Solution.

x	y	z	w_1	$M_1 : \hat{z}$	w_2	$normw_2$
1	0	+1	$\frac{1}{4}$	+1	$\frac{1}{12}$	$\frac{1}{6}$
-1	0	+1	$\frac{1}{4}$	-1	$\frac{1}{4}$	$\frac{1}{2}$
0	1	-1	$\frac{1}{4}$	-1	$\frac{1}{12}$	$\frac{1}{6}$
0	-1	-1	$\frac{1}{4}$	-1	$\frac{1}{12}$	$\frac{1}{6}$

Error $\epsilon_1 = \frac{1}{4}$, $\frac{\epsilon_1}{1-\epsilon_1} = \frac{1}{3}$

x	y	z	w_2	$M_2 : \hat{z}$	w_3	$normw_3$
1	0	+1	$\frac{1}{6}$	+1	$\frac{1}{30}$	$\frac{1}{10}$
-1	0	+1	$\frac{1}{2}$	+1	$\frac{1}{10}$	$\frac{3}{10}$
0	1	-1	$\frac{1}{6}$	+1	$\frac{1}{6}$	$\frac{1}{2}$
0	-1	-1	$\frac{1}{6}$	-1	$\frac{1}{30}$	$\frac{1}{10}$

Error $\epsilon_2 = \frac{1}{6}$, $\frac{\epsilon_1}{1-\epsilon_1} = \frac{1}{5}$

Combined classifier $\log(3)M_1 + \log(5)M_2$

11/1. (5 points) Suppose that you are given a sequential pattern mining interface(black box) that you can call on a given transaction dataset(Input) to generate the frequent **sequential** patterns(Output)

Devise an algorithm that will use the provided interface to generate frequent itemsets from database and also give an approximate support value for such itemsets. Assume the min_sup = 2
Keep in mind you cannot make any change inside the interface.

Solution. The key difference between a frequent itemset and corresponding sequential patterns is the ordering of items. Given a module that can compute sequential patterns, we can combine the resultant sequential patterns to derive frequent itemsets and approximate their support. Notice that, if a sequential pattern is frequent, its corresponding frequent itemset is also frequent. However, the support threshold for mining sequential patterns should be much lower than the support threshold for frequent itemset. This is because the support of a frequent itemset gets divided across multiple sequential patterns, and therefore the support is lower for the corresponding sequential patterns. To merge the sequential patterns corresponding to a frequent itemset, we can use a hash function that returns the same value for any combination of the items in a set.

12/1. (10 points) Consider the construction of a Naive Bayes classifier from the following training dataset. The training examples comprises predicting whether a person has a flu (Yes) or not (No) given four categorical attributes—chills (Yes or No), runny nose (Yes or No), fever (Yes or No) and Headache (No, Mild or Strong) .

Chills	Runny Nose	Fever	Headache	Flu
Y	N	Y	Mild	N
Y	Y	N	No	Y
Y	N	Y	Strong	Y
N	Y	Y	Mild	Y
N	N	N	No	N
N	Y	Y	Strong	Y
N	Y	N	Strong	N
Y	Y	Y	Mild	Y

A. (4 points) Given a patient with the following attributes, what is the probability of the patient having flu.

Chills	Runny Nose	Fever	Headache
Y	N	N	Mild

B. (4 points) Given a patient with the above attributes, what is the probability of the patient **not** having flu.

C. (2 point) Based on the above calculations, what will be you predict about the patient's flu i.e. whether the person has or does not have flu.

Solution.

A. Solve, $P \propto P(Chills = Y|Flu = Y) \times P(RunnyNose = N|Flu = Y) \times P(Fever = N|Flu = Y) \times P(Headache = Mild|Flu = Y) \times P(Flu = Y)$

B. Solve, $P \propto P(Chills = Y|Flu = N) \times P(RunnyNose = N|Flu = N) \times P(Fever = N|Flu = N) \times P(Headache = Mild|Flu = N) \times P(Flu = N)$

C. Whichever of the above two is greater.

13/1. (10 points) For the database given below of 4 transactions, you are to use **CLOSET** to mine the frequent closed patterns. Here min_sup = 2.

customer_id	shopping items
1	abcde
2	bde
3	aef
4	bcde

A. [2] Find the support of all items in the given dataset. Show the output as a sorted list in descending order of support.

B. [4] Suppose your output from the previous step was a list like $x_1, x_2, .., x_m$ where x_i is the i^{th} item.Assuming there are total m items in the transaction database.

Now, Generate the various conditional databases required by CLOSET system.

TDB|x_m, TDB|x_{m-1} but not x_m, ... so on

C. [4] Find closed patterns in the given database from outputs in the previous step.

Solution. Closed Patterns : cbde:2, ae:2, bde:3, e:4