CSCI 5523, Fall 2022 — Homework #3
Due Date: Oct. 26th, 11:59 PM — Submission: Gradescope

Instructions:

- The main PDF submission of HW 3 must be uploaded on **Gradescope**. Your submission will be graded via the **Gradescope** platform.

- Your PDF file should be named as "hw3_lastname_firstname". For example, if your name is Jane Doe and you are submitting HW 3, then name your file as "hw3_doe_jane".

- On the first page of your assignment, please write your name (First Name, Last Name), as it appears on Canvas. Please also include your UMN email, and Student ID.

- **When submitting your PDF file on Gradescope, you are required to match the questions to the correct page(s) of your solutions. If you do not know how to do it, you should watch this short video through the following link: https://www.gradescope.com/get_started#student-submission**

- Type out your solutions on a separate blank file (using Word, Latex, etc.), and don't forget to convert the file to PDF before submitting. **HANDWRITTEN HOMEWORKS WILL NOT BE ACCEPTED.**

- In the HW Questions section, there are two subsets of questions:
  (1) A subset of questions that is graded for correctness. This subset accounts for 75% of the total HW grade. In other words, if there is a question in this subset whose answer is completely wrong, then you would **not** receive any credit.
  (2) A subset of questions that is graded for simply answering them, regardless of your answers being correct or not. This subset accounts for 25% of the total HW grade. In other words, if there is a question in this subset whose answer is completely wrong, then you would receive full credit for attempting it.
  Note that these subsets are not determined a priori, therefore please do your best to answer all the questions correctly.

- Each homework will also include Practice Questions. Please do not submit neither the questions nor the answers for these Practice Questions. The purpose of these questions is to help you gain a deeper understanding of the course concepts and prepare for the midterm exams.

- **Reminder**: all homework, hand-on projects, and exams, must represent your individual effort. Group projects must represent each group members' efforts. Copying from online sources, another's work, or allowing (even negligently) others to copy your work, or possession of electronic computing devices in the testing area, is cheating and grounds for penalties in accordance with the Academic Conduct Policies for Students in Computer Science & Engineering Department Classes (https://www.cs.umn.edu/sites/cs.umn.edu/files/cse-department-academicconductpolicy.pdf).

HW Questions

Question 1.

Answer True or False and briefly explain.
(a) ANN is able to handle redundant attributes.

(b) Both SVM and ANN handle very large training data sets equally well.

(c) SVM and Neural Network always produce the same decision boundary for a given data set with two classes.

Question 2.

You are trying to evaluate four different COVID-19 tests, T1, T2, T3, and T4. These tests have been developed by different organizations, and their evaluations by these organizations have been reported in the following evaluation measures: TPR and FPR as follows

T1: TPR = 0.5 and FPR = 0.01,
T2: TPR = 0.99 and FPR = 0.1,
T3: TPR = 0.99 and FPR = 0.01,
T4: TPR = 0.9 and FPR = 0.05.

Note that T1, T2 and T3 correspond to the classifiers T1, T2 and T3 introduced in the lecture note "chap4_imbalanced_classes.pdf" on pages 25-27

a) Calculate the number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) using the classifier T4 for the following cases:

1. Balanced skew case in which we have 100 positives and 100 negatives (Note: Confusion matrices of T1, T2 and T3 for this case are provided on slide 25)

2. Medium skew case in which we have 100 positives and 1000 negatives (Note: Confusion matrices of T1, T2 and T3 for this case are provided on slide 26)

3. High skew case in which we have 100 positives and 10000 negatives (Note: Confusion matrices of T1, T2 and T3 for this case are provided on slide 27)

b) Compute precision, recall, F-measure, TPR/FPR of T4 for each of the aforementioned three cases

c) Which classifier is strictly better than T4 (irrespective of skew)? Briefly explain why.

d) Which classifier is strictly worse than T4 (irrespective of skew)? Briefly explain why.

e) Which classifiers may be better or worse than T4 (depending on skew)? Briefly explain why.

Question 3.

Consider the market basket transactions shown in the table below:

Transaction ID	Item Bought
1	{Milk,Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies,Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer,Diapers}
9	{Milk,Diaper,Bread, Butter}
10	{Bread,Beer, Cookies}

a. What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

b. What is the maximum size of frequent itemsets that can be extracted (assuming minsup > 0)?

c. Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.

d. What are the support counts for {Bread}, {Milk}, and {Bread, Milk}?

e. What are the confidence of the rules {Bread} -> {Milk} and {Milk}->{Bread}?

f. Find a pair of items, a and b, such that the rules {a} −> {b} and {b} −> {a} have the same confidence.

Question 4.

Suppose ACD is a frequent itemset and AB is NOT a frequent itemset. Given this information, we can be sure that certain other itemsets are frequent and sure that certain itemsets are NOT frequent. Other itemsets may be either frequent or not. Which of the following is a correct classification of an itemset?
Give a one sentence explanation if you believe any statement is incorrect.

a) A is frequent.

b) CD can be either frequent or not frequent.

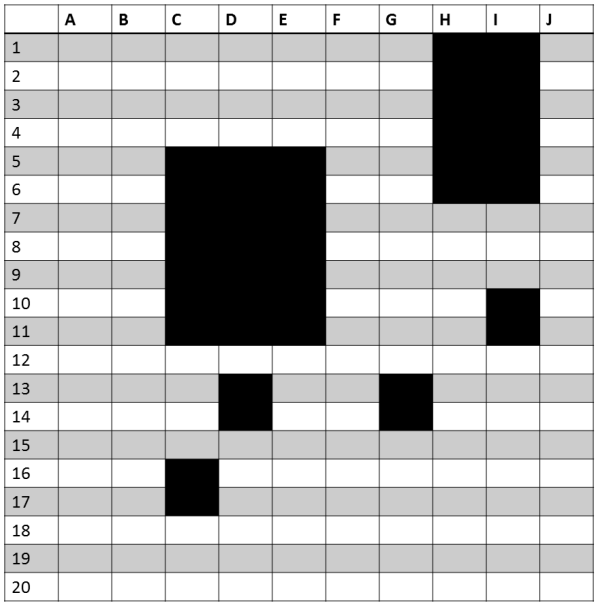
c) ACDE can be either frequent or not frequent.

d) ABCD is frequent.

e) ABCDE can be either frequent or not frequent.

Question 5:

The figure below depicts a transaction matrix with 10 items and 20 transactions. Dark cells indicate the presence of items, and white (or grey) cells indicate the absence of items. We apply the Apriori algorithm to extract frequent itemsets with minsup=20% (i.e., itemsets must be contained in at least 4 transactions). Answer the following questions:



a. List all the maximal frequent itemsets in the dataset.

b. List all the frequent itemsets in the dataset.

c. List all the closed frequent itemsets in the dataset.

Question 6.

Consider the following frequent 3-itemsets:

{a, b, c}, {p, b, c}, {p, a, b}, {p, a, c}, {p, a, w}

The book presents two algorithms for generating candidate 4-itemsets, the Fk−1 x F1 method and Fk−1 x Fk−1 method.

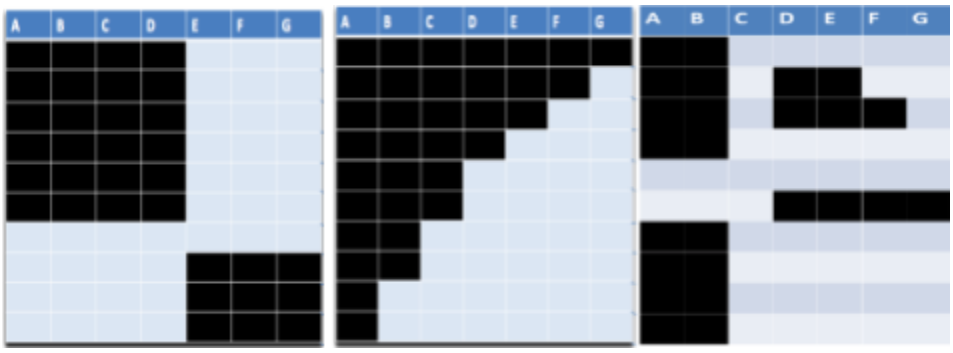
a. List all the 4-itemsets that will be generated by the Fk−1 x F1 candidate generation method and the 4-itemsets that will be selected after the pruning step of the Apriori algorithm.

b. List all the 4-itemsets that will be generated by the Fk−1 x Fk−1 candidate generation method and the 4-itemsets that will be selected after the pruning step of the Apriori algorithm.

c. Based on the list of candidate 4-itemsets generated above, is it possible to generate a frequent 5-itemset? State your reason clearly.

Question 7.

Consider the three datasets below that contain 7 items and 1000 transactions. Each row represents 100 transactions. Dark cells indicate the presence of items and white (and grey) cells indicate the absence of items. We will apply the Apriori algorithm to extract frequent itemsets with minsup = 50% (i.e., itemsets must contain at least 500 transactions).



Dataset 1 Dataset 2 Dataset 3

a) What is the number of frequent itemsets for each dataset? Which dataset will produce the greatest number of frequent itemsets?

b) Which dataset will produce the longest frequent itemset?

c) Which dataset will produce frequent itemsets with highest maximum support?

d) Which dataset will produce frequent itemsets containing items with widely varying support levels (i.e., itemsets containing items with mixed support, ranging from 20% to more than 70%)?

e) What is the number of maximal frequent itemsets for each dataset? Which dataset will produce the greatest number of maximal frequent itemsets?

f) What is the number of closed frequent itemsets for each dataset? Which dataset will produce the greatest number of closed frequent itemsets

Question 8.

Consider a dataset with 10 items (A,B,C,D,E,F,G,H,I,J) and 100 transactions. You are given partial information about the support count of some itemsets as follows:

{A}: support count = 80
{A,B}: support count = 80
{D,E}: support count = 40
{D,E,F}: support count = 30
{D,E,F,G}: support count = 25
{D,E,F,G,H}: support count = 25
{A,B,C,D,E,F,G,H,I,J}: support count = 15
All immediate supersets of {D,E,F} has support counts less than or equal to 27

The information about other support counts are unknown. Based on this, specify whether the following statements are
(a) true, (b) false, or (c) cannot decide based on the given partial information. If you choose (a) or (b), then provide a brief explanation. If you choose (c), give one example when the statement is correct, and another example when the statement is wrong.

1. {A} is a closed itemset

2. {B} is a closed itemset

3. {D,E} is a closed itemset

4. {D,E,F} is a closed itemset

5. {D,E,F,G} is a closed itemset

6. {A,B,C,D,E,F,G,H,I,J} is a closed itemset

Practice Questions:

Question 1

You are asked to evaluate the performance of two classification models, M1 and M2, for a binary classification problem with classes '+' and '–'. For every test instance, x, each of the two models provides a posterior probability of x belonging to class '+'. The table below provides a list of 10 test instances with their true classes, and their posterior probabilities of belonging to class '+', according to M1 and M2. Assume that we are mostly interested in detecting instances from the positive class.

Instance	True Class	P(+|M1)	P(+|M2)
1	+	0.94	0.27
2	+	0.31	0.45
3	+	0.76	0.95
4	+	0.31	0.46
5	+	0.82	0.23
6	-	0.33	0.13
7	-	0.47	0.08
8	-	0.46	0.19
9	-	0.24	0.37
10	-	0.45	0.04

Table 1

a) Plot the ROC curve for both M1 and M2.
(You should plot them on the same graph.)
Which model do you think is better? Explain your reasons.

b) Suppose you choose a cutoff threshold to be t = 0.4 for both the models, M1 and M2. In other words, any test instance whose posterior probability is greater than t will be classified as a positive example. Compute the Precision, Recall, and F-Measure for M1 and M2 after using the cutoff threshold of t. Which model is better using F-measure as the evaluation criterion? Are the results consistent with what you expect from the ROC curve?

c) Repeat part (b) using t = 0.7. Which model is better using F-measure as the evaluation criterion? Are the results consistent with what you expect from the ROC curve?

Question 2

a) Suppose you are given a data set consisting of nominal attributes, such as color, which takes values such as red, blue, green, etc. Can you use this data set directly to train an SVM? If not, how will you transform these attributes into a representation that can be used to train an SVM?

b) List one key similarity and one key difference between support vector machines and artificial neural networks.

c) Answer True or False and briefly explain:
SVM and Neural Network always produce the same decision boundary for a given data set with two classes.

Question 3

Consider a test data of 1000 samples with two classes: + class (100 samples) and – class (900 samples). We have two random classifiers C1 and C2. Classifier C1 classifies test data to + class randomly with a probability p and classifier C2 classifies test data to + class randomly with a probability 2p.

a) What is the expected TPR and FPR for C1 and C2?

b) Is C2 a better classifier than C1?

c) The expected precision for both C1 and C2 is 1/10. Expected recall for C2 is twice that of C1 (2p and p, respectively). If we use precision and recall as the evaluation metrics, C2 appears to be a better classifier than C1. Do you think {precision and recall} correctly indicate the relative performance of C2 and C1?

Question 4

You are trying to evaluate four different blood tests, T1, T2, T3, and T4, that have been developed to detect a particular type of cancer. These tests have been developed by different organizations, and their evaluations by these organizations have been reported in the following confusion matrices, along with the values of the following evaluation measures: TPR, FPR, Precision, the F-measure, and TPR/FPR.

Test T1: TPR: 0.4, FPR: 0.1, Precision: 0.285, F1-Score: 0.33, TPR/FPR: 4

Dataset: (1100 patients)	**Predicted by Blood Test**	
Actual	**Cancer (+ class)**	**No Cancer (- class)**
Cancer (+ class)	40	60
No Cancer (- class)	100	900

Test T2: TPR: 0.2; FPR: 0.05, Precision: 0.4, F1-Score: 0.26, TPR/FPR: 4

Dataset: (200 patients)	**Predicted by Blood Test**	
Actual	**Cancer (+ class)**	**No Cancer (- class)**
Cancer (+ class)	20	80
No Cancer (- class)	5	95

Test T3: TPR: 0.5; FPR: 0.5, Precision: 0.5, F1-Score: 0.5, TPR/FPR: 1

Dataset: (200 patients)	**Predicted by Blood Test**	
Actual	**Cancer (+ class)**	**No Cancer (- class)**
Cancer (+ class)	50	50
No Cancer (- class)	50	50

Test T4: TPR: 0.5; FPR: 0.1, Precision: 0.833, F1-Score: 0.625, TPR/FPR: 5

Dataset: (200 patients)	**Predicted by Blood Test**	
Actual	**Cancer (+ class)**	**No Cancer (- class)**
Cancer (+ class)	50	50
No Cancer (- class)	10	90

a. Between T1 and T3, which test is better?
If you need more information to make a decision, what would it be?

b. Between T1 and T2, which test is better?
If you need more information to make a decision, what would it be?

c. Between T1 and T4, which test is better?
If you need more information to make a decision, what would it be?

d. Between T4 and T3, which test is better?
If you need more information to make a decision, what would it be?

e. Between T4 and T2, which test is better?
If you need more information to make a decision, what would it be?

Question 5

Imagine there are 100 transactions, numbered 1,2,...,100, and 100 items, similarly numbered. Item i is in transaction j if and only if j is divisible by i. For example, transaction 24 is the set of items {1,2,3,4,6,8,12,24}.

a) Describe all the association rules that have 100% confidence.

b) List all 1-itemsets and 2-itemsets that have a support of 15% or more. Use the apriori algorithm to filter candidate itemsets. You may briefly show the computation to explain your selection.

c) Which of the following association rules meet a 100% confidence threshold and a 15% support threshold? Give a one-line **explanation** for each.

i. {3} ->{1,6}

ii. {1,2,6}->{3}

iii. {} ->{1}

iv. {1,4,6} -> {1,2,3,4}

v. {1,2,3,6} -> {3,4}

vi. {3,4} -> {1,2,3,6}

Question 6

Following is a list of all the frequent itemsets of size 3 found for a certain database. {A,B,D}, {A,C,D}, {A,C,E}, {A,D,E}, {B,C,D},{B,C,E}, {C,D,E}

Answer whether the given pairs of size 3 itemsets can be merged to create candidates for frequent itemsets of size 4 by Apriori $F_{k-1} \times F_{k-1}$ method while using lexicographical ordering (as discussed in class and in the textbook)? If the itemset could be generated, list the candidate itemset otherwise write **None.** Also, answer whether the candidate itemset would be pruned in the pruning step or not. Write NA (Not Applicable) if you write **None** in the column of *Candidate Itemset.*

Merging itemsets	Candidate Itemset	Survives Pruning (Yes / No/NA)
{A,C,D} and {A,C,E}		
{B,C,D} and {A,B,D}		
{B,C,D} and {B,C,E}		

Question 7

Consider a market-basket transaction data that only has 4 items: X,
Y, Z, and W. The table on the right shows the support of some itemsets, but the supports of other itemsets are unknown. Label
each itemset listed below with the following letter(s):

• C if it is a closed itemset,

• N if it is not a closed itemset, and

• I if the information is not enough to judge whether or not it is closed.

Itemsets	Support
{X}	20
{Y}	14
{W}	18
{X,Z}	16
{Y,W}	14
{X,W,Z}	14
{X,Y, Z}	14

a) {Z}
b) {Y}
c) {X}
d) {X, Z}

Question 8

The dataset below contains 10 items and 100 transactions. Dark cells indicate the presence of items, while white (and grey) cells indicate the absence of items. We will apply the Apriori algorithm to extract frequent itemsets with minsup = 20% (i.e., itemsets must contain at least 20 transactions). Determine whether or not the following statements are true or false with a brief explanation.

Statement	True/False. Explanation if false
{A, B} is a maximal frequent itemset	
{A, B} is a closed frequent itemset	
{D,E,F} is a frequent itemset	
Confidence of rule {A} → {J} is 100 %	
All subsets of {A,B,D,E} are frequent.	
{A,B,D} is a maximal frequent itemset