CSE 5243: Homework #4

Deadline: **11:59PM on 11/15/2022**.
No late submissions will be accepted.

Instructions.

Please do not code computer programs or call existing packages to solve the problems. Please make sure you understand how to solve them by hand (or with calculators, if necessary).

We currently use a 100-point scale for this homework, but it will take 10% of your final grade.

What you should turn in:

Please submit an electronic copy of your homework (in pdf) on Carmen.

Questions?
Please create a post on Carmen discussion areas to get timely help from other students, TA, and the instructor. Everyone can benefit from first checking what have been asked previously.

Participation takes 5%: If you have been inactive in participation so far, e.g., missed many classes, please try to actively answer questions in Carmen or be active in class.

Bonus points: For this homework (the penultimate one), we don't plan to extend the deadline at this moment; for students that are in really special situations, we can still accommodate your request. For those of you who can turn in your answers by the deadline on time, there will be 5 bullet points as a reward (added to your final grades until you reach 100).

Problem 1 (15 points).

Given the dataset in the following table, use the naive Bayes classifier to classify the new point (T, F,2.0).

Table for Problem 1.

xi	a_1	a_2	a_3	Class
x1	T	T	5.0	N
x2	T	T	7.0	N
x3	T	F	8.0	Y
x4	F	F	3.0	N
x5	F	T	7.0	Y
x6	F	T	4.0	Y
x7	F	F	5.0	Y
x8	T	F	6.0	N
x9	F	T	1.0	Y

Problem 2 (14 points).

Consider the following table. Let us make a nonlinear split instead of an axis parallel split as we did before, given as follows: $AB - B^2 \leq 0$ (i.e., partition the dataset based on whether this condition holds). Compute the information gain of this split based on entropy (use log_2, i.e., log to the base 2).

Table for Problem 2.

	A	B	Class
x1	3.5	4	H
x2	2	4	H
x3	9.1	4.5	L
x4	2	6	H
x5	1.5	7	L
x6	7	6.5	H
x7	2.1	2.5	L
x8	8	4	L

Problem 3 (15 points).

Consider the *distance* matrix for 5 data points in the following table. Use the Single-Linkage method to generate hierarchical clusters. Show the distance between clusters at each step.

Note: There can be multiple possible clustering results and just show one with your intermediate steps.

Dataset for Problem 4.

	A	B	C	D	E
A	0	1	3	2	4
B		0	3	2	3
C			0	1	3
D				0	5
E					0

Problem 4 (10 points).

Let $\mathcal{C}$ be the set of all closed frequent itemsets and $\mathcal{M}$ the set of all maximal frequent itemsets (or, max-patterns) for some database. Prove that $\mathcal{M} \subseteq \mathcal{C}$.

Problem 5 (10 points).

The Apriori algorithm for mining frequent patterns makes use of prior knowledge of subset support properties.

(a) Prove that the support of any nonempty subset s’ of itemset s must be at least as great as the support of s. (2 points)

(b) Prove that all nonempty subsets of a frequent itemset must also be frequent. (2 points)

(c) Given frequent itemset l and subset s of l, prove that the confidence of the rule “s’ => (l - s’)” cannot be more than the confidence of “s => (l - s)”, where s’ is a subset of s. (2 points)

(d) A partitioning variation of Apriori subdivides the transactions of a database D into n non-overlapping partitions. Prove that any itemset that is frequent in D must be frequent in at least one partition of D. (4 points)

Problem 6 (18 points).

Given the database in the following Table and minsup = 3 (i.e, the minimum number of occurrences for an itemset to be frequent), show how the Apriori algorithm enumerates all frequent patterns.

Dataset for Problem 6.

tid	itemset
t1	ABCD
t2	ACDF
t3	ACDEG
t4	ABDF
t5	BCG
t6	DFG
t7	ABG
t8	CDFG

Problem 7 (18 points).

Find all frequent, closed, and maximal itemsets using minsup = 2 in the following database shown in Table 1:

Tid	Itemset
t1	ACD
t2	BCE
t3	ABCE
t4	BDE
t5	ABCE
t6	ABCD

Table 1: Database for Problem 7.