CS 412: Fall'23
Introduction To Data Mining

Take-Home Final

(Due Saturday, December 09, 2023, 06:00 pm)

- The Take-Home Final is due on Saturday, December 09, 2023, 06:00 pm. We will be using Gradescope for the Finals. Please do NOT email a copy of your solution. Contact the TAs if you are having technical difficulties in submitting the assignment. We will NOT accept late submissions!
- Please use Slack first if you have questions about the Finals.
- You will have to answer the questions yourself, you cannot consult with other students in class. It is an open book exam, so you can use the textbook and the material shared in class, e.g., slides, lectures, etc.
- We will be enforcing the following two aspects:
  - You are expected to typeset the solutions, we encourage you to use Latex. **Handwritten solutions will not be graded and you will not get any credit for handwritten solutions**.
  - When submitting the Finals, Gradescope will ask you to select pages for each problem; please do this precisely. **Any pages which are not correctly selected for a problem will not be graded and you will not get any credit for those pages.**
- For each question, you will NOT get full credit if you only give out a final result. Please show the necessary steps and details, including any formulae you use, what the variables mean in the formulae, any derivation or calculation steps, and explanations as appropriate. Consider the following two examples:

Example 1 **Q:** Given a dataset $\mathcal{X} = \{3.1, 4.2, -1\}$, compute the mean.
A: For any set of n numbers $\mathcal{X} = \{x_1, \ldots, x_n\}$, the mean can be computed as $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$. For the given dataset $\mathcal{X}$, the mean is $\mu = \frac{3.1+4.2-1}{3} = 2.1$

Example 2 **Q:** A coin claimed to be unbiased has been tossed 100 times, with 54 heads and 46 tails. What is the χ^2 statistic?
A: For a categorical variable taking k possible values, if the expected values are $e_i, i = 1, \ldots, k$ and the observed values are $o_i, i = 1, \ldots, k$, then the χ^2 statistic can be computed as: $S = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{o_i}$. For the problem, since the coin is claimed to be unbiased, the expected values are $50, 50$. Further, the observed values are $54, 46$. Then, the chi-squared statistic is given by $S = \frac{(54-50)^2}{50} + \frac{(46-50)^2}{50} = 0.64$.

1. (25 points) A sequence database SDB_1 (Table 1) has 5 transactions, and we will consider frequent sequential pattern mining with (absolute) minimum support of 2.

Sequence_ID	Sequence
S_1	$\langle (ac)cb(bd) \rangle$
S_2	$\langle (bf)(fg)b(ce) \rangle$
S_3	$\langle (bf)(ah)abf \rangle$
S_4	$\langle d(ce)(be) \rangle$
S_5	$\langle a(bd)bcb(ade) \rangle$

Table 1: A sequence database SDB_1.

(a) (3 points) What are the length-1 frequent sequential patterns in SDB_1 and what are their supports?

(b) (4 points) What is the projected database for prefix $\langle bb \rangle$? Is $\langle bb \rangle$ a length-2 frequent pattern in SDB_1? What is the support of $\langle bb \rangle$?

(c) (4 points) What is the projected database for prefix $\langle (bd) \rangle$? Is $\langle (bd) \rangle$ a length-2 frequent pattern in SDB_1? What is the support of $\langle (bd) \rangle$?

(d) (4 points) What is the projected database for prefix $\langle b \rangle$?

(e) (10 points) Using the projected database for prefix $\langle b \rangle$, find all frequent subsequences starting with b. Please list your answer in lexicographically ascending order.

2. (25 points) Giving the following transaction database, we will focus on frequent pattern mining with minimum absolute support of 3, i.e., $minsup = 3$.

TID	Items
T_1	A,B,C
T_2	A,D,E
T_3	B,D
T_4	A,B,D
T_5	A,C
T_6	B,C
T_7	A,C,D
T_8	A,B,C,D,E
T_9	B,C,D
T_{10}	A,B,C,E

Table 2: Transaction Database.

(a) (5 points) For an association rule $A \Rightarrow B(s, c)$, calculate its support s and confidence c.

(b) (10 points) Find all frequent itemsets using Apriori algorithm. Please use the alphabetical ordering A, B, C, D, E for the candidate generation (self-join and pruning) phase. Please show all intermediate steps (like how to scan and filter the data, how to generate candidates with the ordering you are referring to, and if any pruning tricks are available, until you reach your final result) to get full credit.

(c) (10 points) What is the FP-tree corresponding to transactions in Table 2?
Requirements: Please insert transactions in the order of $T_1, T_2, \ldots, T_{10}$. You need to draw three FP-Trees after inserting T_1, T_5, and T_{10} to get full credit.

3. (24 points) This question considers optimization for deep learning.

   (a) (6 points) GPT-4 is rumored to have 1.76 trillion parameters. What would be the main computational challenge in training such a model using Newton's method? Clearly explain your answer using suitable notation, equations, and description.

   (b) (6 points) Using suitable and precise notation, clearly describe the parameter update for the Adagrad algorithm. Remark on which parameters have higher vs. lower learning rates for Adagrad.

   (c) (4 points) Professor ScatterBrain claims that the Adagrad algroithm is not practical for situations which need to use a large number of steps of the algorithm as all past gradients have to be stored, leading to significant storage demands. Do you agree with Professor ScatterBrain? Clearly justify your answer.

   (d) (8 points) Using suitable and precise notation, clearly describe the parameter update for the Adam algorithm. What primary concern regarding the Adagrad algorithm does Adam tries to resolve and how?

4. (26 points) This question considers clustering algorithms.

   (a) (6 points) What is the computational complexity of the Partitioning Around Medoids (PAM) k-medoids clustering algorithm? Briefly justify your answer.

   (b) (6 points) Is the k-medians algorithm more computationally demanding than k-means? Briefly explain your answer.

   (c) (14 points) Consider a dataset with 5 data points a, b, c, d, e with pairwise distances given by Table 3.

	a	b	c	d	e
a	0				
b	2.1	0			
c	4	6.1	0		
d	7.8	8.6	6.1	0	
e	7.3	7.2	7.3	3.1	0

Table 3: Pairwise distance matrices between data points. Since the distance is assumed to be symmetric, only the lower diagonal and diagonal entries are shown.

Consider running complete link agglomerative clustering algorithm on the dataset. For each step of the algorithm:

i. (4 points) Show which two clusters will be merged at step based on pairwise distance, and

ii. (10 points) Show the updated pairwise similarity matrix after merging.