CS5600 Data Mining Exam - 2, Fall 2025

(Closed Book, 2 hrs, 35 marks)

Instructions:

- Answers without proper explanation will not receive full marks.
- Be precise and concise in the answers. If you make any additional assumptions, write them clearly.
- The marks for each question is given in square brackets in bold font.

Q1 [5 marks] In the reservoir sampling algorithm, what is the probability that an element present in the sample S after n elements will be kept in S after (n+3) elements. Assume s is the size of S. Write your answer in terms of parameters such as n, s. Consider various cases, such as $n < s$, $n > s$, etc.

Q2 [5 marks] Consider the following modification to Bloom filter: If $B[h_i(x)] = 1$ for any $i = 1,..., k$; then declare that x is in S, where S is the given set of keys and B is the bit array. Assume the size of S and B to be m and n respectively. With this change answer the following:

a) What would be the fraction of 1s in the bit vector?
b) What would be the probability of false positives?
c) What "k" will give the lowest false positive probability?
d) Draw a plot that shows how false positive probability varies with the number of hash functions for the bloom filter and the above modified bloom filter.

Q3 [5 marks] Let C_1 and C_2 be two sets of objects, and let D(x,y) denote a dissimilarity measure between any two objects x and y. Give a mathematical definition of distance function between clusters for single link clustering and for complete link clustering in terms of the given details, such as C_1, C_2, D(x,y), etc.

$D_{single}(C_1,C_2)$ = _____

$D_{complete}(C_1,C_2)$ = _____

Q4 [5 marks] Given a dataset with the following four points in 2D: (1, 1), (2,1), (1, 2), (8,8).
Considering k=2 nearest neighbors, compute the local outlier factor for point (8,8). Is (8,8) an outlier given its LOF score?

Q5 [5 marks] Consider the clustering and ground truth partitions given below:

C/T	T_1	T_2	T_3	Sum
C_1	5	1	0	6
C_2	1	4	1	6
C_3	2	0	3	5
m_j	8	5	4	17

For the above clustering, compute Jaccard index which is defined as follows:

$$\frac{TP}{TP + FN + FP}$$

Q6 [5 marks] Assume you are given records of the form: (StudentID, CourseID, GradePoint, Credits), where GradePoint is an integer between 0-10. We want to use MapReduce to compute CGPA for each student. Write the input and output <key, value> pairs at the mapper and the reducer nodes. Also state the operation that will be performed at the mapper and reducer nodes.

$$\text{CGPA} = \frac{\sum(\text{GradePoint} \times \text{Credits})}{\sum(\text{Credits})}$$

Q7 [5 marks] In the above question, can we use a combiner for CGPA calculation? Justify your choice. If yes, also write what would be the input and output key value pair, and the operation at the combiner.