Practice Questions

Question 1: True/False (Total 8 pts)

For each of the following statements, indicate either **"True" or "False"** and **provide a brief justification of your answer**:

a. Increasing the maximum depth of decision tree reduces bias

b. Decision tree can always reach 100% *training* accuracy if maximum depth is sufficiently large

c. The BFR clustering algorithm does not require any of the data points to be saved in memory across iterations

d. Increasing the number of layers in deep neural network increases variance

e. In K-means, average distance to centroid never increases over iterations

f. LSH may suffer from both false positives and false negatives

g. In decision trees, we may split on the same binary feature in more than one node in the tree.

h. Before training a decision tree on classification dataset, you look at the complete training dataset and observe that feature X_3 is independent from the class label Y, i.e., that the entropy of label Y, H(Y), is equal to the conditional entropy of H(Y|X). Therefore, you can conclude that feature X_3 will not appear in the decision tree.

Question 2: Machine Learning (Total 4 pts)

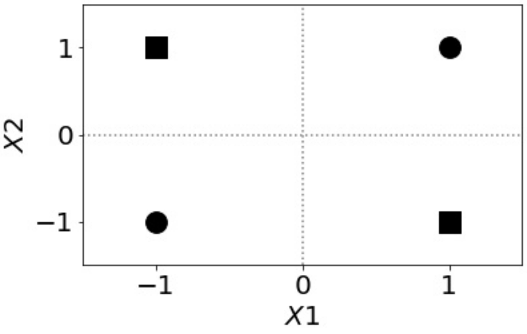
a. [2 pts] You collected a dataset about customers browsing your e-commerce website. The dataset shows for each session how long the customer spent on your website, whether they are new members, whether they were offered a discount, and whether they made a purchase:

Time spent (minutes)	**Discount Offered**	**New Member**	**Purchase**
20	Yes	Yes	Yes
18	Yes	No	Yes
5	Yes	Yes	No
15	No	No	No
8	No	No	No
7	Yes	Yes	No
15	Yes	No	Yes
13	No	Yes	No
22	No	No	No

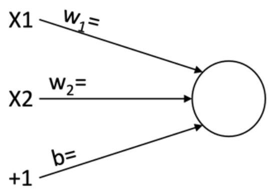
i. You are training a decision tree to predict whether a customer will make a purchase. To decide which of the features "Discount Offered" and "New Member" you should split on at the root of the tree, compute the information gain for both features and determine which would be a better split.

ii. If instead, you would like to split based on the "Time Spent" feature by split of the form "Time spent <= [threshold]". How would you decide the best threshold?
You do not need to compute it, just provide a brief explanation of how you suggest to compute it (there is more than one valid solution).

b. [2 pts] Consider the following training dataset with four points. The dataset has two features X_1 and X_2 and a binary target class attribute Y (either ✖ or ■).
For each of the following types of classifiers, describe the best training accuracy that can be achieved using this classifier and justify your answer by describing the best classifier as explained in the following:



i. Perceptron (please provide the weights (*w_1, w_2, b*) and the best training accuracy):



ii. Decision tree with maximum depth of 1 (i.e., one root note and two leaves). Please draw the decision tree and provide the best training accuracy.

iii. Decision tree with maximum depth of 2. Please draw the decision tree and provide the best training accuracy.

iv. K-nearest neighbors classifier with K=3 (briefly described in class: this is a classifier that chooses the class for a point based on the majority class among its K nearest points in training set). Please only provide the training accuracy and explain your calculation.

Question 3: MapReduce (Total 3 pts)

Consider a very large table with two columns, "Zip Code" and "Age". See on the left an example for the dataset structure.

Zip Code	Age
12345	30.1
12345	40.5
78910	25.8
78910	35.2
12345	19.1
56789	91.9
...	...

(a) **[1 pts]** Design a Map-Reduce program that computes a histogram over ages in buckets of five year, i.e., how many people are between the ages 0 and 5, how many people are between the ages 5 and 10, how many people are between the ages 10 and 15, and so on. You can assume you are reading the file line by line and each line is formatted as "[zipcode],[age]". Provide the algorithm pseudocode (provide both a map function and a reduce function).

(b) **[1 pts]** Design a Map-Reduce program that computes the average age per each zip code. You can assume you are reading the file line by line and each line is formatted as “[zipcode],[age]”. Provide the algorithm pseudocode (provide both a map function and a reduce function).

(c) **[1 pts]** For question (b) regarding the computation of average age per each zip code, can you design a combiner (i.e., a *combine* function) function that would reduce network time by pre-aggregating values in the mapper?
Note: depending on how you implemented (b), you may need to use different map/reduce functions that would be compatible with your proposed combiner. There is no need to change your answer to (b) and you can simply describe these changes here.

Question 4: Frequent Itemsets (Total 2 pts)

a. [1 pts] Assuming L1, the set of frequent itemsets of size 1, has n itemsets. What is the size of C2, the set of candidate itemsets of size 2? Briefly justify your answer.

b. [1 pts] Consider a dataset with 1 million baskets, each basket has 5 items that are chosen from a set of 200,000 different items. Note that the five items in each basket are different, but the same items can appear in multiple baskets (and some items may not appear in any of the baskets). Suppose the support threshold is 100. What are the minimum and maximum numbers of frequent items (i.e., frequent itemsets of size 1)? Briefly explain your answer.

<u>Question 5: Locality-Sensitive Hashing (Total 4 pts)</u>

In this question, we focus on applying LSH to find documents in our dataset that similar to a query document. Documents will be represented using sets of k-shingles and similarity between documents will be computed using Jaccard Similarity.

a. [1 pts] Assuming we have the following Boolean input matrix such that each row corresponds to a k-shingle and each column corresponds to a document in our dataset. In addition, three random permutations of the dataset π_1, π_2, π_3 are provided and one example query document.

π_1	π_2	π_3
3	6	2
2	1	7
7	4	4
1	5	1
5	3	6
6	7	3
4	2	5

Input Matrix

1	0	1	1	0
0	1	0	1	1
1	1	0	0	0
1	0	0	0	1
0	1	1	0	1
0	0	0	1	1
1	0	1	0	1

Query *q*

1
0
0
1
0
1
0

Compute the signature matrix M of our dataset, as well as the signature of the query q, using the ***Min-Hashing*** technique based the provided random permutations.

Matrix M

Query

b. [1 pts] Given the query document q described above, compute both the Jaccard similarity and the signature-to-signature similarity between the query document and documents #1 and #2 from our dataset (document #1 is represented by the most left column, ...)

	Query q and document #1	Query q and document #2
Document-Document similarity		
Signature-Signature similarity		

c. [2 pts] To find the similar documents to any given query you implement the LSH algorithm based on 100 random permutations, with b=10 bands and r=10 rows.
For convenience, you can leave exponents without computing them.

   i. Assuming a query ***q*** and a document ***d_1*** that have a Jaccard similarity sim(***q***, ***d_1***) = 0.75, what is the probability that ***d_1*** will be returned as a candidate match to query ***q***?

   ii. Complete the following statement and justify your answer:
   The **amplified** LSH family described above (with b=10 and r=10) is

   (0.25, 0.75, ______, ______)-sensitive