Question 1 (10 points):

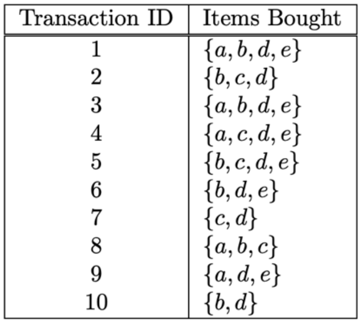
Calculate city block, Euclidean and supremum distances for the below data.

(a). x = (1, -1, 10, 3, 4), y = (10, -1, 4, 5, 2)

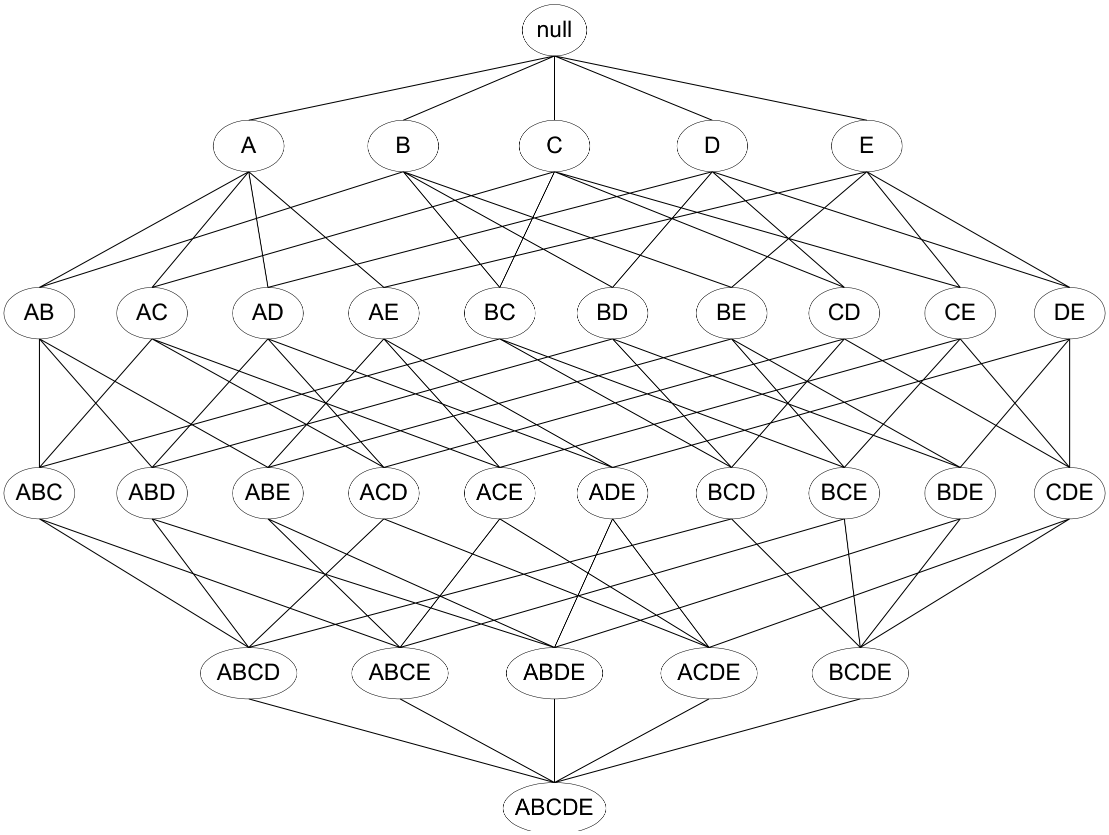
(b). x1 = (5, 4), x2 = (-2, 3)

Question 2 (25 points):

Given the lattice structure in the below picture and the transactions given in the below table, label each node with the following letter(s): I if it is infrequent; F if it is frequent; M if the node is a maximal frequent itemset. Assume that the support threshold (minimum support) is 20%.



Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$



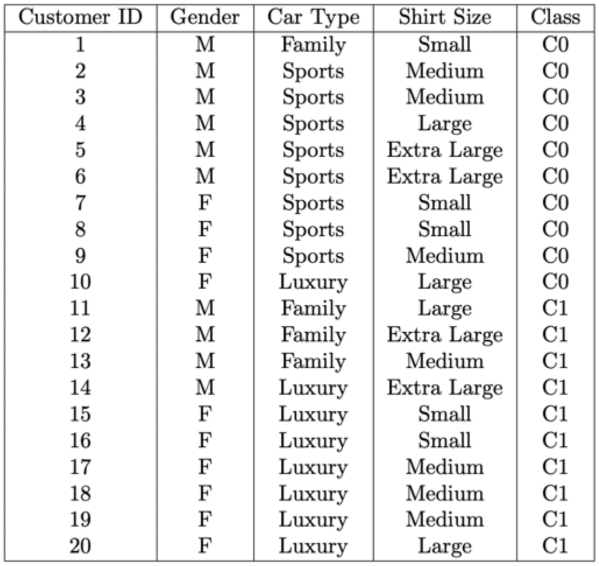
Question 3 (25 points):

Consider the training examples show in the below table for a binary classification problem

(a). Compute the Gini index for the overall collection of training examples.

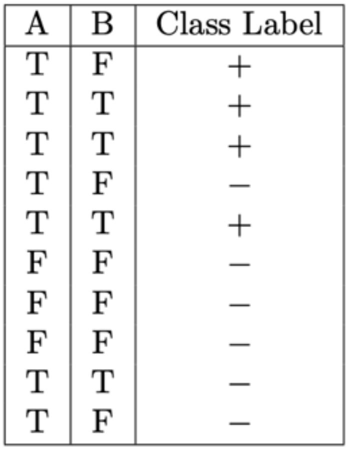
(b). Compute the Gini index for the Customer ID attribute.

(c). Compute the Gini index for the Gender attribute.



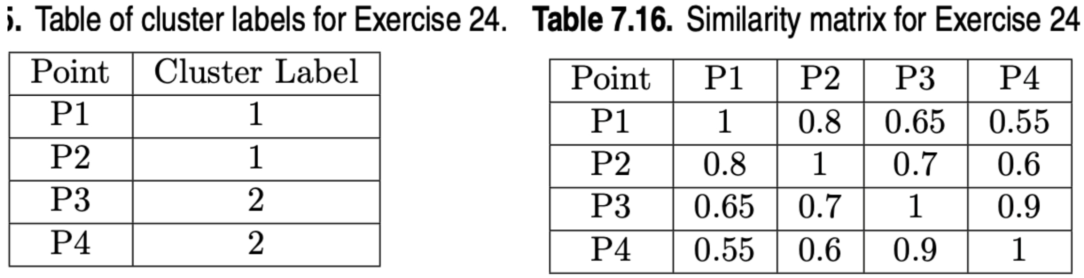
Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

Question 4 (20 poionts):
Consider the following dataset for a binary class problem.
Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?



A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

Question 5 (20 poionts):
Using the below data, compute the silhouette coefficient for each point, each of the two clusters, and the overall clustering.



5. Table of cluster labels for Exercise 24.

Point	Cluster Label
P1	1
P2	1
P3	2
P4	2

Table 7.16. Similarity matrix for Exercise 24

Point	P1	P2	P3	P4
P1	1	0.8	0.65	0.55
P2	0.8	1	0.7	0.6
P3	0.65	0.7	1	0.9
P4	0.55	0.6	0.9	1

Extra points (10 points)
Given the below four faces, darkness or number of dots represents density. Lines are used only to distinguish regions and do not represent points.

For each figure, could you use partition, hierarchical, density, and other algorithms we learned in class to find the patterns represented by the nose, eyes, and mouth? Please list at least 3 different types of algorithms and explain the pros and cons of each.