LEBANESE UNIVERSITY

Faculty of Sciences – Section I

الجامعة اللبنانيّة

كلية العلوم ـ الفرع الأول

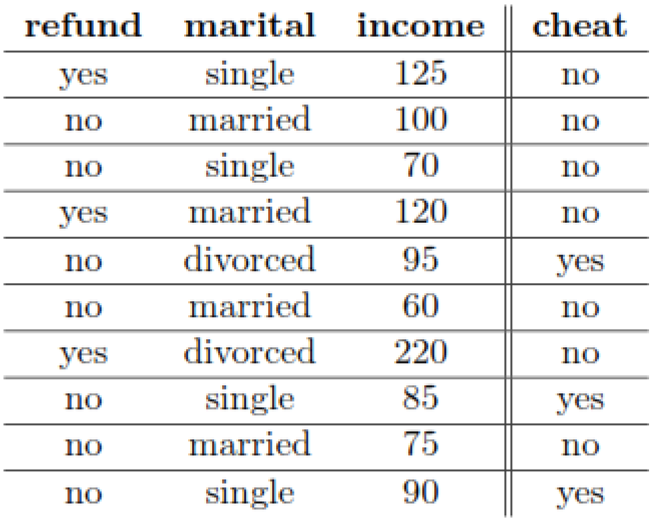
Université Libanaise
Faculté des Sciences 1

Makeup *Spring 2019*: INFO 437 Data Mining (English)

Duration: 2h   Documents not authorized (calculator authorized)

Exercise I (Decision Tree - Gini)   (15pts)

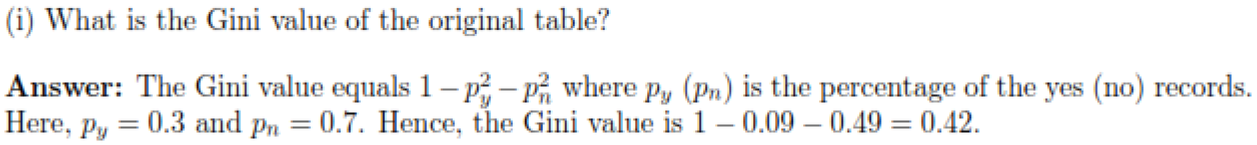
Assume that we have the following training set:



refund	marital	income	cheat
yes	single	125	no
no	married	100	no
no	single	70	no
yes	married	120	no
no	divorced	95	yes
no	married	60	no
yes	divorced	220	no
no	single	85	yes
no	married	75	no
no	single	90	yes

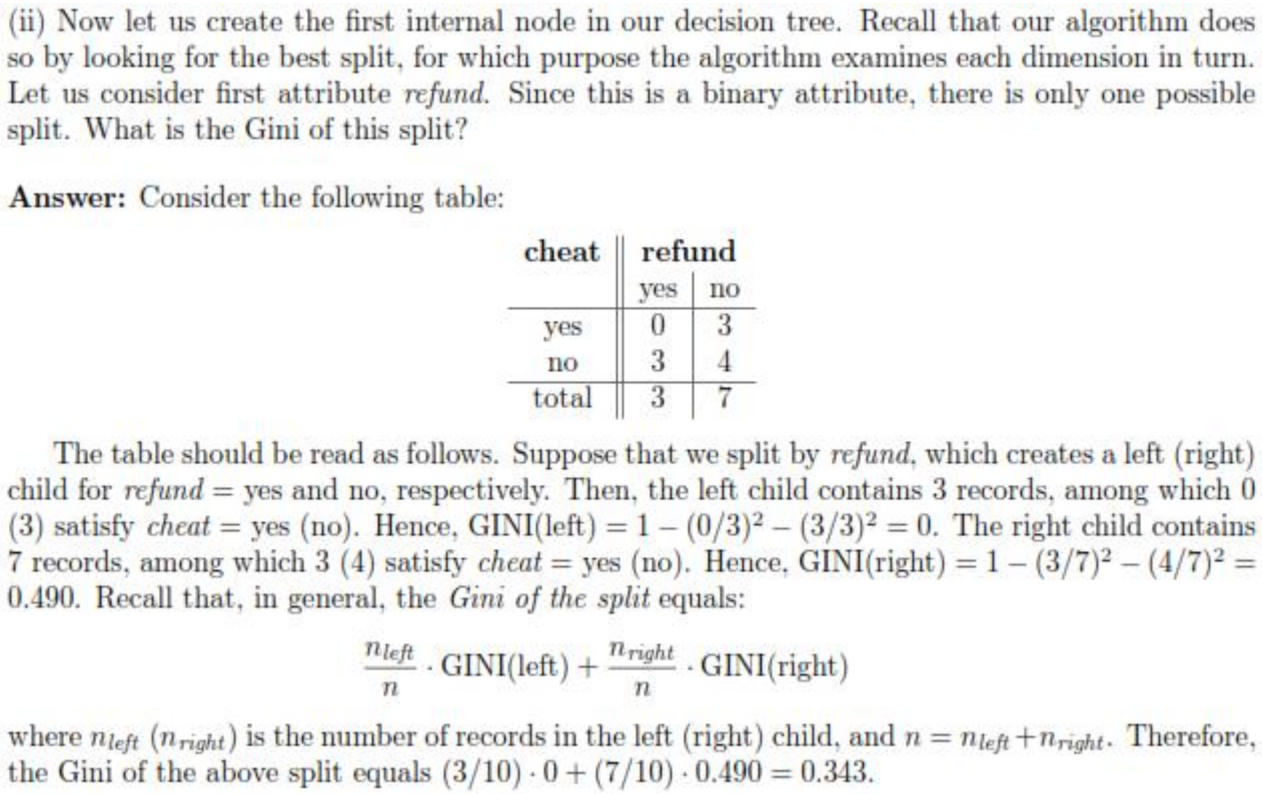
In the above table, cheat is the class label, whereas the other columns are the attributes.

1. What is the Gini value of the original table? 2
2. Now let us create the first internal node in our decision tree. Recall that our algorithm does so by looking for the best split, for which purpose the algorithm examines each attribute in turn. Let us consider first attribute *refund*. Since this is a binary attribute, there is only one possible split. What is the Gini of this split? 2
3. Let us now focus on attribute *marital*. How many splits are possible on this attribute? Which one is the best one, and what is its Gini? 4
4. Repeat the above for attribute *income*. 5
5. Considering all dimensions, which one is the best split? What is its Gain? What is the Root of the decision tree? 2



(i) What is the Gini value of the original table?

Answer: The Gini value equals $1 - p_y^2 - p_n^2$ where p_y (p_n) is the percentage of the yes (no) records. Here, $p_y = 0.3$ and $p_n = 0.7$. Hence, the Gini value is $1 - 0.09 - 0.49 = 0.42$.



(ii) Now let us create the first internal node in our decision tree. Recall that our algorithm does so by looking for the best split, for which purpose the algorithm examines each dimension in turn. Let us consider first attribute *refund*. Since this is a binary attribute, there is only one possible split. What is the Gini of this split?

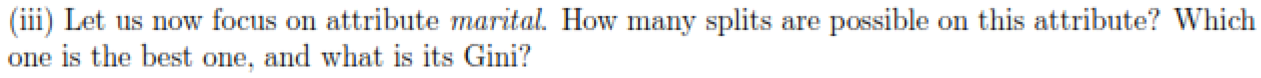
Answer: Consider the following table:

cheat	**refund**	
	yes	no
yes	0	3
no	3	4
total	3	7

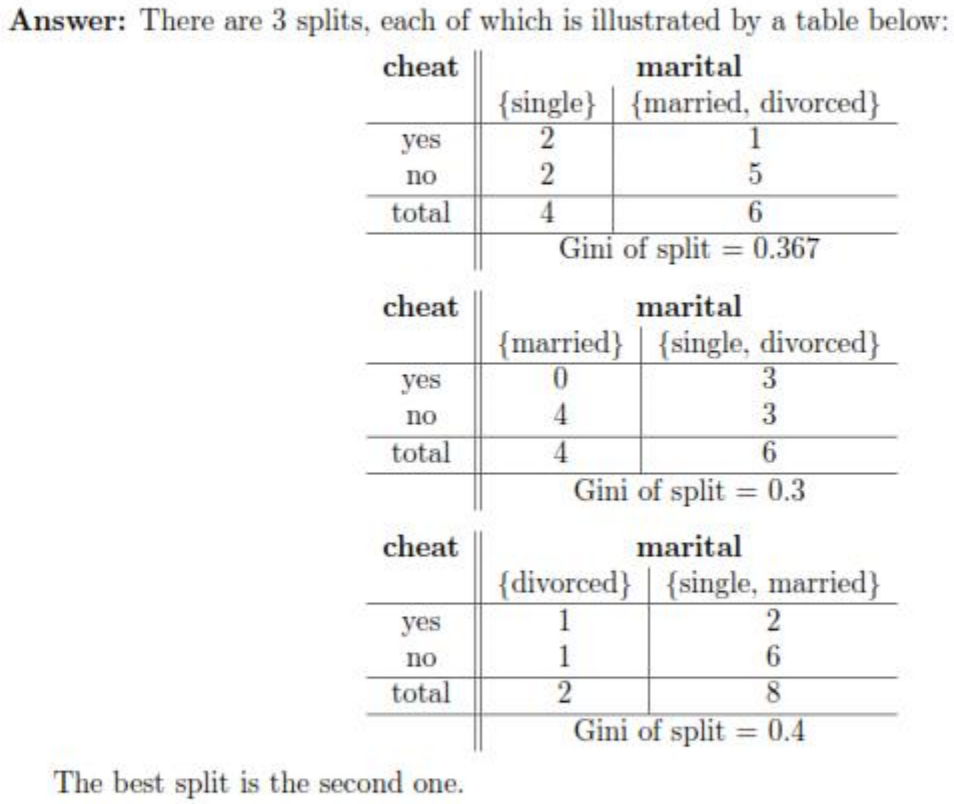
The table should be read as follows. Suppose that we split by *refund*, which creates a left (right) child for *refund* = yes and no, respectively. Then, the left child contains 3 records, among which 0 (3) satisfy *cheat* = yes (no). Hence, GINI(left) $= 1 - (0/3)^2 - (3/3)^2 = 0$. The right child contains 7 records, among which 3 (4) satisfy *cheat* = yes (no). Hence, GINI(right) $= 1 - (3/7)^2 - (4/7)^2 = 0.490$. Recall that, in general, the *Gini of the split* equals:

$$\frac{n_{left}}{n} \cdot \text{GINI(left)} + \frac{n_{right}}{n} \cdot \text{GINI(right)}$$

where n_{left} (n_{right}) is the number of records in the left (right) child, and $n = n_{left} + n_{right}$. Therefore, the Gini of the above split equals $(3/10) \cdot 0 + (7/10) \cdot 0.490 = 0.343$.



(iii) Let us now focus on attribute *marital*. How many splits are possible on this attribute? Which one is the best one, and what is its Gini?



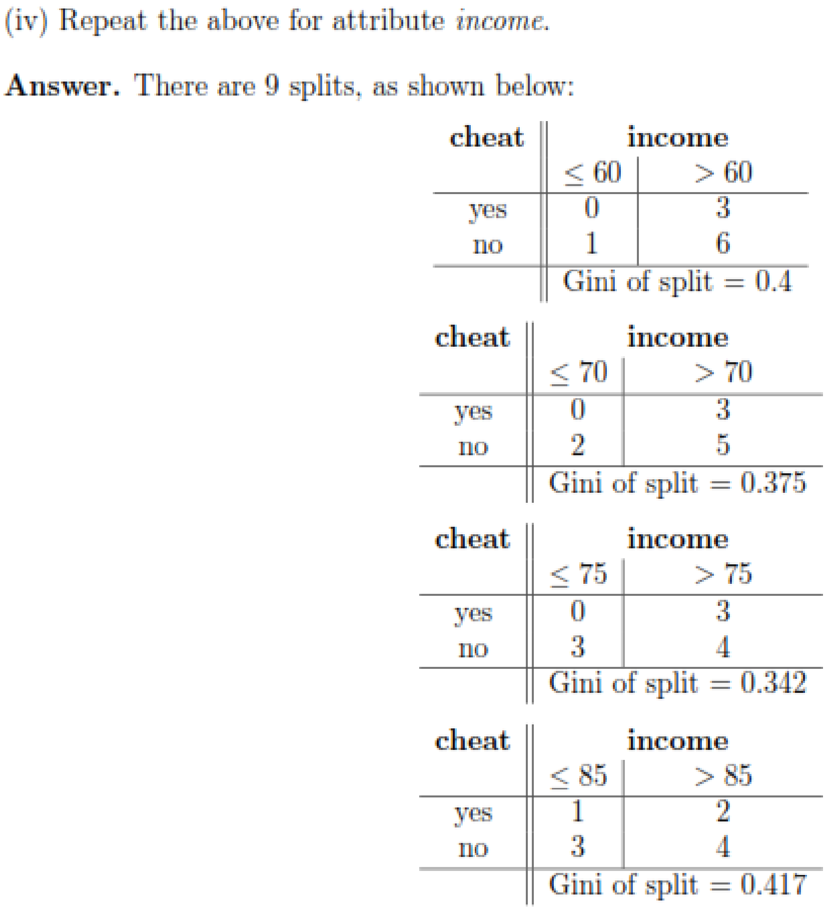
Answer: There are 3 splits, each of which is illustrated by a table below:

cheat	**marital**	
	{single}	{married, divorced}
yes	2	1
no	2	5
total	4	6
	Gini of split = 0.367	

cheat	**marital**	
	{married}	{single, divorced}
yes	0	3
no	4	3
total	4	6
	Gini of split = 0.3	

cheat	**marital**	
	{divorced}	{single, married}
yes	1	2
no	1	6
total	2	8
	Gini of split = 0.4	

The best split is the second one.



(iv) Repeat the above for attribute *income*.

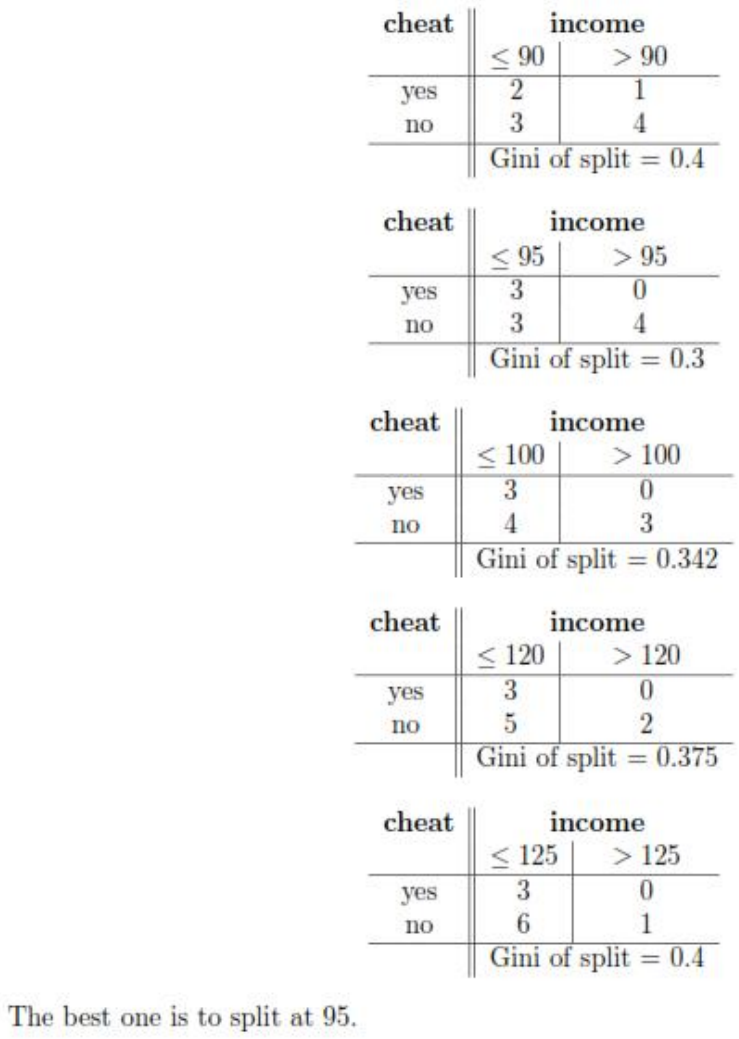
Answer. There are 9 splits, as shown below:

cheat	**income**	
	≤ 60	> 60
yes	0	3
no	1	6
	Gini of split $= 0.4$	

cheat	**income**	
	≤ 70	> 70
yes	0	3
no	2	5
	Gini of split $= 0.375$	

cheat	**income**	
	≤ 75	> 75
yes	0	3
no	3	4
	Gini of split $= 0.342$	

cheat	**income**	
	≤ 85	> 85
yes	1	2
no	3	4
	Gini of split $= 0.417$	



cheat	income	
	≤ 90	> 90
yes	2	1
no	3	4
	Gini of split = 0.4	

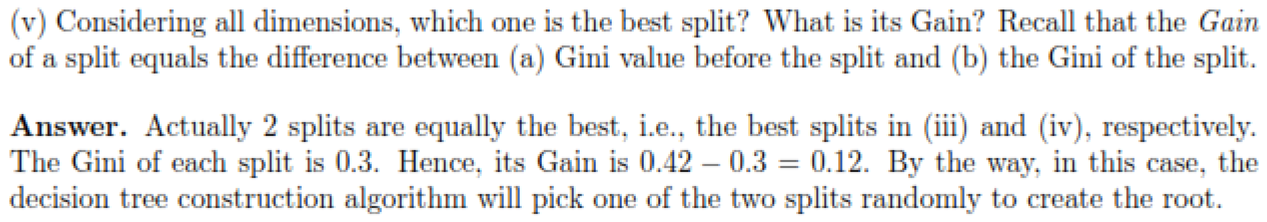
cheat	income	
	≤ 95	> 95
yes	3	0
no	3	4
	Gini of split = 0.3	

cheat	income	
	≤ 100	> 100
yes	3	0
no	4	3
	Gini of split = 0.342	

cheat	income	
	≤ 120	> 120
yes	3	0
no	5	2
	Gini of split = 0.375	

cheat	income	
	≤ 125	> 125
yes	3	0
no	6	1
	Gini of split = 0.4	

The best one is to split at 95.



(v) Considering all dimensions, which one is the best split? What is its Gain? Recall that the *Gain* of a split equals the difference between (a) Gini value before the split and (b) the Gini of the split.

Answer. Actually 2 splits are equally the best, i.e., the best splits in (iii) and (iv), respectively. The Gini of each split is 0.3. Hence, its Gain is 0.42 – 0.3 = 0.12. By the way, in this case, the decision tree construction algorithm will pick one of the two splits randomly to create the root.

Exercise II (Decision Tree - Entropy) (5pts)

Consider the following two possible splits of training records (A and B) during the decision tree induction.

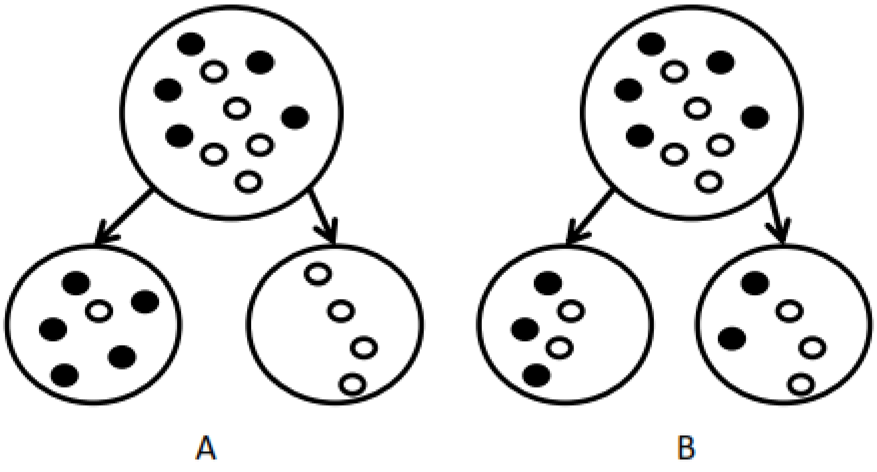


Figure 1. Possible splits of training records according to attributes A and B. Black and white colors represent class labels.

a. What is the entropy of each split? 1.5 + 1.5
b. Which split produces more pure nodes? 2

Entropy(A)=6/10*(-5/6*log(5/6,2)-1/6*log(1/6,2))+4/10*(-4/4*log(4/4,2))=0.39

Entropy(B)=5/10*(-3/5*log(3/5,2)-2/5*log(2/5,2))+5/10*(-2/5*log(2/5,2)-3/5*log(3/5,2))=0.97

Split A produces more pure nodes.

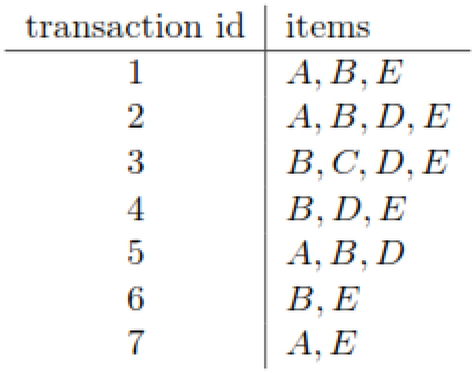
b. What is the information gain of each split? (1 point)
Entropy(All)=-5/10*log(5/10,2)-5/10*log(5/10,2)=1

Info gain (A)=1-0.39=0.61

Info gain (B)=1-0.97=0.03

Exercise III (Association Rules) (20pts)

Consider the mining of association rules on the transactions:



transaction id	items
1	A, B, E
2	A, B, D, E
3	B, C, D, E
4	B, D, E
5	A, B, D
6	B, E
7	A, E

1. What is the support of the itemset {B,D,E}? (1)

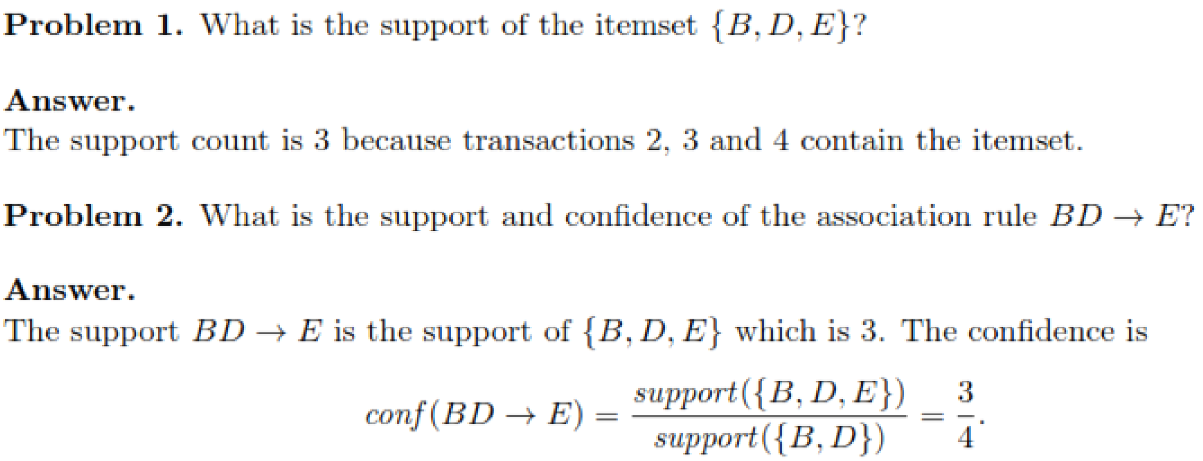
2. Calculate the *support*, *confidence*, and *lift* of the association rule BD → E? (3) Conclude if the items in this association rule are independent of each other or have negative or positive impacts on each other. (1)

3. Consider the application of the Apriori algorithm to find all the frequent itemsets whose counts are at least 3. Recall that the algorithm scans the transaction list a number of times, where the i^{th} scan generates the set F_i of all size-i frequent itemsets from a candidate set C_i. Show C_i and F_i for each possible i. (6)

4. List all *closed frequent* itemsets and *maximal frequent* itemsets (4)

5. Find all the association rules with support at least 3 and confidence at least 3/4. For your convenience, all the itemsets with support at least 3 are {{A}, {B}, {D}, {E}, {A,B}, {A,E}, {B,D}, {B,E}, {D,E}, {B,D,E}}. (5 : each 0.5 Pts + 0.5 bonus if all correct)

Lift = Confidence/Support



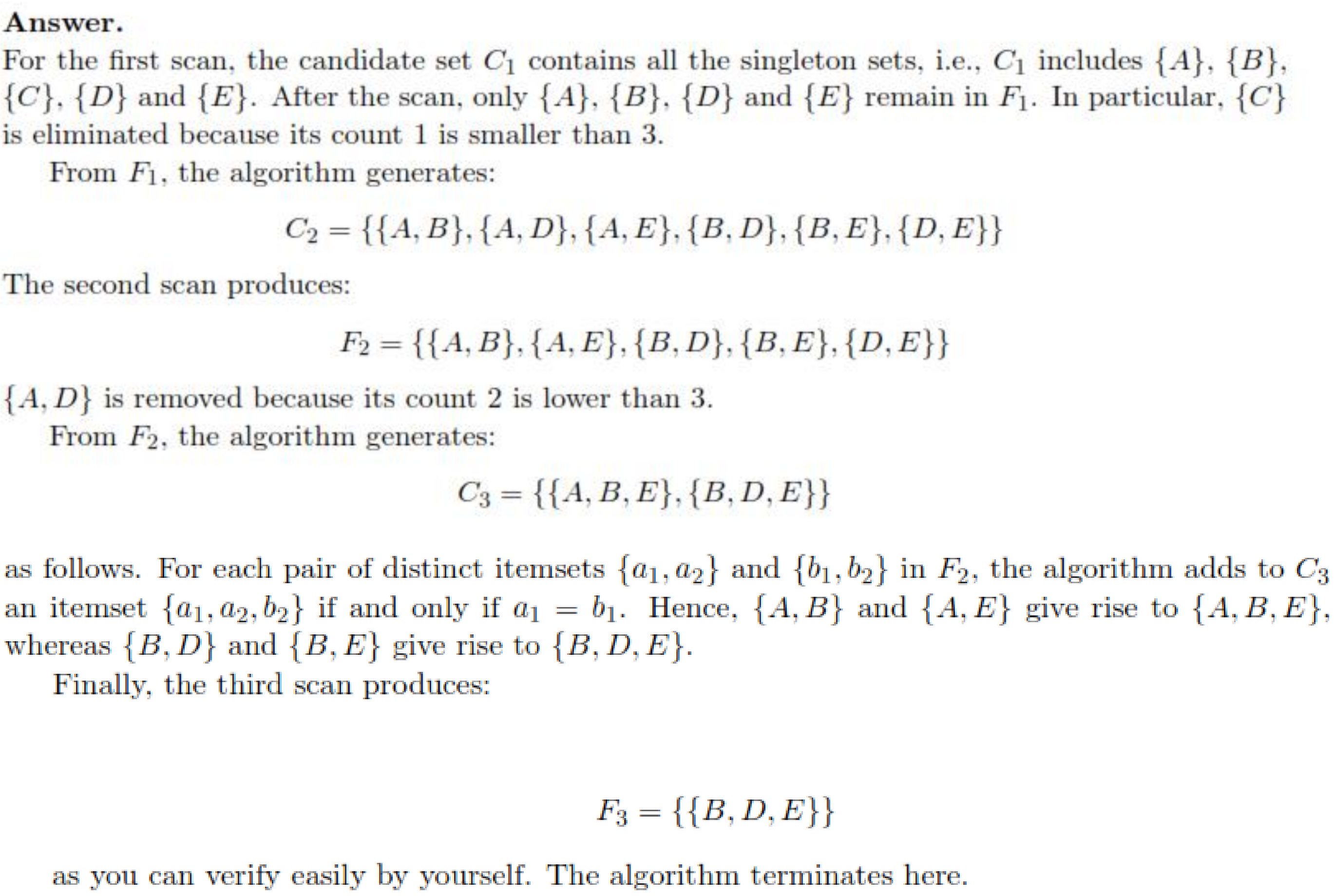
Problem 1. What is the support of the itemset $\{B, D, E\}$?

Answer.
The support count is 3 because transactions 2, 3 and 4 contain the itemset.

Problem 2. What is the support and confidence of the association rule $BD \to E$?

Answer.
The support $BD \to E$ is the support of $\{B, D, E\}$ which is 3. The confidence is

$$conf(BD \to E) = \frac{support(\{B, D, E\})}{support(\{B, D\})} = \frac{3}{4}.$$



Answer.

For the first scan, the candidate set C_1 contains all the singleton sets, i.e., C_1 includes $\{A\}$, $\{B\}$, $\{C\}$, $\{D\}$ and $\{E\}$. After the scan, only $\{A\}$, $\{B\}$, $\{D\}$ and $\{E\}$ remain in F_1. In particular, $\{C\}$ is eliminated because its count 1 is smaller than 3.

From F_1, the algorithm generates:

$$C_2 = \{\{A, B\}, \{A, D\}, \{A, E\}, \{B, D\}, \{B, E\}, \{D, E\}\}$$

The second scan produces:

$$F_2 = \{\{A, B\}, \{A, E\}, \{B, D\}, \{B, E\}, \{D, E\}\}$$

$\{A, D\}$ is removed because its count 2 is lower than 3.

From F_2, the algorithm generates:

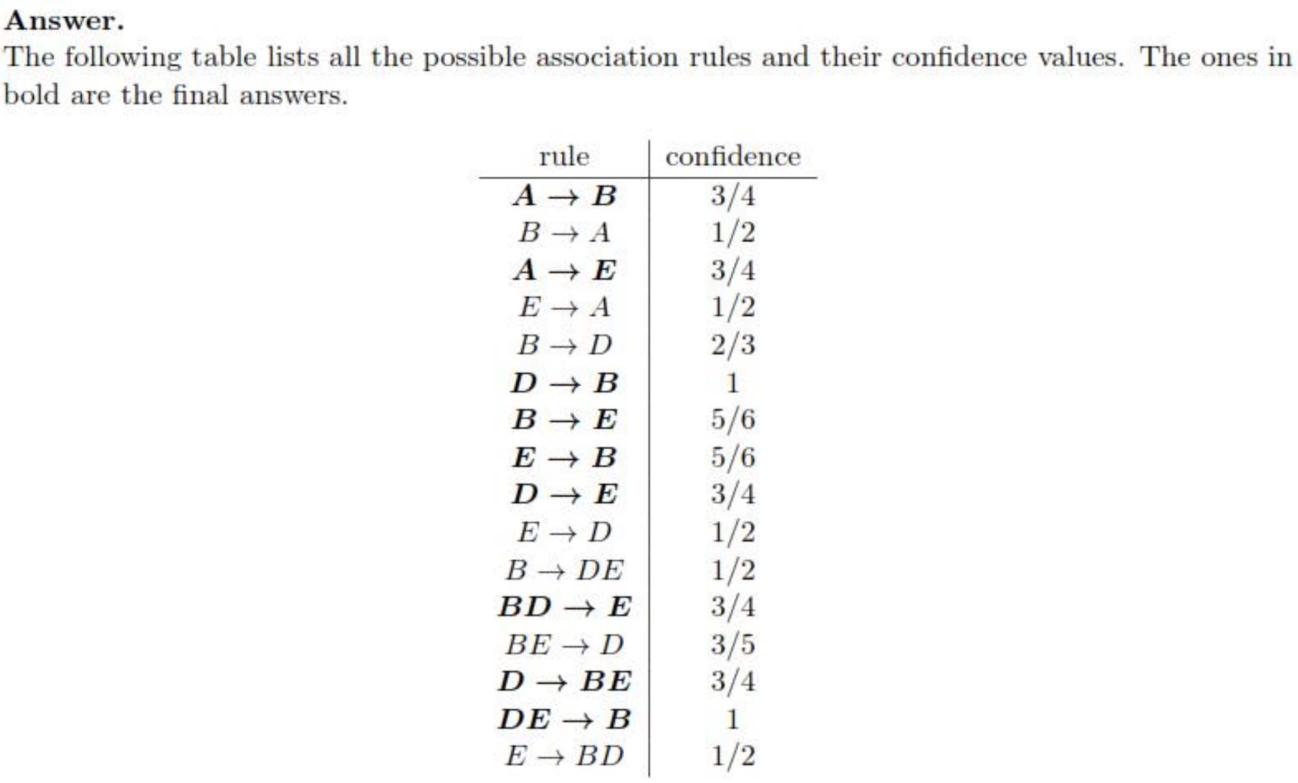
$$C_3 = \{\{A, B, E\}, \{B, D, E\}\}$$

as follows. For each pair of distinct itemsets $\{a_1, a_2\}$ and $\{b_1, b_2\}$ in F_2, the algorithm adds to C_3 an itemset $\{a_1, a_2, b_2\}$ if and only if $a_1 = b_1$. Hence, $\{A, B\}$ and $\{A, E\}$ give rise to $\{A, B, E\}$, whereas $\{B, D\}$ and $\{B, E\}$ give rise to $\{B, D, E\}$.

Finally, the third scan produces:

$$F_3 = \{\{B, D, E\}\}$$

as you can verify easily by yourself. The algorithm terminates here.

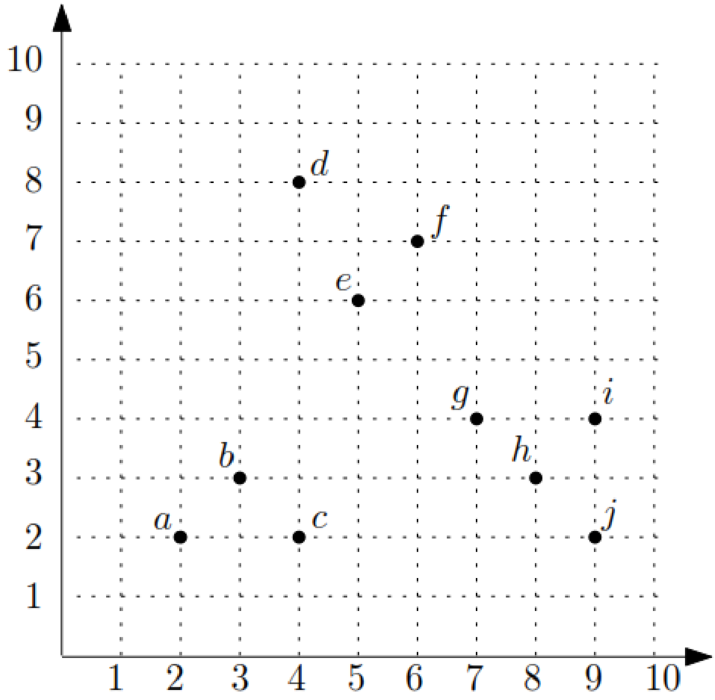


Answer.

The following table lists all the possible association rules and their confidence values. The ones in bold are the final answers.

rule	confidence
$\boldsymbol{A \rightarrow B}$	3/4
$B \rightarrow A$	1/2
$\boldsymbol{A \rightarrow E}$	3/4
$E \rightarrow A$	1/2
$B \rightarrow D$	2/3
$\boldsymbol{D \rightarrow B}$	1
$\boldsymbol{B \rightarrow E}$	5/6
$\boldsymbol{E \rightarrow B}$	5/6
$\boldsymbol{D \rightarrow E}$	3/4
$E \rightarrow D$	1/2
$B \rightarrow DE$	1/2
$\boldsymbol{BD \rightarrow E}$	3/4
$BE \rightarrow D$	3/5
$\boldsymbol{D \rightarrow BE}$	3/4
$\boldsymbol{DE \rightarrow B}$	1
$E \rightarrow BD$	1/2

Exercise IV (K-Means Clustering) (10 pts)



Let P be the set of points given above. Apply the k-means algorithm on P with k = 3 under Euclidean distance. Assume that the algorithm selects a set S = { c, g, h} as the initial centroids.

Pts per iteration (4 + 3 + 3)

Answer.

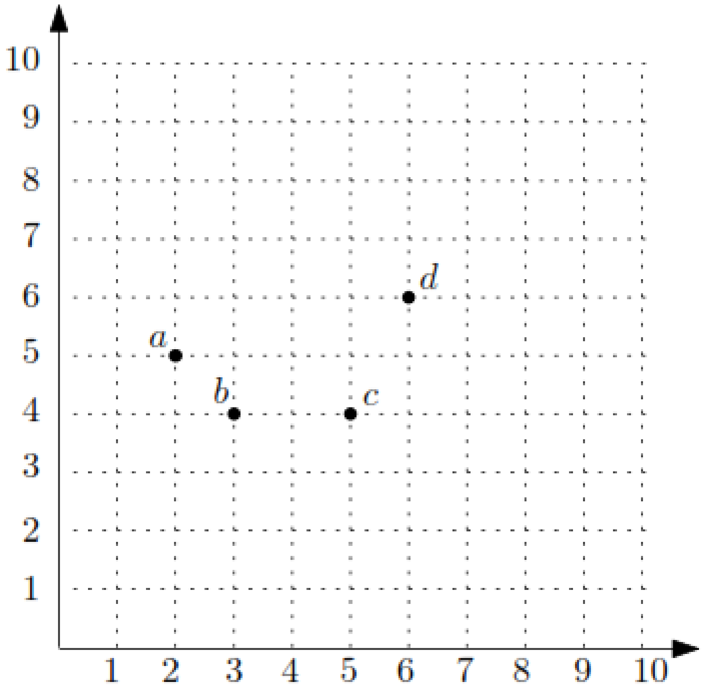
Iteration 1. Let $o_1 = c, o_2 = g, o_3 = h$, namely, the 3 centroids in the initial S. The algorithm divides P into 3 partitions P_1, P_2 and P_3, such that P_i $(1 \leq i \leq 3)$ includes all the points in P that find o_i to be their closest centroids. Specifically, $P_1 = \{a, b, c\}, P_2 = \{d, e, f, g\}$, and $P_3 = \{h, i, j\}$. Then, the algorithm recomputes o_i as the centroid of P_i, for each $1 \leq i \leq 3$, giving $o_1 = (3, 2.33)$, $o_2 = (5.5, 6.25)$, and $o_3 = (8.67, 3)$. $\phi(S)$ is 15.55.

Iteration 2. The algorithm re-divides P into P_1, P_2 and P_3 based on the current centroids. Now, $P_1 = \{a, b, c\}, P_2 = \{d, e, f\}$, and $P_3 = \{g, h, i, j\}$. Accordingly, the centroids are re-computed as $o_1 = (3, 2.33)$, $o_2 = (5, 7)$, and $o_3 = (8.25, 3.25)$. $\phi(S) = 12.17$—the cost is lower than that of the previous iteration.

Iteration 3. After re-dividing P, $P_1 = \{a, b, c\}, P_2 = \{d, e, f\}$, and $P_3 = \{g, h, i, j\}$. The centroids are still $o_1 = (3, 2.33)$, $o_2 = (5, 7)$, and $o_3 = (8.25, 3.25)$, i.e., no change has occurred from the last iteration. The algorithm therefore terminates.

Exercise V (Hierarchical Clustering) (10 pts)

Answer Questions 1-2 based on the following dataset:



Problem 1. Recall that, in discussing hierarchical clustering, we introduced 3 distance metrics on two sets of points: single, complete and average. Let $S_1 = \{a, c\}$ and $S_2 = \{b, d\}$. What is the distance between S_1 and S_2 under those three metrics, respectively (assuming that the distance of two points is calculated by Euclidean distance)? 1 Pt each = 3 Pts

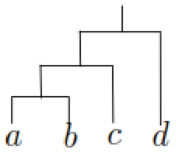
Problem 2. Show the dendrogram returned by the Agglomerative algorithm under the single and complete metrics, respectively. Each dendrogram 3.5 Pts

Answer.
Min: $\sqrt{2}$, as is the distance between a and b.
Max: $\sqrt{17}$, as is the distance between a and d.
Mean: $(\sqrt{2}+\sqrt{17}+2+\sqrt{5})/4$, as is the average of $dist(a,c)$, $dist(a,d)$, $dist(c,b)$ and $dist(c,d)$.

Answer.
Min. At the beginning of the algorithm, each point is regarded as a singleton cluster. In other words, there are 4 clusters, whose mutual distances are given by:

	a	b	c	d
a	-	$\sqrt{2}$	$\sqrt{10}$	$\sqrt{17}$
b	-	-	2	$\sqrt{13}$
c	-	-	-	$\sqrt{5}$

Hence, the algorithm merges S_1 with c into a cluster which we denote as S_2. Now that there are only two clusters left (i.e., S_2 and d), the last merge is trivial. The following dendrogram illustrates the above process.



Max. Repeating the above algorithm with respect to max results in the following dendrogram: