Elements of ML and DS
Exercise 2

Prof. Dr. Wil van der Aalst

Chair of Process and Data Science
RWTH Aachen University

Question 1: Decision Tree

A delivery company asks you to check why some packages are either delivered too early, too late, or on time. They provide you with the data shown in Table 1. You want to create a decision tree to communicate the reasons to the stakeholders. Which attribute should be used in the root node using the entropy-based information gain and why? Do the following:

- Compute the general entropy
- For the descriptive features *Category* and *Logistic Center* compute:
  - The entropy for each feature-value combination
  - The weighted entropy
  - The information gain

Use $\log_2$ for your computations. Round to the third decimal.

Category	Logistic Center	Arrival
Bronze	A	Too early
Silver	A	On-time
Silver	B	Too late
Bronze	B	Too late
Silver	A	Too late
Bronze	A	Too late
Bronze	A	Too early
Silver	A	Too early
Bronze	B	On-time
Silver	B	On-time

Table 1

Question 2: Comparing information gain, gini, and gain ratio

We would like to predict the quality (with values of good and bad) of a bicycle based on two attributes: color (red or blue) and size (small or big). We have a data set of 200 samples, half good and half bad. 50% of good quality bicycles are red, and no bad quality bicycle is red. 75% of the bad quality and 25% of the good quality bicycles are small.

Which attribute should be used as the root of the decision tree when entropy-based information gain is used? Which if we use gini index or information gain ratio?

Question 3: k-Medoids Clustering

Consider the following data set:

Sample	Feature A	Feature B
x_1	5	5
x_2	3	1
x_3	5	2
x_4	1	1
x_5	4	3
x_6	4	1
x_7	1	5
x_8	1	3
x_9	5	1

In this exercise, you are going to perform one iteration of k-medoids clustering with x_1, x_2, and x_3 as initial medoids m_1, m_2, and m_3.

(a) Compute the error values (squared Euclidian distances) between each medoid and non-medoid instance for the initial medoid assignment ($m_1 = x_1$, $m_2 = x_2$, $m_3 = x_3$).

(b) Assign each (non-medoid) instance to a cluster (1, 2, or 3) according to its closest medoid. What is the resulting total error for the given medoids?

(c) Propose another instance in place of x_2 to become the new medoid m_2 so that the total error is reduced. What is the value of the total error after your proposed swap? In your solution, also include your approach.

Question 4: Hierarchical Clustering

Consider the following data set (same as before):

Sample	Feature A	Feature B
x_1	5	5
x_2	3	1
x_3	5	2
x_4	1	1
x_5	4	3
x_6	4	1
x_7	1	5
x_8	1	3
x_9	5	1

In this exercise, you are going to perform agglomerative clustering on the data set above based on the Chebyshev distance. Use the minimum distance between clusters as linkage measure.

(a) Compute the Chebyshev distances between each pair of instances.

(b) Link the (singleton) clusters where the minimum distance between the clusters is minimal across all pairs of clusters. Repeat this step until you joined all instances to a single cluster. Draw the according dendrogram. In your solution, also include your approach.

Total for Question 4: 0

Question 5: Frequent Itemsets

In this exercise, you are going to apply the FP-Growth algorithm to find all frequent itemsets with a minimum support of 0.4 from the following transactions.

Transaction	Items
1	A, B, C
2	B, D
3	A, C, E, F
4	A, C, E, F, G
5	B, C, E
6	A, C, D, E, F
7	A, B, C, G
8	A, E, F, H
9	C, D, E
10	A, B, C, G, H

(a) Compute the support of each item. Drop items below the minimum support threshold and sort the remaining items in descending order of support. Provide the transformed list of transactions.

(b) Construct the FP-tree. In addition, show how you can read the transactions from the FP-tree.

(c) For each frequent item, create a conditional FP-tree and mine all frequent itemset with this item as postfix. Provide all identified frequent itemsets with their support. In your solution, also include your approach.

Question 6: Sequence Mining

Apply the Apriori-All algorithm with a minimum support of 0.25 to find all maximal frequent sequencial patterns in the following set of transaction sequences.

Customer	Customer Sequence
1	$\langle\{C\},\{I\}\rangle$
2	$\langle\{A,B\},\{C\},\{D,F,G\}\rangle$
3	$\langle\{C,E,G\}\rangle$
4	$\langle\{C\},\{D,G\},\{I\}\rangle$
5	$\langle\{I\}\rangle$

(a) Determine all litemsets and provide the transformed set of transaction sequences $\mathcal{X}_T$.

(b) Apply the algorithm to find all frequent sequential patterns $\bigcup_k \mathcal{L}_k$. Which of them are maximal? In your solution, also include your approach.