Data Mining 2

Module 4 - 2020/2021

Name _____ **Surname** _____ **ID:** _____ **Test id. AUTO**

Q1. Letting S_1 be a subsequence of a frequent sequence S_2, refresh why also S_1 is a frequent one.

A1. _____

Q2. Given the following sets of elements, run the GSP algorithm: once you find the candidate 3-sequences, write down which one/s is/are pruned and which one/s is/are the frequent sequences.

$\{DC\}\{CD\}\{D\}\{C\}\{A\}$
$\{A\}\{B\}\{C\}\{E\}$
$\{AD\}\{C\}\{C\}\{CE\}$
$\{C\}\{E\}\{E\}\{A\}$

A2. _____

Q3. Assume that in the following tracking sequence H=home, F=friend's house and X=other, then assume that the elements at time $t > 3$ (highlighted in red) occur after an imposed goverment lockdown aiming to limitate the $\{H\} \rightarrow \{F\}$ sequence. Is it better to impose $gap \geq 3$ or $gap \leq 3$ in order to focus on the forbidden sequence after the lockdown? Explain your answer.

$\{H,F\}\{H\}\{H,F,X\}\{H,X\}$ $\{H\}\{H\}\{H,X\}\{H,F\}$

A3. _____

Q4. Identify the wrong statements about the EM algorithm.

1) Cluster assignment is more flexible than kmeans-like approaches

2) It is not able to cluster points when more than two generative processes are involved

3) Probability of data to belong to each distribution is estimated during the E-step

4) Dependence of data is always assumed

5) It computes the model parameters until convergence is reached

A4. _____

N.B.: this question can have more than one correct answer

Q5. Identify the right statements about the OPTICS algorithm.

1) It extends hierarchical-based algorithms

2) Core distance is updated until all points are comparable to each other

3) Core distance defines the number of minimum MinPts to consider

4) It is not parametric with respect to the radius value

5) It works when heterogenous densities are present in the dataset

A5. _____

N.B.: this question can have more than one correct answer

Q6. Given the following sets of elements, apply the ROCK clustering assuming a similarity threshold of 0.15 and 2 required clusters.

$P_1 = \{cap, sunglasses, shoes\}$
$P_2 = \{pants, shoes, shirt, sunglasses\}$
$P_3 = \{chicken, pants\}$
$P_4 = \{shoes, shirt, cap\}$

A6. _____

Q7. Given the following partitions, evaluate their goodness using the Profit as a fitness function ($r = 2$)

Partition 1
$C_1((c, c), (c, e), (c, c, e, e), (e, e)$
$C_2((d, e), (e, d), (h, e, d), (e, e))$

Partition 2
$C_1((c, e, c), (e, c, e))$
$C_2((d, e, h), (e, e, e))$

A7. _____