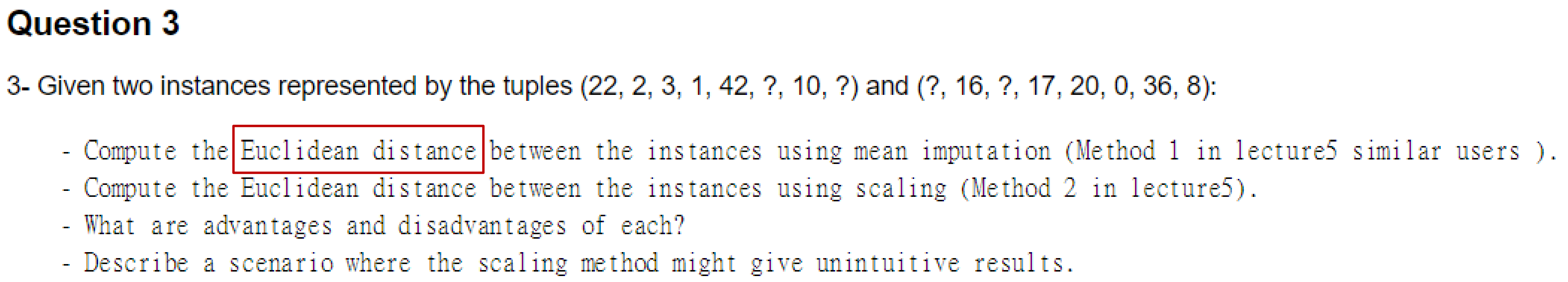
2018 COMP20008 Workshop 4

Follow up on workshop 3



Question 3

3- Given two instances represented by the tuples (22, 2, 3, 1, 42, ?, 10, ?) and (?, 16, ?, 17, 20, 0, 36, 8):

- Compute the Euclidean distance between the instances using mean imputation (Method 1 in lecture5 similar users).
- Compute the Euclidean distance between the instances using scaling (Method 2 in lecture5).
- What are advantages and disadvantages of each?
- Describe a scenario where the scaling method might give unintuitive results.

Scaling method:

Distance measure: (1) Euclidean Distance or (2) Squared Euclidean Distance

- U1 = (2, 1, 42, 36)
- U2 = (16, 17, 20, 10)
- Dissimilarity = **Distance measure(U1, U2)** * Total number of data pairs / Total number of non-missing data pairs

Question 3

(1) Euclidean Distance

- Dissimilarity = **Euclidean Distance**(U1, U2) * Total number of data pairs / Total number of non-missing data pairs
- In the last workshop, we used:

Dissimilarity = **Euclidean Distance**(U1, U2) * **Square root** (Total number of data pairs / Total number of non-missing data pairs)

*If the distance measure is **Euclidean Distance** (instead of **Squared Euclidean Distance**), **Square root** is NOT needed.*

(2) Squared Euclidean Distance

- Dissimilarity = **Squared Euclidean Distance**(U1, U2) * Total number of data pairs / Total number of non-missing data pairs