COMP40370

Tutorial: Data Preprocessing

Prof. Tahar Kechadi

Question 1

Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order)

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

1. What is the mean of the data? What is the median?
2. What is the mode of the data? Comment on the data modality (i.e., bimodal, tri-modal, etc.)
3. What is the midrange of the data?
4. Can you find (roughly) the 1st quartile (Q1) and the 3rd quartile (Q3) of the data?
5. Give the five-number summary of the data
6. Show a boxplot of the data
7. How is the q-q plot different from a quantile plot?

Question 2

Using the data for age given in Question 1, answer the following:

1. Use min-max normalization to transform the value 35 for age onto the range [0.0…1.0].
2. Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years?
3. Use normalization by interquartile range to transform the value 35 for age?
4. Comment on which method you would prefer to use for the given data, giving reasons as to why.