**COMP4433 Data Mining and Data Warehousing**

**FAQ on Data Preprocessing (with suggested answers)**

1. Suppose that the data available for data mining include a list of prices of commonly sold items at an electronics store. The numbers have been sorted as follows: 3.95, 4.12, 5.13, 5.84, 6.63, 7.92, 8.14, 8.78, 9.21, 10.88, 11.11, 12.12.
   a) If the "smoothing by bin means" is used to smooth the above data. What are the results if the equi-depth partitioning with 3 bins is used? List the price values that go to each bin. Is the average of the smoothed numbers the same as the average before smoothing?

   Ans.

   Using equi-depth partitioning, we have:
   Depth of each interval=12/3=4; hence,
   Bin 1: 3.95, 4.12, 5.13, 5.84
   Bin 2: 6.63, 7.92, 8.14, 8.78
   Bin 3: 9.21, 10.88, 11.11, 12.12
   Smoothing by bin means:
   Mean of Bin1 is (3.95 + 4.12 + 5.13 + 5.84)/4=4.76,
   mean of Bin 2 is (6.63 + 7.92 + 8.14 + 8.78)/4=7.8675 and
   mean of Bin3 is (9.21 + 10.88 + 11.11 + 12.12)/4=10.83; hence,
   Bin 1: 4.76, 4.76, 4.76, 4.76
   Bin 2: 7.87, 7.87, 7.87, 7.87
   Bin 3: 10.83, 10.83, 10.83, 10.83
   The average is expected to be the same before and after smoothing

   b) Repeat a) using "smoothing by bin boundaries". Compute the new average price and compare it with the original average and the average computed for a). Explain if there is any difference.

   Ans.

   Smoothing by bin boundaries:
   Bin 1: 3.95, 3.95, 5.84, 5.84
   Bin 2: 6.63, 8.78, 8.78, 8.78
   Bin 3: 9.21, 12.12, 12.12, 12.12
   The average is expected to be different before and after smoothing

   c) Repeat a) using the equi-width partitioning method, also with 3 bins. What are the new partitioning results if 93.95 is added the above list?

   Ans.

   Using equi-width partitioning, we have
   Coverage=12.12-3.95=8.17; Width of each interval=8.17/3=2.72; hence,
   Bin 1 [3.95-6.67]: 3.95, 4.12, 5.13, 5.84, 6.63
   Bin 2 (6.67-9.39]: 7.92, 8.14, 8.78, 9.21
   Bin 3 (9.39-12.12]: 10.88, 11.11, 12.12
   Smoothing by bin means:
   Mean of Bin1 is (3.95 + 4.12 + 5.13 + 5.84 + 6.63)/5=5.134,
   mean of Bin 2 is (7.92 + 8.14 + 8.78 + 9.21)/4=8.5125 and
   mean of Bin3 is (10.88 + 11.11 + 12.12)/4=11.37; hence,
   Bin 1: 5.134, 5.134, 5.134, 5.134, 5.134
   Bin 2: 8.5125, 8.5125, 8.5125, 8.5125

Bin 3: 11.37, 11.37, 11.37
The average is expected to be the same before and after smoothing

When 93.95 is introduced, we have
Coverage=93.95-3.95=90; Width of each interval=90/3=30; hence,
Bin 1 [3.95-33.95]: 3.95, 4.12, 5.13, 5.84, 6.63, 7.92, 8.14, 8.78, 9.21, 10.88, 11.11, 12.12
Bin 2 (33.95-63.95]: empty
Bin 3 (63.95-93.95]: 93.95

2. Suggest an effective method to determine the missing values below. Fill in the missing values accordingly.

| Patient ID | Blood Pressure Level | Sex | Age | Fever | Disease |
|---|---|---|---|---|---|
| 9100123 | 80-120 | Male | 65 | Yes | No |
| 9303034 | 160-200 | Female | 55 | No | Yes |
| 9210126 | 80-120 | Male | 12 | Yes | Yes |
| 9142020 | 120-160 | Female | 35 | No | No |
| 9910111 | 160-200 | Male | 46 | No | Yes |
| 9576732 | 80-120 | Male | 16 | Yes | No |
| 9910115 | 160-200 | Female | | No | Yes |
| 9210120 | 120-160 | | 23 | | |
| 9576737 | | | 28 | Yes | No |

Ans. By using attribute mean/mode for the same class (disease) or other attribute value (e.g. Blood pressure level), one might have the answers filled below.

| Patient ID | Blood Pressure Level | Sex | Age | Fever | Disease |
|---|---|---|---|---|---|
| 9100123 | 80-120 | Male | 65 | Yes | No |
| 9303034 | 160-200 | Female | 55 | No | Yes |
| 9210126 | 80-120 | Male | 12 | Yes | Yes |
| 9142020 | 120-160 | Female | 35 | No | No |
| 9910111 | 160-200 | Male | 46 | No | Yes |
| 9576732 | 80-120 | Male | 16 | Yes | No |
| 9910115 | 160-200 | Female | 37.7 | No | Yes |
| 9210120 | 120-160 | Female | 23 | No | No |
| 9576737 | 80-120 | Male | 28 | Yes | No |

3. a) Normalize the following two time series subsequences so that they can be compared effectively.

| Time | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|
| Subsequence 1 | 130 | 135 | 130 | 125 | 130 |
| Subsequence 2 | 5 | 5.5 | 6 | 5 | 5.5 |

Ans
By using min-max normalization, we have

| Time | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| Subsequence 1 | 0.5 | 1 | 0.5 | 0 | 0.5 |
| Subsequence 2 | 0 | 0.5 | 1 | 0 | 0.5 |

Some would like to use the following

| Time | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|
| Subsequence 1 | 130/130 | 135/130 | 130/130 | 125/130 | 130/130 |
| Subsequence 2 | 5/5 | 5.5/5 | 6/5 | 5/5 | 5.5/5 |

b) Propose a method other than the attribute mean method to fill the missing value of the following time series subsequence with T5-T4=T4-T3=T3-T2=T2-T1, equal time intervals.

| Time | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|
| Subsequence 1 | 135 | 140 | ? | 130 | 130 |

Ans
One may use linear interpolation.
E.g., by taking say 1 neighbor before and after respectively, we can fill the T3 value as (140+130)/2=135.
E.g., by taking say 2 neighbor before and after respectively, we can fill the T3 value as (135+140+130+130)/2=133.75.