COMP4433 Data Mining and Data Warehousing

FAQ on Data Preprocessing (Questions only)

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?

   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.

   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?

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

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

   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