# CENG 414
# Introduction to Data Mining
# Fall 2019-2020
# THE 2

## Question 2

*1- Which method did Discretize use to discretize attribute values?*

Discretize filter is an instance filter. A range of numeric attributes are discretized into nominal attributes via this filter. Equal width binning method is used in this filter, possibly width of first and last bins are not equal to others.(first bin is starting from -inf and last bin is going to +inf). Equal width binning is an unsupervised technique that converts numerical data to categorical data.

*2- When you compare the original distribution of ”preg” attribute with the distribution after applying the function, explain one of the difference you observed.*

Before applying the Discretize filter to “preg” attribute, in the shown graphics for the dataset is divided into 13 bins and width of these bins is ( 17(max value) - 0(min value) ) / 13 = 1.308. Also, these bins seem in a contiguous manner.
However, when Discretize filter is applied, the bin number is reduced to 10(default). Now, bin size is ( 17(max value) - 0(min value) ) / 10 = 1.7. Also, all bins are shown in a separated way.

*3- Assume you are given a hypothetical dataset called ”health” and assume all the values of ”age” attribute from this dataset are given as follows:*
*24,15,25,28,4,21,8,26,9,21,34,29.*
*You are expected to apply binning methods on this attribute:*
*a. Apply equal-width binning method to discretize the values where bin size is 3. Show the resulting bins.*
*b. Apply equal-depth binning method to discretize the values. Show the resulting bins.*

Sorted dataset is 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34.

Equal-width binning means that all bins’ width is same.
(Max element - min element) / bin size = # of bins
(34-4) / 3 = 10.

Then 10 bins should created.

| Bin # | Bin Range | Elements in Bin |
|---|---|---|
| Bin1 | (-inf, 7] | 4 |
| Bin2 | (7 , 10] | 8, 9 |
| Bin3 | (10,13] | |
| Bin4 | (13,16] | 15 |
| Bin5 | (16,19] | |
| Bin6 | (19,22] | 21,21 |
| Bin7 | (22,25] | 24,25 |
| Bin8 | (25,28] | 26,28 |
| Bin9 | (28,31] | 29 |
| Bin10 | (31,34] | 34 |

In equal-depth binning method, all bins(approximately) have same number of elements. In this situation, there are 12 data in the dataset. That means this dataset can be divided into 2,3,4,6,12 bins and bins have exactly same number of item. For other values, some bins will not have same number of item. For example, if 4 bins are created:

| Bin # | Bin Range | Elements in Bin |
|---|---|---|
| Bin1 | (-inf, 9] | 4,8,9 |
| Bin2 | (9 , 21] | 15,21,21 |
| Bin3 | (21,26] | 24,25,26 |
| Bin4 | (26,34] | 28,29,34 |

*4- What is the difference between equal-depth binning and equal-width binning method? Which one of the methods do you prefer to work with numerical attributes?*

In equal-depth binning method, all bins have same number of element as far as possible but in different ranges. In equal-width method, ranges of all bins are same although number of elements in the bins may be different. Equal width binning is the more appropriate way since it shows the overall distribution and most dense bins.

*Alper KOCAMAN - 2169589*