Geraldo Braho
COMP 4353 O.Data Mining
North American University

1. Suppose that the data for analysis include 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.

a. Use smoothing by bin means to smooth the above data, using a bin depth of 3.

First we need to sort the data but since our data tuples is already sorted we do no need to do that. Sort the data.

Then partition the data into equidepth bins of depth 3.

Bin 1	13, 15, 16
Bin 2	16,19,20
Bin 3	20,21,22
Bin 4	22, 25, 25
Bin 5	25,25,30
Bin 6	33,33,35
Bin 7	35, 35, 35
Bin 8	36,40,45

Bin 9	46,52,70

The we will need to Calculate the arithmetic mean of each bin and then we Replace each of the values in each bin by the arithmetic mean calculated for the bin.

Bin 1	142/3, 142/3, 142/3
Bin 2	181/3, 181/3, 181/3
Bin 3	21, 21, 21
Bin 4	24, 24, 24
Bin 5	262/3, 262/3, 262/3
Bin 6	332/3, 332/3, 332/3
Bin 7	35, 35, 35

Bin 8	401/3, 401/3, 401/3
Bin 9	56, 56, 56

(b) How might you determine outliers in the data?

Outliers in the data may be detected by clustering, where similar values are organized into groups, or ‘clusters’. Values that fall outside of the set of clusters may be considered outliers. Alternatively, a combination of computer and human inspection can be used where a predetermined data distribution is implemented to allow the computer to identify possible outliers. These possible outliers can then be verified by human inspection with much less effort than would be required to verify the entire initial data set.

(c) What other methods are there for data smoothing?

Other methods that can be used for data smoothing include alternate forms of binning such as smoothing by bin medians or smoothing by bin boundaries. Alternatively, equiwidth bins can be used to implement any of the forms of binning, where the interval range of values in each bin is constant. Methods other than binning include using regression techniques to smooth the data by fitting it to a function such as through linear or multiple regression. Also, classification techniques can be used to implement concept hierarchies that can smooth the data by rolling-up lower level concepts to higher-level concepts.

Using the data for *age* given in Exercise 3.3, answer the following:

(a) Use min-max normalization to transform the value 35 for *age* onto the range [0.0, 1.0].

Using the corresponding equation with
minA = 13,
maxA = 70,
new minA = 0,

new maxA = 1.0,
Then v = 35 is transformed to v′ = 0.39.

(b) Use z-score normalization to transform the value 35 for *age*, where the standard deviation of *age* is 12.94 years.

Using the corresponding equation where

A = 809/27 = 29.96
σA = 12.94,
v = 35 is transformed to v′ = 0.39

(c) Use normalization by decimal scaling to transform the value 35 for *age*.

Using the corresponding equation where
j = 2,
v = 35 is transformed to v′ = 0.35

(d) Comment on which method you would prefer to use for the given data, giving reasons as to why.

Given the data, one may prefer decimal scaling for normalization as such a transformation would maintain the data distribution and be intuitive to interpret, while still allowing mining on specific age groups. Min-max normalization has the undesired effect of not permitting any future values to fall outside the current minimum and maximum values without encountering an “out of bounds error”. As it is probable that such values may be present in future data, this method is less appropriate. Also, z-score normalization transforms values into measures that represent their distance from the mean, in terms of standard deviations. It is probable that this type of transformation would not increase the information value of the attribute in terms of intuitiveness to users or in usefulness of mining results.