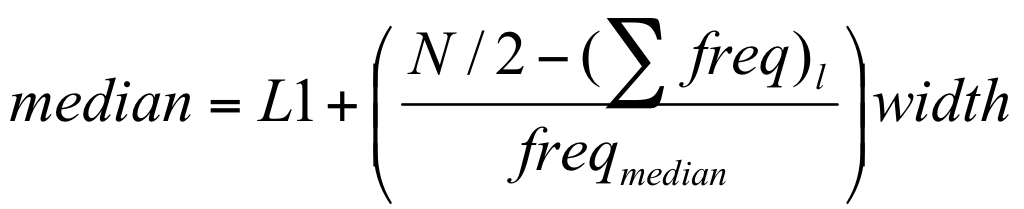
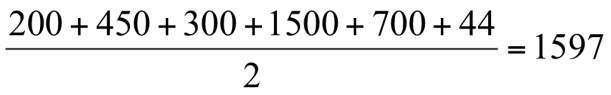
Data Mining Assignment 2

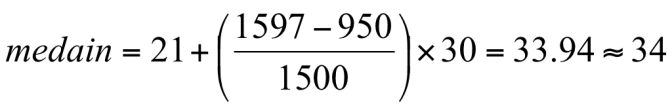
Liqi Zhu

2.3





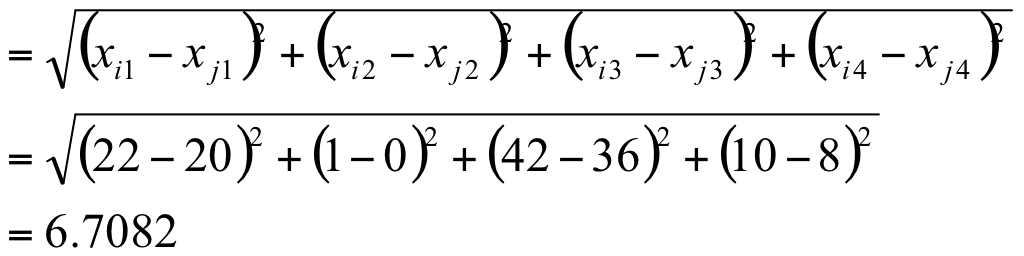
So the median interval is 21-50 age group



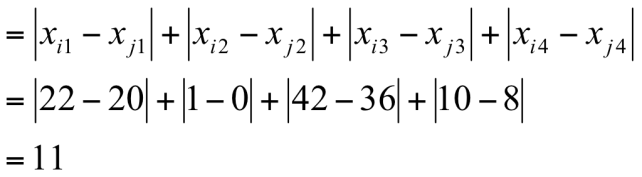
2.6

(22,1,42,10) (20,0,36,8)

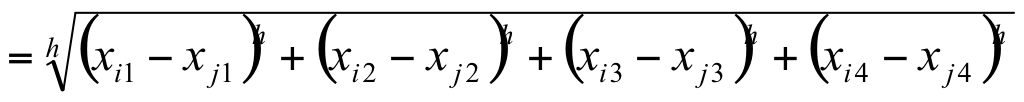
1. Euclidean distance



1. Manhattan distance



1. Minkowski distance

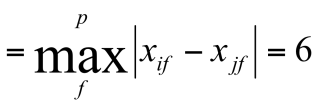




When h=3

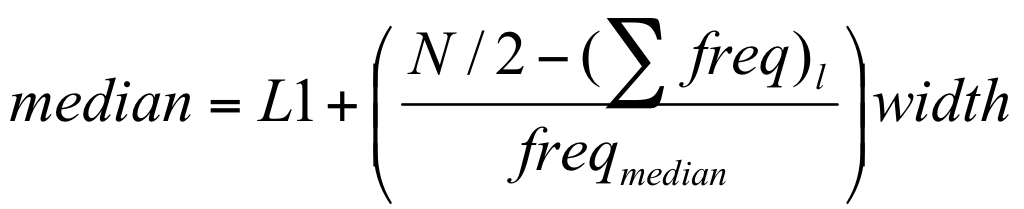
Minkowski distance==

1. Supremum distance



2.7

With all kinds of data sets, the most straight forward method to do approximation is the method used for question 2.3:

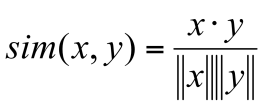


where L1 is the lower boundary of the median interval, N is the number of values in the entire data set, ( freq)l is the sum of the frequencies of all of the intervals that are lower than the median interval, freqmedian is the frequency of the median interval, and width is the width of the median interval.

One other approach is that in a huge data set, divide the data into several intervals, then divide the interval into sub intervals, and another level of sub intervals if needed until you have enough intervals, when doing calculations, use the same formula in question 2.3. Calculate the median of sub intervals from the lowest level to the first level to find the median of the data set.

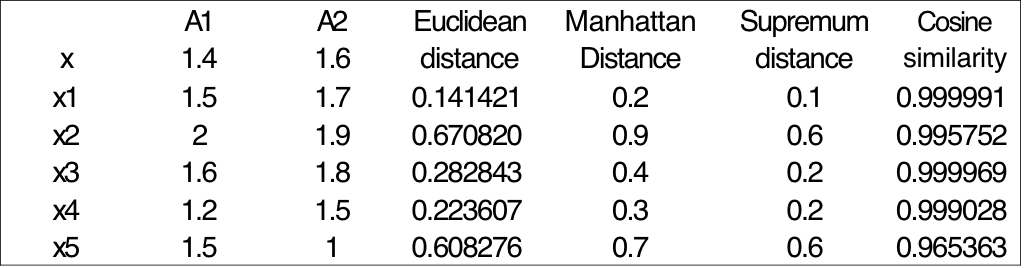
As we know the number of errors decreased as the number of the intervals increases. And obviously time used to process will increase. Since the time used in the calculation is a good index of complexity, the number of errors is a good index for accuracy. So I propose that the product of time used times the number of errors is a good measure to find the most balanced one with acceptable error amount and time.

2.8

1. Cosine similarity 

With the formulas in 2.6

The results:



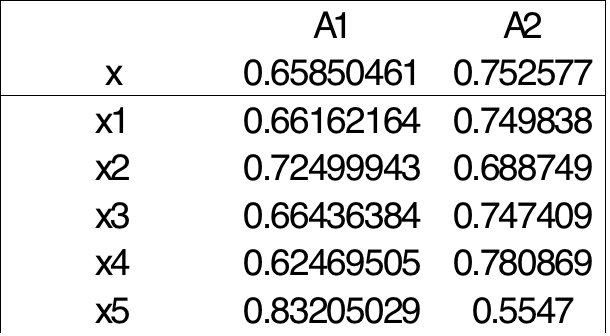
Similarity using Euclidean distance: x1>x4>x3>x5>x2

Manhattan distance: x1>x4>x3>x5>x2

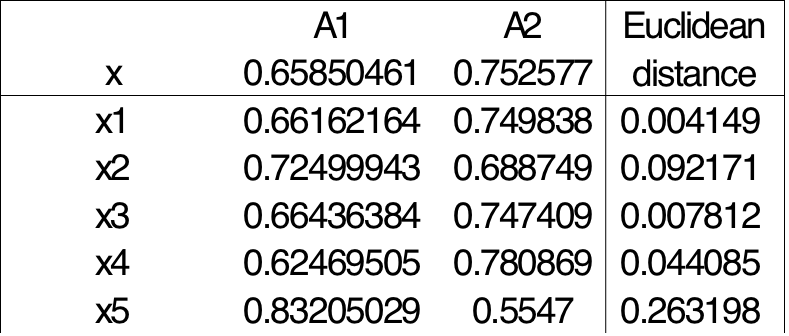
Supremum distance: x1>x4,x3>x5,x2

Cosine similarity: x1>x3>x4>x2>x5

(b) After normalization Data points:



Based on Euclidean distance formula, the results:



So the similarity after normalization: x1>x3>x4>x2>x5

3.1

Accuracy: The degree to which data correctly describes the "real world" object or event being described.

Completeness:The proportion of stored data against the potential of "100% complete".

Consistency: The absence of difference, when comparing two or more representations of a thing against a definition.

Data quality depends on the intended use of the data. For example, for any big company in the fast moving customer industry. It’s so important to be consistent in data values among all departments so that every department is able to accomplish its function properly. Any mistake in database of sales volume or price can cause the false auditing and financial problems. If the data miss values then there must be additional balance sheet issues. Any problem above can cause false marketing strategy and false budget.

And also the customer’s data directly reflects the status of the market. If there’s error in database such as wrong value, missing value or conflicting values then it’s more likely to cause loss since there’s no clear image of the market. Therefore, there’s companies putting so much emphasis on the quality of the data with intended use of data.

Other dimensions that can be used to assess the quality of data include timeliness, believability, value added, interpretability and accessability:

Timeliness: Data must be available within a time frame that allows it to be useful for decision making.

Believability: Data values must be within the range of possible results in order to be useful for decision making.

Value added: Data must provide additional value in terms of information that offsets the cost of collecting and accessing it.

Interpretability: Data must not be so complex that the effort to understand the information it provides exceeds the benefit of its analysis.

Accessibility: Data must be accessible so that the effort to collect it does not exceed the benefit from its use.

3.3

(a)

1.Partition into (equal-frequency) bins:

Bin 1: 13, 15, 16 Bin 4: 22, 25, 25 Bin 7: 35, 35, 35

Bin 2: 16, 19, 20 Bin 5: 25, 25, 30 Bin 8: 36, 40, 45

Bin 3: 20, 21, 22 Bin 6: 33, 33, 35 Bin 9: 46, 52, 70.

1. Calculate the mean value of each bin.
2. Replace the value with the mean value.

Smoothing by bin means:

Bin 1: 14.67, 14.67, 14.67 Bin 4: 24, 24, 24 Bin 7: 35, 35, 35

Bin 2: 18.33, 18.33, 18.33 Bin 5: 26.67, 26.67, 26.67 Bin 8: 40.33, 40.33, 40.33

Bin 3: 21, 21, 21 Bin 6: 33.67, 33.67, 33.67 Bin 9: 56, 56, 56.

(b)

Outliers may be detected by clustering, for example, where similar values are organized into groups, or ‘clusters.’ Intuitively, values that fall outside of the set of clusters may be considered outliers.

(c)

Other methods that can be used include forms of binning such as smoothing by bin medians or smoothing by bin boundaries. Equal-width bins can be used to all forms of binning, interval range of values in each bin is constant.

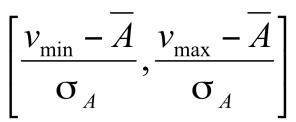
Methods other than binning include using regression techniques to smooth the data by fitting it to a function such as through linear or multiple regression. Classification techniques can be used to implement concept hierarchies that can smooth the data by rolling-up lower level concepts to higher-level concepts.

3.5

(a) min-max normalization

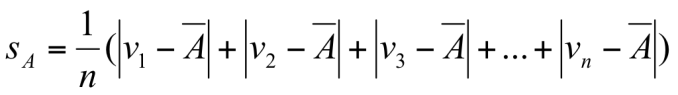
Value range: [new min, new max]

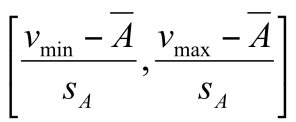
(b) z-score normalization

Value: 

Value Range: (-∞,+∞)

(c) z-score normalization using the mean absolute deviation instead of standard deviation



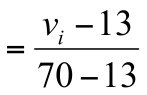
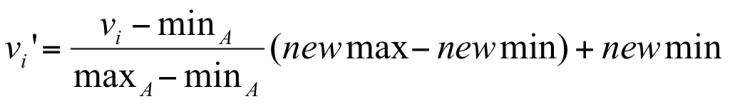
Value: 

Value Range: (-∞,+∞)

(c) normalization by decimal scaling

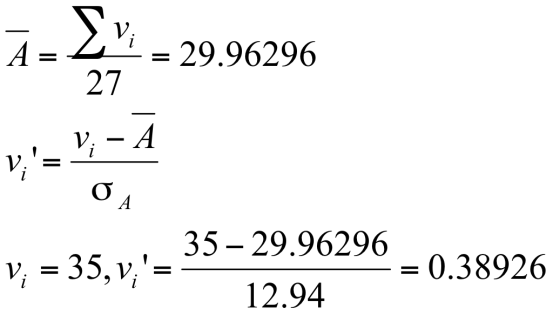
Value range: (-1,1).

3.7

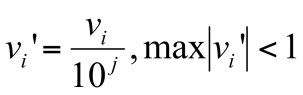


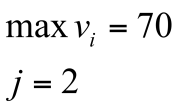
vi=35 vi’=0.385964912

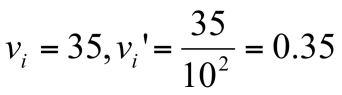
(b)



(c)





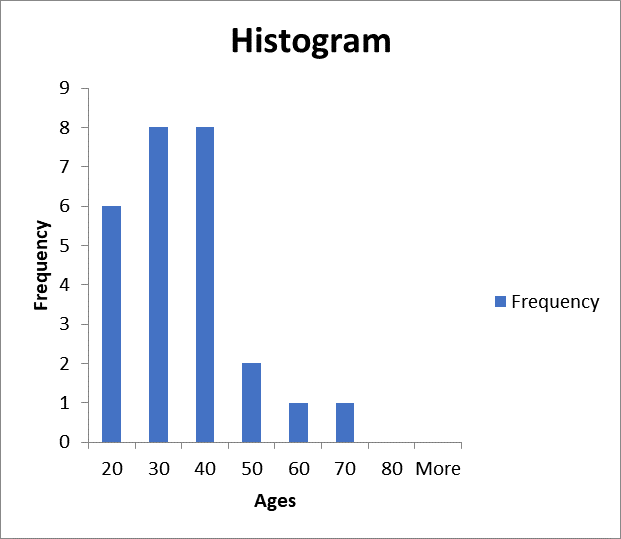


(d)

Because this conversion will keep the data distributed and intuitively interpreted, easy to adjust if there’s new value and allow further mining process in all age groups since it doesn’t affect distribution. Minimal - Maximum Normalization has an undesirable effect that does not allow any future value to fall outside the current minimum and maximum values, there can be new data in the future exceeds the current value range. Z Fractional Normalization’s transformation represents the distance between value and the mean, measured as a standard deviation. Compared to decimal scaling, this kind of normalization may not be intuitive enough.

3.11

(a)



(b)











3.13

Refer to assignment2.py uploaded.