HUMAN-CENTERED INFORMATION AND DATA MINING

HOMEWORK 1

STUDENT:

108368017   ZI-YANG LIN

ADVISOR:

JENQ-HAUR WANG

National Taipei University of Technology

2019

2.4: Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

Age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

Age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) Calculate the mean, and median of age and %fat.

Ans:

Age mean: 46.44444444444444,

Age median: 51.0

fat mean: 28.783333333333328,

fat median: 30.7

2.8: It is important to define or select similarity measures in data analysis. However, there is no commonly accepted subjective similarity measure. Results van vary depending on the similarity measures used. Nonetheless, seemingly different similarity measures may be equivalent after some transformation. Suppose we have the following 2-D data set:

	A1	A2
x1	1.5	1.7
x2	2	1.9
x3	1.6	1.8
x4	1.2	1.5
x5	1.5	1.0

(a) Consider the data as 2-D data points. Given a new data point, x=(1.4,1.6) as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.

Ans:

Euclidean distance: (x2, 0.67) > (x5, 0.608) > (x3, 0.28) > (x4, 0.22) > (x1, 0.14)

Manhattan distance: (x2, 0.89) > (x5, 0.7) > (x3, 0.4) > (x4, 0.3) > (x1, 0.19)

Supremum distance: (x2, 0.6) = (x5, 0.6) > (x3, 0.2) > (x4, 0.19) > (x1, 0.1)

Cosine similarity: (x1, 0.99999) > (x3, 0.99996) > (x4, 0.999) > (x2, 0.995) > (x5, 0.96)