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)

3.8: Using the data for age and body fat given in Exercise 2.4, answer the following:

(a) Normalize the two attributes based on z-score normalization.

(b) Calculate the correlation coefficient (Pearsons product moment coefficient). Are these attributes positively or negatively correlated? Compute their covariance.

Ans:

(a)

-1.825	-1.825	-1.513	-1.513	-0.579	-0.423	0.043	0.198	0.276
0.432	0.588	0.588	0.743	0.821	0.899	0.899	1.055	1.133

(b) Pearsons correlation coefficient: 0.817, so age and %fat are positively correlated.

$$covariance = \begin{bmatrix} 174.732 & 100.0196 \\ 100.0196 & 85.643 \end{bmatrix}$$

3.9: Suppose a group of 12 sales price records has been sorted as follows:

5	10	11	13	15	35
50	55	72	92	204	215

Partition them into three bins by each of the following methods:

(a) equal-frequency (equal-depth) partitioning

(b) equal-width partitioning

Ans:

(a)

equal-frequency partitioning:

$$[5 \quad 10 \quad 11 \quad 13]$$
$$[15 \quad 35 \quad 50 \quad 55]$$
$$[72 \quad 92 \quad 204 \quad 215]$$

(b)

equal-width partitioning:

$$[10 \quad 11 \quad 13 \quad 15 \quad 35 \quad 50 \quad 55 \quad 72]$$
$$[92]$$
$$[204]$$