Educational Data Mining and Applications: HW#1

By J. H. Wang

Oct. 3, 2023

Homework #1

- Chap.2:
  - 2.2(e)(f)
  - 2.8(a)(b)
- Chap.3:
  - 3.3(a)(b)
  - 3.7(a)(b)
  - 3.8(b)
  - 3.11(a)
- Due: 2 weeks (Oct. 17, 2023)

Exercises for Chap.2

- 2.2: Suppose that the data for analysis includes 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.
- (e) Give the *five-number summary* of the data.
- (f) Show a *boxplot* of the data.

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

  (to be continued…)

- (… continued from the previous slide)

  Suppose we have the following 2-D data set:

  - (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.

- (to be continued…)

	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

- (… continued from the previous slide)

- (b) Normalize the data set to make the (Euclidean) norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points.

Exercises for Chap.3

- 3.3: Exercise 2.2 gave the following data (in increasing order) for the attribute *age*: 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 these data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.
- (b) How might you determine *outliers* in the data?

- 3.7: 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].
- (b) Use z-score normalization to transform the value 35 for *age*, where the standard deviation of *age* is 12.94 years.

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

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

  - (b) Calculate the *correlation coefficient* (Pearson's product moment coefficient). Are these two attributes positively or negatively correlated? Compute their covariance.

- 3.11: Using the data for *age* given in Exercise 3.3,
  - (a) Plot an equal-width histogram of width 10.

Homework Submission

- For hand-written exercises, please hand in your homework in class (paper version)
  - Remember to specify your name and student ID
- For those who cannot come to class, please scan or type your answer for the homework in an electronic file and submit it online to iSchool+
  - Under the item [Assignments]\[HW#1]

Thanks for Your Attention!