CENG 414
Introduction to Data Mining
Fall 2019
HW1
Due October 30th, 2019, 23:59

Submit the softcopy to MetuClass.

Answer the following questions briefly (in a few sentences). Show all steps of your calculations, if any.

1. Classify each of the following as nominal-level, ordinal-level, interval-level, or ratio level measurement:

a) Telephone numbers
b) Ratings of fiction books-excellent, good, fair, poor
c) Amounts of money spent on a medical checkup
d) Scores on a data mining exam
e) Blood Types
f) The ages of METU students
g) The movie ratings
h) Colors of METU t-shirts in ODTUDEN Shop
i) Temperatures of hot tubs in local health clubs

2. Why are random numbers used in sampling?

3. Identify the independent variable and the dependent variable in the following:

a) According to the Journal of Health, a regular 30-minute workout could slash your risk of catching a cold by 43%.

b) A research study stated that meditation helps people make more rational decisions.

4. Draw histograms for a positively skewed, a negatively skewed and a symmetric distribution. Show the positions of mode, median and mean for each of them.

5. For the following situations, state which measure of central tendency - mean, median, or mode- should be used.

a) The most typical case is desired.
b) The distribution is open-ended.
c) There is an extreme value in the data set.
d) The data are categorical.
e) Further statistical computations will be needed.

6. What is the value of the mode when all values in the data set are different?

7. Find the mean, median, mode, and midrange for the data
   59, 52, 28, 26, 19, 19, 18, 17, 17, 17.

8. Find the sample variance and the sample standard deviation for the data
   9, 10, 14, 7, 8, 3.

9. Find the range, variance and standard deviation for the data
   33, 10, 62, 132, 123, 316, 123, 133, 18, 150, 26, 138.

10. Let A and B be two mutually exclusive events? Are A and B independent events? Explain your answer.

11. What are the characteristics of a normal distribution? Give the formula and explain briefly what affects the shape and the position of a normal distribution curve?

12. What is the standard normal distribution? What is the total area under the normal distribution curve?

13. What is a z-test? What is a t-test? What is a chi-square test?
    When are they used?

14. Which test is used to compare two variances or standard deviations?

15. What is meant when the relationship between the two variables is called positive/negative?

16. What is meant by the explained variation, the unexplained variation and the total variation?

17. Define the correlation coefficient. What will be the value of the linear correlation coefficient if there is no relationship between the variables?
18. Which of the following measures of central tendency will always change if a single value in the data changes?

    A) Mean

    B) Median

    C) Mode

    D) All of these

19. True/False

[[ ]] Standard deviation can be negative.

[[ ]] Standard deviation is robust to outliers.

[[ ]]The standard normal curve is symmetric about 0 and the total area under it is 1.

[[ ]] Suppose you have been given a variable V, along with its mean and median. Based on these values, you can find whether the variable "V" is left skewed or right skewed for the condition

mean(V) > median(V)

20. We have a linear regression equation (Y = 5X +40) for the below table.

X	Y
5	45
6	76
7	78
8	87
9	79

What is the MAE (Mean Absolute Error) for this linear model?