CSC/DSCI 4740/6740 **Data mining** **D**ue by: Sept. 13

Name:

Assignment 1 (100 points)

Q1: Define an overall pipeline of data mining project step-by-step, using the reference from lecture slides. (10pt)

Q2: Define the following data mining functionalities: Association and correlation analysis, classification, Regression, Clustering, and Outlier analysis. Give examples of what each data mining functionality can do, using a real-life example that you are familiar with. (10pt)

Q3. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) (40pt)

5	11	11	12	13	15	16	16	19	20	20	21	22	22	25	25	25	25	25	26	26	26
27	28	29	31	33	34	34	34	35	35	35	35	35	35	36	36	36	37	40	45	46	52

(a) What is the mean of the data? What is the median?

(b) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).

(c) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

(d) Give the five-number summary of the data.

(e) Show a boxplot of the data.

(f) Show a quantile plot of the data.

(g) Show a quantile-quantile plot against a normal distribution with sample mean and sample standard deviation?

Note: For (e), (f), (g) any programing language can be used to generate plot, even your hand drawing.

Q4:. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): (10pt)

(a) Compute the Euclidean distance between the two objects.

(b) Compute the Manhattan distance between the two objects.

(c) Compute the Minkowski distance between the two objects, using h = 3.

(d) Compute the supremum distance between the two objects.

Q5: 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.

CSC/DSCI 4740/6740 **Data mining** **D**ue by: Sept. 13

Name:

Nonetheless, seemingly different similarity measures may be equivalent after some transformation. Suppose we have the following two-dimensional data set: (10 pt)

	A_1	A_2
x_1	0.66162	0.74984
x_2	0.72500	0.68875
x_3	0.66436	0.74741
x_4	0.62470	0.78087
x_5	0.83205	0.55470

Consider the data as two-dimensional 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. (similar data points rank first)

Q6: Required for graduate students, Bonus points for undergraduate students (10 pt)

Compute Maximum Likelihood Estimation for the following problem;

If you get independent samples $x1$, $x2$, x3, x4 , from an $Exponential(\theta)$ distribution

Pdf (x,θ) $= \theta\, exp(-\theta x) = \theta\, e^{-\theta x}$; $X > 0$

What is the likelihood function?

L(x1,x2,x3,x4, θ) =

What is formular of MLE of θ;

$\Theta =$

Here is data: ($x1$,$x2$,$x3$,$x4$)=(1.3,3.5,1.9, 2.2)., plug in the data values, what is the value of θ;

Q7: Please describe your idea(s) for data mining project. The idea(s) does not need to be final. You can still change, update, and revise it later. The following questions are meant to motivate your thinking, not for following exactly. What question(s) you want to answer? what type of data you are interested in? or which research field(s) you are interested in? or what do you expect your project can do using a certain type of data ? (10 pt)