CSCI 552 (Spring 2021)

Homework #1

Handout: Thursday, Feb. 11, 2021
Due: 11:59 pm, Thursday, Feb. 25, 2021
Total points: 30

All assignments will be submitted through Canvas. Documents will need to be in either Word or PDF format. Images need to be in jpeg format.

1. Given a set of 3D data points: $P_1 = (-1, 0.1285, 1)$, $P_2 = (-0.5, 1, 0.241)$, $P_3 = (0, 0.2557, 0.3508)$, $P_4 = (0.5, 0.399, 0.198)$, compute the Manhattan distances, Euclidean distance, and Cosine distances between each of these points to a point $P = (-0.5, 0.25, 0)$.

2. Prove that the distance formula for categorical data, $d(i, j) = \frac{p-m}{p}$, satisfies the triangle inequality property.

3. The weight (kg) and height (cm) measurements for a group of students are: $(45, 149)$, $(48, 153)$, $(47, 156)$, $(49, 156)$, $(53, 161)$, $(52, 162)$, $(56, 162)$, $(49, 162)$, $(50, 164)$, $(52, 165)$.

   (a) Compute the following height statistics: mean, median, standard deviation, the first quartile, and the third quartile.

   (b) Manually draw a box plot for the height data, and a scatter plot on a weight-height plane.

   (c) Manually draw a linear regression curve, and indicate if there are outliers.

4. Perform an Internet search to find three interesting repositories of publicly available data. Study these data sources and write one or two paragraphs for each repository describing the nature, characteristics and structure of the data. For example, how was the data set collected ? what types of attributes and features are there in the data set? Does it have spatial or temporal attributes? What file format is used? etc.