CS 495/595 –Introduction to Data Mining, HW2

1. (20 pts) Compute the Hamming distance and the Jaccard similarity between the following two binary vectors.

   x = 0101010001
   y = 0100011000

2. (20 pts) For the following vectors, **x** and **y**, calculate the indicated similarity or distance measures.
   a) **x** = (1, 1, 1, 1), **y** = (2, 2, 2, 2) cosine, Euclidean
   b) **x** = (0, 1, 0, 1), **y** = (1, 0, 1, 0) cosine, Euclidean, Jaccard
   c) **x** = (0,−1, 0, 1), **y** = (1, 0,−1, 0) cosine, Euclidean
   d) **x** = (1, 1, 0, 1, 0, 1), **y** = (1, 1, 1, 0, 0, 1) cosine, Jaccard

3. (30 pts) Download the Arrythmia data set from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/machine-learning-databases/arrhythmia/). Write a code (in any language you are comfortable with) to normalize all records to a mean of 0 and a standard deviation of 1.
   a. Create a boxplot of the dataset.
   b. Submit both your code and the boxplot

4. (30 pts) Convert the following weather.xml file to its corresponding weather.json structure.

```
<weatherReport>
        <date>05/29/2002</date>
        <location>
                <city>Philadelphia</city>,
                <state>PA</state>
                <country>USA</country>
        </location>
        <temperature-range>
                <high scale="F">84</high>
                <low scale="F">51</low>
        </temperature-range>
</weatherReport>
```

What to turn in:

Follow the naming convention: **Lastname-hw2.pdf** should contain answers to all above questions (calculations, code, boxplot, json). Make sure your name is printed on top of the pdf document.

Submit your pdf file to Blackboard.