HW1. Due 2/26/19

(i) Question 1 and 2 should be answered by hand (you can use calculator)
(ii) Write separate python program to answer rest of the questions
(iii) Submit your solution (pdf) and python files as single zip

1. (10 points) Suppose we have the following two-dimensional data set:

Data points	X_1	X_2
p1	0.3	0.8
p2	0.7	0.4
p3	1	0.1

(a) Consider the data as two-dimension data points. Give a new data point, p4 = (0.4,0.2) as a query point, rank the data points based on the similarity with respect to the query point using (1) Euclidean distance (2) cosine similarity [round your similarity to three decimal places]

(b) Transform each value in your data set including the query point using the sigmoid function:

$$x\_new = \frac{1}{1 + e^{-x}}$$

And re-rank the data points based on the similarity with respect to the query point using: (1) Euclidean distance (2) cosine similarity [round your similarity to three decimal places]

2. Consider the following two vectors:

x = 0101010011
y = 0100101100

Compute SM and Jaccard Coefficients

3. Implement maximum likelihood classification in python. Input to your program is CSV file, with last attribute assumed to class label. You can assume all non-class attributes are continuous random variables. It should take two input files, training file for constructing model and test file to estimate various accuracy measures.

4. Implement entropy and gain functions. For a given data (CSV file, last attribute is class label), output entropy and gain values for each attribute, and determine root node attribute.