UTSA CS 6243/4593 Machine Learning Fall 2017
Problem Set No. 1

Issued: Tuesday, September 5, 2017
Due: In class, Tuesday, September 12, 2017

This assignment is due at the beginning of class on the due date.

Problem 1.1 (50 points)

1. (40 pt) The k-means algorithm is pretty straight forward. Here is the pseudo-code for it. Please implement k-means on 2-dimensional numerical data by Matlab, java or C++, which should be fairly easy to derive from this.

```
Let n be the number of clusters you want
Let S be the set of feature vectors (|S| is the size of the set)
Let A be the set of associated clusters for each feature vector
Let sim(x,y) be the similarity function
Let c[n] be the vectors for our clusters

Init:
    Let S' = S
    //choose n random vectors to start our clusters
   for i=1 to n
        j = rand(|S'|)
       c[n] = S'[j]
      S' = S' - {c[n]} //remove that vector from S' so we can't choose it again
   end

  //assign initial clusters
   for i=1 to |S|
     A[i] = argmin(j = 1 to n) { sim(S[i], c[j]) }
   end

Run:
    Let change = true
   while change
       change = false //assume there is no change
      //reassign feature vectors to clusters
     for i = 1 to |S|
         a = argmin(j = 1 to n) { sim(S[i], c[j]) }
     if a != A[i]
        A[i] = a
        change = true //a vector changed affiliations -- so we need to
                      //recompute our cluster vectors and run again
```

```
            end
        end

    //recalculate cluster locations if a change occurred
     if change
         for i = 1 to n
         mean, count = 0
         for j = 1 to |S|
             if A[j] == i
                 mean = mean + S[j]
                 count = count + 1
        end
     end
     c[i] = mean/count
  end
end
```

Use your code to cluster the following eight points (with (x, y) representing locations) into three clusters A1(2, 10) A2(2, 5) A3(8, 4) A4(5, 8) A5(7, 5) A6(6, 4) A7(1, 2) A8(4, 9). The distance (similarity) function between two points $a=(x1, y1)$ and $b=(x2, y2)$ is defined as: $\rho(a, b) = |x2 - x1| + |y2 - y1|$.

To be simple, you can just set n=3 and set S equals to the above eight points. Your program needs output the final clustering result (members in each cluster).
For example: 1: {A1, A4, …}
2: {A3, A5, ..}
3: {A2, …}

Note for machine problem:

For grading, you should hand in a printout of your MATLAB (or C++) files and a concise report which should include all the necessary texts, figures, and labels, etc.

IMPORTANT:

1) Submit your printed source code and output, and a short note of how to execute it.

2) You should also upload your source (e.g., MATLAB) files to Blackboard by the beginning of the class the due date. You should give a Subject title: CS 6243 Homework 1.