Winter 2016 **COSC 4P76: Assignment 2**
Instructor: B. Ombuki-Berman
TA: Justin Maltese

Assigned: Monday 7th, March 2016
Due date: Wednesday, 23rd, March 2016

Objective: Implement and analyze the k-means algorithm for clustering problems

Your task is to implement the k-means algorithm and apply it to the clustering of several datasets. The data that you are to use is available at the following link:

http://cs.joensuu.fi/sipu/datasets/

Note: Use the S1, S2, S3 and S4 data sets. Feel free to use any additional data sets from the above link.

To gauge the desirability of the clusters generated by your algorithm, you are required to use the Dunn index. Of course, you can utilize any additional validity measures for comparison purposes if you choose to. A good resource on some popular validity measures can be found at:

http://www.biomedcentral.com/content/supplementary/1471-2105-9-90-s2.pdf

You are also required to provide a detailed empirical study on the k-means algorithm. Specifically, you are required to analyze the following:

1. For each of the datasets, test different values of k. How does the k-value impact performance? What do you notice when a too large or too small k value is employed? **Note: The optimal k-values are provided at the above dataset link. Test reasonably larger and smaller k-values in additional to the optimal ones.**

2. Test three different distance measures, such as Euclidean, Chebyshev, Minkowski, etc. How does the distance measure impact performance for each of the datasets?

3. Provide a visualization of the data and generated cluster centroids by plotting them in Excel. For several runs, compare the respective Dunn index values produced to the data plots. Visually, do you find that more desirable Dunn Index values correlate to better clustering?

Since the programming portion required is significantly smaller than A1 (k-means is relatively simple), you should focus mostly on writing a clear, concise report, which demonstrates that you are able to independently produce and analyze results. However, this report will be much smaller than that given in assignment 1.

Assignment Requirements and Grading:

The results are to be handed in via a technical paper with the following headings:

1) **Abstract, Introduction & problem definition**
    - What is the goal of this work? Outline of rest of paper, Etc.
    - Clear definition of the problem and data being used, applicability of k-means algorithm for this problem

2) **K-means algorithm**

    - Here, describe the basics of the k-means algorithm, including relevant formulas and equations
    - Include a description of each of the distance measures

3) **Results and Discussion**
    - Here you will outline the parameters used and show the plots of your clustering results
    - Also, explain what the results lead you to conclude
    - This should be the biggest section, take special care in the analysis of your results
    - Ensure that all required items are tested
    -

4) **Conclusions**
    - In 2-3 paragraphs clearly outline your conclusions and why they are valid.

You must use the IEEE style found on the 4P76 website. The report must be written in a clear and concise manner. As a general guideline:

MINIMUM REPORT MARKS:

- Does not adhere to guidelines, tests only one k-value
- Experiments are non-intuitive, comparisons are redundant
- Write-up is not in technical format, has spelling errors, incorrect explanations, etc.
- No plots are shown in write-up
- No references, or not properly referenced
- Conclusion not well supported by results

MAXIMUM REPORT MARKS

- Report is very clear, concise, well-written, correct explanations, etc.…
- Shows plots as required, compares distance values, tests multiple k-values
- Conclusions are well supported

* Failure to submit working code will result in a 0 (even if other sections of the report are completed correctly).

**** Academic misconduct according to the departmental regulation (plagiarism, cheating, false results, etc.) is NOT tolerated, and will result in a 0 for this class.**