CSE471: Statistical Methods in AI

Monsoon 2016

Assignment #4: Data Clustering

Due: Before 05:00pm on 12th November 2016 **Total Marks:** 40 (Mapped to 4% of course credits)

General Instructions:

- Assignment can be implemented in Matlab/Octave, Python, C/C++, R .
- Ensure that submitted assignment is your original work. Please do not copy any part from any source including your friends, seniors and/or the internet. If any such attempt is caught then serious actions including an **F grade in the course** is possible.
- A single pdf file needs to be uploaded to the Courses Portal. The file should contain your answers as well as the code you have written and its output (Or as directed by the TA's).
- Include the assignment number, your name and roll number at the top-left of the first page of your submission.
- Your grade will depend on the correctness of answers and output. In addition, due consideration will be given to the clarity and details of your answers and the legibility and structure of your code as well viva based oral examination done by TA's.

Problem

Derive formulation for the Kernel K-means clustering algorithm. (10 Marks)

Implement and provide the pseudo code, assumptions made, mean accuracy results and your observations in a report format along with the code for the following clustering methods on any two dataset from the list of 48 UCI ML repository datasets. You can randomly sample a subset of data points/instances of relatively smaller size (order of 1000) for experimentation, in case when the the size of original dataset is huge in terms of number of data points.

a) Agglomerative Clustering with two cluster merging criterions. (5x2 Marks)
b) Spectral Clustering. (10 Marks)
c) Kernel Kmeans Clustering with RBF and polynomial Kernels. (5x2 Marks)