**MST 690 – Data Science Mathematics** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Session 4 - Clustering**

**Exercise**

Instructions: Use the jupyter notebook to solve each problem---array already in the notebook. Use markdown to write answers to problems.

1. You are an OSINT analyst for the Marine Corps Intelligence Activity. You’ve collected a significant number of tweets from a particular geographic region of interest, and are interested in developing an algorithm to illuminate differences that may exist within your data set. You know that some of the tweets are from known military personnel, and you hope to demonstrate that a clustering technique can be used to highlight these differences within the larger data set. You’ve engineered quantifiable features from your data, which you intend to use to build a supervised clustering algorithm.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **User ID** | **Feature 1** | **Feature 2** | **Feature 3** | **Military (1=YES, 0=NO)** |
| 1001 | 8 | 22 | 62 | 1 |
| 1002 | 15 | 51 | 85 | 0 |
| 1003 | 9 | 44 | 121 | 0 |
| 1004 | 8 | 51 | 136 | 0 |
| 1005 | 8 | 20 | 93 | 1 |
| 1006 | 15 | 64 | 124 | 0 |
| 1007 | 14 | 56 | 101 | 0 |
| 1008 | 5 | 10 | 80 | 1 |
| 1009 | 5 | 18 | 73 | 1 |
| 1010 | 9 | 26 | 79 | 1 |

Consider the above data set. You have determined the three features that you believe have the greatest correlation with military status.

1. Write the code for k-means clustering on the above data set using the algorithms in the slides. Use the coordinates **(10, 20, 80)** and **(10, 50, 110)**, corresponding to **(Feature 1, Feature 2, Feature 3)**, as your initial “best guess” clusters.
2. Determine if convergence occurred after **two, five** iterations of k-means.
3. How well did your algorithm cluster military personnel versus non-military personnel? Construct a confusion matrix, and calculate the Matthews’ Correlation Coefficient (write the code vs using numpy---feel free to check with numpy).
4. You selected three features to use in this computation because you determined that they are the three most correlated features with “military” status. While adding additional features up to a certain point will enhance clustering model accuracy, adding too many features diminishes accuracy. Explain why this is true.