# PES University, Bangalore

## UE15CS322 - Data Analytics

**Session: Aug – Dec 2017**
**Week 9 – Assignment 7**

**Date of Submission:** **Max Marks:** 20

***NOTE:*** *In your assignment books, write the question, R commands required to get the desired output and the obtained output. For questions that require you to plot graphs, you will have to also print the obtained output (graphs/charts) and attach it in your assignment book. The solutions to the assignment must be* ***hand-written****.*

### TOPIC: Unsupervised Learning and Association Rules

*The rise of artificial intelligence technology along with machine and deep learning are opening up almost limitless possibilities. There is also an element of fear towards this exponential growth. However, the humanization of our machines and devices is seldom discussed or reported. Turning to technology and innovation for solutions to non-existing problems often leads to duplication of our work. Once again, we need to ask ourselves if this is the most efficient method of managing our workload in this digital age. Technology now enables us to snatch back an hour or two from those pesky time stealers that have crippled our productivity for years. Maybe it's time we started to embrace the benefits.*

Excerpts from 'Why it's Time to Take AI Seriously', Inc.com

### Question 1 (12 points)

***NOTE:*** *Clustering is an unsupervised learning algorithm – i.e. none of the instances in the dataset are labelled to belong to a certain class – you create clusters and then label each cluster based on your investigation. However, in the current example, labelling a cluster will involve studying multiple images for which we do not have the bandwidth (and time) as a part of this course. Hence, we will use the labels already present for each instance and label the cluster based on polling.*

The dataset, optdigits.csv contains the pre-processed data obtained by processing 32 X 32 bitmap images of handwritten digits provided by 30 people. The dataset contains normalized bitmap values of handwritten digits from a pre-printed form. The bitmaps were divided into 4 X 4 non-overlapping blocks and the number of dark pixels in each of these blocks were counted. This generated an 8 X 8 input matrix where each element in the matrix was an integer in the range, 0 to 16. The elements of this 8 X 8 matrix has been presented before you as 64 input features and the task in hand is to the recognize the digit based on these features. (You will learn how to perform such processing on images next semester in the course, 'Digital Image Processing').

(**Note:** use set.seed(10))

a) Use k-means clustering to cluster the dataset into multiple clusters. Use 200 iterations. How many instances of each digit do each of the clusters contain (give complete details)? Also, label each cluster based on the maximum number of instances of a digit that is present in it.

b) One of the clusters seem to have an almost equal distribution of two digits in it. Perform hierarchical clustering only on the instances present in this cluster. Print the dendrogram displaying branches only above a height of 50. Bring down the number of clusters formed via hierarchical clustering to 2 and print the number of instances of each digit contained in each cluster.

c) Load the test data and calculate the distance between each instance in the test data with the centres of the clusters formed in question (a). Based on the distances, identify the number written in each image. Mention the image number along with its classification (cluster number and digit in the image).

d) The test data instances that were classified to belong to the cluster used in question (b) will need to be further classified. Unfortunately, the model built by agglomerative clustering cannot be used for classification but can be used as a pre-processing step for classification. Label each of the data points in the clusters formed by hierarchical clustering, by the dominant label present in each cluster. Print the number of data points present under each label. Now, we'll use a supervised learning algorithm to predict the label of the data that was classified to belong to this cluster in question (c). Use k-nearest neighbour with k = 7 to classify each of these data points that were classified to belong to the cluster in question and mention their new labels.

**Question 2 (8 points)**

The given data, handwriting_recognition.csv specifies how the gender and profession of a person can possibly affect the recognition of his/her handwriting. What, if any, are the association rules with the default setting?

Find the support, confidence and lift of the following association rules:

(i) {Artist, Female} => Recognized
(ii) {Engineer} => Male
(iii) {Actor, Recognized} => Female
(iv) {Doctor, Male} => Unrecognized