CS1762: Experiment 3

Maximum Marks: 10

Submission deadline: 07 September, 2020

Write a program to explore the use of unsupervised learning methods for clustering data, and also for obtaining lower dimensional representations.

To explore the use of unsupervised learning methods for clustering data, and also for obtaining lower dimensional representations.

1. Data Visualization

Randomly generate the data sets using *make_blobs* and *make_moons* (for limitations of k-means). Try to visualize on 2D feature space.

2. **Binary classification:**

Investigate *K*-means algorithm on the randomly generated data sets *make_blobs* and *make_moons* using various hyper-parameters. Study the effects of changing the different parameter values, including the type of kernel function being used in kernelized *K*-means. How do they affect the accuracy?

3. **Multiclass classification:**

(i) Now, consider the data sets with multiple classes. Using an implementation of your choice, run *K*-means with $K = 10$ on your randomly generated handwritten digits data set. Assess the clusters obtained: do they actually correspond to the digits 0–9? If you label each cluster with the digit that occurs most frequently within it, then what is your classification accuracy with this unsupervised method? What kinds of misclassifications are happening, and why? Now try re-running *K*-means with $K = 5$. Do your clusters make any sense in this case? Why or why not?

(ii) (Marks is not associated with this section and it is only for practice)

For the previous assignment, you were provided a low dimensional representation of a data set of images. We now provide the corresponding original data of the images containing the pixel intensities.Using an implementation of your choice, run *K*-means with $K = 10$ on your randomly generated handwritten digits data set. Assess the clusters obtained: do they actually correspond to the digits 0–9? If you label each cluster with the digit that occurs most frequently within it, then what is your classification accuracy with this unsupervised method? What kinds of misclassifications are happening, and why? Now try re-running *K*-means with $K = 5$. Do your clusters make any sense in this case? Why or why not?