Indian Institute of Technology Jodhpur
Machine Learning I: Fractal 2
Programming Assignment 1
Maximum Marks: 30

Instructions:

- Submit a .zip file with named "RollNo_PA1.zip". Before compressing using zip, please name your folder as "RollNo_PA1".
- Submit a report containing a description of algorithm, results, and evaluation. This should be named "RollNo_PA1.pdf". Also, name the code file for each question as "RollNo_PA1_QNo.py".
- Your code should be executable by the command `python RollNo_PA1_QNo.py <path to the dataset file>` and should plot the results and print the accuracy.
- Inbuilt functions can not be used to solve the problem. For example, any inbuilt function for spectral clustering from any library can not be used. It should be implemented by you. You can use any inbuilt function for finding the Eigenvalue decomposition of a matrix.
- Plagiarism will not be tolerated and will result in severe consequences if found.
- The preferred programming language is `python`.
- Please attempt both the questions.

1. Implement the k-means and spectral clustering algorithms for clustering the points given in the datasets: `http://cs.joensuu.fi/sipu/datasets/jain.txt`. Plot the obtained results. In order to evaluate the performance of these algorithms, find the percentage of points for which the estimated cluster label is correct. Report the accuracy of both the algorithm. The ground truth clustering is given as the third column of the given text file. [**15 Marks**]

2. Implement the Principal Component Analysis algorithm for reducing the dimensionality of the points given in the datasets: `https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data`. Each point of this dataset is a 4-dimensional vector ($d = 4$) given in the first column of the datafile. Reduce the dimensionality to 2 ($k = 2$). This dataset contains 3 clusters. Ground-truth cluster IDs are given as the fifth column of the data file. In order to evaluate the performance of the PCA algorithm, perform clustering (in 3 clusters) before and after dimensionality reduction using the Spectral Clustering algorithm and then find the percentage of points for which the estimated cluster label is correct. Report the accuracy of the Spectral Clustering algorithm before and after the dimensionality reduction. Report the reconstruction error for $k = 1, 2, 3$. [**15 Marks**]