CS3390/CS5590/AI5000 Foundations of Machine Learning
Assignment 2

Marks : 75, Submission Deadline: Nov 19, 2023

Instructions:

- The submission should be a single ZIP file with the name **ROLL_NUMBER-foml-assignment-1.zip**
- Datasets used should be present in the same zip file
- Before submitting, students should ensure that the code works and there are no issues with file paths of the dataset
- Every programming question should be in a separate ipynb notebook named by question number
- All the codes need to be written in python and you may use various packages available in python such as numpy, scipy, scikit-learn, matplotlib etc.

In this assignment, you need to conduct experiments using the SVHN dataset, which consists of digit images. It has 10 classes, 1 for each digit. Digit '0' has label ‘0’, '1' has label 1, '9' has label 9. 73257 digits for training, 26032 digits for testing. For the experiments here, you may consider a subset of the training data with randomly chosen 2500 samples from each class from the original training set, totaling to 25,000 samples as your new training data. Test data remains the same with 26032 samples.

Links to dataset:http://ufldl.stanford.edu/housenumbers/
(use format 2: Cropped Digits: train_32x32.mat, test_32x32.mat)

Q1. Principal component analysis

(a) Perform PCA on SVHN data set. Find how many top eigenvectors are required to keep the proportion of variance above 0.9. Plot PoV against number of eigenvectors.

(b) Visualize top 10 eigenvectors and provide reconstruction of 10 SVHN samples (one from each class) using top 10 eigenvectors.

(c) Run k-NN (for k=5 and k=7) on raw data and data obtained after PCA dimensionality reduction for dimension as found in part (a) and for dimension 10 as in part (b). Provide the accuracy of the predictions on the test data set for these various cases and discuss your observations.

Marks : 10 + 10 + 15 = 35

Q2. K-Means clustering

(a) Perform k means clustering with k=10 on the raw data and on data obtained after PCA dimensionality reduction for dimension as found in question 1 (a) and for dimension 10 as in question 1 (b). You may use only training data of SVHN to perform this. Do you observe images from the same class to be clustered together ? For each cluster, provide the image closest to the centroid.

(b) Find the sum squared error for each of these different clustering obtained in 2 (a). Make use of the label information associated with the images to evaluate your clustering.Specifically, evaluate the goodness of your clusters for various cases in 2 (a) by using two evaluation metrics purity and rand-index.

(c) If you label each cluster with the digit that occurs most frequently within it, then what is your classification accuracy with this unsupervised method for various cases ? How does this compare with the accuracy you got with the K-NN classifier in Q1 (c) ?

(d) Perform k means clustering with k=5 on data obtained after PCA dimensionality reduction for dimension as found in question 1 (a). Find the sum squared error for each of these cases. Do you observe images from different classes to be clustered together ? which all classes do you find are getting clustered together.

Marks : 15 + 10 + 5 + 10 = 40