Assignment 5
CIS530 Advanced Data Mining **Spring 2023**

Due: 4/16/2022 23:59PM Submit to myCourses

Q1. (40 pts) In this question, we will use SVM for a few experiments. There are many good implementations of SVM, e.g., MATLAB built-in functions, and LIBSVM. You can use either of them, or other implementations found by yourself to solve the following questions.

a. (20 pts) USPS is a hand-written digit database including ten digits, i.e., 0, 1, 2, …, 9. We will use a subset (USPS_sub.mat) including 1K images in this experiment. The resolution of each image is 16×16, and thus the length of each vector is 256. In this question, you will use SVM for the digit classification. For each digit, you will use the first 5, 10, or 15 images as the training data, and the rest as test data to finish this multi-class classification problem. In each setting, you will be required to use: (1) linear kernel, (2) polynomial kernel, (3) radial basis function (RBF) kernel. As there are a few model parameters for each kernel, e.g., band-width γ in RBF kernel, and the common slack variable C, you will need to use cross-validation on the training set to select the best model parameters. You can use "grid search" to find the best values for them. So, in total there are ***9 different results (recognition accuracy)*** from 3 different training setting and 3 different kernels.

b. (20 pts) Please implement k-nearest-neighbor (kNN) classifier by yourself and then use the USPS dataset above for the same experiments. First, you will use 5, 10, or 15 images as the training data and the rest as test data to finish this multi-class classification problem. Second, for each experiment, please try k=1, 3, and 5 for your kNN classifier. Therefore, ***3*3=9 experimental results will be reported***.

[**Submissions**] Source codes and experimental results report (in PDF)
[**Hints**] You may want to apply PCA, centralization, or normalization before running SVM and kNN classifier, which may speed up your learning process. A sample code (SVMSample.m) has been provided for RBF kernel. A guidance for beginners of LIBSVM can be found here: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

[**Provided files list**] USPS_sub.mat, libsvm-3.11, SVMSample.m

Q2. (20 pts) Consider the points:

$$x_1 = (0, 16), \quad x_2 = (0, 9), \quad x_3 = (-4, 0), \quad x_4 = (4, 0).$$

Suppose we wish to separate these points into two clusters and we initialize the K-Means algorithm by randomly assigning points to clusters. This random initialization assigns x_1 to cluster 1 and all other points to cluster 2.

With these points and these initial assignments of points to clusters, what will be the final assignment of points to clusters computed by Kmeans? Explicitly write out the steps that the Kmeans algorithm will take to find the final cluster assignments. In this case, does Kmeans find the assignment of points to clusters that minimizes the following Equation?

$$\sum_{i=1}^{m} d(x^{(i)}, \mu_{\ell_i}),$$

where μ_c is the centroid of all points $x^{(i)}$ with label $l_i = c$ and $d(x, y)$ is the distance between points x and y. Usually d is the Euclidean distance, given by

$$d(x, y) = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{1/2}$$

Q3. (40 pts) We've learned Kmeans and Hierarchical Clustering in the class. In this question, you will apply both of them for data clustering tasks. For Kmeans, you could use either the MATLAB build-in function or a fast implementation here. For Hierarchical Clustering, please refer to the MATLAB built-in-function.

In this question, you will run Kmeans and Hierarchical Clustering on USPS database. To evaluate your results, you will report ***clustering accuracy (Acc), NMI, and running time***. Acc and NMI are standard metrics in clustering tasks and details of their meaning can be found here. The purity in the link has the similar meaning with Acc. Both of their implementations are provided in the sample codes. In addition, a sample code (KmeansSample.m) for Kmeans and its evaluations on USPS dataset have been provided for your reference. For Hierarchical Clustering, please refer to MATLAB documents: link1, link2. You will mainly leverage function "linkage" for this purpose. There are a few options that you may consider for different criteria, including (1) single (Min), (2) complete (Max), (3) weighted, (4) ward. Only Euclidean distance will be used throughout the question.

a. (15 pts) For Kmeans, please evaluate on USPS dataset, and report: (1) clustering accuracy (Acc), (2) NMI, and (3) end-to-end running time **for 10 clusters**. Please **run five trials** and report average and standard deviation for each.
b. (25 pts) For Hierarchical Clustering, please evaluate on USPS dataset, and report: (1) clustering accuracy (Acc), (2) NMI, and (3) end-to-end running time **for 10 clusters**. **Four different criteria** will be evaluated in this experiment: (1) single (Min), (2) complete (Max), (3) weighted, (4) ward. ***Therefore, 3*4 results will be reported.***

[**Submissions**] Source codes and experimental results report (in PDF)

[**Hints**] You may want to apply PCA, centralization, or normalization before running any algorithms.

[**Provided files list**]: litekmeans.mat, USPS_sub.mat, KmeansSample.m, hungarian.m, accuracy.m, nmi.m