Fall 2016
CSCE 666 Pattern Analysis
Homework #3

Due date: 10/31/2016

In recognition of the Texas A&M University policies of academic integrity, I certify that I have neither given nor received dishonest aid in this homework assignment.

Name: _____ *Signature:* _____

PLEASE FOLLOW THESE GUIDELINES:

1. *Download the compressed file 'hw3.zip' from the course web page*
2. *Submit your solutions as a report, with each problem being a separate section of the report –you may use this assignment as a template*
3. *Please show your work and discuss your findings. This ensures full credit if your results are correct, and allows me to give you partial credit otherwise*
4. *Sign and return this page with your finished assignment.*
5. *Submit your code as a ZIP file through csnet, each problem on a separate subfolder*

Problem 1 (20%)

Download the dataset 'hw3p1_data.mat', which contains frontal images of senior faculty in the TAMU Computer Science Department; each image is represented by a row vector in data matrix 'x'. You are to generate a PCA decomposition of these faces using the 'snapshot' approach.

(a) Generate an image of the average face
(b) Generate images of the first six eigenvectors (i.e., "eigenfaces")
(c) Generate a 2D PCA scatter plots of the corresponding principal components
(d) DISCUSS YOUR RESULTS.

NOTE: You are expected to write your own implementation of the snapshot PCA method.

HINT: You may use the function 'reshape' to convert the 2D array of row vectors into images, and the command 'colormap' to change the default "false color" scheme in MATLAB

Problem 2 (20%)

(a) Download the dataset 'us_city_distance.mat', which contains the shortest geodesic distance (in meters) between pairs of cities in the United States. You are to compute the embedding manifold, and generate 2D and 3D scatter plots for these cities. How many dimensions are needed to capture the manifold?

(b) Repeat (a) on the dataset 'world_city_distance.mat', which contains the distance between pairs of cities worldwide. How many dimensions are needed to capture the manifold?

(c) DISCUSS YOUR RESULTS.

NOTE: You are expected to write your own implementation of the manifold learner.

Problem 3 (20%)

Download the image 'hw3p3_im.jpg', and perform k-means clustering to vector-quantize pixels according to their RGB color.

(a) Reconstruct the image using the colors in the codebook for values of $k=1,2,\ldots,10$.
(b) Generate a color JPEG image for each of the reconstructed images, and save it with the filename *'c1.jpg', c2.jpg', c3.jpg', … 'hw3p2c10.jpg'*.
(c) Can you explain which codewords emerge from the image as the codebook length increases?
(d) Generate a plot that shows the sum-squared-error (SSE) between the reconstructed image and the original image as a function of k, the number of clusters. To do so, repeat part (a) several times and, for each k, record the clustering that gives the minimum SSE. Can you make sense of this plot? NOTE: You may sub-sample the image (say, 2:1) in order to speed up computations in this part.
(e) DISCUSS YOUR FINDINGS.

HINTS: Use the functions 'imread' and 'imagesc' to load and display the image. You may use the function 'reshape' to convert the image into a 2D array of row vectors. Use the function 'dist' to compute the distance between cluster centers and the data points.

NOTE: You are expected to code your own implementation of the k-means method.

Problem 4 (40%)

You are given a dataset containing nutritional content for a large number of food products from the 10 different groups shown in Table 1. Table 2 shows the nutritional information contained on each of the 46 features in the dataset. When you load the data, you will notice that a significant number of entries have a value of -1; this value indicates that the corresponding nutritional feature was not available for the sample in question.

As in Homework 2, you are given the following datasets:

- Dataset '*hw3p4_train.mat*' containing the following matrices:
  - x1: training set (row vectors)
  - clab1: training set labels
- Dataset '*hw3p4_test.mat*' contains the following matrices:
  - x2: test set (row vectors)
  - clab2: test set labels

You are to:

(a) Perform feature subset selection to determine a reduced number of features ($D \leq 10$) that provide good discriminatory information. You may employ any feature subset selection technique, and any of the classifiers that you have developed in previous homework assignments.
(b) To evaluate your final classifier, I will use a similar approach as in Homework #2. Prepare a MATLAB program called '*hw3p4.m*' that will load '*hw3p4_test.mat*' and classify each of the examples in the dataset x2. Once you submit your code through email, I will run your '*hw3p4.m*' program with a separate 'hw3p4_test.mat' file containing my own test data. Your grade will be based on the performance of your classifier on my test data, which will contain a very large number of examples so I can approximate the true error rate. All datasets will obviously be generated from the same distribution.

(c) DESCRIBE YOUR APPROACH AND DISCUSS YOUR RESULTS. What search technique(s) did you try? What objective function(s) did you try? What classifier(s) did you employ? Which features were selected? How did you determine how many features to settle for? …

NOTES:

- Please submit your code using the "turnin" utility at https://csnet.cs.tamu. You should submit a single ZIP file (**your_last_name.zip**). Please refer to the syllabus for the late submission policy.
- When your code loads my separate test set, the class label vector clab2 will obviously have dummy values, so your program should not attempt to use them. On the other hand, the class label vector on the test set in 'hw3.zip' does have correct values, so you can use them for validation.
- Make sure your code can handle ANY number of examples in x2. The number of classes and dimensions will obviously be the same that you trained on.
- Make sure your code works!!! Are all required files included? I will run your code on unix.cs.tamu.edu, so please test it on that machine. Submissions that do not run will receive a 15% penalty over the entire homework grade.
- **Your program 'hw3p4.m' should not perform feature subset selection. It should only classify new data using a feature subset you will have previously selected off-line.**
- To facilitate grading, your '*hw3p4.m*' file should create a COLUMN VECTOR called '*uclab*' containing the predicted class labels for each of the rows in x2:

  uclab = [1 3 2 1 1 3 3 2 …]';

  These are the class predictions that I will compare against the true class labels of my separate test set, which I have kept aside.

Table 1. Class descriptions

Class label	Description
1	Baked Products
2	Vegetables and Vegetable Products
3	Soups, Sauces, and Gravies
4	Sweets
5	Fast Foods
6	Fruits and Fruit Juices
7	Breakfast Cereals
8	Poultry Products
9	Beef Products
10	Lamb, Veal, and Game Products

Table 2. Feature descriptions

Feature	Description	Feature	Description
1	Water (g/100 g)	24	Vitamin B6 (mg/100 g)
2	Food energy (kcal/100 g)	25	Total Folate (µg/100 g)
3	Protein (g/100 g)	26	Folic acid (µg/100 g)
4	Total lipids (fat) (g/100 g)	27	Food Folate (µg/100 g)
5	Ash (g/100 g)	28	Folate (µg diet folate equivalent/100 g)
6	Carbohydrate (g/100 g)	29	Vitamin B12 (µg/100 g)
7	Total dietary fiber (g/100 g)	30	Vitamin A (IU/100 g)
8	Total sugars (g/100 g)	31	VitA (µg retinol activity equivalents/100g)
9	Calcium (mg/100 g)	32	Retinol (µg/100 g)
10	Iron (mg/100 g)	33	Vitamin E (alpha-tocopherol) (mg/100 g)
11	Magnesium (mg/100 g)	34	Vitamin K (phylloquinone) (µg/100 g)
12	Phosphorus (mg/100 g)	35	Alpha-carotene (µg/100 g)
13	Potassium (mg/100 g)	36	Beta-carotene (µg/100 g)
14	Sodium (mg/100 g)	37	Beta-cryptoxanthin (µg/100 g)
15	Zinc (mg/100 g)	38	Lycopene (µg/100 g)
16	Copper (mg/100 g)	39	Lutein + zeazanthin (µg/100 g)
17	Manganese (mg/100 g)	40	Saturated fatty acid (g/100 g)
18	Selenium (µg/100 g)	41	Monounsaturated fatty acids (g/100 g)
19	Vitamin C (mg/100 g)	42	Polyunsaturated fatty acids (g/100 g)
20	Thiamin (mg/100 g)	43	Cholesterol (mg/100 g)
21	Riboflavin (mg/100 g)	44	1st household weight from Weight file
22	Niacin (mg/100 g)	45	2nd household weight from Weight file
23	Pantothenic acid (mg/100 g)	46	Percent refuse