

Fall 2008, CPSC 689-607
Special Topics in Pattern Classification
Homework #3
Due date: 10/28

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: _____

NOTES:

1. Start each problem on a separate page
2. Provide hardcopies of all plots, MATLAB code and derivations.
3. Download the compressed file '[hw3.zip](#)' from the course web page
4. Sign and return this page with your finished assignment.

Problem 1 (20%)

Download the dataset '[hw3p1_data.mat](#)', which contains mug shots of pioneers in computer science and engineering; 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.

Problem 2 (20%)

Download the dataset '[hw3p2_data.mat](#)' from the course webpage.

- (i) Perform k-means clustering for values of k between 2 and 15.
 - a. For each value of k, run the clustering n=20 times with different initial conditions.
 - b. Store the clustering results for each value of k and n.
- (ii) For each value of k and n, compute the mean-squared-error (MSE) of the clustering result, which is the average of the Euclidean distance from each point in the dataset to its assigned cluster center.
 - a. For each value of k, find the clustering with the minimum MSE (out of n=20 repetitions).
 - b. For this minimum MSE clustering, generate a 2D scatterplot of the principal components of the data, and label each point by the index of the cluster to which it was assigned.
- (iii) Generate a plot of the minimum MSE as a function of k. From this plot, can you determine an appropriate number of clusters for this problem? Generate a labeled 2D PCA scatterplot for it.

DISCUSS YOUR RESULTS.

Problem 3 (20%)

The file '[hw3p3_data](#)' contains a database of 67 animals, where each animal is identified by its name and a feature vector containing sixteen descriptors. The variable 'readme' saved in that file contains the definitions of these descriptors.

- (a) Perform hierarchical clustering using the 'ward' metric, and generate a dendrogram, labeling each leaf by the name of the animal. Can you identify any meaningful groupings of animals?
- (b) A partition of the data can be obtained by pruning the hierarchical tree at a given depth. Generate a plot of the mean squared error (MSE) as a function of the number of clusters in the partition; you can do this by pruning the tree at different depths (i.e., from the root towards the leaves). Can you identify any trends (i.e., gaps, convergence) in the MSE that allow you to assign a cutoff for the tree? Can you reconcile these results with those in part (a) above?
- (c) DISCUSS YOUR RESULTS.

NOTE: In clustering, the MSE is the average of the Euclidean distance from each point in the dataset to its assigned cluster center.

HINT: You may want to use the MATLAB functions 'linkage', 'dendrogram' and 'cluster'.

Problem 4 (40%)

The goal of this problem is for you to perform feature subset selection. 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 that (hopefully) perform better than in the original space. 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, 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 as those you trained on.
- Make sure your code works!!! Are all required files included? You will receive no credit if the program returns with an error. I will run your code on `unix.cs.tamu.edu`.
- **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.