Homework 3

Statistics W4240: Data Mining

Columbia University

Due Tuesday, October 14 (All Sections)

For your .R submission, submit a file for questions 3 and 4 labeled `hw03_q3.R` and `hw03_q4.R`, respectively. The write up should be saved as a .pdf of size less than 4MB. **DO NOT** submit .rar, .tar, .zip, .docx, or other file types.

Problem 1. (20 Points) James 3.7.3

Problem 2. (20 Points) James 3.7.5

Problem 3. (20 Points) James 3.7.13

Problem 4. (40 Points) In this problem, we will use 1NN classification and PCA to do facial recognition.

a. (5 Points) Load the views `P00A+000E+00`, `P00A+005E+10`, `P00A+005E-10`, and `P00A+010E+00` for all subjects in the CroppedYale directory. Convert each photo to a *vector*; store the collection as a matrix where each row is a photo. Give this matrix the name `face_matrix_4a`. For each image, record the subject number and view in a data frame. The subject numbers will be used as our data labels.

Use the following commands to divide the data into training and testing sets:

```
fm_4a_size = dim(face_matrix_4a)
# Use 4/5 of the data for training, 1/5 for testing
ntrain_4a = floor(fm_4a_size[1]*4/5)
ntest_4a = fm_4a_size[1]-ntrain_4a
set.seed(1)
ind_train_4a = sample(1:fm_4a_size[1],ntrain_4a)
ind_test_4a = c(1:fm_4a_size[1])[-ind_train_4a]
```

Here `ind_train_4a` is the set of indices for the training data and `ind_test_4a` is the set of indices for the testing data. What are the first 5 files in the training set? What are the first 5 files in the testing set?

b. (5 Points) Do PCA on your training set and use the first 25 scores to represent your data. Specifically, that means creating the mean face from the training set, subtracting off the mean face, and running `prcomp()` on the resulting image matrix. Project your testing data onto the first 25 loadings so that it is also represented by the first 25 scores. Do not rescale the scores. Use 1NN classification in the space of the first 25 scores to identify the subject for each testing observation. In class we discussed doing kNN classification by majority vote of the neighbors; in the 1NN case, there is simply one vote. How many subjects are identified correctly? How many incorrectly? Plot any subject photos that are misidentified next to the 1NN photo prediction.

c. (10 Points) Rerun parts (a) and (b) using the views `P00A-035E+15`, `P00A-050E+00`, `P00A+035E+15`, and `P00A+050E+00` for all subjects in the CroppedYale directory. Give this matrix the name `face_matrix_4c`. For each image, record the subject number and view in a data frame. Use the following commands to divide the data into training and testing sets:

```
fm_4c_size = dim(face_matrix_4c)
# Use 4/5 of the data for training, 1/5 for testing
ntrain_4c = floor(fm_4c_size[1]*4/5)
ntest_4c = fm_4c_size[1]-ntrain_4c
set.seed(2)
ind_train_4c = sample(1:fm_4c_size[1],ntrain_4c)
ind_test_4c = c(1:fm_4c_size[1])[-ind_train_4c]
```

Do PCA on your training set and use the first 25 scores to represent your data. Project your testing data onto the first 25 loadings so that it is also represented by the first 25 scores. Use 1NN in the space of the first 25 scores to identify the subject for each testing observation. Do not rescale the scores. How many subjects are identified correctly? How many incorrectly? Plot any subject photos that are misidentified next to the 1NN photo prediction.

d. (5 Points) Rerun part (c) with 10 different training and testing divides. Display the number of faces correctly identified and the number incorrectly identified for each. What do these numbers tell us?

e. (10 Points) Compare the results for parts (b) and (c). Are the testing error rates different? What does this tell you about PCA?

f. (5 Points) What happens if we use uncropped photos? Why? Some examples are included in the Resources tab of Piazza. If you would like to try PCA/kNN on the uncropped photos (not required to answer this question, but recommended), you will need to reduce the image sizes. Photos for subjects 1 to 10 do not currently exist in the uncropped database.