Final Project, Introduction to Data Mathematics

Due Date and Place: Monday, May 8th at the final exam time slot, 11:30AM in the scheduled final exam room, Sage 5510. This is where the final presentations will be given.

The Contract: The face company, McStraightface, has heard about your consulting company's great work for Chems-R-Us and has decided to hire you to consult on data analytics. McStraightface's goal is to create a model to predict whether a given image of a face is well centered on the frame and looking straight at the camera. The company wants your consulting company to develop a computational model to predict whether a given image is centered on the frame from left to right and also so the persons head is position so it is aligned straight with the camera. To accomplish this they have provided the data file **facedata.csv and faceclass.csv.** There are 400 rows in both facedata.csv and faceclass.csv and the rows correspond. Each of the 400 rows of the facedata.csv data set consists of a face id with 4096 pixels, each pixel in its own column. The file faceclass.csv has 400 rows with one column consisting of a 1 or -1. It is listed as 1 if the McStraightface visualization team has determined that the corresponding face is looking straight at the camera and is centered in the image and -1 if it is not.

Initial Consultation: Your company had an initial consultation with the folks at McStraightface and it was discussed how there were 4096 attributes and only 400 samples. It was agreed upon that this will mean that you will apply dimensionality reduction with PCA before applying the Fisher method.

Each company should turn-in one paper copy of the project report and R scripts/files and give an 8-10 minute project presentation. Details are as follows.

1. Please use the same consulting company on Chems-R-Us. Each consulting company must consist of 2 to 3 students. You should submit one hard copy project report per company. Provide the name of your company and the names of the students involved in your project report.
2. Produce a clearly-written grammatically correct report which includes the following items.
   (a) An introduction giving an overview of the report.
   (b) A basic description of the data. Describe the size of the data (number of attributes, number of points in each class). For each of the two classes in faceclass.csv, show the mean face of the images in facedata.csv Describe any observations that you have.
   (c) Split the data into training and testing sets where 90% is used for training and 10% is used for testing.
      - Run PCA on the training data.
      - Project the training data onto the first K eigenvectors to get the training matrix.
      - Project the testing data onto the first K eigenvectors to get the testing matrix.
      - Run Fisher LDA on the training matrix and plot histograms and find the errors for the training and testing matrices.
   (d) Do a study comparing using different numbers of eigenvectors K. Determine which value of K your company will suggest if the Fisher LDA will be used. Include in your report a discussion of the optimal number of eigenvectors to use, how you reached this determination, and any supporting evidence you may have. Also include the normal and threshold of the separating hyperplane, and analysis of the testing errors.
   (e) Describe the predictive model you suggest for predicting whether a particular image is a straight and centered face. This could be the model that you have made in the previous step or a new one that you think would be better. Describe the process you use to make the model. Specify the model in full detail. For Fisher or other linear models this is done by specifying the hyperplane normal and threshold. If you do some other type of model, please provide an appropriate equation, description, and R code for the final model.

(f) Report how well your model does in terms of class 1 error, class -1 error, and total error on training and testing data.

(g) Report how well you estimate your model will do on future data. Describe your procedure for estimating this. Give your estimates of class 1 error, class -1 error, and total error.

(h) Provide a conclusion which summarizes your results briefly and adds any observations/suggestions that you have for McStraightface about the data, model, or future work. Part of your conclusion could be about how well the visualization team did in making its classifications of faces.

(i) Optional extra credit: If you think it is appropriate, you could tell McStraightface to fire its visualization team and redo their work.

(j) Optional extra credit: Provide any additional analysis or visualizations that may be insightful to McStraightface or any extra steps you came up with for improving your final predictive model. (use your imagination, extra credit for creativity here).

3. Provide hard copies of one or more R scripts that execute all the results that you gave in your report. If you write additional code, please submit that code as well.

4. Each team should give an 8 to 10 minute oral presentation summarizing your results.

5. McStraightface will evaluate your teams' work using the following criteria:

   (a) (30 pts) Was item 2.c, the study of the number of eigenvectors and the use of Fisher LDA with the 90% train and 10% test set successfully done?

   (b) (10 pts) Was the procedure for constructing the final predictive model well thought-out, described, and executed?

   (c) (10 pts) Was the procedure for evaluating the final model well thought-out, described, and executed?

   (d) (20 pts) Were all of the remaining (non-optional) items above included?

   (e) (10 pts) What is the grammatical quality and clarity of the written report? Did you communicate your results effectively in written form.

   (f) (20 pts) The effectiveness of your 8 to 10 minute pitch of your methods and results at conveying your message. Things to think about: ability to convey message to audience, quality of visual aids, visual aids prepared in advance, clear delivery, and well rehearsed.

   (g) (Extra credit up to 10 pts) Extra Credit will be given for creativity. So feel free to experiment with how you produce the final model and to include additional analysis that may be helpful to McStraightface or that supports your model.