ENGR 421 / DASC 521

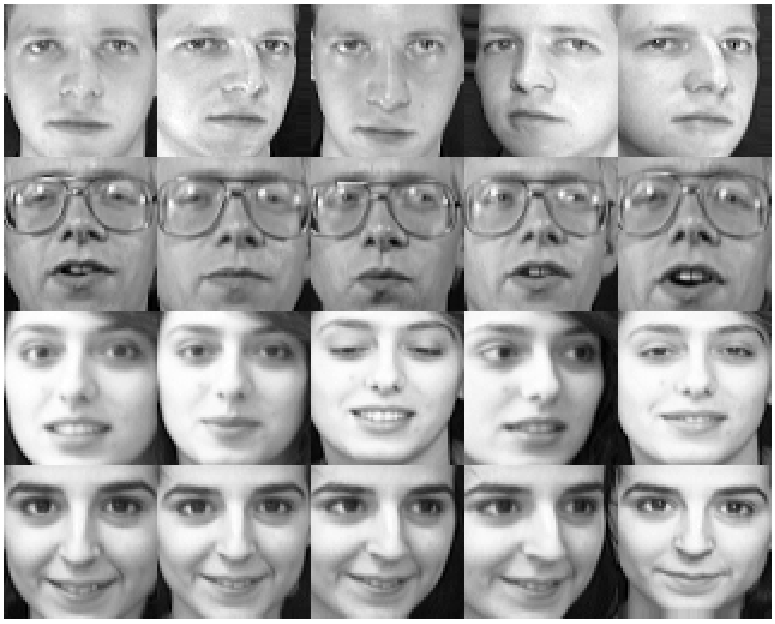
Homework 01: Naïve Bayes' Classifier

Deadline: October 14, 2019, 11:59 PM

In this homework, you will implement a naïve Bayes' classifier in R, Matlab, or Python. Here are the steps you need to follow:

1. Read Section 5.7 from the textbook.

2. You are given a multivariate classification data set, which contains 400 face images of size 64 pixels × 64 pixels (i.e., 4096 pixels). These images are from 40 subjects, where we have 10 images from each subject. The figure below shows ten sample face images from each gender. You are given two data files:
   a. `hw01_images.csv`: face images,
   b. `hw01_labels.csv`: corresponding gender labels (1: female, 2: male).



3. Divide the data set into two parts by assigning the first 200 images to the training set and the remaining 200 images to the test set.

4. Estimate the mean parameters $\hat{\mu}_{1,1}, \hat{\mu}_{1,2}, \ldots, \hat{\mu}_{1,4096}, \hat{\mu}_{2,1}, \hat{\mu}_{2,2}, \ldots, \hat{\mu}_{2,4096}$, the standard deviation parameters $\hat{\sigma}_{1,1}, \hat{\sigma}_{1,2}, \ldots, \hat{\sigma}_{1,4096}, \hat{\sigma}_{2,1}, \hat{\sigma}_{2,2}, \ldots, \hat{\sigma}_{2,4096}$, and the prior probabilities $\hat{P}(y = 1), \hat{P}(y = 2)$ using the data points you assigned to the training set in the previous step. Your parameter estimations should be similar to the following figures. Please note that, in Section 5.7, the naïve Bayes' classifier is derived for binary input features. However, in this homework, the input features are continuous.

```
> print(means[,1])
[1] 0.3796078 0.3982353 ...
> print(means[,2])
[1] 0.3902179 0.3944227 ...
> print(deviations[,1])
[1] 0.1442828 0.1488465 ...
> print(deviations[,2])
[1] 0.1705677 0.1728641 ...
> print(priors)
[1] 0.1 0.9
```

5. Calculate the confusion matrix for the data points in your training set using the parametric classification rule you will develop using the estimated parameters. Your confusion matrix should be similar to the following matrix.

```
        y_hat
y_train   1   2
      1  18   2
      2  24 156
```

6. Calculate the confusion matrix for the data points in your test set using the parametric classification rule you will develop using the estimated parameters. Your confusion matrix should be similar to the following matrix.

```
       y_hat
y_test   1   2
     1  15   5
     2  19 161
```

What to submit: You need to submit your source code in a single file (.R file if you are using R, .m file if you are using Matlab, or .py file if you are using Python) and a short report explaining your approach (.doc, .docx, or .pdf file). You will put these two files in a single zip file named as ***STUDENTID.zip***, where ***STUDENTID*** should be replaced with your 7-digit student number.

How to submit: Submit the zip file you created to Blackboard. Please follow the exact style mentioned and do not send a zip file named as ***STUDENTID.zip***. Submissions that do not follow these guidelines will not be graded.

Late submission policy: Late submissions will not be graded.

Cheating policy: Very similar submissions will not be graded.