CS 479/679 Pattern Recognition
Spring 2023 – Prof. Bebis
Programming Assignment 4 - Due: 5/8/2023 @ 11:59pm

In this assignment, you will experiment with two different classifiers on the problem of **gender classification**: SVMs and Bayesian classifier.

Data Set: The dataset to be used in your experiments contains 400 frontal images from 400 distinct people, representing different races, different facial expressions, and different lighting conditions. The 400 images are equally divided between males and females. The data, which can be downloaded from the course's webpage, contains images of two different sizes: **16x20** and **48x60**; you would need to experiment with both image sizes and compare your results. Each classifier will be evaluated using a **three-fold cross-validation** procedure to account for potential bias due to using a specific training or test set. For this, we have randomly divided the dataset three times as follows:

Fold 1: Training (69M, 65F), Validation (73M, 60F), Test (58M, 75F)
Fold 2: Training (62M, 72F), Validation (58M, 75F), Test (80M, 53F)
Fold 3: Training (71M, 63F), Validation (67M, 66F), Test (62M, 71F)

The validation data should be used for finding an **optimum** set of parameters for SVMs (see below). For each fold, you would need to compute and report the classification error (i.e., percentage of miss-classifications) on the test set of that fold. You should also report the **average** classification error over all three folds as it would be more representative than the classification error on each fold separately.

PCA will be considered again for feature extraction. To avoid discrepancies in your classification results due to incorrectly computing the projection coefficients **Ω** (i.e., eigen-coefficients), we have already pre-computed them for you. Specifically, for each fold, we have used the training data in that fold to compute the covariance matrix and its eigenvectors/eigenvalues. Since there are 134 training images in each fold, there will be **134** eigenvalues/eigenvectors for each fold. These are stored in files ***EigenValues_xx*** and ***EigenVectors_xx*** where **xx** corresponds to the specific fold (see "README" for details). For each fold, we have projected the images in the training/validation/test sets on the eigenvectors of that fold and stored the eigen-coefficients in ***trPCA_xx***, ***valPCA_xx*** and ***tsPCA_xx*** where **xx** corresponds to the specific fold again. Therefore, each of these files contains N lines (i.e., N corresponds to the number of images in each set) with each line containing **134** eigen-coefficients. The class labels (1 for male and 2 for female) are stored in **T*trPCA_xx***, **T*valPCA_xx*** and **T*tsPCA_xx*** while the actual images are stored in ***tr_xx***, ***val_xx***, and ***ts_xx*** folders.

Use only the first **30** eigen-coefficients for each image in your experiments (i.e., the ones corresponding to the projection on the first **30** principal components). Please note that you would **not** need to use the eigenvalues/eigenvectors in your experiments; they have only been provided for completeness (e.g.. in case you need to perform any reconstructions).

Experiment 1: Apply Support Vector Machines (SVMs) for gender classification. For this, you will use the C-SVM classifier from the ***LibSVM*** which you can download using the link provided on the course's webpage. Experiment both with **polynomial** and **RBF** kernels as well as with different C values. In the case of the polynomial kernel, you need to try d=1, 2, and 3. It should be noted that LibSVM has two extra parameters (γ and c_0) for the polynomial kernel; set γ=1 and c_0=0. In the case of the RBF kernel, you need to try γ=0.1., 1, 10, and 100. For the C parameter, try C=0.1, 1, 10, 100. To find the best set of (γ, C) parameters, **you need to use the validation set** in each fold. The idea is to train the SVM on each combination of (γ, C) parameters (i.e., there are 16 possible combinations) and use the validation set to find out which set of parameters yields the least number of misclassifications; let's call (γ_{opt}, C_{opt}) the optimum set of parameters. Using the SVM model trained on (γ_{opt}, C_{opt}), compute the classification error on the test set. This process must be repeated for each fold separately to compute the classification error for each fold as well as the **average** classification error over all folds as described earlier. Show your results both for 16x20 and 48x60 size images.

<u>Warning:</u> read carefully the "README" file included in the LibSVM software download as well as the paper "A Practical Guide to Support Vector Classification" to fully understand how to train/test the SVM classifier (i.e., using "***svm-train***" and "***svm-predict***"), how to scale the data (i.e., using "***svm-scale***"), and how to store the data in the required format. You should also try to run some of the examples provided to make sure that you use the software correctly. Improper scaling of the data, for example, will not yield the highest possible classification accuracy. Going though these steps carefully is very important before you perform your own experiments.

Experiment 2: For comparison purposes, apply the Bayes classifier on the same problem. Model the male and female classes using the training data sets assuming a Gaussian distribution for each fold (i.e., the validation data will not be used in this experiment). Use ML estimation to estimate the parameters for each class. Since PCA de-correlates the data, you can assume that the covariance matrix for each class is diagonal (set the off-diagonal elements to zero) and use the Mahalanobis distance for classification. For a fair comparison, use the same eigen-coefficients (i.e., corresponding to the top 30 principal components) as in the case of SVMs but do not scale them as required by LibSVM (i.e., scaling is accomplished in this case by the Mahalanobis distance which divides the distances by the eigenvalues as we have seen in class). Assume equal prior probabilities (e.g., $P(\omega_1) = P(\omega_2)$) for the male and female classes. Compare your results with those obtained using SVM and discuss your findings.