August-November 2024 Semester
CS616: Statistical Pattern Recognition
CS612: Statistical Pattern Recognition Laboratory
Programming Assignment 2

Date: 12th September 2024

Deadline for submission of code and report: **Sunday, Oct. 13, 2024, 10:00 PM**

Datasets:

Dataset 1: Nonlinearly separable classes: 2-dimensional data of 2 or 3 classes that are nonlinearly separable. (Same data used in Assignment 1.)

Dataset 2: Real world data set:

(a) Two-dimensional speech dataset (vowel data) used in Assignment 1

(b) 3 class scene image datasets

(c) Cervical cytology (cell) image dataset

Data of each class is given separately. For Dataset 1 and Dataset 2(a), divide the data from each class into training, and test data. From each class, train, and test split should be 70% and 30% respectively. For Dataset 2(b) and Dataset 2(c), training and test sets are given.

Note: Each batch of students must use the datasets identified for that batch.

1. Classification task:

Build the Bayes classifier using GMM on Dataset-1, Dataset-2(a) and, Dataset-2(b) Parameters of GMM are to be initialized using K-means clustering.

Note:

i. **Perform the experiments on different number of mixtures of GMM (For e.g. 1, 2, 4, 8, 16, 32, 64).**

ii. **Perform the experiments on Dataset 2(b) using set of 24-dimentional colour histogram feature vectors and 32-dimentional bag-of-visual-words (BoVW) feature vector separately. Report the results for both the representations on different number of mixtures (For e.g. 1, 2, 4, 8, 16).**

2. Segment the cell images by clustering the local feature vectors from cell image datasets into 3 groups using (a) K-means clustering and (b) Modified K-means clustering (using Mahalanobis distance).

Note: Both the K-means clustering methods are initialized by the same initial centres.

Report should include the results of studies presented in the following forms for each classifier and for each dataset:

1. Classification accuracy, precision for every class, mean precision, recall for every class, mean recall, F-measure for every class and mean F-measure on test data (for each of the different parameters).
2. Confusion matrix based on the performance for test data (for the best GMM model).
3. Constant density contour plot for all the classes with the training data superposed **(only for Dataset-1** and **Dataset 2(a) on best model)**.
4. Decision regions plot with the training data superposed **(only for Dataset-1** and **Dataset 2(a) on the best model)**
5. Comparison with the results from the Assignment 1**.**
6. Result should also consist of plot of 3 clusters on training data of **Dataset 2(c)** and the result of cluster projected on test images (i.e., segmentation).
7. Report should also include the graph of ***iterations vs log likelihood*** for all the datasets with different number of components.

Features to be extracted from images of Dataset 2(b) and Dataset 2(c):

1. Features from images of Dataset 2(b):

1. **Colour histogram feature:**
   - Consider 32 x 32 nonoverlapping patches on every images (from training and test sets). For example, if image size is 256 x 256, there will be 32 number of 32 x 32 nonoverlapping patches.
   - Extract 8-bin colour histogram from every colour channel (R, G and B) from a patch. It results in 3, 8-dimentional feature vectors. Concatenate them to form 24-dimentional feature vector.
   - Similarly extract 24-dimentional feature vector from every patch.
   - Stack the 24-dimentional feature vectors corresponding to every patch in an image and save them as a file in the corresponding class folder.
   - Thus, an image is represented as set (collection) of 24-dimentional colour histogram vectors representation
   - Repeat the above steps to all the images in training and test sets of all the classes.

Colour histogram computed as follows from a colour channel:

NOTE: Colour histogram is computed at every patch

   - When the given image is read, it will be read as 3-dimentioanl matrix of pixel values. Each patch of an image is also a 3-dimentioanl matrix of pixel values. Each dimension is corresponding to a colour channel. The pixel values in each colour channel are in the range 0 to 255.
   - For a colour channel,
     - Divide this range into 8 equal bins.
     - Count the number of pixels falling into each bin. This results in a vector of 8 values.
     - This is the 8-dimentional colour histogram (from a colour channel) feature vector.
     - Normalise this vector by dividing it by the number of pixels in that patch.
   - Do the same for other colour channels. Concatenate those three 8-dimentional colour histogram vectors to form 24-dimentional vector.

2. **Bag-of-visual-words (BoVW) feature using K-means clustering:**
   - Take the 24-dimentional colour histogram feature vectors of all the training examples of all the classes.
   - Group them into 32 clusters using K-means clustering algorithms.
   - Now take an image, assign each 24-dimentional colour histogram feature vector to a cluster.
   - Count the number of feature vectors assigned to each of the 32 clusters.
   - This results in a 32-dimentional BoVW representation for that image.
   - Normalise this vector by dividing it by the number of 24-dimentional histogram feature vectors in that image.
   - Repeat this for every image in training and test set.

2. Features from images of Dataset 2(c):

- Consider 7 x 7 overlapping patches with a shift of 1 pixel on every training cell images.
- Compute mean and standard deviation of intensities of pixels in the 7 x 7 patch.
- Thus a 7 x 7 patch is represented as 2-dimentional feature vector.
- In the similar way compute 2-dimentional feature vector from every patch from every training image.
- Stack all the 2-dimentional feature vectors in a file.
- ***For test images:*** *Each test image is represented as a separate file of stacked 2-dimentional feature vectors.*

Each group of students must use the dataset identified for that group only.

Expectation of the assignment is to implement from scratch using Python or MATLAB or any other programming language.
Note: You are not supposed to use libraries of Bayes classifier, multivariate Gaussian distribution, likelihood, K-means clustering, GMM etc.

Report should be in PDF form and report by a team should also include the observations about the results of studies.

Instruction:

Upload in Moodle all your codes in a single zip file.

- **Give the name of the code folder as Group<*number*>_Assignment2_code**
  Example: Group01_Assignment2_code.
- **Give the name of the zip file as Group<*number*>_Assignment2_code.zip**
  Example: Group01_Assignment2_code.zip

Upload the report as PDF file.

- **Give the name to the report file as Group<*number*>_Assignment2_report.pdf**
  Example: Group01_Assignment2_report.pdf

We will not accept the submission if you don't follow the above instructions.