STATISTICAL PATTERN RECOGNITION (FALL 2021)
HOMEWORK#5: K-MEANS, GMM AND SVM

Due date: 30th Jan 2022

In order to do this homework, you have to go through K-means, SVM and GMM classification theories and concepts.

K-means

Dataset : bird.tiff.zip

In this exercise, you will implement the K-means clustering algorithm and apply it to compress an image. You will use your K-means algorithm to select the 16 colors that will be used to represent the compressed image. Concretely, you will treat every pixel in the original image as a data example and use the K-means algorithm to find the 16 colors that best group (cluster) the pixels in the 3-dimensional RGB space. Once you hav computed the cluster centroids on the image, you will then use the 16 colors to replace the pixels in the original image.

- Implement K-means algorithm.
- Explain How you would select initial center points in K-Means?
- Use K-means to perform image compression on given data.
- Plot your original image alongside to the reconstructed one.
- Choose the best K and explain why you did chose that specific K?

In a straightforward 24-bit color representation of an image,1 each pixel is represented as three 8-bit unsigned integers (ranging from 0 to 255) that specify the red, green and blue intensity values. This encoding is often refered to as the RGB encoding. Our image contains thousands of colors, and in this part of the exercise, you will reduce the number of colors to 16 colors. By making this reduction, it is possible to represent (compress) the photo in an efficient way. Specifically, you only need to store the RGB values of the 16 selected colors, and for each pixel in the image you now need to only store the index of the color at that location (where only 4 bits are necessary to represent 16 possibilities). The original image required 24 bits for each one of the 128 x 128 pixel locations, resulting in total size of 128 x 128 x 24 = 393,216 bits. The new representation requires some overhead storage in form of a dictionary of 16 colors, each of which require 24 bits, but the image itself then only requires 4 bits per pixel location. The final number of bits used is therefore 16 x 24 + 128 x 128 x 4 = 65,920 bits, which corresponds to compressing the original image by about a factor of 6.

STATISTICAL PATTERN RECOGNITION (FALL 2021)
HOMEWORK#5: K-MEANS, GMM AND SVM

Gaussian Mixture Model Classifier (GMM)

Dataset：User Knowledge Modeling Data Set (UKM.xls), Iris, Vehicle.dat, Health.dat

UKM: https://archive.ics.uci.edu/ml/datasets/User%20Knowledge%20Modeling

Iris: https://archive.ics.uci.edu/ml/datasets/Iris

In this part, Gaussian Mixture Model (GMM) is used as a generative classifier. You can use the GMM toolbox in MATLAB or scikit-learn library from python which uses the Expectation Maximization (EM) to train a GMM model. A GMM model can be employed to estimate the PDF of some samples (like a parametric density estimator). Here, you should train an individual GMM model (with K Components) for each class. Therefore, N GMM models will be created where N shows the number of classes. The label of a sample can be determined using Maximum Likelihood (ML) criteria. In another words, you should find the likelihood of a sample in all classes and then select the class with the maximum likelihood as the label of the sample.

- Plot the training data (Different colors for each class).
- Construct a GMM classifier, with K = 1,5,10, Gaussian components and train on Train Data.
- For each k, plot the test data classified by the GMM classifier.
- Use five-time-five-fold cross validation to determine the best K.
- Report the train and test accuracy for the best K.

Support Vector Machine (SVM)

- Linear SVM

  Dataset: Use “Dataset1.mat” which is a 2D and 2-class dataset to do this part.
  - Train the SVM using two different values of the penalty parameter, i.e., C=1 and C=100.
  - Plot the data and the decision boundary.
  - Report the train accuracy for both C=1 and C=100.
- Kernel SVM for two-class problem

  In general, SVM is a linear classifier. When data are not linearly separable, Kernel SVM can be used. Here, you will utilize SVM with RBF kernel for non-linear classification. Perform the following step for “Dataset2.mat” and “Health.dat” datasets.

STATISTICAL PATTERN RECOGNITION (FALL 2021)
HOMEWORK#5: K-MEANS, GMM AND SVM

- Train SVM with the penalty parameter C and the standard deviation for RBF kernel. Determine the best value C by ten-time-ten-fold cross validation. Note: It is better to test the values in multiplicative steps such as 0.01, 0.04, 0.1, 0.4, 1, 4, 10 and 40. Therefore, you should evaluate 64 (82) different models to select the best model.
- Plot train and test accuracies and their corresponding variances of five-time-five-fold cross validation for different values of C and σ.
- Plot the data and the decision boundary for "Dataset2.mat" (for best model)
- Report the test accuracy using the selected model (best C and σ)

Notes:

- ✓ Pay extra attention to the due date. It will not extend.
- ✓ Be advised that submissions after the deadline would not grade.
- ✓ Prepare your full report in PDF format and include the figures and results.
- ✓ Submit your assignment using a zipped file with the name of "StdNum_FirstName_LastName.zip"
- ✓ Feel free to use your preferred programming languages.
- ✓ Using other students' codes or the codes available on the internet will lead to zero grades.