**ECE 8527: Introduction to
Machine Learning and Pattern Recognition**

# HW No. 4: Gaussian Mixture Distribution Parameter Estimation

You might find this tutorial useful:

*https://www.geeksforgeeks.org/gaussian-mixture-model/*

Let’s start with 1D data that can be easily visualized. The tasks to be accomplished in this homework assignment are:

1. Generate 10,000 data points for three 1D Gaussian distributions with means of -2, 0, 2 and variances of 1.0, 0.5 and 1.0, respectively. Pool the data together (30,000 points) and estimate the mean and covariance. Plot a pdf of the original data estimated using a histogram and overlay a plot of the estimated Gaussian distribution (consider this a single mixture component). Comment on the nature of the fit.
2. Next, estimate the coefficients of a two-mixture and three-mixture component distribution. Generate a second plot by overlaying plots of these pdfs on the original pdf in (1). Comment on what you observe.
3. Estimate the coefficients for a four-mixture distribution and plot the original pdf along with the two-mixture and four-mixture model.
4. Compute the log probability of the data given the model as you vary the number of mixture components over the range N = [1,10]. Plot the log probability as a function of the number of mixtures. Explain what you observe. Does it make sense?

For the next part of this assignment, you will use the data set located here:

*https://www.isip.piconepress.com/courses/temple/ece\_8527/resources/data/set\_13/*

We will focus on the files *train.csv* and *eval.csv*, which contain training and evaluation data. You can assume the loss function weighs all errors equally, and that the priors are equal.

1. Consider each class separately for the training data. Using a Gaussian mixture model trained on each class independently, plot the log probability of the data given the model, as we discussed in class, as a function of the number of mixture components to the range $N = [1,10]$. Create one plot with two lines, one for each class. Explain what you observe. What is the optimal number of mixture components for each class?
2. For one of the classes, estimate the density function using a histogram technique, and plot it using a 3D plot. Then create a similar plot for the GMM model for $N = 1, 2, 4$. Compare these four plots and discuss what you observe. You might find this tutorial useful:

*https://blog.finxter.com/how-to-plot-a-3d-normal-distribution-in-python/*

1. Using these mixture models, classify both the training data and the evaluation data as a function of the number of mixture components. You must use a posterior probability calculation for this part. Assume equal priors. Report **the errors rates, as a percentage,** in a table using the template below.
2. In your table below, include your results for your Python implementation of QDA and your IMLD baseline from previous homework assignment. Analyze your results.

Results should be reported using this table:

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Data** | **Train** | **Eval** |
| IMLD - QDA | Set No. 13 | ??.??% | ??.??% |
| Python - QDA | Set No. 13 | ??.??% | ??.??% |
| GMM (N=1) | Set No. 13 | ??.??% | ??.??% |
| GMM (N=2) | Set No. 13 | ??.??% | ??.??% |
| ... | Set No. 13 | ??.??% | ??.??% |

Be sure you use two decimal points of precision in your table. This is very important from a statistical significance point of view, which we will discuss later in the class.