# Assignment 8

Machine Learning COMS 4771

Spring 2014, Itsik Pe'er

Assigned: March 31st  Due: Wednesday, April 9th 1:10pm

Submission: Your submission folder on Courseworks. Submit folders for `Assignment08.Problem01,` and `Assignment08.Problem02`

1) This question pertains to the variables `X` and `y_int` in the midterm data file `data.mat`. Apologies to those of you who have already done this or similar analysis during the midterm. Consider the entire set of $N$=10000 input datapoints (rows of `X`) as well as the two classes of datapoints (rows of `X` corresponding to the different values of `y_int`) hereby defined `X1` and `X2`. Use principal component analysis for eigenvector decomposition of each of these three sets. Create three scatter-plots, each in 3D, and each of all $N$ datapoints, color coded by class. The first scatter-plot will be along the coordinate systems defined by the top principal components of `X`. The second and third scatter-plots will be along the coordinate systems defined by the top principal components of `X1` and `X2`. Submit three MatLab figure files for the three plots (`scatterX.fig`, `scatterX1.fig`, `scatterX2.`
   [20pt]
2) Consider the exponential distribution defined in Assignment 7 Problem 2. A mixture of exponentials is a random variable whose distribution has a parameter $\lambda$ chosen at random among $\{\lambda_i\}$ , $i$=1,..,$K$ with respective prior probabilities $\{\pi_i\}$ . This would model, e.g., your pile of pistachios being selected at random from among $K$ varieties, each with its own rate of spontaneous combustion. Write a function `SimMixExps` to simulate data from this distribution per the attached prototype. Assume, w.l.o.g. $\lambda_i$ are in increasing order.
   [20pt]
3) Develop EM for inferring $\{\lambda_i\}$ and $\{\pi_i\}$ from data.
   a) Define the hidden variables, mixture proportions, responsibilities. Write down the log likelihood, and the expected log likelihood. Develop the update equations for each E-step and M-step.
      [15 pt]
   b) Implement (a) in `EMExps` per the attached prototype
      [15pt]
   c) Choose particular $\{\lambda_i\}$ and $\{\pi_i\}$ values, for which you will benchmark the performance of EM as a (plotted) function of $N$. Measure performance in two ways: root-sum-of-squared-differences for $\{\lambda_i\}$ and root-sum-of-squared-differences $\{\pi_i\}$. Choose a range for $N$ that would take you from poor to great performance. This will depend on the values you choose. Submit the MatLab figures for both (`PlotRMSDLambda.fig`, `PlotRMSDPi.fig` ) and the code to do this: a script `MakePlotsRMSD.m`
      [15pt]

Good luck!