EE 511 Simulation Methods for Stochastic Systems
Project #3: Clustering… In many Ways

[Testing Faith]

Download the “old faithful” data set from blackboard. This contains samples of a 2-D random variable: the first dimension is the duration of the old faithful geyser eruptions. The second is the waiting time between eruptions. Generate a 2-D scatter plot of the data. Run a k-means clustering routine on the data for k=2. Show the two clusters in a scatterplot.

[EM]

- Write a 2-dimensional RNG for a Gaussian mixture model (GMM) pdf with 2 sub-populations. Use any function/sub-routine available in your language of choice.
- Implement the expectation maximization (EM) algorithm for estimating the pdf parameters of 2-D GMMs from samples (Refer to the Noisy Clustering Paper linked on blackboard for the relevant update equations).
- Compare the quality and speed your GMM-EM estimation on 300 samples of GMM distributions featuring each of the following: a) spherical covariance matrices, b) ellipsoidal covariance matrices, and c) poorly-separated subpopulations.
- Apply your GMM-EM algorithm to fit the “old faithful” data set to a GMM pdf with two components.

[Clusters of Text]

Download the “nips-87-92” data set from blackboard. This contains a *bag-of-words* data set for NIPS papers from 1987-1992. Columns in this bag-of-words model represent the (scaled) number of times a specified word appears in the different documents. The first column specifies a document id for each paper. Each row has over 11000 dimensions. So scatterplots are not useful.

- Run your k-means clustering method to cluster the row vectors of the data set. Try different values of k and pick the best value
- Report your clusters by listing the document ids for each cluster.

Turn in:

- A summary of your experiments including histograms of your data, any relevant plots, and relevant numerical results
- brief discussions of the results
- a print out of your code.