CS 445/545: Machine Learning, Spring 2020

Programming Assignment #3

A. Rhodes

Note: This assignment is **due by Wednesday, 6/3 by 1000pm**; you will turn in the assignment by email to our grader, as instructed below.

Data set: The data set is 2d data (for ease of visualization) simulated from 3 Gaussians, with considerable overlap. There are 500 points from each Gaussian, ordered together in the file. The dataset is posted on D2L with this assignment.

Assignment #1: K-Means

Implement the standard version of the K-Means algorithm as described in lecture. The initial starting points for the K cluster means can be K randomly selected data points. You should have an option to run the algorithm *r* times from *r* different randomly chosen initializations (e.g., *r* = 10), where you then select the solution that gives the lowest sum of squares error over the *r* runs. Run the algorithm for several different values of K and report the sum of squares error for each of these models. Please include a 2-d plot of several different iterations of your algorithm with the data points and clusters.

Assignment #2: Fuzzy C-Means

Implement the standard version of the fuzzy c-means (FCM) algorithm as described in lecture. As shown in lecture, the update formulae for the centroids and membership weights are as follows:

$$\mathbf{c}_k = \frac{\sum_x w_k(\mathbf{x})^m \mathbf{x}}{\sum_x w_k(\mathbf{x})^m} \qquad w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| \mathbf{x}_i - \mathbf{c}_j \right\|}{\left\| \mathbf{x}_i - \mathbf{c}_j \right\|} \right)^{\frac{2}{m-1}}}$$

Where *m* > 1 is a "fuzzifier" parameter (just fix this value during the algorithm – you are welcome to experiment by trying different values for different runs of FCM if you wish). Begin by initializing the centroids randomly, then compute the weights, update the centroids, recompute the weights, etc. As before, you should have an option to run the algorithm *r* times from *r* different randomly chosen initializations (e.g., *r* = 10), where you then select the solution that gives the lowest sum of squares error over the *r* runs. Run the algorithm for several different values of K (where K is the number of clusters) and report the sum of squares error for each of these models. Please include a 2-d plot of several different iterations of your algorithm with the data points and clusters.

Report: Your report should include a short description of your experiments, along with the plots and discussion paragraphs requested above and any other relevant information to help shed light on your approach and results.

Here is what you need to turn in:

- Your report.
- Readable code.

How to turn it in (read carefully!):

- Send these items in electronic format to our TA by106pm on the due date. No hard copy please!
- The report should be in pdf format and the code should be in plain-text format.
- Put "[CS 545] PROGRAMMING #3: your_name" in the subject line.