Rachael Kenney

Phone (513) 746-6219

Algorithms for Data Science

Homework 4

Due 7/29/2019

1) Create an Expectation Maximization Method (Collaborative)

   Code is in python file hw4.py. To run entire file, need to provide an argument for the location of the iris data set as an arff file for problem 2.

   Used https://github.com/mcdickenson/em-gaussian/blob/master/em-gaussian.py to create EM method, made adjustments move method into a function that could take in variables for the dataframe, guess (initial mu and sigma), and optionally the maximum number of iterations to perform and how closely the convergence of the last two iterations must be to stop.

   Analysis: Method is converging after 2 iterations no matter what the initial guess is. The separations observed do not match what was in the expectation maximization slides though. My 2 clusters both cluster around the points [2, 4]. However, the shape of each cluster appears to be different.

2) Use EM from problem 1 on iris data set (Collaborative)

   Code is in python file hw4.py. To run entire file, need to provide an argument for the location of the iris data set as an arff file for problem 2.

3) Consider three mean values of μ = [μ1, μ2, μ3] = [4.5, 2.2, 3.3] with a corresponding covariance matrix as follows: Σ = [[0.5 0.1 0.05] [0.1 0.25 0.1] [0.05 0.1 0.4]] The respective minimums are min = [3.5, 1.7, 2.5] and maximums are max = [5.5, 2.7, 4.1]. Generate 300 observations. Using the EM algorithm from Problem 1 and the generated date estimate the unknown parameters μk, σk, pk.

   Code is in python file hw4.py. To run entire file, need to provide argument for the location of the iris data se as an arff file for problem 2.