SPRING SEMESTER 2017
BIS 692B, CBB 645B, STAT 645B
Statistical Methods in Genetics and Bioinformatics

Homework #3
Due 2/9/2017 (submit homework in class, printed or handwritten, with source code attached)

(1) Convergence of *k*-means clustering

Show that the objective function of *k*-means clustering algorithm is non-increasing in each step you try to reassign a sample. You only need to show this for one dimension case.

(The monotone convergence theorem suggests that a monotone bounded sequence will converge. For *k*-means clustering, the objective function is bounded by 0 and it is non-increasing in each step. So the objective function will converge to local optimal.)

(2) *k*-means clustering and model-based clustering

Simulate 500 samples (group label 1) from N(0, 1) and 500 samples (group label 2) from N(2,1).

1. Use the EM algorithm for two components Gaussian mixture model to estimate the means, variances and pi. Write your own code for the algorithm. Compare the estimated parameters with the true value.
2. Calculate the posterior probability of group assignment (based on the estimated parameters by the EM algorithm). To assign group membership, you need to select a cut-off for the posterior probability: 0.5 is a reasonable choice. Calculate the misclassification error.
3. Use *k*-means clustering (R function “kmeans”) with k=2 on the simulated data. Calculate the misclassification error.

Note: be cautious on the switch of labels.