

# CG168 Homework 6

Mark Johnson

due 18th November 2008

Please construct an EM clusterer for the 3 data sets that you used for  $k$ -means clustering (all available in `/course/cog168/asgn/data`).

## Homework:

1. Write an EM clustering algorithm and apply 20 EM iterations for  $m = 2, \dots, 20$  clusters to each of these data sets 20 times with different random jittered initial values. Print out the negative log likelihood at each iteration, and ensure that it decreases at each iteration. For each data set and each  $m$ , report the average and the lowest value of the negative log likelihood. Does EM seem to get trapped in local minima? How many clusters do you think are actually present in each of the data sets?
2. Have your program print out the most probable cluster for each data item. For your “best” clustering with 10 clusters for `motherese.txt` and `brown.txt`, report the 10 most salient words in each cluster.
3. (200-level) In our discussion of Naive Bayes, we pointed out that smoothing with a Dirichlet prior with parameter  $\alpha$ , and raising the  $P(X_j|Y)$  components of the Naive Bayes model to a power  $\beta$  sometimes improved the performance of the classifier. Write out the formula for the EM updates where the Naive Bayes model that underlies our classifier has been modified to include these. Then modify your EM code, and experiment to see if these make any difference to the clusters produced by the EM clusterer.