Lab 3 - Stat215A, Fall 2014

Due: Tuesday, October 21, 4:00 PM

1 Parallelizing k-means

We will be investigating the stability of k-means using a popular procedure outlined in Ben-Hur et al. [2001], which uses stability as a guide for picking k. The procedure is outlined in algorithm 1. You should consult the paper for more details, particularly regarding the similarity measures you can use.

In this part of the lab, you will be implementing this method on the binary-coded linguistic data from Lab 2. Pleas use the Rdata file included in this lab to ensure consistency. Set the $k_{max} = 10$, $n = 100$, and m as large as you can get it while also running in a reasonable time (no less than .2, no more than .8).

This will take a lot of computation, so you should do the following things to deal with this:

1. Parallelize the outer loop of this method using `foreach`. Run your job on the SCF cluster using `qsub -pe smp` k_{max}. Look here for detailed instruction on using the cluster: `http://statistics.berkeley.edu/computing/servers/cluster`

2. To compute the similarity of clusterings, you will (in theory) be dealing with a $k \times k$ matrix, where k is the size of intersect (if $m = .8$, k will be approximately 29,000, meaning C will be a 6GB matrix). This could be prohibitively large. You can use any similarity measure mentioned in the paper - correlation, Jaccard, matching.

   (a) Write a function to calculate the similarity in R that will actually complete in reasonable time up to inputs of size 5000.

   (b) Write a memory-efficient version of this function to calculate similarity in C++. Can you avoid storing the $k \times k$ matrix? Compare the timing of this function to your R version.

3. Make a plot similar to Figure 3 in Ben-Hur et al. [2001]. What would you pick as k for this dataset? Do you trust this method?

In this lab, you will graded primarily on your code. As always, please follow the Google R style guide. Make sure your variable names are meaningful and write comments liberally. So that your code can be run on any computer, please encode the working directory in a single variable that can be easily changed and refer to all data files using this variable. The writeup can be a page or two containing the running times from 2, the picture from 3, and your answer to question 3.

References

Asa Ben-Hur, André Elisseeff, and Isabelle Guyon. A stability based method for discovering structure in clustered data. In *Pacific symposium on biocomputing*, volume 7, pages 6–17, 2001.

Algorithm 1 Calculation of clustering similarities in k-means

for $k = 2$ **to** k_{max} **do**
  for $i = 1$ **to** n **do**
    $\text{sub}_1 = \text{subsample}(X, m)$, a subsample of fraction m of dataset X
    $\text{sub}_2 = \text{subsample}(X, m)$, a subsample of fraction m of dataset X
    $L_1 = \text{cluster}(\text{sub}_1)$
    $L_2 = \text{cluster}(\text{sub}_2)$
    $\text{intersect} = \text{sub}_1 \cap \text{sub}_2$
    $S(i,k) = \text{similarity}(L_1(\text{intersect}), L_2(\text{intersect}))$
  end for
end for