Lab 3 - Stat215A, Fall 2017

Due: Monday, October 23, 9:00 PM (NOTE DIFFERENT DUE DATE)

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 lab, you will be implementing this method on the binary-coded linguistic data from Lab 2. Please use the .Rdata file included in this lab to ensure consistency. Set the maximum number of clusters to consider: $k_{max} = 10$; the number of repeated iterations: $N = 100$; and the sampling proportion, m as large as you can get it while also running in a reasonable time (m should be no less than 0.2, no more than 0.8).

This will require a decent amount of computation, so you should try to run this in parallel on the SCF cluster:

1. Parallelize the outer loop of this method using `foreach`. Run your job on the SCF cluster using `sbatch`. See the following reference for detailed instruction on using the cluster: `http://statistics.berkeley.edu/computing/servers/cluster` (specifically see the section called SCF: Parallel jobs).

2. To compute the similarity of clusterings, you will (in theory) be dealing with a $q \times q$ matrix, where q is the the number of data points that are common to each subsample (if $m = 0.8$, q will be approximately 29,000, meaning that the similarity matrix, 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 between two membership vectors in R that will actually complete in reasonable time for inputs up to size 5000.

   (b) Write a memory-efficient version of this function to calculate similarity in C++. Can you avoid storing the $q \times q$ 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 more carefully than usual on your code. As always, please follow the Google R style guide. Make sure your variable names are meaningful and write comments liberally.

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
```