Heidelberg University
Winter semester 2017/2018

Group Parallel and Distributed Systems (PVS)
Artur Andrzejak, Diego Costa

Problem Set 2 for lecture Mining Massive Datasets

Due November 6, 2017, 11:59 pm

Exercise 1 (2 points)

In agglomerative hierarchical clustering one successively merges clusters (lecture 3). At each step the *closest* pair of clusters will be merged. There are several metrics to express cluster distance.

Consider the following six points in $\mathbb{R}^2$: $A = (0,0), B = (10,10), C = (21,21), D = (33,33), E = (5,27), F = (28,6)$.

Perform agglomerative hierarchical clustering on these points for each of the four metrics below. Use dendrograms to visualize your results.

1.1 Centroid distance: the distance between the centroids of the clusters.

1.2 Single linkage: the shortest distance between any pair of points, one from each cluster.

1.3 Complete linkage: the longest distance between any pair of points, one from each cluster.

1.4 Average distance: the average distance between any pair of points, one from each cluster.

Exercise 2 (2 points)

The Spark implementation of the subroutine *merge* shown in lecture 3 (part "Hierarchical Clustering in Spark") operates only on the centroids, and does not maintain information which data points are in which cluster. Write pseudocode of steps 3 & 4 which not only computes the new centroids, but also updates the sets of points corresponding to each cluster.

Exercise 3 (6 points)

In lecture 3, we introduced the metrics between-cluster variation (BCV) and within-cluster variation (WCV). The quotient of these two metrics is a measure for the quality of a clustering. Also, the sum of squared errors (SSE) between each point and its cluster's centroid can be used to assess the clustering quality.

1. Revisit the k-means implementation from lecture 2[1] and make some modifications. Change the stopping criterion to a maximal number of iterations (or until the centroids do not change any more, whichever occurs first). Implement (as additional routines) the metrics BCV/WCV and SSE for evaluating the quality of a clustering (use RDDs `closest` and `centroids`). Submit as your solution the source code and the results of program runs.

[1] Available on IMMD Public Code or on the Moodle at K-Means Source Code - Lecture 2

2. Modify the k-means from above such that it loads data from the file dataset-problemset2-ex3[2]. You can read from the dataset using the numpy load function[3]. Run the k-means for the values $k = 1, \ldots, 20$ with a maximum number of iterations of 50 on this data and output both BCV/WCV and SSE for the final clusterings. Make sure, that you use the dataset each time, so that the results are comparable. Submit as your solution the source code and the results of program runs.

3. Plot the values of BCV/WCV and SSE for each k. I.e., create two line diagrams with k on the horizontal axis and the quality metrics on the vertical axes. Can you conclude what is the appropriate k value to cluster this dataset? For this task, submit your plot and your explanation.

Exercise 4 (Bonus, 4 points)

Read chapter 7.4 in the book *Mining of Massive Data Sets*[4] about the CURE clustering algorithm. Write a pseudocode implementation using Spark transformations and actions.

[2] Available on Moodle at Dataset for Problem Set 2 - Ex 3.2 and 3.3

[3] https://docs.scipy.org/doc/numpy-1.9.0/reference/generated/numpy.load.html

[4] available for free online: http://www.mmds.org/#ver21