Due Date
February 20, 2020

Assignment 3

CS 6665
Spring 2020

k-means on Spark

This problem requires you to implement the k-means algorithm on Spark. The input is a set $\mathcal{X}$ of n data points in the d-dimensional space $R \in d$, the given number of clusters k, and the set of k initial centroids $\mathcal{C}$. The distance between any two points is computed using the Euclidean distance.

The corresponding cost function Φ that is minimized when we assign points to clusters using the Euclidean distance metric is given by:

$$\Phi = \sum_{x \in \mathcal{X}} \min_{c \in \mathcal{C}} ||x - c||^2$$

Task. Implement the k-means algorithm using Spark. In this task, we use $k = 10$. Please use *data.txt* as the data points and *centroid.txt* as the initial centroids.

- data.txt contains the dataset with 1000 rows (i.e., n = 1000) and 20 columns (i.e., d = 20).
- centroid.txt contains 10 initial cluster centroids.

For the convergence condition, you can either set the number of iterations to 20 or use a threshold 0.01.

Run the k-means on data.txt and centroid.txt. Generate a graph where you plot the cost function as a function of the number of iterations.

What to submit

1. The source code (.java or .py files).
2. The plot of cost vs. iteration.