Data Mining: Learning from Large Data Sets - Fall Semester 2015

caifa.zhou@geod.baug.ethz.ch
pungast@student.ethz.ch
llara@student.ethz.ch

December 1, 2015

Extracting Representative Elements

To extract 100 clusters of representative elements from a large dataset, we used k-means ++ and k-means in the mapper and reducer, where the mapper outputs 500 points (cluster centers) for each batch and the reducer finds the final 100 based on the output of the mappers.

The **mapper** is constructed as follows. Data is processed in batches of 10000 data points.

The cluster centers of each stream are initialised using k-means++ to find a solution in reasonable time, and then sequential k-means is run. Weights are assigned to all points, whereby points farther away from the existing centers get a higher weight and thus have a higher probability to be selected. Iteratively the subsequent centers are chosen from the remaining data points with probability proportional to its squared distance ($\|x_i - \mu_j\|_2^2$) from the closest existing cluster center μ_j.

Third, the sequential k-means algorithm is implemented to compute representative elements (μ_j) one at a time. The centers (means) are given by the vector μ with $\mu_1, ..., \mu_k$, the algorithm calculates $\partial L/\partial \mu$ with $\partial L(x, \mu) = min \, \|x_i - \mu_j\|_2^2$. If μ_i is closest to x, μ_i is replaced by

$$\frac{1}{k_i} \sum_{j=1}^{k_i} x_j$$

where k_i is the number of data points assigned to i^{th} cluster. Last, the emit function prints the centers into stdout.

The **reducer** receives the cluster centers from all mappers. As in the mapper, the reducer finds runs k-means++ in this reduced space to find initial cluster centers. Sequential k-means algorithm is then run on these initial centers to find the final representative elements.