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

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

November 28, 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 ouputs 500 clusters and the reducer the final 100.

The **mapper** is constructed as follows. First, it initializes the batch matrix, by splitting the input data with the dimension = 500. Then a stream of size = 10'000 is compiled of the input data. In the second step, the cluster centers of each stream are initialized with use of k-means ++ to find a solution in reasonable time. The first cluster center is thereby chosen uniform at random and weights are assigned to the points, whereby points farer away get a higher weight and thus have a higher probabillity 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 point's x_i 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 $\mu_i + \alpha \times (x - \mu_i)$ with step size $\alpha \in (0, 1)$. Last, the emit function prints the mean vector from each stream.

The **reducer** receives the weight vectors from the mapper, it initializes the batch matrix with the dimension = 500. Then it parses each line into weight vectors of dimension = 500 and generates a stream of size = 100 with the weight vectors. As in the mapper, the reducer finds the initial cluster centers by use of k-means ++. Subsequently, the initialized means are then used for the sequential k-means algorithm to compute k = 100 clusters of representative elements.