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

fgmehlin@student.ethz.ch
matteopo@student.ethz.ch
piusv@student.ethz.ch

December 13, 2015

Extracting Representative Elements

This project is about extracting representative elements from a large image data set. Each image is represented by a set of 500 features. The idea is to extract a subset that best represents the whole dataset.

1 Mapper

In the mapper, the following two approaches were tested :

Coresets via Adaptive Sampling : We tried to implement this method at first by following the algorithm depicted in the lecture. However, the computation was very expensive. Indeed, computing the distance between points and cluster centers as well as the **q** distribution took too much time.

MiniBatchKmeans :
In the final implementation of the mapper we run a batched version of K-means (provided by scipy as MiniBatchKMeans) on the whole data set accessible to the mapper. We try to find a bigger number of cluster centers as are needed in the end. The idea is that this will hopefully summarize small local clusters decently well and approximate something like a core set.

The optimal solution included computing the batched k-means with 600 clusters in each mapper, batch size of 1000 samples and the number of restarts to 10.

2 Reducer

In the reducer we compute regular K-means on the output of all mappers.

3 Participation

As usual, we started by approaching the problem individually, so each of us could deeply understand the requirements of the project. During the few meetings we have had, we tried the aforementioned approaches together and came up with the final version.