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 8, 2015

Large Scale Image Classification

To classify the images into the categories nature and people, using a Support Vector Machine as classifier and Parallel Stochastic Gradient Descent procedure, a Map and Reduce function was implemented.

Summary: we used a primal version of the Support Vector Machine (implemented in sklearn.svm.LinearSVC) and transformed the original features into Random Fourier Features.

The **mapper** is constructed as follows. First, it reads each line, extracts the features and transforms those. Then it adds the transformed features to the batch and calculates the weight vector on that batch, processes the batch and starts a new one until all input is processed. If the batch size is infinite, each mapper treats its whole input as a single batch.

The particular parameters used in the best submission were: infinite batch size and 800 random features.

A batch of examples is processed by first calculating the weight vectors, with $\eta = \frac{1}{\sqrt{t+1}}$, such that if $y\mathrm{w}^{\mathrm{T}}\mathrm{x} < 1$ then update the weights, such that $\mathrm{w}'_{\mathrm{t}} \leftarrow \mathrm{w} + \eta_{\mathrm{t}} y\mathrm{x}$ and set $\mathrm{w}_{\mathrm{t+1}} = \min\left\{1, \frac{1/\sqrt{\lambda}}{\|\mathrm{w}'_{\mathrm{t}}\|}\right\}$.

As the last step of the processing of one batch, the emit function prints the weight vector from the batch into standard output.

The **reducer** receives the weight vectors from the mapper. A weight vector is thereby represented by one input line, which is parsed into a vector. Then the reducer calculates the average of all the vectors received from mappers and prints the space-separated coefficients of the model.

The groupwork was done as following. First we sat together and discussed the assignment and overall structure of the project. Taivo Pungas created a first draft of the mapper and reducer, Caifa Zhou continuously contributed to the structure of the mapper and both tested different versions. Lara Lingelbach wrote the report.