

# 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.

We used a primal Version of the Support Vector Machine and added 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 batch are proceeded. If the batch size is infinite, each mapper treats its whole input as a single batch.

The dimension used by the mapper is set to 400, the batch size is infinite and the transformed dimension equals 800.

The mapper thereby uses the following functions.

The `permute_data` function returns two randomly permuted arrays, which is later used in the `permutation_iter` function. The transform function then converts the original features to random Fourier Features according to  $z_{\omega,b}(x) = \sqrt{2}\cos(\omega^T x + b)$ .

Third, the `emit` function prints the weight vector from one batch of stochastic gradient descent. A batch of examples is processed by first calculating the weight vectors, with  $\eta = \frac{1}{\sqrt{t+1}}$ , such that if  $yw^T x < 1$  then update the weights, such that  $w'_t \leftarrow w + \eta_t yx$  and set  $w_{t+1} = \min \left\{ 1, \frac{1/\sqrt{\lambda}}{\|w'_t\|} \right\}$ . Each row in matrix 'batch' holds an example. Each element in vector 'labels' holds the corresponding label.

With the `permutation_iter` function ... **(im not exactly sure, what that function does)**

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 these vectors 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.