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

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

November 8, 2015

Large Scale Image Classification

This project consists of training a model for classification of two images, namely : Nature and People. From a set of extracted feature, we applied made use of a Stochastic Gradient Descent (SGD) classifier.

1 Mapper

In order to make a justified choice of the several parameters we could tune, we used an estimation of the out-of-sample accuracy of the classification algorithm. To obtain this estimation, every time we run the classifier we randomly split the provided data set into a training set (80% of the dataset) and a smaller test set (20%). This split is made on-line as we stream the data. We then train the classifier on the training set and evaluate it on the test set, so that we obtain an estimate of the out-of-sample accuracy.

The mapper is divided into two subsequent parts :

1. Feature transformation
2. Classification with (SGD)

1.) The method transforms applies the following transformation on the sample features :

- $\widetilde{x_1} = \sqrt{|x|}$
- $\widetilde{x_2} = cosh((\frac{\pi}{2}) * x) - 1$
- $\widetilde{x_3} = sin((\frac{\pi}{2}) * x)$

The output feature is given by the following concatenation: $x_{out} = [1, \widetilde{x_1} + \widetilde{x_2} + \widetilde{x_3}]$. The leading 1 is used as intercept.

NOTE: We also tried to implement Random Fourier Feature transformation but they gave worse results that the aforementioned transformations.

2.) We use a Stochastic Gradient Descent Classifier algorithm with l1 regularizer provided by the sklearn library. We use *hinge* as loss-function which is the default linear SVM.

Finally the mapper will output the tuple (1, feature_weights) for the learned dataset.

2 Reducer

The reducer aggregates the weights from the different mappers to produce the final feature weights.

3 Participation

We started by approaching the problem individually, so each of us could deeply understand the project and try to solve it. After some time, we defined one of the models as the most competent one and tried to improve it individually.