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

eholmer@student.ethz.ch
melisr@student.ethz.ch
pmichal@student.ethz.ch

November 8, 2015

Large Scale Image Classification

Mapper

We decided to use the **Linear Support Vector Classificatier** in `sklearn.svm.LinearSVC` instead of **Stochastic Gradient Classification** because it produces better results within reasonable time. The mapper is straightforward. After all input lines are read, transformed and stored in a matrix of data points, **LinearSVC** fits the model coefficients to the data.
We use default values for all parameters of **LinearSVC**, except for the penalty parameter C of the error term. Cross validation showed that $C = 20$ gives the best results.

Transform

The transforming of the features is the key to our good results. First we use the **Additive Chi Squared Kernel** in `sklearn.kernel_approximation.AdditiveChi2Sampler` to transform the features. This kernel is given by

$$k(x, y) = \sum_i \frac{2x_i y_i}{x_i + y_i}.$$

These created features are used as input for the **RBFSampler** in `sklearn.kernel_approximation.RBFSampler`. This sampler constructs an approximate mapping for the radial basis function kernel, also known as *Random Kitchen Sinks*. We decided empirically to create 16 times as much samples as the original features.

Reducer

The reducer outputs the average of all received model coefficients.

Member contribution

Michal Porvazník	Implemented first version of algorithm
Erik Holmer	Introduced AdditiveChi2Sampler and RBFSampler to the algorithm
Rik Melis	Experimented to find optimal parameters