Data Mining: Learning from Large Data Sets - Spring Semester 2014
Instructor: Prof. Andreas Krause

Course Project
Part II

Large-Scale Image Classification

1 Introduction

In this part of the project your task is to classify images in one of two classes according to their visual content. We will provide a labeled dataset containing two sets of images: **Nature** and **People**.

Although appropriate feature selection is an important part of image classification, it is out of the scope of this course and won't be part of this task. Instead, we will provide a set of features that has been extracted from each image. Your task then is to train a model that works well given the feature representation.

The goal is to solve this classification problem using Support Vector Machines and implement the solution using Parallel Stochastic Gradient Descent.

2 Dataset Description

A set of 400 features has been extracted from each picture. We provide $100k$ images, $50k$ from each category. We only provide the features for each picture, from which the actual image cannot be reconstructed.

The training set is given in the file "training". Each line in the file corresponds to an image and is formatted as follows:

- Elements are space separated.
- The first element in the line is the class $y \in \{+1, -1\}$ which correspond to Nature and People class, respectively.
- Next 400 elements are real numbers which represent the feature values $x_0 ... x_{399}$

3 Evaluation and Grading

Your task is to provide a Map function and a Reduce function written in Python.

The output of the Reduce function should be the set of space-separated coefficients of your final model. The model produced by the Reducer will be used for evaluation on a separate dataset. The prediction of your model on a test instance $\mathbf{x}$ will be calculated as $\hat{y} = \mathbf{w}^T\mathbf{x}$.

If you decide to apply any transformation to the given features your predictions will be given by $\hat{y} = \mathbf{w}^T\phi(\mathbf{x})$. By default $\phi(\mathbf{x}) = \mathbf{x}$.

If you apply transformations to the original features you have to change the "transform" function in the given mapper template, otherwise we will not be able to transform the evaluation data using the same function, and thus, to evaluate your submission. The evaluation code that we are using is provided in the supporting material.

Each submission will be scored according to the classification accuracy which is the percentage of the test instances that were classified correctly.

We will compare the score of the submission with two baseline predictions: a weak one (called "baseline easy") and a strong one (called "baseline hard"). These will have a score of PBE and PBH respectively, calculated as described above. Both baselines will appear in the rankings together with the score of your submitted predictions.

Performing better than the weak baseline will give you 50% of the grade, and matching or exceeding the strong baseline on the **test set** will give you 100% of the grade. This allows you to check if you are getting at least 50% of the grade by looking at the ranking. If your prediction performance on the **test set** ($\mathsf{P}_{\mathsf{test}}$) is in between the baselines, the grade is computed as:

$$\mathsf{Grade} = \left(\frac{\mathsf{PBH}_{\mathsf{test}} - \mathsf{P}_{\mathsf{test}}}{\mathsf{PBH}_{\mathsf{test}} - \mathsf{PBE}_{\mathsf{test}}} \right) \times 50\% + 50\%$$

The number of submissions per team is limited to 50. Time limit per submission is 15 minutes. The number of mappers that will be used is 10. There will be only one reducer. The submission with the highest accuracy on the evaluation set will be used for grading.

3.1 Visual Test

We provided the actual images for a small subset of the training dataset, together with a program that evaluates the predictions made on it and produces a visualization of the classification accuracy of your model. This is provided in the support material under the folder "visual_test" in which you can find more information.

3.2 Report

At the end of the course you will be requested to provide a team report which describes your solution for all project tasks. You should update the template given with the first task and add a section titled "Large-Scale Image Classification". The maximum length of a task report is 2 pages (which means that the final report will contain a maximum of 8 pages). You will be required to upload the reports **after the last project of the course.**

3.3 Deadline

The submission system will be open from **Tuesday, 25.03.2014, 17:00** until **Monday, 14.04.2014, 23:59:59**.