

550.640 Project 4. Due on Monday April 28.

Preparation This project uses three training sets that are available on the class web site. They are dataEasy.txt, dataMedium.txt and dataHard.dat.

The training sets contain one array with 3000 rows and 1025 columns. Each row is a sample. The first column is the class. The remaining 1024 ones are the predictive, binary, variables.

Use the 2000 first rows for training and the 1000 last rows for testing.

Part 1

Provide, for each of the training sets, the classification rates obtained by the following classifiers

- Linear Discriminant Analysis.
- Support Vector Machines. (Use a Gaussian kernel associated to $\exp(-\gamma|x|^2)$ with $\gamma = 0.1$. Describe how the weight on the slack variables is estimated.)
- Nearest neighbors. (Use the standard sum of squares distance)
- Adaboost, where the weak classifiers are given by weighted LDA.

Part 2

Each sample is in fact a binary 32 by 32 image that you can visualize in matlab using `imagesc(reshape(sample(k,2:end), 32, 32))`.

This is an open question: using the information you have obtained by visualizing images belonging to different classes, transform the predictive variables in a way that will simplify the learning task.

Your answer should be written as follows:

(1) Introduction: summarize the structure of the dataset and what (visually) separates the classes.

(2) Data preprocessing: carefully describe and explain how you transformed the original data, and the heuristics of this transformation. The transformation may depend on the dataset.

(3) Results: provide classification results with at least two types of classifiers for each of the training sets.

(4) Analysis: try to explain the difference (or absence of difference) in the performance of the classifiers before and after processing.

Although trying to get the best possible classification rate is fun and encouraged, this rate will only have a minor impact on the grade. Important factors will be the theoretical soundness, the originality (try to work alone) of the approach, and the clarity of the explanations.