Programming Exercise 1

Due on October 26th

From the following repository:

`http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`

Select four datasets:

- A regression dataset with a small training set (`training size`$< 10^3$)
- A regression dataset with a larger training set (`training size`$> 10^3$)
- A classification dataset with a small training set (`training size`$< 10^3$)
- A classification dataset with a larger training set (`training size`$> 10^3$)

In this task we will focus on the training set only. For each regression dataset, apply linear regression on a random subset of the training set of increasing size, i.e. you should randomly select training sets that include more and more data points:

- Plot the approximation error (square loss) on the training set and the required cpu-time as a function of the number of samples (i.e. data points in the training set).
- Explain the behaviour of both curves.
- Plot the learned weights for two different number of training samples. Can you find an interpretation for the learned weights?

For each classification dataset, apply logistic regression on a random subset of the training set of increasing size, and answer the same questions as above.