Programming Exercise 1

Due on October 30th

In this task we will focus on a training set only. Consider the datasets in:

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

1 Regression Task

Choose a regression dataset and apply linear regression on a random subset of the training set of increasing size. You should select training sets that include more and more data points.

1. Plot the approximation error (square loss) on the training set as a function of the number of samples N, i.e., data points in the training set.
2. Plot the cpu-time as a function of N.
3. Explain in detail the behaviour of both curves: what is the trend you observe? does it stabilize? why?
4. Explore how the learned weights change as a function of N. For this, you can make separate stem plots for several values of N. Can you find an interpretation for the learned weights?

2 Classification Task

Choose a classification dataset and apply logistic regression. Repeat the previous four steps using as error the mean accuracy.

3 Remarks

- Feel free to use your favourite software. We recommend Python.
- You can obtain smoother curves by averaging over several permutations of the dataset. Ideally you could also plot the variance, not only the mean.
- Check that your results are in agreement with the theory.
- Submit the report (pdf) with the answers and the code separately.
- Useful links:
  `http://scikit-learn.org/stable/user_guide.html`
  `http://jupyter.org/`
  `https://twitter.com/zacharylipton/status/1167298276686589953?lang=en`