Computational Data Analysis, spring 2017.

Case 1

Hand in March 21 the latest (on CampusNet).

We have generated a synthetic data set with 100 variables and a scalar response. You have 100 observation with (y,x) and an additional 1000 observations of x only. The data is in the file Case1_Data.xls, read it with `readtable('Case1_Data.xls')`.

Your tasks are

- Build a prediction model that can predict y for new features x.
- Make a prediction of outputs for the 1000 x observations with unknown y.
- Give an estimate for your prediction error for the 1000 predictions, as relative RMSE,

  ```
  sqrt(mean((y-yhat).^2)) / sqrt(mean((y-mean(y)).^2))
  ```

- Write a short report (seriously, just a few pages!), describing
  - The model you choose to work with
  - How you handled missing data
  - How you handle the different kinds of features
    - Feature X100 is different...
  - How you made sure that you obtained the best possible model
  - How you made sure that you have the best possible estimate of prediction error

We will have a prize for the best prediction so have your predictions available when we go through the case on March 23.

Work in groups of 1-3 persons.

Trouble? Make a choice and work with that. Are you in doubt, chose a simple solution. Keep It Simple Smarty…