Computational Data Analysis, spring 2018.

Case 1

Hand in March 20 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.csv, read it with `readtable('Case1_Data.csv')`.

Your tasks are

- Build a prediction model that can predict y for new features x.
- Make a prediction of outputs for the 1000 observations with unknown y. You should bring this with you to Lecture 8 where we compare results.
- Calculate an estimate for your prediction error for the 1000 predictions, as relative RMSE,

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

  (You can not calculate the correct RMSE before you have the right “y” but you can make an estimate of what you expect it to be.)

- 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
    - State what rRMSE you expect to have on the 1000 predictions
  - Not more than 2 - 3 pages!

On Lecture 8, March 22, you will be handed the 1000 correct y values such that you can calculate your actual prediction error for the unknown observations. We will have a prize for best prediction and another prize for those who came closest to what they expected.

Work in groups of 1-3 persons.

Hint, it is a good idea to work with linear models and regularization.

Trouble? Make a choice and work with that. Are you in doubt, chose a simple solution.