IST Austria: Data Science and Scientific Computing
Track Core Course 2015/16 – Segment 2: "Predictive Models"

Instructor: Christoph Lampert <chl@ist.ac.at>
TAs: Anna Levina <alevina@ist.ac.at>, Srdjan Sarikas <ssarikas@ist.ac.at>

Exercise Sheet 1

Literature

1) Read this paper: [L. Breiman, *"Statistical Modeling: The Two Cultures"*, Statistical Science 16 (3), 199–231 (2001)] as well as as the comments and replies following it at `http://projecteuclid.org/download/pdf_1/euclid.ss/1009213726`. Answer the following questions in one or two sentences each.

a) what does Breiman identify as the biggest problem of the "statistical community"?

b) how did the "statistical community" react to Breiman's claims?

c) what aligns better with your own interests: data modeling or algorithmic modeling? (or something else?)

d) if you were in charge of IST's Data Science program, how much of which aspects would you have taught?

Hands-On

2) Reproduce the figures about *model selection* for ridge regression shown in the lecture:

- download the `diabetes.txt` dataset from the course website
  use the first 10 columns as features and the last column as target value
- repeat the following experiment 50 times with different random seeds
  - pick a random subset of 100 example for training and 100 examples for validation
  - a train least squares regression model without regularization (using matrix operations)
  - for each regularization parameters $\lambda \in \{2^{-20}, \dots, 2^{10}\}$
    - train a ridge regression model with regularization λ (using matrix operations)
  - evaluate the trained models on the validation set
- plot the results in a style that you find most appropriate/informative

Additional exercises (voluntary, to get more practice)

3) In the above experiments, create two validation sets instead of just one and plot the average absolute difference in error between both of them. Interpret your findings.

4) In the experimental setting above, also compute and plot the error estimate of from 10-fold cross-validation. For each repeat, identify the 'best' regularization constant according to its CV error and evaluate the corresponding error on the validation set. Interpret your findings.