Programming Project: Ridge Regression and Model Selection

Project Definition

Work in groups of two students. Exceptions have to be arranged with the instructor.

1. Choose a nonlinear function $f(x)$ of a scalar variable x. Generate a data set of N data points $t_n = f(x_n) + \epsilon$, where ϵ is drawn from a normal distribution with variance σ^2.

2. Implement regularized linear regression using an M-th order polynomial regression function.

3. Generate a plot like that on Slide 16 of Bishop Ch. 1[1], evaluating E_{RMS} as a function of λ by K-fold cross-validation.

4. Using the full data set of N points, generate plots like those shown on Slides 14–15, including for $\lambda = 0$ and for three representative values of $\lambda > 0$, including overfitting, underfitting, and the best value of λ you identified. Combine plots in the same axes where it makes sense.

Hint

In order to demonstrate the effect of λ, do not choose $N >> M$.

Implementation

You are free to choose almost any programming language. Since this project involves mostly maths and graphics, you are advised to implement your solution in an appropriate, high-level scripting language such as R[2], Python SciTools[3], Octave[4][5], or similar. Non-free programming languages (Matlab) are discouraged. Your code needs to be able to run on the instructor's Linux machine without installing substantial extra software. If in doubt, check with the instructor.

Your program should start from a command line (shell, R or Octave interpreter), should not accept any required parameters on the command line, and and should not read anything from standard input.

Submission

Prepare a report of a few pages with the following content:

- Give the names of all members of the group. If the members' contributions differ substantially, explain their respective contributions.
- Discuss any algorithmic choices you made. Explain any deviations from the methods discussed in class.
- Give illustrations, numbers and results as described above.
- Discuss your results and any interesting insights and observations.

One member of the group should submit the report in PDF format as well as the complete source code as indicated on the course Web page.

[1] https://www.microsoft.com/en-us/research/wp-content/uploads/2016/05/prml-slides-1.pdf
[2] https://www.r-project.org/
[3] https://pypi.org/project/SciTools/
[4] http://www.gnu.org/software/octave/
[5] a free Matlab clone