Computational Statistics
Lab II - linear regression

Luca Bortolussi

Department of Mathematics and Geosciences
University of Trieste

Office 238, third floor, H2bis
luca@dmi.units.it

Trieste, Winter Semester 2016/2017

Outline

TASK 1

- Consider the 1 dim non-linear data in Moodle (regression lab), and split the dataset into training, testing and validation.
- Implement the non-regularised ML linear regression, with a polynomial model of degree $M = 12$. Compute the solution by solving the linear system of equations.
- Implement a regularised linear regression (ridge regression) with a polynomial model of degree $M = 20$.
- Set λ by n-fold cross-validation (choose a reasonable n), and by using a validation dataset. Compare the two approaches.

TASK 2

- Get the 2-dimensional data from moodle (lab2), and split the dataset into training, testing and validation.
- Fit a model using gaussian basis functions:

$$\phi_i(\mathbf{x}) = \exp(\|\mathbf{x} - \mathbf{x_i}\|^2/\gamma^2)$$

  Consider basis functions centred in a $k \times k$ regular grid in $[0, 1]^2$ (choose e.g. $k = 5$).
- Fit the model using ML, and use a validation set to identify the best γ.
- Fit the model using ridge regression, identifying both γ and λ by cross validation.

TASK 3

- Consider the 1D big dataset in moodle. Split it in several small datasets (say 100 points each), and run regression many times, evaluating bias and variance (approximate the integral by sampling).
- Compare the regression results obtained by averaging over 100 regressed curves for datasets of 20 points each, and a single regression over a dataset of 2000 points. Which is better? Which is faster?

SUGGESTIONS

- Try to implement the regression algorithm in a way which is independent on the dimension of the input and on the basis function model selected (the latter should be a parameter of the regression routine).
- Try to implement a matlab function that generates a 1D polynomial basis functions of degree M, for arbitrary M and the 2D Gaussian basis functions, for arbitrary k.

Outline

TASK 3

- Implement Bayesian regression, with type II likelihood optimisation of α and β.
- For the 1d non-linear dataset, use polynomial model of degree 12.
- Plot predictions and 95% confidence intervals, from the predictive distribution.
- For the 2d non-linear dataset, use the Gaussian functions models. How can we set the lengthscale γ?