T-61.3050 Exercise session 2/2013

You should attend only one problem session (Wednesday or Thursday) during a week.

The problems are divided into demonstration and home assignments. The deadline for the home assignments is noon (11:59am) the next Monday after the corresponding exercise session. Please note that late submissions will not be graded. Please, submit your answer to the course e-mail address (t613050@james.hut.fi) as pdf. Alternatively, exercises can be returned on paper to a box (labeled with the course name) in the 3rd floor B-corridor of the Computer Science Building, but we prefer electronic submission. Your answers will be graded, and returned to you at exercise sessions later on.

See `https://noppa.aalto.fi/noppa/kurssi/t-61.3050/etusivu`
for up-to-date information.

Demonstration

1. In comparing classification or regression models, it is often useful to have a *dummy model* as a baseline. If your model beats the dummy model, then it may be of some use. By definition, the prediction of a dummy model is constant with respect to the covariates. Any reasonable classification or regression model (for a given data) should be able to beat the dummy model (sometimes the dummy model is surprisingly good, for example, in classification tasks when the class distribution is strongly skewed). What is a good dummy model for a (i) classification and a (ii) regression task?

2. Why is it better to use the average of S and G (for definitions, see Section 2.1 in the course book) than just any consistent hypothesis as the final hypothesis? You can assume that there is no noise and use family car classification as an example.

3. Derive the solution for linear regression given by Equation (2.16) in the book (in the second edition it is Equation (2.17)). The equations are also given in slide 25 of the third lecture.

Home assignments

1. Derive the general least squares solution for the linear regression problem, i.e. solve for W when W is a $1 \times k$ vector and the solution is given by $\mathrm{W} = \mathrm{R}\Phi^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})^{-1}$ (slide "Least Squares Solution to Regression with Linear Basis Functions" of the third lecture).

2. Download the data set (T-61_3050_data_set.zip) for the problem session from the course webpage. The data is exchange rates of euro versus dollar from 1999 to 2013, with some additional noise. The data set has 3766 data items (rows) and 2 variables x and y (columns) in ASCII format; the first column is the time index (in decades since 1999) and the second column is the euro-dollar exchange rate. There are 50 data items in the the training set, and 3716 in the test set. Your task is to train a regressor (predicting the exchange rate from the time index) using the training set only such that the prediction error on the test set is as small as possible. Train some polynomial regressors of different orders and choose the best one by validation. You need to leave out part of the training set for validation. Write a short report (1 page at most) explaining your method and be sure to compute and include the average squared error on the test set. (Matlab hint: Build the matrix Φ and take its pseudoinverse (function pinv). Check the Matlab files of the first exercise session as well.)

3. Your boss has taken an interest in properties of cars. He wants you to write a report in which you describe the essential properties of the *Auto MPG* data set, available from the UCI Machine Learning Repository. The boss does not understand any advanced methods that you are going to learn during the course, so scatterplots, histograms, means etc. have to suffice. He has a short attention span, so your report should be short (at *most* one page). Perform the analysis with some data analysis software (e.g., R, S, Octave or Matlab) and submit your report to the boss as an answer to this problem.