732A95 Introduction to Machine Learning /TDDE01 Machine Learning
Division of Statistics and Machine Learning
Department of Computer and Information Science

Computer lab 2

Instructions

- Create a report to the lab solutions in PDF.
- Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments.
- **Include all your codes as an appendix into your report.**
- **Use set.seed(12345) for every piece of code that contains randomness**
- A typical lab report should 2-4 pages of text plus some amount of figures plus appendix with codes.
- The lab report should be submitted via LISAM before the deadline.

Assignment 1. Feature selection by cross-validation in a linear model.

1. Implement an R function that performs feature selection (best subset selection) in linear regression by using k-fold cross-validation without using any specialized function like lm() **(use only basic R functions**). Your function should depend on:
    - X: matrix containing X measurements.
    - Y: vector containing Y measurements
    - Nfolds: number of folds in the cross-validation.

You may assume in your code that matrix X has 5 columns. The function should plot the CV scores computed for various feature subsets against the number of features, and it should also return the optimal subset of features and the corresponding cross-validation (CV) score. Before splitting into folds, the data should be permuted, and the seed 12345 should be used for that purpose.

2. Test your function on data set **swiss** available in the standard R repository:
    - Fertility should be Y
    - All other variables should be X
    - Nfolds should be 5

Report the resulting plot and interpret it. Report the optimal subset of features and comment whether it is reasonable that these specific features have largest impact on the target.

732A95 Introduction to Machine Learning /TDDE01 Machine Learning
Division of Statistics and Machine Learning
Department of Computer and Information Science

Assignment 2. Linear regression and regularization

The Excel file **tecator.xlsx** contains the results of study aimed to investigate whether a near infrared absorbance spectrum can be used to predict the fat content of samples of meat. For each meat sample the data consists of a 100 channel spectrum of absorbance records and the levels of moisture (water), fat and protein. The absorbance is -log10 of the transmittance measured by the spectrometer. The moisture, fat and protein are determined by analytic chemistry.

1. Import data to R and create a plot of Moisture versus Protein. Do you think that these data are described well by a linear model?
2. Consider model M_i in which Moisture is normally distributed, and the expected Moisture is a polynomial function of Protein including the polynomial terms up to power i (i.e M_1 is a linear model, M_2 is a quadratic model and so on). Report a probabilistic model that describes M_i. Why is it appropriate to use MSE criterion when fitting this model to a training data?
3. Divide the data into training and validation sets($50\%/50\%$) and fit models $M_i, i = 1 \ldots 6$. For each model, record the training and the validation MSE and present a plot showing how training and validation MSE depend on i (write some R code to make this plot). Which model is best according to the plot? How do the MSE values change and why? Interpret this picture in terms of bias-variance tradeoff.

   Use the entire data set in the following computations:

4. Perform variable selection of a linear model in which *Fat* is response and *Channel1-Channel100* are predictors by using stepAIC. Comment on how many variables were selected.
5. Fit a Ridge regression model with the same predictor and response variables. Present a plot showing how model coefficients depend on the log of the penalty factor λ and report how the coefficients change with λ.
6. Repeat step 6 but fit LASSO instead of the Ridge regression and compare the plots from steps 6 and 7. Conclusions?
7. Use cross-validation to find the optimal LASSO model, report the optimal λ and how many variables were chosen by the model and make conclusions. Present also a plot showing the dependence of the CV score and comment how the CV score changes with lambda.
8. Compare the results from steps 4 and 7.