Computer lab 4

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.**
- 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. Uncertainty estimation

The data file **State.csv** contains per capita state and local public expenditures and associated state demographic and economic characteristics, 1960, and there are variables

- MET: Percentage of population living in standard metropolitan areas
- EX: Per capita state and local public expenditures ($)

1. Reorder your data with respect to the increase of MET and plot EX versus MET. Discuss what kind of model can be appropriate here. Use the reordered data in steps 2-5.
2. Use package **tree** and fit a regression tree model with target EX and feature MET in which the number of the leaves is selected by cross-validation, use the entire data set and set minimum number of observations in a leaf equal to 8 (setting *minsize* in *tree.control*). Report the selected tree. Plot the original and the fitted data and histogram of residuals. Comment on the distribution of the residuals and the quality of the fit.
3. Compute and plot the 95% confidence bands for the regression tree model from step 2 (fit a regression tree with the same settings and the same number of leaves as in step 2 to the resampled data) by using a non-parametric bootstrap. Comment whether the band is smooth or bumpy and try to explain why. Consider the width of the confidence band and comment whether results of the regression model in step 2 seem to be reliable.
4. Compute and plot the 95% confidence and prediction bands the regression tree model from step 2 (fit a regression tree with the same settings and the same number of leaves as in step 2 to the resampled data) by using a parametric bootstrap, assume $Y \sim N(\mu_i, \sigma^2)$ where μ_i are labels in the tree leaves and σ^2 is the residual variance. Consider the width of the confidence band and comment

whether results of the regression model in step 2 seem to be reliable. Does it look like only 5% of data are outside the prediction band? Should it be?

5. Consider the histogram of residuals from step 2 and suggest what kind of bootstrap is actually more appropriate here.

Assignment 2. Principal components

The data file **NIRspectra.xls** contains near-infrared spectra and viscosity levels for a collection of diesel fuels. Your task is to investigate how the measured spectra can be used to predict the viscosity.

1. Conduct a standard PCA by using the feature space and provide a plot explaining how much variation is explained by each feature. Does the plot show how many PC should be extracted? Select the minimal number of components explaining at least 99% of the total variance. Provide also a plot of the scores in the coordinates (PC1, PC2). Are there unusual diesel fuels according to this plot?
2. Make trace plots of the loadings of the components selected in step 1. Is there any principle component that is explained by mainly a few original features?
3. Perform Independent Component Analysis with the number of components selected in step 1 (set seed 12345). Check the documentation for the fastICA method in R and do the following:
    a. Compute $W' = K \cdot W$ and present the columns of W' in form of the trace plots. Compare with the trace plots in step 2 and make conclusions. What kind of measure is represented by the matrix W'?
    b. Make a plot of the scores of the first two latent features and compare it with the score plot from step 1.
4. Fit a PCR model in which number of components is selected by cross validation to the data, use seed 12345. Provide a plot showing the dependence of the mean-square predicted error on the number of the components in the model and comment how many components it is reasonable to select.

Submission procedure

Assume that X is the current lab number, Y is your group number.

If you are neither speaker nor opponent for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members submits the group report using *Lab X group report* in the *Submissions* folder before the deadline

If you are a speaker for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members does the following before the deadline:
    - submits the group report using *Lab X group report* in the *Submissions* folder before the deadline
    - Goes to Study room *Group Y→Documents* and opens file *Password X.txt*. Then the student should put your group report into ZIP file *Lab X_Group Y.zip* and protect it with a password you found in Password X.txt
    - Uploads the file to *Collaborative workspace →Lab X* folder

If you are opponent for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members submits the group report using *Lab X group report* in the *Submissions* folder before the deadline
- After the deadline for the lab has passed, go to Collaborative workspace→*Lab X* folder and download the appropriate ZIP file. Open the PDF in this ZIP file by using the password available in *Course Documents→Password X.txt*, read it carefully and prepare (in cooperation with other group members) **at least three questions/comments/improvement suggestions per lab assignment** in order to put them at the seminar.