

Computer lab 2 block 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.
- Some exercises that require implementations of methods are provided with a skeleton of the code (see LISAM), you are free to use this code or implement your own code from scratch.
- **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. Using GAM and GLM to examine the mortality rates

The Excel document **influenza.xlsx** contains weekly data on the mortality and the number of laboratory-confirmed cases of influenza in Sweden. In addition, there is information about population-weighted temperature anomalies (temperature deficits).

1. Use time series plots to visually inspect how the mortality and influenza number vary with time (use Time as X axis). By using this plot, comment how the amounts of influenza cases are related to mortality rates.
2. Use `gam()` function from `mgcv` package to fit a GAM model in which Mortality is normally distributed and modelled as a linear function of Year and spline function of Week, and make sure that the model parameters are selected by the generalized cross-validation. Report the underlying probabilistic model.
3. Plot predicted and observed mortality against time for the fitted model and comment on the quality of the fit. Investigate the output of the GAM model and report which terms appear to be significant in the model. Is there a trend in mortality change from one year to another? Plot the spline component and interpret the plot.
4. Examine how the penalty factor of the spline function in the GAM model from step 2 influences the estimated deviance of the model. Make plots of the predicted and observed mortality against time for cases of very high and very

- low penalty factors. What is the relation of the penalty factor to the degrees of freedom? Do your results confirm this relationship?
5. Use the model obtained in step 2 and plot the residuals and the influenza values against time (in one plot). Is the temporal pattern in the residuals correlated to the outbreaks of influenza?
 6. Fit a GAM model in R in which mortality is be modelled as an additive function of the spline functions of year, week, and the number of confirmed cases of influenza. Use the output of this GAM function to conclude whether or not the mortality is influenced by the outbreaks of influenza. Provide the plot of the original and fitted Mortality against Time and comment whether the model seems to be better than the previous GAM models.

Assignment 2. High-dimensional methods

The data file **data.csv** contains information about 64 e-mails which were manually collected from DBWorld mailing list. They were classified as: 'announces of conferences' (1) and 'everything else' (0) (variable Conference)

1. Divide data into training and test sets (70/30) without scaling. Perform nearest shrunken centroid classification of training data in which the threshold is chosen by cross-validation. Provide a centroid plot and interpret it. How many features were selected by the method? List the names of the 10 most contributing features and comment whether it is reasonable that they have strong effect on the discrimination between the conference mails and other mails? Report the test error.
2. Compute the test error and the number of the contributing features for the following methods fitted to the training data:
 - a. Elastic net with the binomial response and $\alpha = 0.5$ in which penalty is selected by the cross-validation
 - b. Support vector machine with “vanilladot” kernel.Compare the results of these models with the results of the nearest shrunken centroids (make a comparative table). Which model would you prefer and why?
3. Implement Benjamini-Hochberg method for the original data, and use `t.test()` for computing p-values. Which features correspond to the rejected hypotheses? Interpret the result.

Submission procedure

First read ‘Course Information.PDF’ or ‘Course Information- PhD.PDF’ at LISAM, folder ‘Course documents’

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.