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

Computer lab 1, 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.
- **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 – Implementation of basis expansions

The csv-file **cube.csv** contains points in a two-dimensional plane.

1. Implement a function ***myspline*** that depends on vectors X, Y and $knots$ and that fits a piecewise linear model to the data with target Y and feature X and knot positions given by $knots$. The original data and the fitted data should be then visualized in one graph. In your implementation, use the following steps:
    a. Compute the values of the piecewise-linear basis functions (which you implement yourself) for each observation of X; by doing this you will obtain a new set of features $h_1(X), \dots h_m(X)$, where m is the amount of basis functions.
    b. Apply a linear regression of Y against the new features to estimate the optimal coefficients of the basis function expansion, and compute the predicted values of the target for the current observations.
2. Use the provided data set and knots vector $knots = \begin{Vmatrix} 2 \\ 4 \end{Vmatrix}$ to test your function from step 1. Comment on the quality of fit. Does the fitted model seem to be a continuous function? Should it be?
3. Use smooth.spline() to fit the same data, provide a plot showing the predicted and the original data, and compare this result with the plot from step 2.

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

Assignment 2. 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.