31st October, 2019

STK-IN4300 / STK-IN9300
Statistical learning methods in Data Science

Mandatory assignment 2 of 2

Submission deadline

Tuesday 19th November 2019, 14:30 at Canvas.

Instructions

You can choose between scanning handwritten notes or typing the solution directly on a computer (for instance with LaTeX). The assignment must be submitted as a single PDF file. Scanned pages must be clearly legible. The submission must contain your name, course and assignment number.

It is expected that you give a clear presentation with all necessary explanations. Remember to include all relevant plots and figures. Students who fail the assignment, but have made a genuine effort at solving the exercises, are given a second attempt at revising their answers. All aids, including collaboration, are allowed, but the submission must be written by you and reflect your understanding of the subject. If we doubt that you have understood the content you have handed in, we may request that you give an oral account.

In exercises where you are asked to write a computer program, you need to hand in the code along with the rest of the assignment. It is important that the submitted program contains a trial run, so that it is easy to see the result of the code.

Application for postponed delivery

If you need to apply for a postponement of the submission deadline due to illness or other reasons, you have to contact the Student Administration at the Department of Mathematics (e-mail: studieinfo@math.uio.no) well before the deadline.

All mandatory assignments in this course must be approved in the same semester, before you are allowed to take the final examination.

Complete guidelines about delivery of mandatory assignments:

uio.no/english/studies/admin/compulsory-activities/mn-math-mandatory.html

GOOD LUCK!

Problem 1. In a study by Ihorst et al. (2004), the long- and medium-term effects of ozone on the forced vital capacity and on the forced expiratory volume of 2153 school children are investigated. Consider the subset of the data used in Buchholz et al. (2008) available at `https://www-adm.uio.no/studier/emner/matnat/math/STK-IN4300/h19/ozone_496obs_25vars.txt`. It contains information on 496 children and 24 variables potentially affecting the (continuous) outcome "forced vital capacity" `FFVC`. The independent variables consist in 9 variables containing general information:

- `ALTER`: age;
- `ADHEU`: allergic coryza diagnosed by a physician;
- `SEX`: gender (0-male, 1-female);
- `HOCHOZON`: patient lives in a villages with high ozone values;
- `AMATOP`: maternal atopy;
- `AVATOP`: paternal atopy;
- `ADEKZ`: neurodermatitis diagnosed by a physician;
- `ARAUCH`: smoking (yes/no);
- `AGEBGEW`: weight at birth;

and 15 predictors for time point F (leading "F"):

- `FSNIGHT`: cough at night or in the morning;
- `FLGROSS`: height (cm) at lufu;
- `FMILB`: assured sensitivity against house dust;
- `FNOH24`: maximal NO_2 value of last 24h before lufu (micg/m3);
- `FTIER`: assured sensitivity against animal hair;
- `FPOLL`: assured sensitivity against pollen;
- `FLTOTMED`: total number of medis/lufu;
- `FO3H24`: maximal O_3 value of last 24h before lufu (micg/m3);
- `FSPT`: assured response to an allergen;

- `FTEH24`: max. temperature of last 24h before lufu (°C);
- `FSATEM`: shortness of breath, laboured breathing;
- `FSAUGE`: itchy eyes;
- `FLGEW`: weight (kg)/lufu;
- `FSPFEI`: whistling or wheezy breath;
- `FSHLAUF`: cough at effort.

Tasks:

1. Divide the dataset into a training and a test set, and standardize the data. For the former task, allocate 1/2 of the observations to the training set, 1/2 to the test set, paying some attention to the categorical variables (i.e., avoid splits in which there is a categorical variable with only one modality in the training set). For the latter task, explicitly report the R code you use and comment in details all the steps you do, justifying, in particular, how you decide to proceed with the categorical variables.

2. Only using the training set, estimate a linear Gaussian regression model to relate the forced vital capacity to the independent variables. Which covariate has the strongest association with the forced vital capacity? Report the coefficient estimates, their standard error and the associated p-value and comment on them.

3. Perform backward elimination and forward selection to find a reduced model, and contrast the models obtained in the two cases. For both, repeat the analysis using a different stopping criterion and add the two additional models on the comparison above. Are the selected models the same? Comment on that. For all 4 models, report the coefficient estimates, their standard error and the associated p-value and comment on the results. Do you expect these models to predict better or worse than the full model estimated at the previous point in terms of mean square prediction error, i.e., the mean square error computed on the test set? Why?

4. Use both a bootstrap procedure and a cross-validation procedure (choose the number of folds you prefer) to find the best (in term of deviance minimization) complexity parameter of a lasso regression among a grid of your choice. Provide a plot in which the results of the two procedures are contrasted and comment on them.

5. Fit a Generalized Additive Model in which the possible non-linear effects of specific variables are modelled by splines. Can the possible non-linearity(-ies) be captured by adding polynomial terms to the linear model? Fit such a model and comment on the two solutions.

6. Fit a component-wise boosting model, using as base learners for the continuous explanatory variables: (i) linear models; (ii) splines; (iii) trees. Report the variables selection frequencies in all three cases and the regression coefficients for the first model.

7. For each approach (full model of point 1, 4 reduced models of point 2, lasso of point 4, GAM with spline and linear model with polynomial terms of point 5, 3 boosting models of point 6) report the training and the test error. Comment on them.

Problem 2. The Pima Dataset is a publicly available dataset analysed many times in the literature (Royston & Sauerbrei, 2008). It contains information about 768 women of a population (Pima) particularly susceptible to diabetes. The response `diabetes` identifies which of the persons involved in the study (268 women, `diabetes = 'pos'`) developed the disease. Eight continuous independent variables contain information on:

- `pregnant`: number of pregnancies;
- `glucose`: plasma glucose concentration at 2 h in an oral glucose tolerance test;
- `pressure`: diastolic blod pressure (mm Hg);
- `triceps`: triceps skin fold thickness (mm);
- `insulin`: 2-h serum insulin (μU/mL);
- `mass`: body mass index (kg/m^2);
- `pedigree`: diabetes pedigree function;
- `age`: age (years);

Import the data from the R package `mlbench`, using the command `data(PimaIndiansDiabetes)`, and divide the dataset in a training (approximately 2/3 of the sample size) and a test set, keeping the proportion of women with diabetes and those without similar in the two sets.

1. Classify the patients using k-NN, selecting the best number of neighbours both via a 5-fold and a loo cross-validation procedure. Plot the two estimated error for each possible value of k. Add to the plot the corresponding test errors (i.e., the test error you would have obtained fitting k-NN with the same k) and comment on the results.

2. Fit a generalized additive model with splines and use a variable selection (subset selection) to find the best model.

3. Use a classification tree, bagging (both with "probability" and "consensus" votes), random forest, neural network and AdaBoost, to classify the persons between positive and negative to diabetes.

4. Which method would you choose if someone asks you to analyse these data? Why?

5. Looking more closely at the data, it has been noted that several values are implausible (e.g., a body mass index equal to 0). This means that some observations are actually not zeros, but missing values. Use the correct data (by using `data(PimaIndiansDiabetes2)`) and compare all the methods implemented in the previous points (consider only one k for the first point) by removing the observations with missing values. Do you obtain similar results as before? Comment the results.

Bibliography

Buchholz, A., Holländer, N. & Sauerbrei, W. (2008). On properties of predictors derived with a two-step bootstrap model averaging approach: a simulation study in the linear regression model. *Computational Statistics & Data Analysis* **52**, 2778–2793.

Ihorst, G., Frischer, T., Horak, F., Schumacher, M., Kopp, M., Forster, J., Mattes, J. & Kuehr, J. (2004). Long-and medium-term ozone effects on lung growth including a broad spectrum of exposure. *European Respiratory Journal* **23**, 292–299.

Royston, P. & Sauerbrei, W. (2008). *Multivariable Model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables.* Wiley, Chichester.