13th October, 2022

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

Mandatory assignment 2 of 2

Submission deadline

Thursday, 3rd of November 2022, 14:30 at Canvas.

Instructions

The assignment must be submitted as a single PDF file. 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. Note that you have one attempt to pass the assignment. This means that there are no second attempts. It is important that you (try to) answer every part of every problem. If the assignment is on the edge of pass or failure, missing exercises will have a negative effect on the evaluation. 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.

You need to hand in the code along with the rest of the assignment .This can be done by including the code in the text of the assignment, by adding the code in an appendix, or by adding a link to an on-line repository. 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. Regression

Consider the data from the study of Cassotti et al. (2014) on aquatic toxicity. The goal is to predict its effect towards Daphnia Magna, a small planktonic crustacean that lives in fresh water. Aquatic toxicity is measured through the variable `LC50`, namely the concentration that causes death in 50% of crustacean over a test duration of 48 hours.
For the prediction, 8 variables have been considered:

- `TPSA`: the topological polar surface area calculated by means of a contribution method that takes into account nitrogen, oxygen, potassium and sulphur;
- `SAacc`: the Van der Waals surface area (VSA) of atoms that are acceptors of hydrogen bonds;
- `H050`: the number of hydrogen atoms bonded to heteroatoms;
- `MLOGP`: expresses the lipophilicity of a molecule, this being the driving force of narcosis;
- `RDCHI`: a topological index that encodes information about molecular size and branching;
- `GATS1p`: information on molecular polarisability;
- `nN`: the number of nitrogen atoms present in the molecule;
- `C040`: the number of carbon atoms of a certain type, including esters, carboxylic acids, thioesters, carbamic acids, nitriles, etc.;

The data can be found at: `https://archive.ics.uci.edu/ml/machine-learning-databases/00505/qsar_aquatic_toxicity.csv`: the first 8 columns contain the independent variables (in the order they are described here), the last one the response.

1. Split the data into a training and a test set, with approximately 2/3 and 1/3 of the observations, respectively. Look at the count variables, would you model them with a linear effect or dichotomize them in 0 (absence of specific atoms) and 1 (presence of the specific atoms)? Fit two models with the count variables codified in these ways (linear effects and dichotomous variables), computing training and test errors. Comment on the results, both in terms of significance of the regression coefficients and test error.

2. Repeat the procedure of point 1 for 200 times, each time changing the split in training and test data and compute the average test errors. Do you obtain the same result? Try to explain why one often obtains, like in this case, a worse result by dichotomizing the variables.

3. Using the first training/test split, compare the results of different variable selection procedures (at least backward elimination and forward selection) with different stopping criteria (at least AIC and BIC). Do you obtain the same model? Why does this happen here?

4. Implement ridge regression (hereafter, use the first training/test set split you computed at point 1). In particular, 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 in a grid of your choice. Provide a plot in which the results of the two procedures are contrasted and comment on them.

5. Consider non linear effects for the variables. In particular, try to fit a generalised additive model in which the effects of the variables are fitted with smoothing splines. Try smoothing splines with (at least two) different levels of complexity, report the results and comment on them.

6. Fit a model using a regression tree. Draw the tree and report how the cost-complexity pruning led to the selected tree size.

7. Compare all the model implemented in the previous points both in term of training and test error and comment on the results.

Problem 2. Classification

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)`. If you are using a different programming language, see on the course web page the instructions to get a .cvs file. 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 errors 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. Report the model found and comment on it, describing the effects of the variables on the response.

3. Use a classification tree, bagging (both with "probability" and "consensus" votes), random forest and AdaBoost, to classify the persons between positive and negative to diabetes. Report the training and the test error for each model.

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 contrast the reuslts of all methods implemented in the previous points (consider only one k for the first point) to those obtained after having removed the observations with missing values. Do you obtain similar results as before? Comment the results.

Bibliography

Cassotti, M., Ballabio, D., Consonni, V., Mauri, A., Tetko, I. V. & Todeschini, R. (2014). Prediction of acute aquatic toxicity toward daphnia magna by using the ga-k nn method. *Alternatives to Laboratory Animals* **42**, 31–41.

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