26th October, 2026

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

Mandatory assignment 2 of 2

Submission deadline

Tuesday 10th November 2026, 14:30 in Canvas.

Instructions

This assignment follows the policy "submit to pass", so it will be approved even if you only submit a blank page. You will receive feedback only if you try to solve the exercises yourself. Assignments only generated by Large Language Models will not be evaluated (it means that you can still use them, but you need to re-elaborate the answers yourself if you want any feedback).

You can choose between scanning handwritten notes or using a typesetting software for mathematics (e.g. LaTeX). Scanned pages must be clearly legible. The assignment must be submitted as a single PDF file. Remember to include any relevant programming code and resulting plots and figures, in the PDF-file.

All aids, including collaboration, are allowed, but the submission must be written by you and reflect your understanding of the subject.

Application for postponed delivery

If you need to apply for a postponement of the submission deadline due to illness or other reasons, see the part about extensions in the following page:

`https://www.uio.no/english/studies/examinations/compulsory-activities/mn-math-mandatory.html`

which also contains general guidelines for mandatory assignments at the Department of Mathematics.

Complete guidelines about delivery of mandatory assignments:

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

GOOD LUCK!

Problem 1. Comparison of several approaches

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 dataset contains 546 observations and it is available 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.

(a) Split the data into a training and a test set, with approximately 2/3 and 1/3 of the observations, respectively. In terms of the count variables, consider the two following modelling options:

  (i) model each of them directly as a linear effect;

  (ii) transform each of them using a 0/1 dummy encoding where 0 represents absence and 1 the presence of the atom.

Fit two separate linear regression models using all the available predictors as input and two options above for modelling the count variables (i.e., one model with option (i), one model with option (ii)). Compute training and test errors and comment on the results, both in terms of significance of the regression coefficients and prediction error.

(b) Repeat the procedure described in (a) 200 times, such that each time: you do a new training/test split (with same proportions as in (a)), fit the models with option (i) and (ii), and record the test errors. Make a plot that illustrates the empirical distributions of the test error for each modelling option and compare the average test error. What is the point of repeating the experiment in this way before drawing any conclusions? Try to explain why one often obtains, like in this case, a worse result by using option (ii).

For the remaining sub-problems, use the training/test split from (a) and use the original encoding of the count variables, that is, do not transform them into binary presence-absence variables

(c) Apply different variable selection procedures (at least backward elimination and forward selection) with different stopping criteria (at least AIC and BIC) and compare the results. Do you obtain the same model? Explain the reason.

(d) Apply ridge regression and use both a bootstrap procedure and cross-validation (choose the number of folds you prefer) to find the optimal complexity parameter in a grid of candidate parameter values of your own choice. Provide a plot in which the results of the two procedures are contrasted and comment on them.

(e) To also allow for nonlinear effects, fit a generalised additive model (GAM) in which the effects of the variables are fitted using smoothing splines. Try at least two different levels of complexity for the smoothing splines. Report the results and comment on them.

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

(g) Fit a random forest and a gradient boosting algorithm, commenting on the choice of the tuning parameters and, for boosting, on the choice of the base learner.

(i) Compare all the model implemented in the previous points both in term of training and test error and comment on 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.