15th October, 2024

STK-IN4300/STK-IN9300: Statistical Learning Methods in Data Science

Mandatory assignment 2 of 2

Submission deadline

Thursday 31th OCTOBER 2024, 14:30 in Canvas (canvas.uio.no).

Instructions

Note that you have **one attempt** to pass the assignment. This means that there are no second attempts.

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. 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.

When producing results from some analysis, **include a discussion about this**. We will be a much stricter on this compared to the first assigment!

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) no later than the same day as 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

Regression and Classification

The assignment consists of two problems (Problems 1-2), which are the same for both STK-IN4300 and STK-IN9300. However, for PhD candidates in STK-IN9300, there is one additional assignment given in Problem 3, which students in STK-IN4300 can disregard. To pass the assignment, the presented solutions do not have to be entirely correct, but you are expected to make a genuine effort at solving the problems, and this should also be reflected by the submitted report.

For writing the report, it is highly recommended to use R Markdown or some similar file format. For those that want to use R Markdown but prefer Python over R, the reticulate package can be a useful. However, most importantly, when writing the report, make sure to include the code you used to produce the results along with the resulting tables and figures, and of course any relevant comments and discussion. As an alternative to adding the code directly to the main report, you may also share it through an online repository, to which you add a link in the report.

Problem 1. Regression

For this problem, we are going to consider data from a study where the goal was to predict acute aquatic toxicity for different organic molecules based on various descriptors of the molecule. Aquatic toxicity is measured through variable `LC50`, as the concentration that causes death in 50% of the planktonic crustacean *Daphnia magna* over a test duration of 48 hours. The study identified the following 8 molecular descriptors as being most informative in terms of predicting `LC50`:

- `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 contains 546 observations and it is available at the UCI machine learning repository and can be directly downloaded from this page. Note that the first 8 columns contain data of the input variables (following the above order) and the last column contains the response values.

(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 of the specific atom and 1 represents presence of the specific atoms.

Fit two separate linear regression models using option (i) and (ii), and compute the training and test errors. Comment on the results, both in terms of significance of the regression coefficients and test 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)**.

(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?

(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). Compare all the models implemented in the previous points in terms of both training and test error and comment on the results.

Problem 2. Classification

The Pima Indians Diabetes Database is a publicly available dataset that has been used extensively to test and assess statistical learning methods. It contains information about 768 women from a population (Pima indians) that is particularly susceptible to diabetes. The two-class response variable, `diabetes`, identifies whether a person involved in the study has developed the disease (`diabetes = 'pos'`) or not (`diabetes = 'neg'`). In addition, there are 8 numeric input variables:

- `pregnant`: number of pregnancies;
- `glucose`: plasma glucose concentration at 2 h in an oral glucose tolerance test;
- `pressure`: diastolic blood 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(PimaIndiansDiabetes2)`. Alternatively, you can download a csv-file of the data through the course semester page. Randomly split the dataset into a training set (approximately 2/3 of the sample size) and a test set, such that the class distributions (i.e. the empirical distribution of `diabetes`) is similar in the two sets.

(a). Fit a k-NN classifier on the data and select the optimal number of neighbours, k, using both 5-fold and leave-one-out cross-validation. Plot the estimated errors of each cross-validation procedure for the considered range of candidate values of k in a single figure. Add also to the figure the corresponding test errors, that is, the test error you would have obtained when fitting k-NN for the different values of k. Comment on the results.

(b). Fit a generalized additive model (GAM) with splines and use a variable selection method to find the best model. Report the selected model and comment on the results, in particular, describe briefly the effects that the variables have on the response.

(c). Fit (i) a classification tree, (ii) an ensemble of bagged trees and (iii) a random forest to the data. Report the training and the test error for each method and comment on results.

(d). Fit a neural network to the data. Specify what procedure you have used in order to choose the size of the network.

(e). Which method would you choose if someone asks you to analyse the data? Motivate why you would choose that particular method.

Problem 3. Oral Presentation (Only for STK-IN9300)

As part of the mandatory assignment, each PhD candidate is expected to give a 15-minute oral presentation on a topic related to statistical learning methods. The presentation will be given sometime during the two weeks after the submission deadline. As part of this report, write down a title for the presentation along with a brief abstract (a few sentences) describing the planned presentation. Feel free to confer with the lecturer if you are uncertain about the topic.

GOOD LUCK!