6th February, 2020

STK-2100

Mandatory assignment 2 of 2

Submission deadline

Thursday 16th April 2020, 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. Consider the dataset from Golub et al. (1999). It contains 7128 gene expressions of 72 patients with leukaemia and can be found at `http://web.stanford.edu/~hastie/CASI_files/DATA/leukemia_big.csv` (pay attention to the number of rows and columns).

(a) The dataset actually contains data of patients with two types of the disease, acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML). As a graphical investigation, we are interested in looking at the two first principal components, to see if there is the chance to separate the two types of patients. Derive the first two principal components and produce a scatter plot in which the observations are coloured differently based on the type of leukaemia (look at the column labels in the original dataset).

(b) In addition, a continuous response has been generated (without any meaning, just for this exercise), that can be found at the following link: `https://www.uio.no/studier/emner/matnat/math/STK2100/v20/eksamen/response_train.csv` (note that the first column refers to the observation, the second is the actual response).

Since the number of covariates (gene expressions) exceeds the number of observations, one cannot use ordinary linear regression, and should rely on a regularized approach like lasso. Explain the role of the penalty parameter in a lasso model and estimate it with K-fold cross-validation. In particular, try different version of cross-validation, namely 3-fold, 10-fold and leave-one-out. Report the optimal λ in term of minimization of the cross-validation error for all three cases, and comment on the results: did you expect to find different values?

(c) Rename the covariates `gene1`, ..., `gene7128` and fit a lasso model for each of the penalties found at point (b). Report which of the regression coefficient estimates are different from 0 and their estimated values.

(d) We have seen in class that another regularized regression method is ridge regression. Fit a ridge regression model and report the estimates for the first eleven regression coefficients (you can choose how to select the best penalty parameter).

(e) Alternatively, one can fit a prediction model by using principal components, i.e., by using the principal component regression method. Fit a model using this approach, justifying your choice for the number of principal components used.

(f) The file available at https://www.uio.no/studier/emner/matnat/math/STK2100/v20/eksamen/test_set.csv contains a test set for this problem. Use it to compute the mean square prediction error for the three lasso models, the ridge regression model and the principal component regression model. Report the results.

(g) Some authors (see, e.g., Hastie et al., 2001; Krstajic et al., 2014) argue that, in a regularized regression method, "the largest value of λ such that error is within 1 standard error of the minimum" is a better estimate for the penalty parameter (in `R`, using the function `cv.glmnet` of the package `glmnet` you can find this value by selecting `$lambda.1se` instead of `$lambda.min`).

Repeat the analyses performed at points (c) and (d) using this strategy instead of the classical approach of selecting the value of λ that minimizes the cross-validation error. Compare the prediction error with those obtained in point (f) and comment on the results: did you obtain better (in term of prediction error) models?

Problem 2. Consider again the dataset from Luke et al. (1997) used in the exercise 1 of the first mandatory assignment (and available at https://www.imbi.uni-freiburg.de/imbi/Royston-Sauerbrei-book/Multivariable_Model-building/downloads/datasets/res_bodyfat.zip).
The aim is still to find how well `bmi` can be used to predict `pbfm`.

(a) Try a local regression method to model the relationship. Explain in details how you have chosen the tuning parameter and provide three plots to highlight the sensitivity of the method to this choice (so plot one case in which the estimated function is too smooth, one in which is fine and one in which is too "wiggly").

(b) Repeat the procedure above with the optimal value of the tuning parameter and show that, in contrast, the choice of the kernel is not very influential (try three different kernels).

(c) Model the relationship between `bmi` and `pbfm` by using regression splines: place 4 knots (justify your choice for their location) and fit splines with polynomials of order 1, 2 and 3. Report the result for all three cases in a plot.

(d) Repeat the analyses of point (a) with smoothing splines: describe how you have chosen the tuning parameter and show graphically its importance.

(e) Now that you have tried all the methods, suppose that your goal was to compare them using this dataset. Ignoring all the previous results, conduct a study in which you compare "simple linear regression", "polynomial regression", "local regression" and "splines" for the goal of predicting `pbfm` with `bmi`. Explain how you conducted the study justifying your choices. Finally, choose one method and explain why in your opinion is the best in this case.

Bibliography

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. & Lander, E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer, New York.

Krstajic, D., Buturovic, L. J., Leahy, D. E. & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics* **6**, 1–15.

Luke, A., Durazo-Arvizu, R., Rotimi, C., Prewitt, T. E., Forrester, T., Wilks, R., Ogunbiyi, O. J., Schoeller, D. A., McGee, D. & Cooper, R. S. (1997). Relation between body mass index and body fat in black population samples from nigeria, jamaica, and the united states. *American Journal of Epidemiology* **145**, 620–628.