Final project STK4030/9030 - Statistical learning: Advanced regression and classification

Available: Wednesday, November 8th 2017.
Deadline: Monday, November 20th 2017 at 9.00.

November 8, 2017

This is the problem set for the project part of the final exam of STK4030/ STK9030 (fall 2017). The report shall be individually written and include the code used in an appendix. The report should be handed in through Devilry (`https://devilry.ifi.uio.no/`). Remember that the project must not be mark by name, but by the candidate number.

All data sets are available on the course webpage.

Riccardo De Bin

Problem 1: Boston housing

Harrison and Rubinfeld (1978) reported the results of a study on the house prices in Boston. In particular, the logarithm of the median value of owner-occupied houses in thousand of dollars (`y`) was recorded for 506 tracts of the Boston Standard Metropolitan Area (from 1970), together with 14 variables that were considered potentially correlated to the house prices:

- `crim`: crime rate by FBI 1970 report;
- `zn`: proportion of residential land zoned for lots greater than 25,000 square feet;
- `indus`: proportion nonretail business acres
- `chas`: close/not close to the Charles river;

- `nox`: nitrogen oxide concentrations in pphm (annual average concentration in parts per hundred million);
- `part`: particulate concentrations in mg/hcm^3 (annual average concentration in milligrams per hundred cubic meters);
- `rm`: average number of rooms in owner units;
- `age`: proportion of owner units built prior to 1940;
- `dis`: weighted distances to five employment centers in the Boston region;
- `rad`: index of accessibility to radial highways;
- `tax`: full value property tax rate ($/$1O,OOO);
- `ptratio`: pupil-teacher ratio of the school district;
- `bk`: proportion of black population;
- `lstat`: logarithm of the proportion of population that is lower status (based on education and kind of job).

The data has been arbitrarily divided in equally sized training (`train = TRUE`) and test (`train = FALSE`) sets.

1. Estimate a linear Gaussian regression model including all 14 independent variables by (ordinary) least squares (OLS) on the training set. Report the estimated coefficients. Which covariates have the strongest association with y? In particular, the study focused on the effect of air pollution, measured through the concentrations of nitrogen oxide pollutants (`nox`) and particulate (`part`). Do they have any effect on the house price? If yes, which kind of effect?

2. The model above can be also used to predict the price for the other tracts (test set). Compute the prediction error on the test data. Moreover, derive two reduced models by applying a backward elimination procedure with AIC and $\alpha = 0.05$ as stopping criteria, respectively. For both models, report the estimated coefficients and the prediction error estimated on the test data. Comment the results.

3. Estimate a principal component regression model, selecting the number of components by 10-fold cross-validation. How many components have been selected? What does it mean?

4. Repeat the procedure to choose the number of components by using the .632 bootstrap procedure. Does the number of selected components change? Report the estimate of the prediction error for each possible number of components.

5. Estimate the regression model by ridge regression, where the optimal tuning parameter λ is chosen by 10-fold cross-validation. Report the estimated coefficients, the obtained value of lambda and the prediction error computed on the test data.

6. Repeat the same procedure by using lasso and component-wise L_2Boost. Use 10-fold cross-validation to find the optimal value for λ (lasso) and m_{stop} (L_2Boost), while set the boosting step size ν equal to 0.1. Report, for both models, the estimated coefficients, the obtained value of lambda and the prediction error computed on the test data.

7. It has been argued that the predictors `rm` and `dis` do not have a linear effect on the outcome. Substitute the former with its cube and the latter with its inverse ($\texttt{dis}^{-1}$) in the first model (OLS) and refit the model. Compute the prediction error on the test set and compare the result with that obtained at point 1.

Problem 2: oral cancer

From a large population-based case–control study on Oral cancer conducted in the US (Day et al., 1993), the data related to the African American population (194 cases, here `ccstatus` = 1, and 203 controls, here `ccstatus` = 0) has been selected. The aim of the study is to evaluate the risk of Oral cancer based on the variables `drinks` (number of 1oz ethanol-equivalent drinks consumed per week), `sex`, `age` and `cigs` (number of cigarettes smoked per day).

The original dataset has been again arbitrarily split into a training (199 observations, `train = TRUE`) and a test (198, `train = FALSE`) set.

1. Use the k-nearest-neighbour algorithm to classify cases and controls. Draw in the same plot the training and test error and write your considerations.

2. Back to the k-nearest-neighbour, select the optimal value of k via 5- 10- and LOO- cross validation. Plot k versus the prediction error estimates. Do the three cross-validation procedure provide similar results? Repeat the selection by using a simple bootstrap procedure and the

.632 bootstrap procedure and plot the results as well. Are the results similar to the previous ones?

3. Use LDA and QDA to classify cases and controls. Which of the two algorithms should be preferred in this case?

4. Repeat the analyses by using logistic regression, penalized versions of it based on L_1 and L_2 penalties, and L_2Boost. Report the regression coefficients and compute training and test errors.

5. Implement the algorithm AdaBoost with a tree as a classifier and plot train and test errors as a function of the number of boosting steps (iterations). Would have AdaBoost profited by an early stop in this specific case?

Problem 3: forest fire

Cortez and Morais (2007) analysed the forest fire data from the Montesinho natural park, from the Trás-os-Montes northeast region of Portugal. They collected several quantities, including:

- `X`: x-axis spatial coordinate (from 1 to 9) within the Montesinho park map;
- `Y`: y-axis spatial coordinate (form 2 to 9) within the Montesinho park map;
- `FFMC`: fine fuel moisture code;
- `DMC`: duff moisture code;
- `DC`: drought code;
- `ISI`: initial spread index;
- `temp`: temperature in Celsius degrees;
- `RH`: relative humidity in %;
- `wind`: wind speed in km/h;
- `rain`: outside rain in mm/m^2;

The goal of this exercise is to create a model to predict the burned area of the forest, that Cortez and Morais (2007) suggested to logarithmically transform due to its skewness toward 0. The dataset already contains the transformed variable (`logArea`). Try some of the approaches seen in the course and provide a final model that can be used to predict the logarithm of the burned area. Describe your work and the reasons behind the choice of the final model.

References

- P. Cortez and A. Morais (2007). A data mining approach to predict forest fires using meteorological data. In *J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence*, Guimarães, Portugal, pp. 512-523.
- Day GL, Blot WJ, Austin DF, Bernstein L, Greenberg RS, Preston-Martin S, Schoenberg JB, Winn DM, McLaughlin JK and Fraumeni Jr JF (1993). Racial differences in risk of oral and pharyngeal cancer: alcohol, tobacco, and other determinants. *Journal of the National Cancer Institute*, 85:465 – 473.
- Harrison D and Rubinfels, DL (1978). Hedonic house prices and the demand of clear air. *Journal of Enviromental Economics and Management*, 5: 81 – 102.