Tilastollisen oppimisen jatkokurssi 2018–2019
Advanced statistical learning 2018–2019

Round 1, 26.11.–9.12.

Topics:

- Model selection: Subset selection, forward and backward selection
- Recall the use of information criteria (AIC, BIC), adjusted R^2 and the cross validation methods, together with the ideas of using training and test samples and corresponding training and test errors.
- Shrinkage methods and sparse modelling: Ridge and lasso regressions (including different stages of ridge and lasso modelling such as the determination of tuning parameters)

Material:

- ISLR chapters 6.1, 6.2 and partly 6.4 (including the corresponding R labs as throughout the course)
- ISLR videos 6.1–6.8
- FDP videos: 12 and 13
- ESLR chapters 3.3–3.4
- CASI chapters 7.3 and 16

Exercises:

1. ISLR 6.9, parts (a)–(d) and (g), excluding PCR and PLS methods in (e) and (f). Dataset College is included in ISLR package.[1]

2. ISLR 6.11. Exclude PCR method which will be considered later on. In other words, consider best subset selection, ridge and lasso as covered in this round. Dataset Boston is included in ISLR package.

3. Varian (2014, JEP): Replicate the Lasso analysis in connection to Table 4 in terms of finding the Lasso estimates. Dataset is `FLS-data.csv` and can be found (with R code) in the supplementary material of the article published in the JEP (see the link in Moodle).

Report and comment on the estimation results (and different stages of lasso modelling) more generally than just reporting the rank of the predictors as in Table 4 in Varian (2014). In other words, which variables lasso excludes from the model, how to specify the tuning parameter and present the estimation result of the final model. [2]

4. ISLR 6.5.

5. CASI, related to chapter 16 and SPAM example: Forward stepwise selection.

See the attached R file `Ex1.5_stepwise-search.r` and dataset `SPAM.csv` related to CASI chapter 16 (pages 298–323, especially Figure 16.2). The idea in this exercise is to replicate the

[1] Throughout the course, if working with a different program than R, load the dataset first in R environment and move it to another software.

[2] You can also find additional details in Sala-i-Martin (1997).

analysis and, in particular, explain by words and augment the attached R program to clarify how the analysis is executed. In particular, you should provide and explain statistical formulae behind different stages in your solution.

This exercise was used by Sangita Kulathinal and Kari Auranen at the University of Helsinki in the spring (2018) in the course of CASI book. Their instructions are below.

The data consist of 4601 email messages sent to “George” at HP-Labs. He labeled 1813 of these as spam, with the remainder being good email (ham). The data matrix has 59 columns: spam Logical variable, TRUE is spam, FALSE is ham (good email), testid Logical variable. An optional split into train (FALSE) and test (TRUE) data (as used in, for example, in ESLR). The remainder of the columns are features used to build a prediction model. They are relative frequencies of the chosen words in each email (standardised by the length of the email).

The goal is to predict whether the future emails are spam or ham using these keywords, to build a spam filter for George!

1. Read the spam data (see the CASI website)

2. Fit the logistic regression to all data using 57 covariates after standardising the covariates. (Standardise all covariates so that each columns of have mean 0 and sums of squares 1).

3. Perform forward-stepwise logistic regression using the training SPAM data and use test data for computing misclassification error. You may try to program the Algorithm 16.2 or use the ready function (such as step) from R.

4. Try forward-stepwise linear regression using the same data and plot the two misclassification errors. (Figure 16.2).