Assignment 3 880260, 2019/20

**Due at 23:55 on Sunday, 08-12-2019**

Requirements: Assignments have to be handed in, one for each group of students. Please write all group member names with corresponding student (SNR) and administrative (ANR) numbers on the top of the first sheet that you hand in. Remember that assignments have to be typeset in English and submitted in Canvas as a single PDF file.

Always justify your answers! When asked to use R, always make sure you provide your code at the end of the assignment; the results you obtained should be reproducible.

**Exercise:** [15 points in total] Systems biology of the influenza vaccine

In this exercise, we consider data that result from a systems biology study on vaccination against influenza (Nakaya et al., 2011). One of the aims of this study was to predict vaccine efficacy with micro-array gene expression data obtained soon after vaccination. The authors made data for two seasons, 2007 and 2008, publicly available on the NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) with accession numbers GSE29614 and GSE29617. For both seasons, a micro-array analysis was performed on the genomewide expression in peripheral blood mononuclear cells collected just before and 3 days after vaccination for all participants (24 in 2008 and 9 in 2007). Two different array platforms for measuring gene expression were used but the first 54,675 of 54,715 probe sets of the 2008 season are shared with the 2007 season. Hence, the 2007 data can be used as an independent test sample.

The gene expression data for the 2008 season are stored in the file TIV2008affyPLM.txt and corresponding vaccine efficacy scores – obtained four weeks after vaccination - are stored in TITER2008\_centered.txt. The independent test set data are stored in TIV2007affyPLM.txt and TITER2007\_centered.txt.

Recommended reading material: Sections 6.2 (shrinkage methods) and 6.6 of the ‘Introduction to Statistical Learning’ (James et al., 2017).

Part I: [6 points in total] Shrinkage methods on the full set of 54,675 variables

1. [3 points] Determine a regression model to predict vaccine efficacy using the elastic net with ** fixed at 0.10. Build the regression model using the 2008 data. Use cross-validation to tune the lasso penalty parameter. Which type of cross-validation will you use? Give a brief motivation for your choice. Make a plot of the cross-validation errors against the different values of the lasso tuning parameter. Discuss how you chose the optimal value of the lasso tuning parameter and report this optimal value and the corresponding number of non-zero coefficients.
2. [2 points] What is the estimated prediction error derived from the 2008 data. Report the estimate and explain how you obtained it.
3. [1 point] Now use the 2007 data as an independent test set to obtain an estimate of the prediction error. Report the value you found as an answer to this question.

Part II: [9 points in total] Subset of 20 variables: Best subset selection

From the full set of variables, a subset of 20 variables have been chosen (corresponding to genes with the following official symbols: RFC2, HSPA6, PAX8, GUCA1A, THRA, PTPN21, CCL5, CYP2E1, EPHB3, ESRRA, IL17RA, SERPIND1, IFNGR2, MAPK1, IL25, CTSG, FOXP3, STAT3, CCL2, IRF8). The subset obtained from the 2008 data can be found in subset2008.txt while the subset for the 2007 data can be found in subset2007.txt. The aim of this exercise is to obtain the best possible multiple regression model containing *at most* five predictors.

1. [1 point] Among how many regression models will you have to select the best model? Show how you obtained this number.
2. [4 points] Implement a 5 fold cross-validation procedure to perform the selection of the best model. Make a histogram of the cross-validation errors obtained for each of the models and report the set of predictor variables that – together - have the lowest cross-validation error.
3. [1 point] Fit the selected model to the data. Report the regression coefficients.
4. [2 points] Use the selected model to predict the vaccine efficacy scores on the subset of the 2007 data. Report the predicted score for each case and the overall MSETest.
5. [1 point] Suppose you would use the best subset selection approach to the full set of variables (54 675 here) in order to find the best subset (in terms of cross-validation error) of *k* variables (with *k* any possible number from 1 to 54 675). Name two crucial drawbacks of this approach.

If you fail to implement the cross-validation procedure in step 2, you can use a model with some set of at most 5 predictors to answer questions 3 and 4. In this way, you can still obtain the 3 points related to these two questions.

References

Nakaya, H.I., Wrammert, J., Lee, E.K., Racioppi, L., Marie-Kunze, S., et al.: Systems biology of vaccination for seasonal influenza in humans. *Nature immunology 12(8),* 786–795 (2011). doi:10.1038/ni.2067