Project Task 2 - BAN404

Jonas Andersson

From March 20, 2019, 09.00 until March 27, 2019, 14.00

Introduction

The report should be no longer than totally 15 pages (including everything). In addition, you are required to submit your well documented R-code; the code should be possible to run for the examiners. To the extent that you are making your own choices in the project, you should choose methods from James et al. (2015).

The data might have been previously analysed by other authors; be careful to cite studies that you draw inspiration from.

Asking for help

You are required to solve the project task indepently. Therefore, note the following rules concerning this:

- There is no opportunity to ask questions about the choices you are supposed to make yourself during the project period.
- You can, however, to a reasonable extent, ask for help concerning technical issues, such as data handling or R-coding, that you have, unsuccessfully, tried to solve yourself first.
- All questions should be sent by email to the teaching assistants Benjamin Fram (benjamin.fram@nhh.no) or Rabia Masood (rabia.masood@nhh.no). If they need to, they will ask the lecturer for input.

The project

In this project you shall classify a household into one of two classes, income below $50,000 or $50,000 or above, based on 14 predictors. The dataset is called marketing and you find it in the R-package ElemStatLearn, Halvorsen (2019). Note that you might need to deal with missing observations, e.g. with the R-function na.omit and with transforming predictors to other data types, e.g. with the R-function as.factor.

a. Describe relevant features of the input and output variables with descriptive statistics. Create a variable high which is one for Income equal to $50,000 or more and zero otherwise.

b. Use logistic regression, linear discriminant analysis, and classification trees with pruning to predict high. Compare the predictions with cross-validation. Motivate your choice of cross-validation method and the choice of fitting function used to measure the quality of the predictions.

c. Choose two additional prediction methods from the chapters 7-10 in James et al. (2015) and compare their predictions with the other methods.

d. Elaborate on what your analysis has to say about the different predictors association with high? Does a high income household have any particular features?

e. Summarize your results in a conclusion.

References

Halvorsen, Kjetil B. 2019. *ElemStatLearn: Data Sets, Functions and Examples from the Book: "The Elements of Statistical Learning, Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani*

and Jerome Friedman. Material from the book's webpage and R port and packaging by the author, https://CRAN.R-project.org/package=ElemStatLearn.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2015. *An Introduction to Statistical Learning.* Springer.