Statistical Learning, Homework #1

Veronica Vinciotti, Marco Chierici

Released: 28/03/2022. Due: 11/04/2022

This homework deals with classification methods. You should submit an RMarkdown file and a pdf file of the report. The RMarkdown file should reproduce exactly the pdf file that you will submit. The pdf file should be rendered directly from the RMarkdown (using `output: pdf_document`) and not converted from any other output format.

Note that:

- your code should run without errors (except for minor adjustments such as file paths);
- in your report, you should: introduce the analysis, discuss/justify each choice that you make, provide comments on the results that you obtain and draw some conclusions.

Description of dataset

The data set for this homework is available at breastfeed.Rdata. The data come from a study conducted at a UK hospital, investigating the possible factors affecting the decision of pregnant women to breastfeed their babies, in order to target breastfeeding promotions towards women with a lower probability of choosing it.

For the study, 135 expectant mothers were asked what kind of feeding method they would use for their coming baby. The responses were classified into two categories (variable **breast** in the dataset): the first category (coded 1) includes the cases "breastfeeding", "try to breastfeed" and "mixed breast- and bottle-feeding", while the second category (coded 0) corresponds to "exclusive bottle-feeding". The possible factors, that are available in the data, are the advancement of the pregnancy (**pregnancy**), how the mothers were fed as babies (**howfed**), how the mother's friend fed their babies (**howfedfr**), if they have a partner (**partner**), their age (**age**), the age at which they left full-time education (**educat**), their ethnic group (**ethnic**) and if they have ever smoked (**smokebf**) or if they have stopped smoking (**smokenow**). All of the factors are two-level factors.

Homework tasks

In your report, you should:

1. Explore the data: what is the distribution of the response variable (`breast`)? Are there potential issues with any of the predictors? Do you need pre-processing or can you proceed with the data as is?
2. Split the data into (reproducible) training and test sets. Given the class imbalance, you could aim for sets that have the same imbalance with respect to the outcome variable. In order to do this, you could either perform the splitting manually on each class, or use dedicated functions (for example, `caret::createDataPartition(labels, p=train_size)`, with `train_size` a number between 0 and 1 representing the percentage of data you would like to use for training.
3. Fit the following GLM model:

$$\begin{aligned}\text{logit}(\text{E}(\text{breast})) = \beta_0 &+ \beta_1 \text{pregnancy} + \beta_2 \text{howfed} + \beta_3 \text{howfedfr} \\ &+ \beta_4 \text{partner} + \beta_5 \text{age} + \beta_6 \text{educat} + \beta_7 \text{ethnic} + \beta_8 \text{smokenow} + \beta_9 \text{smokebf}\end{aligned}$$

   Discuss the `summary` and the interpretation of the model in the context of the study.
4. Fit a k-nn classifier, by performing a careful selection of the tuning parameter k.
5. Fit a Naïve Bayes classifier.
6. Evaluate the performance of the methods and compare the results.

Hints

In R, the Naïve Bayes (NB) classifier is included in the package `e1071`. The syntax is `naiveBayes(formula, data)` for model fitting, and the usual `predict(fit, newdata)` for predicting on new data. A fitted `naiveBayes` object stores the conditional probabilities for each feature, together with the *a priori* probabilities. To compute the posterior probabilities, you call `predict()` with the argument `type="raw"`.