Exercise 5

Advanced Methods for Regression and Classification

December 1, 2022

1. Use the data from
   `http://archive.ics.uci.edu/ml/datasets/Bank+Marketing`,
   which are also available on our TUWEL course. Load the smaller data set using `d <- read.csv2("bank.csv")`. The data contain information about direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit or not. This information is contained in the binary variable y (last one).

   (a) Select randomly a training set with 3000 observations, and use logistic regression (function `glm()` with `family="binomial"`). Look at the inference table (with `summary()`) and interpret the outcome.

   (b) Use the model to predict the group label of the remaining test set observations (what does the function actually predict by default?). Compute the misclassification rate for every group separately.

   (c) Since the data set is heavily imbalanced, i.e. we have many more "no" clients, we might have a problem with high misclassifications for the "yes" clients, which are in fact the interesting ones, since the bank does not want to lose potential customers. A way to consider this problem is to assign a weight to every observation, by using the `weights` argument in the `glm()` function. How do you have to select the weights, and what are the resulting misclassification rates?

   (d) Based on the model from 1(c), use stepwise variable selection with the function `step()` to simplify the model. Does this also lead to an improvement of the misclassification rates?

2. Use the data set `data(Khan)` from the package `ISLR`. The data set consists of a number of tissue samples corresponding to four distinct types of small round blue cell tumors. For each tissue sample, 2308 gene expression measurements are available (see also help file). The task is to train a classifier based on the training data (`Khan$xtrain`, `Khan$ytrain`), to use the trained classifier for predicting the class of the test data (`Khan$xtest`), and to evaluate the predictions using the group information of the test data (`Khan$ytest`).

   (a) Why would LDA or QDA not work here? Would RDA work (you can either try it out, or simply argue)?

   (b) Use the function `cv.glmnet()` from the package `glmnet`, with the argument `family="multinomial"`, to build a model for the training set (the response might need to be converted to a factor). Plot the outcome object. What do you conclude? What could be the objective function to be minimized?

(c) Which variables contribute to the model? To see this, you can use `coef()` for the output object. You obtain an object with 4 (= number of groups) list elements, containing the estimated regression coefficients. Thus, this is different from our approach to logistic regression with K groups in the course notes, where you would only obtain $K - 1$ coefficient vectors.

(d) Select one of the variables from 2(c) which is relevant e.g. for the first group, and plot this variable against the resonse (using the training data). What you should see is that the values of the first group clearly differ from those of the other groups.

(e) Now use the trained model and predict the group membership of the test data. Be careful, `predict()` yields predictions for each observations to each class, and you need to select the appropriate class. Report the confusion table and the misclassification error for the test data.