Statistical Learning, Homework #1

Veronica Vinciotti, Marco Chierici

Released: 16/03/2021. Due: 30/03/2021

This homework deals with classification methods. Your report should be an RMarkdown file with the code you used for the exercises as well as commentaries on what you did.

Note that:

- your code should run without errors (except for minor adjustments such as file paths);
- you results should be reproducible (eg make sure to set a random seed if you draw a random sample!)
- you should discuss/justify each choice you make.

In this homework, you will analyse medical record data for a number of patients: some of them have tested positive to a disease, some have not. The objective will be to predict the onset of this disease from a number of clinical measurements.

What is in the database?

- 392 observations
- 8 attributes (clinical measurements) plus the class (disease 0/1)

The dataset is stored in the comma-separated text file `disease.txt` which you can find on Moodle.

Exercise 1

- Load the data
- Perform basic data exploration
- Check the distributions of attributes
- Plot a scatterplot matrix between all the independent variables, coloring the data by disease status
- Through plotting, try to understand which attributes are most related to the outcome

For all the subsequent exercises:

- Split the data randomly into reasonably sized train and test sets
- Evaluate whether the re-scaling of predictors has an effect on your analyses/conclusions

Exercise 2

- Perform a classification of the data into the two classes using a logistic regression model on all the predictors.
- Evaluate the output to identify the most significant predictors.
- Evaluate the model performance in terms of confusion matrix and train/test accuracy.
- Using the fitted model, predict the probability of having the disease for someone who has C2=31 and all other predictors set to their average value.

Exercise 3

- Perform a classification via a k-nn model using all of the available variables and exploring different values of k;
- Discuss the results and reach a conclusion on the optimal value for k.

Exercise 4

Explore the use of LDA, QDA, and Naive Bayes to predict disease onset using all the predictors. For each method:

- Train the model on the training data
- Apply the fitted model to the test set; compute the confusion matrix and prediction accuracy

Exercise 5

- Compare all the methods considered so far on the test data (logistic regression, k-nn, LDA, QDA, NB).
- Draw the ROC curve, combining all of the ROC curves in a single plot. Compute also the AUC for each method. Discuss whether there is any method clearly outperforming the rest.
- Reflect on what is not ideal on this comparative analysis, which could bias the results in favor of one of the methods (which one?)? Do you see this on the results?

Exercise 6

Since the predictors are continuous, there could be also the option of including polynomial terms.

In the context of logistic regression

- Consider only the predictor C2 and fit a model which includes higher orders of this variable in the model;
- Explore different degrees and check the performance on test data to reach a conclusion on the optimal degree;
- Write the formula of your optimal model.