# Machine learning and statistical learning

## Logistic regression: Goodness of fit for a classifier

S. Hué

## 1 Objective and context

In this exercise, you will first estimate a logit model on bank data, using an implemented method from your statistical software.
These data are freely available at the following url: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing. Make sure to download the sample that contains only 10% of the observations.
The objective of the first step is to predict the categorical response variable $y$, which takes the value `yes` when a client has subscribed to a term deposit and `no` otherwise. You will use the variable `duration` as a predictor, *i.e.*, the last contact duration, in seconds.
The model writes:

$$y_i = \beta_0 + \beta_1 \text{duration}_i + \varepsilon_i,$$

where we assume that the error term $\varepsilon$ is logistically distributed with zero mean and variance $\pi^2/3$.
In a second step, based on the probabilities predicted to belong to either the class 0 (the client has not subscribed to a term deposit) or 1 (the client has), you will assign a predicted probability:

$$\hat{y} = \begin{cases} 1 & \text{if } \hat{\mathbb{P}}(Y = 1 \mid X = x_0) \geq \tau \\ 0 & \text{if } \hat{\mathbb{P}}(Y = 1 \mid X = x_0) < \tau \end{cases},$$

where $\tau$ will vary in the exercise.

## 2 Questions

1. Load the data in your software. Assign the value 1 for a client that has subscribed to a term deposit and 0 if he or she has not.
2. Split your data into a training and a test datasets.
3. Using the appropriate routine, fit a logit model in the train set to predict the probability of subscribing to a term deposit using the following predictors: duration, education, and campaign.
4. In a table, store the true observed value in a first column, and the predicted probability estimated by the logit model.
5. Using a threshold value of of .5, assign a predicted class to each individual.

6. Compute a confusion matrix based on that threshold.
7. Create a function that computes, given a value $\tau$, the true positive rate, the false positive rate, the true negative rate, the false negative rate and the overall error of your predictions based on $\tau$.
8. Apply this function to multiple values of $\tau$ ranging from 0 to 1.
9. Plot the ROC curve.
10. Comment the results.