6 Logistic Regression

Exercise 6.1 (Logistic Regression with a Normal Regressor)
Let Y be a binary random variable taking on its value in $\{0, 1\}$ and let $p = \Pr(Y = 1)$.
Let X be a random variable such that the distribution of X given $Y = j$ is an univariate normal distribution with mean m_j and standard-deviation σ. The pdf of X given $Y = j$ is denoted $f_j(x)$.

1. Calculate the pdf $f_X(x)$ of X in function of $f_0(x)$, $f_1(x)$ and p.
2. Calculate $\Pr(Y = 1|X = x)$ in function of $f_0(x)$, $f_1(x)$ and p.
3. Show that $\Pr(Y = 1|X = x) = g(a + bx)$ where $g(\cdot)$ is the logistic function and a, b are some real values depending on m_0, m_1, σ and p.

Exercise 6.2 (Logistic Regression with Gradient Ascent)
Create your own R code to generate some samples following a logistic regression model and to estimate this model from the samples. The tasks are the followings :

1. Create $n = 1000$ samples $x_1, \ldots, x_n$ of a random variable X following a normal distribution with mean $m_0 = 0.5$ and standard deviation 1.2.
2. Create n samples $x_{n+1}, \ldots, x_{2n}$ of a random variable X' following a normal distribution with mean $m_1 = 1.1$ and standard deviation 1.2.
3. Generate some binary (0 or 1) response samples $y_1, \ldots, y_n$ based on the realizations $x_1, \ldots, x_n$ of X with $\beta_0 = 0.3$ and $\beta_1 = 1.7$.

   Hints : for a fixed value $X = x$, the probability to obtain $Y = 1$ must satisfy $\Pr(Y = 1|X = x) = P(x)$ with $P(x) = f(\beta_0 + \beta_1 x)$ where $f(\cdot)$ is the logistic function. The mechanism generating the response samples y_i's is crucially important for the rest of the work.
4. Generate some binary (0 or 1) response samples $y_{n+1}, \ldots, y_{2n}$ based on the realizations $x_{n+1}, \ldots, x_{2n}$ of X' with the same parameters β_0 and β_1.
5. From all the samples $(y_1, x_1), \ldots, (y_{2n}, x_{2n})$, estimate the logistic regression model by using the maximum likelihood principle and the gradient ascent algorithm as follows :
   (a) Compute the gradient (at each step of the loop),
   (b) Update the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the parameters (at each step),
   (c) Compute the maximum likelihood function (at each step). Do not forget to verify that this cost function is increasing at each step (plot the cost as a function of the iterations and plot also the norm of the gradient as a function of the iterations).

   The step γ of the gradient ascent should be tuned. You can divide the gradient by its norm to facilitate the choice of γ.
6. Study numerically the convergence and the quality of the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. You can modify the step, the number n of samples, the starting point of the gradient ascent, etc.
7. Estimate the probabilities $\Pr_{m_0}(Y = 1)$ and $\Pr_{m_1}(Y = 1)$ where $\Pr_m(Y = 1)$ is the probability to obtain $Y = 1$ when the X used to generate the response Y follows the normal distribution with mean m. The estimates must be based on the samples $(y_1, x_1), \ldots, (y_{2n}, x_{2n})$ generated above.

Exercise 6.3 (Logistic Regression with GLM)

1. Load the data set “bank-additional.csv”. Its content is described on the web page
   https ://archive.ics.uci.edu/ml/datasets/bank+marketing
   The goal is to predict the variable “y” from the subset of variables “age”, “job”, “marital”, and “duration”.
2. Prepare the dataset to use a k-fold cross-validation with $k = 10$.
3. Use the command “glm” with the option “(...,family="binomial",...)” to estimate the logistic regression model from each train dataset.
4. Exploit the estimated model to predict the class of each test dataset.
5. Estimate the false positive rate and the false negative rate for each fold of the k-fold cross-validation.
   Hints in R : you can use the library “caret” and its function “confusionMatrix”.
6. Plot the false positive rate and the false negative rate with respect to the fold number.