Logistic Regression
using gmpy2 arithmetic

1 The model

The logistic regression model or (logit model) is a binary classificatin model in which the conditional probability is

$$p(y_i = 1|x_i) = \frac{1}{1 + e^{\beta_0 + \sum_{i=1}^{N} \beta_i x_i}}$$

where $(x_i, y_i), i = 1 \cdots N$ is the observed sample of data and $\beta_{i=0 \cdots N}$ is the vector of parameters.We note $X = (x_i)_{i=1 \cdot N}), Y = (y_i)_{i=1 \cdot N})$. It is assumed that y_i is a Bernouilli random variable. We he also

$$p(y_i = 0|x_i) = 1 - \frac{1}{1 + e^{\beta_0 + \sum_{i=1}^{N} \beta_i x_i}}$$

The likelihood of the observed sample $(x_i, y_i), i = 1 \cdots N$ is

$$L(X, Y, \beta) = \prod_{i=1}^{N} S(\beta.X)^{y_i} (1 - S(\beta.X))^{1-y_i}$$
$$S(\beta.X) = \frac{1}{1 + e^{\beta_0 + \sum_{i=1}^{N} \beta_i x_i}}$$

The log likelihood is

$$l(X, Y, \beta) = \sum_{i=1}^{N} y_i \log S(\beta.X) + (1 - y_i) \log(1 - S(\beta.X))$$

The maximum likelihood estimator solves $\widehat{\beta} = \arg\max_\beta l(X, Y, \beta)$, it is obtained when it possible of solving equation

$$\nabla_\beta l(X, Y, \beta) = 0$$

The first order condition above has no explicit solution . In most statistical software packages it is solved by using the the Newton-Raphson Technique. The method is pretty simple: we start from a guess of the solution $\widehat{\beta_0}$, (e.g. $\widehat{\beta_0} = 0$), and then we recursively update the guess with the equation

$$\widehat{\beta_n} = \widehat{\beta_{n-1}} \nabla_{\beta\beta} l(X, Y, \widehat{\beta_{n-1}})^{-1} \nabla_\beta l(X, Y, \widehat{\beta_{n-1}})$$

until numerical convergence (of $\widehat{\beta_n}$ to the solution $\widehat{\beta}$). Here we use the gmpy2 library for arbitrary-precision arithmetic.

2 Python computation

Our dataset is made up of a column X of 100 random integer in the range [55000..78000], and a (boolean) column Y of 100 value.

```
import pandas as pd
z ={'col1':np.random.randint(55000,78000,size =100),
 'col2':np.random.randint(2, size=100)}
pd.DataFrame(z).to_csv("data.csv")
```

We specialize the case of a vector of 2 parameter $\beta = [\beta_0, \beta_1]$. The file "newton.py" is the newton raphson method, it returns the vector β solution, when starting from initial vector $[\beta_0 = 15.1, \beta_1 - 0.4]$. The file "graph.py" plots graph of a function of two variable "log likelihood" $l(X, Y, \beta)$ as function of β.