SPLEX TME 4

La régression logistique

Implementer l'algorithme de la régression logistique et tester le sur les données simulées et les données "obeses_temoins" pour classifier les patients en deux classes. Comparez vous resultats avec ceux obtenus avec un package R.

1. Simulez un jeu de données

```
set.seed(666)
x1 = rnorm(1000)
x2 = rnorm(1000)
z = 1 + 2*x1 + 3*x2
pr = 1/(1+exp(-z))
y = rbinom(1000,1,pr)
```

et tester la régression logistique (la fonction *glm()*) sur ce jeu de données

```
df = data.frame(y=y,x1=x1,x2=x2)
glm( y~x1+x2,data=df,family="binomial")
```

2. La régression logistique binaire

Soit une ensemble "training set" de N observation $\{X_n, Y_n\}_{n=1}^N$. Dans le cadre de la régression logistique binaire, la variable Y prend deux modalités possibles $\{1, 0\}$. Les variables X sont exclusivement continues ou binaires.

La régression logistique est un modèle parametrique donc la log-vrasemblance est donnée par

$$\ell(Y|X;\theta) = -\sum_{i=n}^{N} \Big(y_n \theta^T x_n - \log(1 + \exp(\theta^T x_n)) \Big) \qquad (1)$$

où θ est le vecteur de paramètres à estimer.

Pour classifier une nouvelle observation X, on calcule et comapre les probabilités de classes sachant l'observation

$$p(Y = 1|X) = \frac{\exp \theta^T X}{1 + \exp \theta^T X} \tag{2}$$

$$p(Y = 0|X) = \frac{1}{1 + \exp \theta^T X}. \tag{3}$$

La méthode de Newton-Raphson

La méthode de Newton-Raphson qu'on utilise pour maximiser la log-vraisemblance et pour estimer les paramètres θ du modèle est une procedure itérative du gradient. Pour la régression logistique binaire la procedure est la suivante:

Initialize $\theta = (0, \ldots, 0)$

for $t = 1 : T$ // Faire plusieurs itérations ou jusqu'à la convergence **do**

Compute the first derivative
$\frac{\partial \ell(\theta)}{\partial \theta} = \sum_{n=1}^{N} x_n(y_n - p(y = 1|x_n))$ // dimension = 1×nb of parameters

Compute the Hessian matrix
$\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta'} = \sum_{n=1}^{N} x_n x_n^t p(y = 1|x_n)(1 - p(y = 1|x_n))$
//dimension = nb of parameters × nb of parameters

Update the parameters
$\theta = \theta - \frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta'}^{-1} \frac{\partial \ell(\theta)}{\partial \theta}$
end for