Logistic regression hypothesis function is :

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

$$= P(y=1 | x; \theta)$$

The cost function $J(\theta)$ is :

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[-y^i \log(h_\theta(x^i)) - (1-y^i)\log(1-h_\theta(x^i))\right]$$

We will minimize the function. From Netwton-Raphson method, we know the update rule is

$$\theta^{t+1} = \theta^t - H^{-1}\nabla_\theta J$$

We know, the gradient and hessian are,

Gradient, $\nabla_\theta J = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)\, x^i$

Hessian, $H = \frac{1}{m}\sum_{i=1}^{m}\left[h_\theta(x^i) \cdot (1-h_\theta(x^i)) \cdot x^i \cdot (x^i)^T\right]$

The implementation steps are follows :

1. There are 16 ~~some~~ missing values in dataset. They are filled up by mean of that column.

2. The output column contains 2 for benign and 4 for malignent. We replaced the value as 0 for benign and 1 for malignant.

3. The dataset is randomly sampled to create training and test dataset. Training dataset contains 80% and test dataset contains 20% of ~~datas~~ the total data.

4. The sigmoid function is defined.

5. Then Newton-Raphson method is used iteratively for 15 times to find the optimal solution. Gradient, hessian, ~~and~~ updated parameters and cost function is calculated in every iteration.

6. After getting the optimal parameter values, they were ~~all~~ applied on the test dataset. If the value is greater than 0.5, then it is considered as malignant or 1. Otherwise it is considered as benign or 0.

7. Then sum of squared error and accuracy of classification are calculated.

$$\text{error} = \frac{1}{2}\sum_{i=1}^{m} (p^i - y^i)^2$$

$$\text{accuracy} = \frac{\text{correct output}}{\text{total predictions}}$$

8. Then average error and average classification accuracy are calculated for 10 trials.

Results:

Average error : 2.5

Average accuracy : 96.43% (0.9643)

Checklist :

1. Files : (i) assignment3.py
   (ii) data.csv

2. Report

~~To~~ The whole assignment code is in assignment3.py file. Dataset file data.csv is needed to be in same directory.