FUNDAMENTALS OF MACHINE LEARNING
ASSIGNMENT-1

DEEPSHIKHA-CS21BTECH110116
TEJAL-CS21BTECH11058

Question 4

Logistic regression sigmoid function :

$$h_w(x) = \frac{1}{1 + e^{-w^T x}}$$

Here, w denotes the parameter vector.
For a model containing n features, we have w = [w_0, w_1, ..., w_n] containing n + 1 parameters.
Using logarithmic function to represent the cost of logistic regression,

$$cost(h_w(x), y) = \begin{cases} -log(h_w(x)) & \text{, if } y = 1 \\ -log(1 - h_w(x)) & \text{, if } y = 0 \end{cases}$$

For m observations, we can calculate the cost as:

$$J(w) = -\sum_{i=1}^{m} \left[y^{(i)} \times log(h_w(x^{(i)})) + (1 - y^{(i)}) \times log(h_w(x^{(i)})) \right]$$

Part(a):

- Gradient of log likelihood function wrt w is
$$\frac{\partial J(w)}{\partial w} = \sum_{i=1}^{m} [h_w(x) - y] X_n$$
- The Hessian matrix is the second derivative of the cost function with respect to the parameters, given by
$$H = X^T D X$$
where, D is the diagonal matrix of weights, where $W_{ii} = h_w(x)$(1-$h_w(x)$), where $h_w(x)$ is the i_{th} row of the feature matrix.
- The parameter update in the Newton-Raphson optimization scheme is given by:
$$w_{new} = w_{old} - (H^{-1}) \nabla J$$
where,
 w_{new} is the updated parameter vector,
 w_{old} is the current parameter vector,
 H^{-1} is the inverse of Hessian matrix,
 ∇J is the gradient of cost function
- Algorithm :
  (1) Initialise w with some initial value.
  (2) Repeat until the change in the parameter values becomes very small or after a fixed number of iterations i.e until convergence
    (a) Compute gradient ∇J .
    (b) Compute Hessian matrix H.

Date: 8th October 2023.

(c) Update parameter value using update equation

(3) Once convergence is achieved, parameter vector w will contain the optimal parameter values.

Part(b): The Newton-Raphson update formula for the logistic regression model then becomes

$$\begin{aligned} w_{new} &= w_{old} - H^{-1}\nabla J \\ &= w_{old} - (X^T D X)^{-1} X^T (h_w(x) - y) \\ &= (X^T D X)^{-1} \left\{ X^T D X w_{old} - X^T (h_w(x) - y) \right\} \\ &= (X^T D X)^{-1} X^T D \left\{ X w_{old} - D^{-1} (h_w(x) - y) \right\} \end{aligned}$$

This update formula takes the form of a set of normal equations for a weighted least-squares problem. Because the weighing matrix D is not constant but depends on the parameter vector w, we must apply the normal equations iteratively, each time using the new weight vector w to compute a revised weighing matrix R. For this reason, the algorithm is known as iterative reweighted least squares,

Part(c): For a function to be convex, its second derivative should be positive. The second derivative of the error function is the Hessian function .

We have ,

$$H = X^T D X$$

We know that the entries of D are positive $h_w(x)$(1-$h_w(x)$) is the derivative of the sigmoid function , so it is positive which follows from the form of the logistic sigmoid function, we see that $u^T H u > 0$ for an arbitrary vector u, and so the Hessian matrix H is positive definite. It follows that the error function is a convex function of w and hence has a unique minimum.