Part 4 - Hessian of L2 penalized binary logistic regression
Show that the Hessian or second derivative of $J(\theta)$ can be written as

$$H = \frac{1}{m}(X^T SX + \lambda I)$$

$$S = diag(h_\theta(x^{(1)})(1 - h_\theta(x^{(1)})), ..., h_\theta(x^{(m)})(1 - h_\theta(x^{(m)})))$$

Show that H is positive definite. You may assume that $0 < h_\theta(x^{(i)}) < 1$ so the elements of S are strictly positive and that X is full rank.

Solution
S is a diagonal matrix with the elements $h_\theta(x^{(i)})(1 - h_\theta(x^{(i)}))$ on the diagonal. This can be represented as:

$$S = \text{diag}(h_\theta(X)(1 - h_\theta(X)))$$

Where $h_\theta(X)$ is a vector containing the predicted values for all training examples. Let us rewrite the gradient $\nabla J(\theta)$ using H, S, X, and θ:

$$\nabla J(\theta) = \frac{1}{m}X^T(h_\theta(X) - y) + \frac{\lambda}{m}\theta = \frac{1}{m}X^T S(X\theta - y) + \frac{\lambda}{m}\theta$$

where $X\theta$ represents the predictions $h_\theta(X)$ for all training examples. Let compute the Hessian of $J(\theta)$:

$$\begin{aligned} H &= \frac{\partial}{\partial\theta}\left(\frac{1}{m}X^T S(X\theta - y) + \frac{\lambda}{m}\theta\right) \\ &= \frac{1}{m}\left(\frac{\partial}{\partial\theta}\left(X^T SX\theta - X^T Sy\right) + \frac{\partial}{\partial\theta}\lambda\theta\right) \\ &= \frac{1}{m}\left(X^T SX + \lambda I\right) \end{aligned}$$

A matrix is positive definite if it's symmetric and all its eigenvalues are positive. First, let us prove that H is symmetric. By definition, a symmetric matrix is equal to its transpose:

$$\begin{aligned} H^T &= \left(\frac{1}{m}(X^T SX + \lambda I)\right)^T \\ &= \frac{1}{m}(X^T SX + \lambda I)^T \\ &= \frac{1}{m}(X^T(SX)^T + (\lambda I)^T) \\ &= \frac{1}{m}(X^T SX + \lambda I) = H \end{aligned}$$

Given that $0 < h_\theta(x^{(i)}) < 1$ for all i, then all elements on the diagonal of the matrix S are strictly positive since they are of the form $h_\theta(x^{(i)})(1 - h_\theta(x^{(i)}))$.

Now, let's prove that all eigenvalues of H are strictly positive. Consider $H = \frac{1}{m}(X^T SX + \lambda I)$. Let $A = X^T SX$ and $B = \lambda I$. Since X is full rank, all of its columns of linearly independent; thus, $X^T X$ is positive definite (all of its eigenvalues are strictly positive). Similarly, since all elements of S are strictly positive, S is also positive definite.

Now, consider λ_A and λ_B, the eigenvalues of A and B, respectively. Then, the eigenvalues of H are:

$$\lambda_H = \frac{1}{m}(\lambda_A + \lambda_B)$$

Since both λ_A and λ_B are positive, then λ_H is also positive. Therefore, all eigenvalues of H are strictly positive.
Since $H = H^T$ and $\lambda_H > 0$, then we can conclude that H is positive definite.