National Economics University, Vietnam

Faculty of Mathematics Economics

Data Science in Economics and Business

Machine Learning 1

Homework Week 5: Logistic Regression

Student: Nguyễn Anh Tú - ID: 11207333

1 Problem 1

Calculate the gradient vector for loss function of Logistic Regression.
The cross entropy loss function of Logistic Regression is given as:

$$\mathcal{L} = -\log p(t|w) = -\sum_{i=1}^{m}\{t^{(i)}\log y^{(i)} + (1-t^{(i)})\log(1-y^{(i)})\}$$

where $y^{(i)} = \sigma(x^{(i)}w)$ and $t^{(i)} \in \{0, 1\}$.
Before going to find the gradient vector, we need to prove a property of the sigmoid function which will be useful for the later calculation, that is:
For $\sigma(a) = 1/(1+e^{-a})$, then:

$$\sigma(a)' = \sigma(a)(1-\sigma(a))$$

Proof:

$$\begin{aligned}
\sigma(a)' &= \left(\frac{1}{1+e^{-a}}\right)' \\
&= (e^{-a})' \cdot \left(\frac{1}{1+e^{-a}}\right)' \\
&= -e^{-a} \cdot \left(-\frac{1}{(1+e^{-a})^2}\right) \\
&= \frac{e^{-a}}{(1+e^{-a})^2} \\
&= \frac{1}{1+e^{-a}} \cdot \frac{e^{-a}}{1+e^{-a}} \\
&= \frac{1}{1+e^{-a}}\left(1-\frac{1}{1+e^{-a}}\right) \\
&= \sigma(a)(1-\sigma(a))
\end{aligned}$$

The partial derivative of $\mathcal{L}$ respected to w is given by:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w_j} &= -\frac{\partial}{\partial w_j}\sum_{i=1}^{m} t^{(i)}\log y^{(i)} + (1-t^{(i)})\log(1-y^{(i)}) \\
&= -\sum_{i=1}^{m} t^{(i)}\frac{\partial \log(\sigma(x^{(i)}w))}{\partial w_j} + (1-t^{(i)})\frac{\partial \log(1-\sigma(x^{(i)}w))}{\partial w_j} \\
&= -\sum_{i=1}^{m} t^{(i)}\frac{\sigma(x^{(i)}w)(1-\sigma(x^{(i)}w))}{\sigma(x^{(i)}w)}\cdot\frac{\partial x^{(i)}w}{\partial w_j} - (1-t^{(i)})\frac{\sigma(x^{(i)}w)(1-\sigma(x^{(i)}w))}{1-\sigma(x^{(i)}w)}\frac{\partial x^{(i)}w}{\partial w_j} \\
&= -\sum_{i=1}^{m} t^{(i)}(1-\sigma(x^{(i)}w_j))x_j^{(i)} - (1-t^{(i)})\sigma(x^{(i)}w_j)x_j^{(i)} \\
&= -\sum_{i=1}^{m} t^{(i)}x_j^{(i)} - t^{(i)}x_j^{(i)}\sigma(x^{(i)}w_j) - \sigma(x^{(i)}w_j)x_j^{(i)} + t^{(i)}x_j^{(i)}\sigma(x^{(i)}w_j) \\
&= -\sum_{i=1}^{m} t^{(i)}x_j^{(i)} - \sigma(x^{(i)}w_j)x_j^{(i)} \\
&= \sum_{i=1}^{m}(\sigma(x^{(i)}w_j) - t^{(i)})x_j^{(i)}
\end{aligned}$$

The above result can be re-written in the form of vector calculus to derive the gradient vector for the loss function as:

$$\frac{d\mathcal{L}}{dW} = X^T(\sigma(XW) - T)$$

2 Problem 2-4

1. Implement a Logistic Regression class which uses Gradient Descent algorithm to find the optimal w.
2. Fit the model above in the dataset in file dataset.csv then find the model coefficient.
3. Plot the decision boundary for the dataset.

Solution.

1. Choosing initial $w = \overrightarrow{0}$, the learning rule of Gradient Descent is given by:

   $$w_{k+1} = w_k - \alpha\frac{d\mathcal{L}}{dw_k}$$

   where α is learning rate. Iteration is terminated when w_{k+1} and w_k is close enough, i.e. $||w_{k+1} - w_k||$ is smaller than some tolerance ϵ.

2. The coefficient found by fitting the dataset to Logistic Regression model is:

Intercept	-27.53284233
Salary	2.75525136
Experience	11.87527319

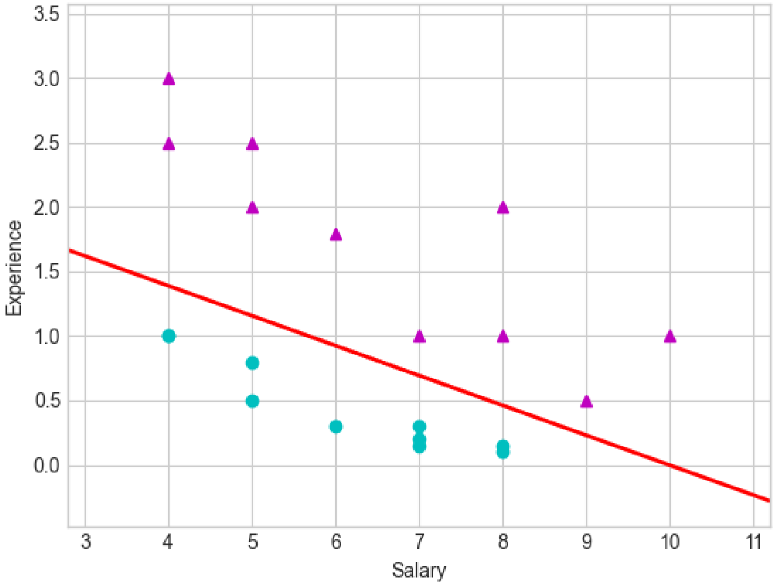
3. It's important to know how the decision boundary is derived, so we can implement it later. We often choose the separation for two classes is where the probability for each class equals to 0.5 (an input x is predicted to be in class 1 if $p(C1|x) < 0.5$ and in class 2 if $p(C1|x) > 0.5$). The posterior probability for a class (in the case of two classes) is $p(C1|x) = \sigma(a)$ where:

$$a = \log \frac{p(x|C1)p(C1)}{p(x|C2)p(C2)}$$

Therefore the decision boundary is the equation $p(C|x) = 0.5$ or $\sigma(xw) = 0.5$

$$\begin{aligned} \sigma(xw) = 0.5 &\Leftrightarrow \frac{1}{1+e^{-xw}} = \frac{1}{2} \\ &\Leftrightarrow e^{-xw} = 1 \\ &\Leftrightarrow -xw = 0 \\ &\Leftrightarrow w_0 + w_1x_1 + w_2x_2 = 0 \\ &\Leftrightarrow x_2 = -\frac{w_0}{w_2} - \frac{w_1}{w_2}x_1 \end{aligned}$$

Data points and the decision boundary is plotted as:



3 Problem 3

Prove that in Logistic Regression Model:

1. if loss function takes the form of binary cross entropy function, it is convex.
2. if loss function takes the form of mean square error, it is non-convex.

Solution. To prove that a function is convex, we need to show that its Hessian matrix is positive semi definite (second order derivative characterization of convexity).

1. Proof of convex loss cross entry function:
   As shown in Problem 1, the first order partial derivative of $\mathcal{L}$ is:

$$\frac{\partial \mathcal{L}}{\partial w_j} = \sum_{i=1}^{m} (\sigma(x^{(i)} w_j) - t^{(i)}) x_j^{(i)}$$

   Then the Hessian matrix entries or the second order derivative of $\mathcal{L}$ is:

$$\begin{aligned}\mathbf{H}_{jk} = \frac{\partial^2 \mathcal{L}}{\partial w_j \partial w_k} &= \sum_{i=1}^{m} \frac{\partial(\sigma(x^{(i)} w_j) - t^{(i)}) x_j^{(i)}}{\partial w_k} \\ &= \sum_{i=1}^{m} x_j^{(i)} \sigma(x^{(i)} w_j)(1 - \sigma(x^{(i)} w_j)) \frac{\partial x^{(i)} w_j}{\partial w_k} \\ &= \sum_{i=1}^{m} x_j^{(i)} \sigma(x^{(i)} w_j)(1 - \sigma(x^{(i)} w_j)) x_k^{(i)}\end{aligned}$$

   Therefore $\mathbf{H} = X^T \sigma(XW)(1 - \sigma(XW))X$
   To show that $\mathbf{H}$ is positive semidefinite, we need to show that the quadratic form of $z^T \mathbf{H} z \geq 0$ for any $z \in \mathbb{R}^n$.

$$\begin{aligned}z^T H z &= \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{n} \sigma(x^{(i)} w_j)(1 - \sigma(x^{(i)} w_j)) x_k^{(i)} x_j^{(i)} z_j z_k \\ &= \sum_{i=1}^{m} \sigma(x^{(i)} w_j)[1 - \sigma(x^{(i)} w_j)][(x^{(i)})^T z]^2\end{aligned}$$

   Since $\sigma(a) = 1/(1 + e^{-a})$ then $0 \leq \sigma(a) \leq 1$.

   Therefore:

$$\sum_{i=1}^{m} \sigma(x^{(i)} w_j)[1 - \sigma(x^{(i)} w_j)][(x^{(i)})^T z]^2 \geq 0 \Leftrightarrow z^T \mathbf{H} z \geq 0 \quad \square$$

2. Proof of non-convex loss MSE:

If loss function takes the form of MSE, it will be:

$$\mathcal{L} = \frac{1}{m}\sum_{i=1}^{m}(t^{(i)} - \sigma^{(i)})^2$$

The first order derivative of $\mathcal{L}$ is:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w_j} &= \frac{1}{m}\sum_{i=1}^{m}\frac{\partial (t^{(i)} - \sigma^{(i)})^2}{\partial w_j} \\
&= -\frac{2}{m}\sum_{i=1}^{m}(t^{(i)} - \sigma^{(i)})\frac{\partial \sigma^{(i)}}{\partial w_j} \\
&= -\frac{2}{m}\sum_{i=1}^{m}(t^{(i)} - \sigma^{(i)})\sigma^{(i)}(1 - \sigma^{(i)})\frac{\partial x^{(i)}w}{\partial w_j} \\
&= -\frac{2}{m}\sum_{i=1}^{m}(t^{(i)} - \sigma^{(i)})\sigma^{(i)}(1 - \sigma^{(i)})x_j^{(i)} \\
&= -\frac{2}{m}\sum_{i=1}^{m}(t^{(i)}\sigma^{(i)} - t^{(i)}(\sigma^{(i)})^2 - (\sigma^{(i)})^2 + (\sigma^{(i)})^3)x_j^{(i)}
\end{aligned}$$

The second order derivative of $\mathcal{L}$ or the entry of Hessian matrix is:

$$\begin{aligned}
\mathbf{H}_{jk} = \frac{\partial^2 \mathcal{L}}{\partial w_j \partial w_k} &= -\frac{2}{m}\sum_{i=1}^{m}\frac{\partial (t^{(i)}\sigma^{(i)} - t^{(i)}(\sigma^{(i)})^2 - (\sigma^{(i)})^2 + (\sigma^{(i)})^3)x_j^{(i)}}{\partial w_k} \\
&= -\frac{2}{m}\sum_{i=1}^{m}x_j^{(i)}(t^{(i)} - 2t^{(i)}\sigma^{(i)} - 2\sigma^{(i)} + 3(\sigma^{(i)})^2)\frac{\partial \sigma^{(i)}}{\partial w_k} \\
&= -\frac{2}{m}\sum_{i=1}^{m}x_j^{(i)}(t^{(i)} - 2t^{(i)}\sigma^{(i)} - 2\sigma^{(i)} + 3(\sigma^{(i)})^2)\sigma^{(i)}(1 - \sigma^{(i)})\frac{\partial x^{(i)}w_j}{\partial w_k} \\
&= -\frac{2}{m}\sum_{i=1}^{m}x_j^{(i)}(t^{(i)} - 2t^{(i)}\sigma^{(i)} - 2\sigma^{(i)} + 3(\sigma^{(i)})^2)\sigma^{(i)}(1 - \sigma^{(i)})x_k^{(i)} \\
&= -\frac{2}{m}\sum_{i=1}^{m}x_j^{(i)}x_k^{(i)}\sigma^{(i)}(1 - \sigma^{(i)}) \cdot A
\end{aligned}$$

for A $= t^{(i)} - 2t^{(i)}\sigma^{(i)} - 2\sigma^{(i)} + 3(\sigma^{(i)})^2$.
The quadratic form $z^T\mathbf{H}z$ for any $z \in \mathbb{R}^n$ is:

$$z^T\mathbf{H}z = -\frac{2}{m}\sum_{i=1}^{m}\sigma(x^{(i)}w_j)[1 - \sigma(x^{(i)}w_j)][(x^{(i)})^T z]^2 \cdot A$$

As shown in part 1, $\sum_{i=1}^{m}\sigma(x^{(i)}w_j)[1 - \sigma(x^{(i)}w_j)][(x^{(i)})^T z]^2 \geq 0$. Then we need to examine the value of A.

Case 1: $t^{(i)} = 0$

$$A = 0 - 2 \cdot 0 \cdot \sigma^{(i)} - 2\sigma^{(i)} + 3(\sigma^{(i)})^2 = -2\sigma^{(i)} + 3(\sigma^{(i)})^2$$

Since $\sigma^{(i)} \in [0, 1]$ then $A \in [-\frac{1}{3}, 1] \Rightarrow z^T\mathbf{H}z$ is not positive semidefinite. (1)

Case 2: $t^{(i)} = 1$

$$A = 1 - 2\sigma^{(i)} - 2\sigma^{(i)} + 3(\sigma^{(i)})^2 = 1 - 4\sigma^{(i)} + 3(\sigma^{(i)})^2$$

Since $\sigma^{(i)} \in [0, 1]$ then $A \in [-\frac{1}{3}, 1] \Rightarrow z^T\mathbf{H}z$ is not positive semidefinite. (2)

From (1) and (2) we can conclude that the Hessian matrix of $\mathcal{L}$ is not positive semidefinite so loss MSE is non-convex for Logistic Regression. □