

Review Logistic Regression

Tran Giang Son, tran-giang.son@usth.edu.vn

ICT Department, USTH



Review

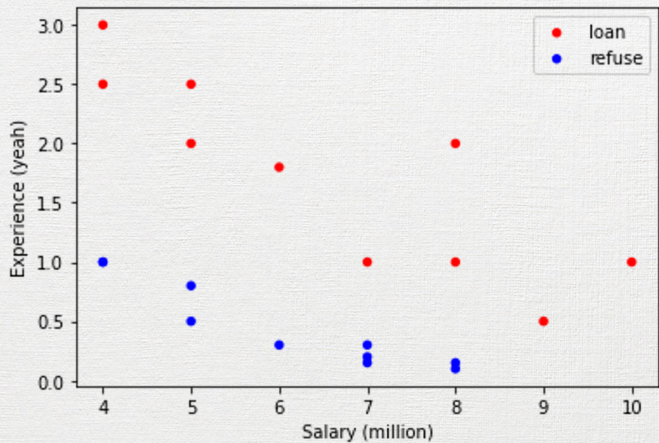


Example

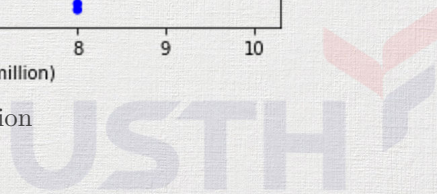
- Problem: decision supporting for loan program
- Input: Having data about salary and working time of employees
- Output: Loan or not?

| Salary (M VND) | Experience (year) | Loan |
|----------------|-------------------|------|
| 10 | 1 | 1 |
| 9 | 0.5 | 1 |
| 5 | 2 | 1 |
| ... | ... | ... |
| 8 | 0.1 | 0 |
| 6 | 0.3 | 0 |
| 7 | 0.15 | 0 |
| ... | ... | ... |

Example



Visualization



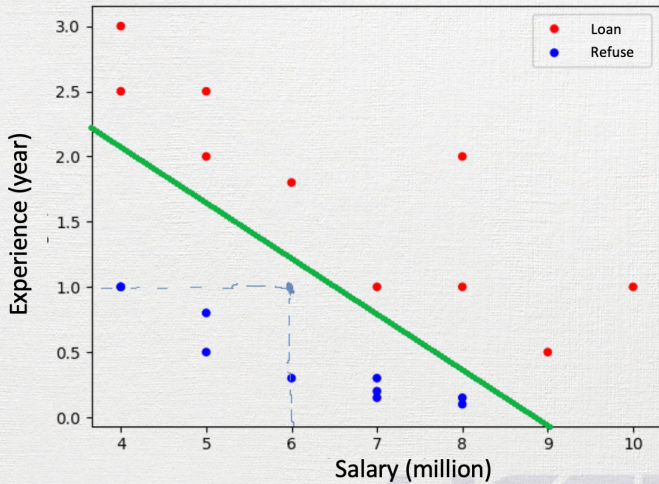
Formulation

- Let $x_i^{(1)}$ be the salary and $x_i^{(2)}$ be the working time of the profile i
- Prediction model is defined as follows:

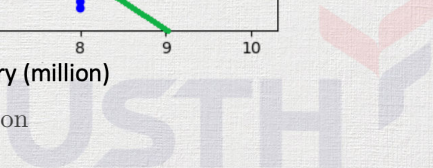
$$\hat{y}_i = w_1 x_i^{(1)} + w_2 x_i^{(2)} + w_0$$



Visualization



Visualization



Prediction

- Requirement: estimate the probability that a new profile should be loaned or not
- Output:
 - If the estimated loan probability \geq threshold t , then the new profile should be loaned
 - Otherwise, it should be refused



Sigmoid

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (1)$$

- Continuous function with real values in the interval $(0, 1)$
- Derivative at every point (for applying gradient descent)



Model definition

- Estimated loan probability \hat{y}_i is therefore:

$$\hat{y}_i = \sigma(\hat{y}_i) = \sigma(w_1x_i^{(1)} + w_2x_i^{(2)} + w_0) \quad (2)$$

- But $\sigma(x) = \frac{1}{1+e^{-z}}$, therefore:

$$\hat{y}_i = \frac{1}{1 + e^{-(w_1x_i^{(1)} + w_2x_i^{(2)} + w_0)}} \quad (3)$$



Loss function

- Consider the probability that the model predicts that the profile i will be **loaned** as follows:

$$p(x^{(i)} = 1) = \hat{y}_i \quad (4)$$

- Consider the probability that the model predicts that the profile i will be **refused** as follows:

$$p(x^{(i)} = 0) = 1 - \hat{y}_i \quad (5)$$

- In total:

$$p(x^{(i)} = 1) + p(x^{(i)} = 0) = 1 \quad (6)$$

Loss function

- For each data point $(x^{(i)}, y_i)$, loss value L_i is defined as:

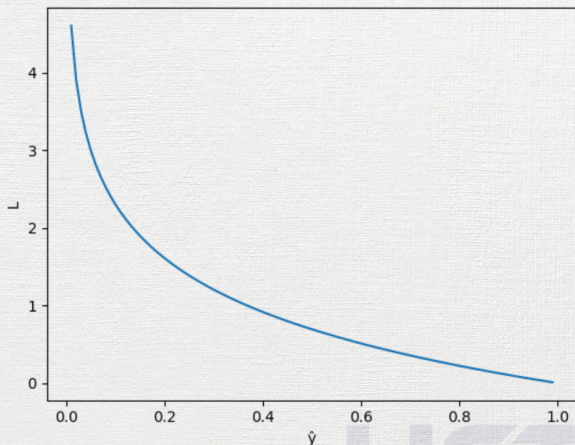
$$L_i = \begin{cases} -\log(\hat{y}_i) & \text{if } y_i = 1 \\ -\log(1 - \hat{y}_i) & \text{if } y_i = 0 \end{cases} \quad (7)$$

- To combine:

$$L_i = -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \quad (8)$$

Loss function

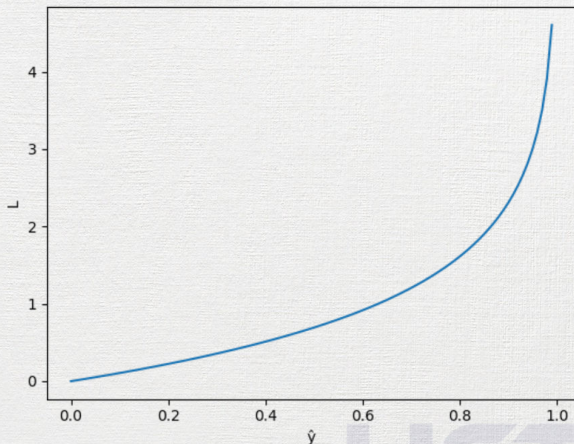
$$L_i = -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i))$$



$y_i = 1$

Loss function

$$L_i = -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i))$$



$y_i = 0$

Loss function

- For all data points, loss function J is defined as:

$$J = -\frac{1}{N} \sum_{i=1}^N (y_i \log(\hat{y}_i) + (1 - y_i) * \log(1 - \hat{y}_i)) \quad (9)$$

- Binary cross entropy loss
- Q: Why not MSE?



Training

- Function

$$J = -\frac{1}{N} \sum_{i=1}^N (y_i \log(\hat{y}_i) + (1 - y_i) * \log(1 - \hat{y}_i)) \quad (10)$$

- Apply gradient descent algorithm to find parameters $\{w_0, w_1, w_2\}$ which minimize J



Prediction

- Given a new profile (x_{new}, y_{new}) calculate predicted loan probability \hat{y}_{new} using found parameters $\{w_0, w_1, w_2\}$
- Loan decision is defined as follows:
 - If $\hat{y}_{new} > t$, loaned
 - Else, refused



Practice!



Labwork 3: Logistic Regression

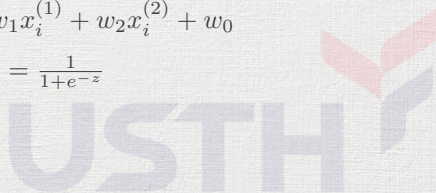
- Implement (from scratch!) logistic regression using previous gradient descend code to optimize w_0, w_1, w_2
 - Input: a CSV file with 3 columns
 - Output: w_0, w_1, w_2
 - Print the intermediate iterative steps
- Try experimenting with the loan decision example
- Write a report (in $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$):
 - Name it « Report.3.Logistic.Registration.tex »
 - How you implement the algorithm
 - Analyze the effect of different learning rate r w.r.t. convergence
- Push your code and report to your forked repository

Labwork 3: Logistic Regression (extras)

- Gradient descend for linear regression with binary cross entropy error loss
- Remind: single loss value with sigmoid

$$L_i = -(y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log(1 - \sigma(\hat{y}_i))) \quad (11)$$

- Where
 - \hat{y}_i is predicted output, $\hat{y}_i = w_1 x_i^{(1)} + w_2 x_i^{(2)} + w_0$
 - $\sigma(z)$ is sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$



Labwork 3: Logistic Regression (extras)

$$L_i = -(y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log(1 - \sigma(\hat{y}_i))) \quad (12)$$

- We have

$$y_i \log \sigma(\hat{y}_i) = y_i \log \frac{1}{1 + e^{-\hat{y}_i}} = -y_i \log(1 + e^{-\hat{y}_i}) \quad (13)$$

$$\begin{aligned} (1 - y_i) \log(1 - \sigma(\hat{y}_i)) &= (1 - y_i) \log\left(1 - \frac{1}{1 + e^{-\hat{y}_i}}\right) \\ &= (1 - y_i) \log\left(\frac{e^{-\hat{y}_i}}{1 + e^{-\hat{y}_i}}\right) \\ &= (1 - y_i)(\log(e^{-\hat{y}_i}) - \log(1 + e^{-\hat{y}_i})) \\ &= (1 - y_i)(-\hat{y}_i - \log(1 + e^{-\hat{y}_i})) \end{aligned} \quad (14)$$

Labwork 3: Logistic Regression (extras)

- Single loss value

$$\begin{aligned}L_i &= -(y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log(1 - \sigma(\hat{y}_i))) \\ &= -(-y_i \log(1 + e^{-\hat{y}_i}) + (1 - y_i)(-\hat{y}_i - \log(1 + e^{-\hat{y}_i})))\end{aligned}\quad (15)$$



Labwork 3: Logistic Regression (extras)

$$\begin{aligned}L_i &= -(-y_i \log(1 + e^{-\hat{y}_i}) + (1 - y_i)(-\hat{y}_i - \log(1 + e^{-\hat{y}_i}))) \\&= -(y_i \hat{y}_i - (\hat{y}_i + \log(1 + e^{-\hat{y}_i}))) \\&= -(y_i \hat{y}_i - \log e^{\hat{y}_i} - \log(1 + e^{-\hat{y}_i})) \\&= -(y_i \hat{y}_i - \log((e^{\hat{y}_i})(1 + e^{-\hat{y}_i}))) \\&= -(y_i \hat{y}_i - \log(1 + e^{\hat{y}_i}))\end{aligned}\tag{16}$$



Labwork 3: Logistic Regression (extras)

- Loss of all data points

$$\begin{aligned} J &= \frac{1}{N} \sum_{i=1}^N L_i = -\frac{1}{N} \sum_{i=1}^N (y_i \hat{y}_i - \log(1 + e^{\hat{y}_i})) \\ &= -\frac{1}{N} \sum_{i=1}^N (y_i (w_1 x_i^{(1)} + w_2 x_i^{(2)} + w_0) - \log(1 + e^{w_1 x_i^{(1)} + w_2 x_i^{(2)} + w_0})) \end{aligned} \quad (17)$$

- Minimize this function w.r.t w_0, w_1, w_2 ,

- Needs partial derivatives $\frac{dJ}{dw_0}, \frac{dJ}{dw_1}, \frac{dJ}{dw_2}$

Labwork 3: Logistic Regression (extras)

- Remind: gradient descend
 - Initial value x_0
 - Function $f(x)$
 - First order derivative $f'(x)$
 - Learning rate r
 - Threshold t



Labwork 3: Logistic Regression (extras)

- 3D gradient descend for linear regression weights w_0, w_1, w_2
 - Initial value $w_0^{(0)}, w_1^{(0)}, w_2^{(0)}$
 - Function $f(w_0, w_1, w_2) = L_i$
 - First order partial derivatives $\frac{df}{dw_0}, \frac{df}{dw_1}, \frac{df}{dw_2}$
 - Learning rate r
 - Threshold t



Labwork 3: Logistic Regression (extras)

- Function to calculate single loss value:

$$f(w_0, w_1, w_2) = -y_i(w_1x_i^{(1)} + w_2x_i^{(2)} + w_0) + \log(1 + e^{w_1x_i^{(1)} + w_2x_i^{(2)} + w_0}) \quad (18)$$

- Therefore

$$\begin{aligned} \frac{df}{dw_0} &= -y_i + \frac{e^{w_1x_i^{(1)} + w_2x_i^{(2)} + w_0}}{1 + e^{w_1x_i^{(1)} + w_2x_i^{(2)} + w_0}} \\ &= 1 - y_i - \frac{1}{1 + e^{w_1x_i^{(1)} + w_2x_i^{(2)} + w_0}} \\ &= 1 - y_i - \sigma(-(w_1x_i^{(1)} + w_2x_i^{(2)} + w_0)) \end{aligned} \quad (19)$$

Labwork 3: Logistic Regression (extras)

- Function to calculate single loss value:

$$f(w_0, w_1, w_2) = -y_i(w_1x_i^{(1)} + w_2x_i^{(2)} + w_0) + \log(1 + e^{w_1x_i^{(1)} + w_2x_i^{(2)} + w_0}) \quad (20)$$

- Therefore

$$\begin{aligned} \frac{df}{dw_1} &= -y_i x_i^{(1)} + \frac{x_i^{(1)} e^{w_1 x_i^{(1)} + w_2 x_i^{(2)} + w_0}}{1 + e^{w_1 x_i^{(1)} + w_2 x_i^{(2)} + w_0}} \\ &= -y_i x_i^{(1)} + x_i^{(1)} \left(1 - \frac{1}{1 + e^{w_1 x_i^{(1)} + w_2 x_i^{(2)} + w_0}} \right) \\ &= -y_i x_i^{(1)} - x_i^{(1)} (1 - \sigma(-(w_1 x_i^{(1)} + w_2 x_i^{(2)} + w_0))) \end{aligned} \quad (21)$$

Labwork 3: Logistic Regression (extras)

- Function to calculate single loss value:

$$f(w_0, w_1, w_2) = -y_i(w_1x_i^{(1)} + w_2x_i^{(2)} + w_0) + \log(1 + e^{w_1x_i^{(1)} + w_2x_i^{(2)} + w_0}) \quad (22)$$

- Therefore

$$\begin{aligned} \frac{df}{dw_2} &= -y_i x_i^{(2)} + \frac{x_i^{(2)} e^{w_1 x_i^{(1)} + w_2 x_i^{(2)} + w_0}}{1 + e^{w_1 x_i^{(1)} + w_2 x_i^{(2)} + w_0}} \\ &= -y_i x_i^{(2)} + x_i^{(2)} \left(1 - \frac{1}{1 + e^{w_1 x_i^{(1)} + w_2 x_i^{(2)} + w_0}}\right) \\ &= -y_i x_i^{(2)} - x_i^{(2)} (1 - \sigma(-(w_1 x_i^{(1)} + w_2 x_i^{(2)} + w_0))) \end{aligned} \quad (23)$$

Labwork 3: Logistic Regression (extras)

- 3D gradient descend for linear regression weights w_0, w_1
 - Step 1: Random initialization $w_0 = 0, w_1 = 1, w_2 = 2$
 - Step 2: descend...
 - $w_0 = w_0 - r * \frac{dL}{dw_0}$
 - $w_1 = w_1 - r * \frac{dL}{dw_1}$
 - $w_2 = w_2 - r * \frac{dL}{dw_2}$
 - Step 3: compute $f(w_0, w_1, w_2)$. Still big?
 - Move to point (w_0, w_1, w_2)
 - Repeat Step 2