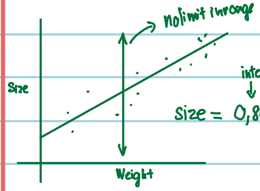
Generalized Linear Model (GLM) — Logistic Regression / Linear Models

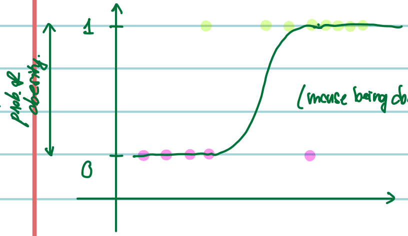
Review on Linear Regression



intercept ↓ slope ↓

size = 0,86 + 0,7 . weight → predicted_size = 0,86 + 0,7 . same_weight → not limit the equation to specific domain and range.

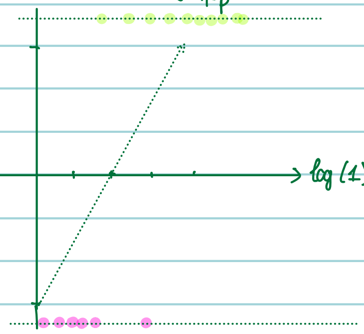
Logistic Regression : y-axis is confined to prob values btw 0 and 1



(mouse being obey)

+) y-axis : probability of obesity (0,1)
↓
log (odds of obesity) : $(-\infty, \infty)$

$\log(\text{odd of obesity}) = \log\left(\frac{p}{1-p}\right)$ → p: prob of mouse being obese. comes. to value $y \in [0,1]$ // $p = 0.5 \Rightarrow \log(\text{odd}) = 0$

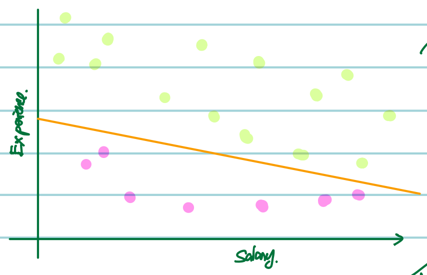


$\log(1) = 0$ $(p = 0.5)$

Lecture: Linear Regression:

Example: Classification Problem
- Input: Dataset
- Output: One of specify class

(Salary - Experience) → allow loan or refuse.



we assume there is a line to devide to 2 area (refuse and allow)

→ Let $x_i^{(1)}$ be salary, $x_i^{(2)}$ be exp } profile i

→ Prediction model is defined as follows:

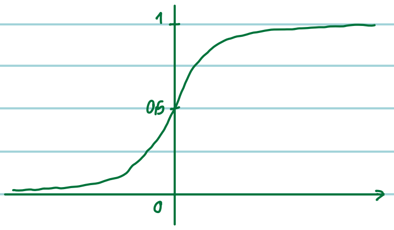
$\hat{y}_i = \omega_1 * x_i^{(1)} + \omega_2 * x_i^{(2)} + \omega_0$ — If $\hat{y}_i > 0.5$ → accept, $\hat{y}_i < 0.5$ → refuse || estimate a profile to be allow for loan or refuse.

if we find $(\omega_1, \omega_2, \omega_0)$ → we can predict label of any value

transform to sigmoid

→ Output: Estimate loan prob > threshold

⊛ Sigmoid: $\sigma(x) = \frac{1}{1+e^{-x}}$
- (continuos with value (0,1))
- Derivative at every point (for applying gradient descent)

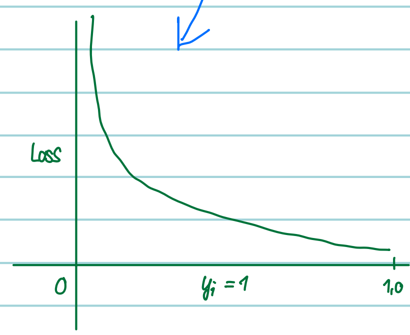


→ Estimate loan prob: $\hat{y}_i = \sigma(\hat{y}_i) = \sigma(\omega_1 * x_i^{(1)} + \omega_2 * x_i^{(2)} + \omega_0)$
(using sigmoid)

$\hat{y}_i = \frac{1}{1+e^{-(\dots)}}$

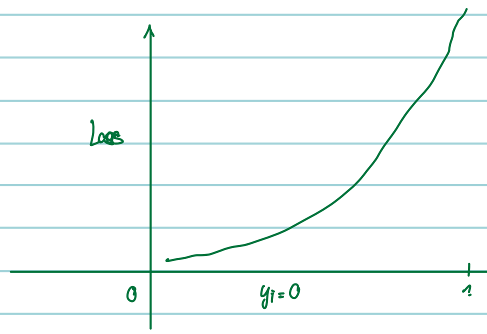
- case allow: $p(x^{(i)} = 1) = \hat{y}_i$ (profile allow)
- refuse: $p(x^{(i)} = 0) = 1 - \hat{y}_i$ (profile refuse)

→ Loss ~~function~~ value: $L_i = \begin{cases} -\log(\hat{y}_i) & \text{if } y_i = 1 \\ -\log(1-\hat{y}_i) & \text{if } y_i = 0 \end{cases}$



→ Combine: $L_i = -(\underbrace{y_i * \log(\hat{y}_i)}_{y_i = 1} + \underbrace{(1-y_i) * \log(1-\hat{y}_i)}_{y_i = 0})$

for 1 datapoint (1 profile)



for all datapoint: $J = \frac{-1}{N} * \sum_{i=1}^{N} (y_i * \log(\hat{y}_i) + (1-y_i) * \log(1-\hat{y}_i))$ (sigmoid inside)

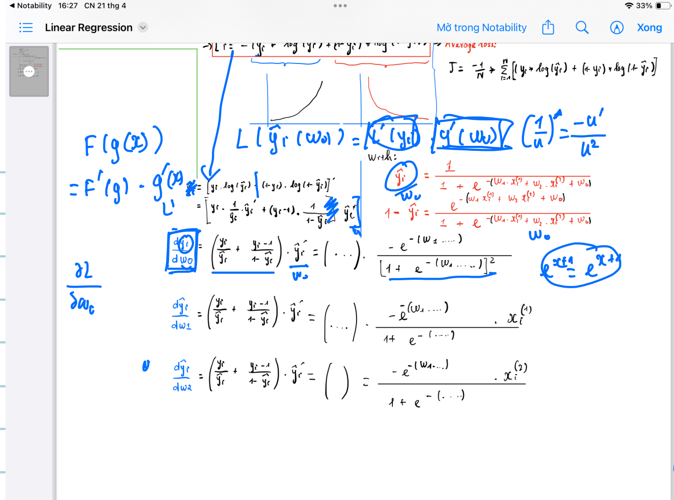
Binary cross entropy loss. (Not MSE)
- 2 diff possible case output — How far it is from 1 or from 0
- only one output

Apply gradient descent algorithm to find $(\omega_0, \omega_1, \omega_2)$ which minimize J

after train model.

give a test (x_{new}, y_{new}) ⟶ $\hat{y}_{predicted}$, if $\hat{y}_{predicted}$ > thres → accept, else: refuse.

Loss



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

$$\hat{y}_i = \frac{1}{1+\exp[-(\omega_1 x_i^{(1)} + \omega_2 x_i^{(2)} + \omega_0)]}$$

$$1-\hat{y}_i = \frac{\exp[-(\omega_1 x_i^{(1)} + \omega_2 x_i^{(2)} + \omega_0)]}{1+\exp[-(\omega_1 x_i^{(1)} + \omega_2 x_i^{(2)} + \omega_0)]}$$

$$\frac{\partial L_i}{\partial \omega_0} = -y_i \underbrace{\frac{\partial}{\partial\omega_0}\log(\hat y_i)}_{A} - (1-y_i)\frac{\partial}{\partial \omega_0}\log(1-\hat y_i)$$

$$\frac{\partial A}{\partial\omega_0} = -y_i \cdot \frac{\partial}{\partial \omega_0}\log\left(\frac{1}{1+\exp[-(\omega_1 x_i^{(1)} + \omega_2 x_i^{(2)} + \omega_0)]}\right)$$

$$= -y_i \cdot \left(1+\exp[-(\omega_1 x_i^{(1)}+\omega_2 x_i^{(2)}+\omega_0)]\right) \cdot \frac{\partial}{\partial\omega_0}\left(1+\exp[-(\omega_1 x_i^{(1)}+\omega_2 x_i^{(2)}+\omega_0)]\right)^{-1}$$

$$u^{-1} = \frac{-u'}{u^2}$$

$$= -y_i \cdot \left(1+\exp[-(\dots)]\right) \cdot \frac{1}{[1+\exp[-(\dots)]]^2}\cdot \frac{\partial}{\partial\omega_0}\exp[-(\omega_1 x_i^{(1)}+\omega_2 x_i^{(2)}+\omega_0)]$$

$$(e^u)' = u' \cdot e^u$$

$$= \frac{y_i}{1+\exp[(\dots)]}\cdot(-1)\cdot\exp[-(\dots)] = \frac{-y_i \cdot \exp[-(\dots)]}{1+\exp[(\dots)]}$$

$$\frac{\partial L}{\partial \omega_0} = \frac{\partial L}{\partial \hat y_i}\cdot\frac{\partial \hat y_i}{\partial \omega_0}$$

$$\frac{\partial L}{\partial \hat y_i} = -\left[y_i \cdot\frac{1}{\hat y_i} + (1-y_i)\frac{1}{(1-\hat y_i)}\right]$$

$$\frac{\partial \hat y_i}{\partial\omega_0} = \frac{\partial}{\partial\omega_0}\left(1+\exp[-(\omega_1 x_i^{(1)} + \omega_2 x_i^{(2)}+\omega_0)]\right)^{-1}$$

$$= \frac{1}{[1+\exp[-(\dots)]]^2}\cdot\frac{\partial}{\partial\omega_0}\exp[-(\omega_1 x_i^{(1)}+\omega_2 x_i^{(2)}+\omega_0)]$$

$$= \frac{1}{[1+\exp[(\dots)]]^2}\cdot(-1)\cdot\exp[-(\dots)]$$

$$\frac{\partial L}{\partial \hat y_i} = -\left[\frac{y_i - y_i\hat y_i + \hat y_i - y_i \hat y_i}{\hat y_i(1-\hat y_i)}\right]$$

$$= \frac{y_i+\hat y_i - 2y_i\hat y_i}{\hat y_i(1-\hat y_i)}$$

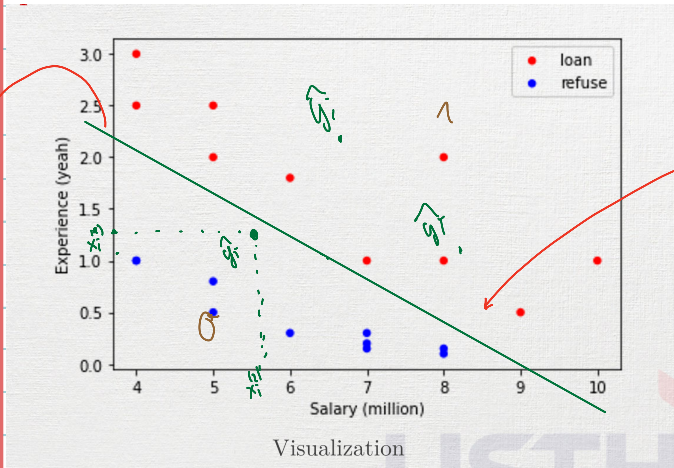
$$\frac{\partial L}{\partial\omega_0} = \frac{y_i+\hat y_i-2y_i\hat y_i}{\hat y_i(1-\hat y_i)} \cdot \frac{1}{(\hat y_i)^2}\cdot\left[-\exp[-(\omega_1 x_i^{(1)}+\omega_2 x_i^{(1)}+\omega_0)]\right]$$

$$\frac{\partial L}{\partial\omega_1} = \frac{y_i+\hat y_i-2y_i\hat y_i}{\hat y_i(1-\hat y_i)} \cdot \frac{-x_i^{(1)}}{(\hat y_i)^2}\cdot\left[-\exp[-(\omega_1 x_i^{(1)}+\omega_2 x_i^{(2)}+\omega_0)]\right]$$

$$\frac{\partial L}{\partial\omega_2} = \frac{y_i+\hat y_i-2y_i\hat y_i}{\hat y_i(1-\hat y_i)} \cdot \frac{-x_i^{(2)}}{(\hat y_i)^2}\cdot\left[-\exp[-(\omega_1 x_i^{(1)}+\omega_2 x_i^{(2)}+\omega_0)]\right]$$

Example :

attribute $\begin{cases} \text{salary } x_i^{(1)} \\ \text{working time } x_i^{(2)} \end{cases}$ $\Longrightarrow$ Label: Loan / Not Loan



Visualization

$0 = w_1^{*} x_i^{(1)} + w_2^{*} x_i^{(2)} + w_0$

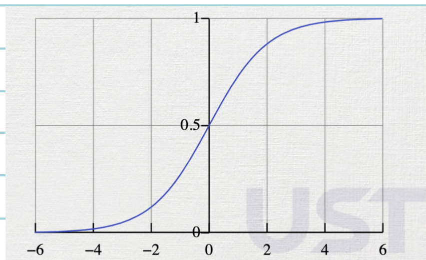
⊛ Profile i : $x_i^{(1)}, x_i^{(2)}$

⊛ Prediction model is defined as follows:

$z \;\; \hat{y}_i = w_1 * x_i^{(1)} + w_2 * x_i^{(2)} + w_0$ → estimate a new profile should be loaned or not

- estimate loan probability >= threshold → loaned (1)
- < threshold → not loaned (0)

linear → means to probability

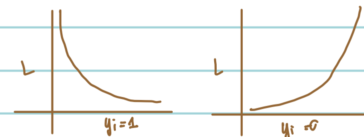


Sigmoid and model definition : $\sigma(\hat{y}_i) = \sigma(w_1 * x_i^{(1)} + w_2 * x_i^{(2)} + w_0) = \hat{y}_i$ → possibility to get label

$$= \frac{1}{1 + e^{-(w_1 \cdot x_i^1 + w_2 \cdot x_i^2 + w_0)}}$$

Loss function : $L_i = \begin{cases} -\log(\hat{y}_i) & \text{if } y_i = 1 \\ -\log(1-\hat{y}_i) & \text{if } y_i = 0 \end{cases}$ by definition

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

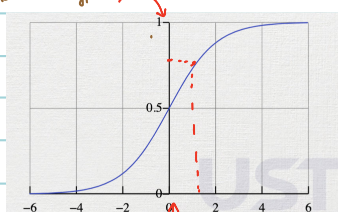


Or for all data point : $J = -\frac{1}{N} * \sum_{i=1}^{n} (y_i \cdot \log(\hat{y}_i) + (1-y_i)\cdot \log(1-\hat{y}_i))$

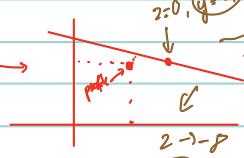
$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^{N} [\, y_i \cdot \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\,]$$

predicted result

actual label $y_i = 1$ actual label $1 - y_i = 0$



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



$z=0$, $\hat{y}=0.5$ $= \frac{1}{1+e^{-z}}$ $(y=1)$ $z \to +\infty$

$z=+\infty$; $\hat{y} = 1$

$z=-\infty$, $\hat{y} = 0$

$z \to -\infty$

$y=0$

$e^{+\infty}$

$\frac{1}{1 + \frac{1}{e^{z}}}$