Logistic Regression

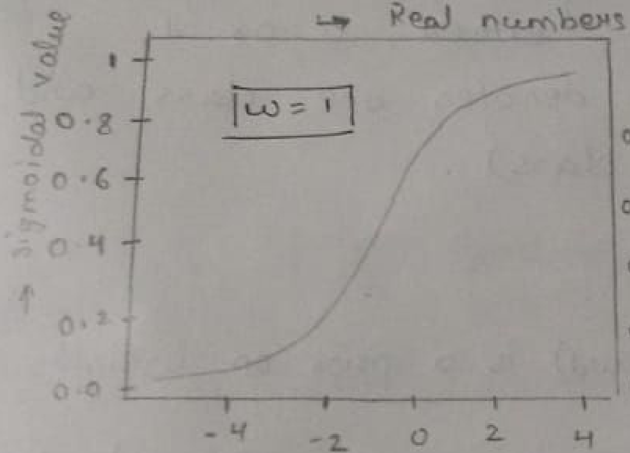
Q1)

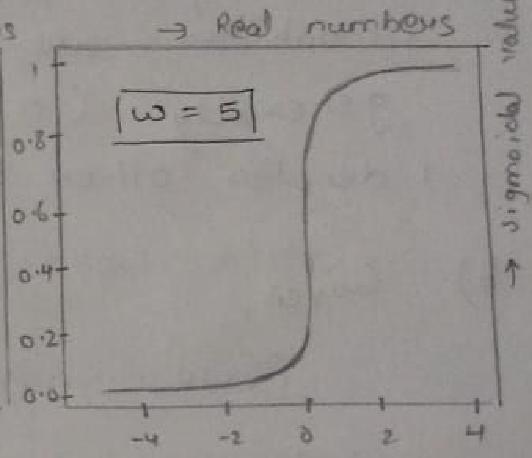
a) Given,

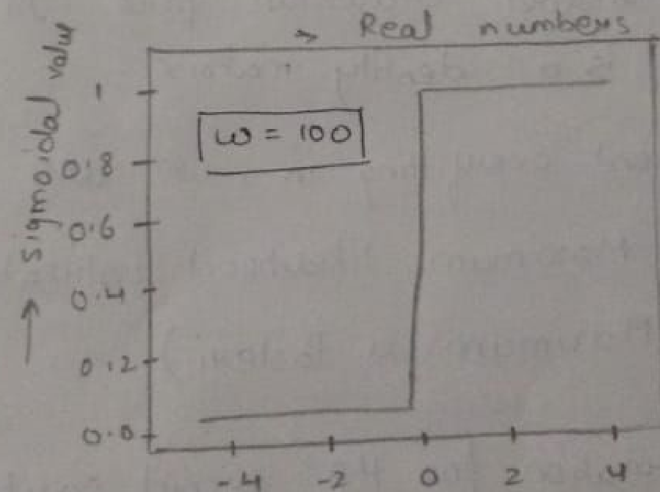
$$\omega = \{1, 5, 100\}$$

To plot the sigmoid function $1/(1+e^{-\omega x})$ vs $x \in R$ or ω

For $\omega = 1$







Observations from the above graphs :-

- Like you can see that as the weight value is increasing from 1 to 100 the line in between become almost vertical or at 90° to the x-axis.

→ As ~~#~~ with higher weight values, when we change the input values slightly then also overall change will be high or more.

→ Thus high weights leads to overfitting as ~~hig~~ because of high variation in output

→ with higher weights, the output probabilities are either 0 or 1 generally. (0 denotes one class and 1 denotes other class).

b) Given,

$P(w_0, \ldots, w_d)$ is a prior on weights.

→ Standard Gaussian prior $\mathcal{N}(0, I)$ where I is a identity matrix.

→ To prevent overfitting in, we are replacing MLE (Maximum likelihood Estimation) with MAP (Maximum a Posteri)

→ The derivation for the second part of the question proceeds as follows,

Here, $w = [w_0, \ldots, w_d]^T$

Log conditional posterior is

$$L(\omega) = \log\left(p(\omega) \prod_{j=1}^{n} p(y^j \mid x^j, \omega)\right) \quad \text{— (1)}$$

$$p(\omega) = \prod_{i=0}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\omega_i^2}{2}\right) \quad \text{— (2)}$$

Now putting this $p(\omega)$ value of equation (2) in the equation (1)

the MAP estimate is

$$\omega^* = \underset{\omega}{\operatorname{argmax}}\, L(\omega) = \underset{\omega}{\operatorname{arg\,max}} \left[\sum_{j=1}^{n} \log(p(y^j \mid x^j, \omega)) - \sum_i \frac{\omega_i^2}{2}\right]$$

Now applying gradient ascent update rule

$$\omega_i^{t+1} \leftarrow \omega_i^{t} + \eta \left.\frac{\partial L(\omega)}{\partial \omega_i}\right|_t \quad \text{— (3)}$$

ω_i^{t+1} ↑ weight at time $t+1$

ω_i^{t} ↑ weight at time t

η → step size

Here,

In equation (3)

$$\frac{\partial L(\omega)}{\partial \omega_i} = \frac{\partial}{\partial \omega_i} \log p(\omega) + \frac{\partial}{\partial \omega_i} \log\left(\prod_{j=1}^{n} p(y^j \mid x^j, \omega)\right)$$

Now putting $p(\omega)$ value from equation (2) into equation (3)

$$\frac{\partial \log p(\omega)}{\partial \omega_i} = \frac{\partial}{\omega_i} \log\left(\prod_{i=0}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\omega_i^2}{2}\right)\right)$$

$$= \frac{\delta}{\delta w_i}\left(\sum_{i=0}^{d} \frac{1}{\sqrt{2\pi}} + \frac{\delta}{\delta w_i} \log \exp\left(-\frac{w_i^2}{2}\right)\right)$$

$$= \frac{\delta}{\delta w_i}\left(\sum_{i=0}^{d} \frac{1}{\sqrt{2\pi}}\right)^{0} + \frac{\delta}{\delta w_i}\left(-\frac{w_i^2}{2}\right)$$

$$= 0 + -2 \times \frac{1}{2} \times w_i$$

$$= -w_i$$

$$\Rightarrow \frac{\delta}{\delta w_i} \log(P(w)) = -w_i$$

And, $\frac{\delta}{\delta w_i} \log\left(\prod_{j=1}^{n} P(y^j / x^j, w)\right) = \sum_{j=1}^{n} x_i^j \left(y^j - P(y=1 / x^j, w^{(t)})\right)$

Final update rule :-

$$\boxed{w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta\left(-w_i^{(t)} + \sum_{j=1}^{n} x_i^j \left(y^j - P(Y=1 / x^j, w^{(t)})\right)\right)}$$

c) Extending logistic regression to multi - class (k class labels) :-

Since sum of all probabilities must sum to 1, we should have

$$P(Y = y_K / X) = 1 - \sum_{k=1}^{K-1} P(Y = y_k / X)$$

Defining logistic regression as we defined in binary classification,

$$P(Y = y_k | X) = \frac{1}{1 + \sum_{k=1}^{K-1} \exp -(w_{k0} + \sum_{i=1}^{d} w_{ki} X_i)}$$

and for $k = 1, \dots K-1$

$$\boxed{P(Y = y_k | X) = \frac{\exp(w_{k0} + \sum_{i=1}^{d} w_{ki} X_i)}{1 + \sum_{k=1}^{K-1} \exp(w_{k0} + \sum_{i=1}^{d} w_{ki} X_i)}}$$

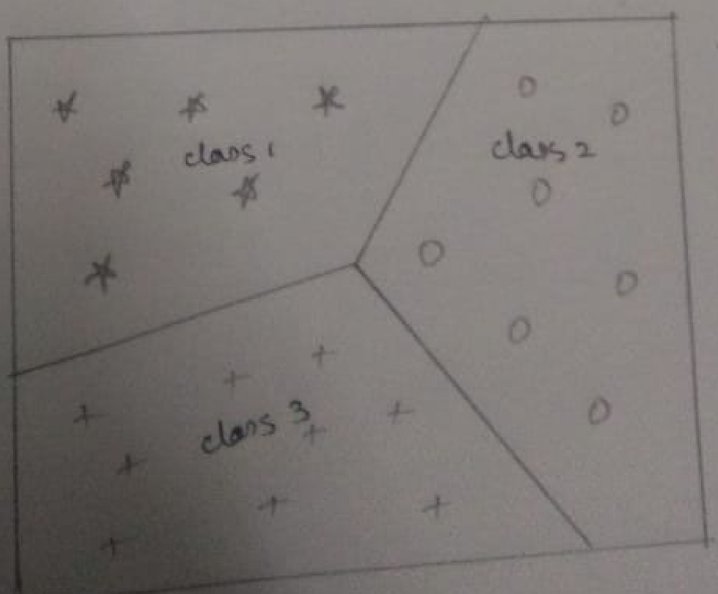
Classification Rule :-

$$y = y_k$$

where $k = \underset{k \in \{1, \dots K\}}{\arg\max} P(Y = y_k | X)$

thus, it picks the class which has highest probability.

d) 3 - class logistic regression Decision boundary



In multi-class logistic regression, the decision boundary is decided like we do in one vs all SVM, here also we find like that only.

→ Decision boundary between each pair of class will be linear, Hence the overall decision boundary will be piece-wise linear.