LOGISTIC REGRESSION

Logistic Regression is used when the outcome is categorical (usually binary). Unlike Linear Regression, Logistic Regression try to find the best sigmoid function to classify examples.

Motivation Example:

$$P(y=1 \mid x) = h(w^T x) = \frac{1}{1+\exp(-w^T x)}$$

Dataset: $\{(x_1, y_1), \dots, (x_m, y_m)\}$

How to find w? We use Maximum Likelihood

- Likelihood function : $L(w \mid data) = \prod_{i=1}^{m} (\underbrace{h(w^T x)}_{P(y=1 \mid x)})^{y_i} (\underbrace{1 - h(w^T x)}_{P(y=0 \mid x)})^{(1-y_i)}$

- Log Likelihood function:

$$\log L(w \mid data) = \sum_{i=1}^{m} y_i \log(h(w^T x_i)) + (1-y_i) \log(1 - h(w^T x_i))$$

$$= \sum_{i=1}^{m} y_i \log\left(\frac{1}{1+\exp(-w^T x_i)}\right) + (1-y_i) \log\left(\frac{\exp(-w^T x_i)}{1+\exp(-w^T x_i)}\right)$$

$$= \sum_{i=1}^{m} y_i \cdot (-\log(1+\exp(-w^T x_i))) + (1-y_i)\left[\log(\exp(-w^T x_i)) - \log(1+\exp(-w^T x_i))\right]$$

$$= \sum_{i=1}^{m} y_i (-\log(1+\exp(-w^T x_i))) + (1-y_i)(-w^T x_i) - (1-y_i)\log(1+\exp(-w^T x_i))$$

$$= \sum_{i=1}^{m} \left[(1-y_i)(-w^T x_i) - \log(1+\exp(-w^T x_i)) \right]$$

Goal:

$$\boxed{\max_{w} \sum_{i=1}^{m} \left[(1-y_i)(-w^T x_i) - \log(1+\exp(-w^T x_i)) \right]} \quad y \in \{1, 0\}$$

- Since the second term is log, you won't find a nice approx by setting gradient $= 0$ and solve for w
- So we use gradient "ascent" to slowly converge to w

$$\nabla \log L(w \mid data) = \sum_{i=1}^{m} (1-y_i)(-x_i) - \frac{\exp(-w^T x_i)}{1+\exp(w^T x_i)} \cdot (-x_i)$$

$$= \sum_{i=1}^{m} -x_i + y_i x_i + x_i \frac{\exp(-w^T x_i)}{1+\exp(-w^T x_i)}$$

$$= \sum_{i=1}^{m} y_i x_i - x_i \frac{1}{1+\exp(-w^T x_i)}$$

$$= \sum_{i=1}^{m} x_i \left(y_i - \frac{1}{1+\exp(-w^T x_i)} \right)$$

Gradient Update Rule:

$$w_{t+1} = w_t + \eta \nabla \log L(w \mid data)$$
$$= w_t + \eta \left(\sum_{i=1}^m x_i \left(y_i - \underbrace{\frac{1}{1+\exp(-w^T x_i)}}_{} \right) \right)$$

actual label : $\{0;1\}$

$h(w^T x_i)$: chance point x_i be labeled as 1

Some Observations:

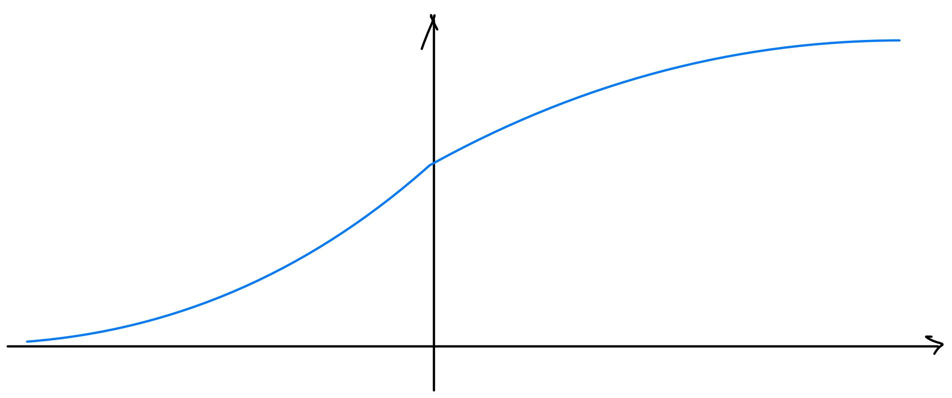
- Gradient is a linear combination of data points :
$$x^T b = \sum_{i=1}^m x_i \beta_i \quad , \quad \beta_i = y_i - \frac{1}{1+\exp(-w^T x_i)}$$
- If $h(w^T x_i)$ is close to 1, then $y_i - h(w^T x_i)$ is close to 0, meaning the guess on point x_i won't contribute much to updating w_{t+1}. Vice versa, if $h(w^T x_i)$ is close to 0, then $y_i - h(w^T x_i)$ is close to 1, meaning the guess on point x_i contribute significantly to updating w_{t+1}. In conclusion, w_{t+1} converges faster on mistakes.

Once w converged:

Calculate $\hat{y} = \text{sign}(\hat{w}^T x)$

why? Since we'll use this for binary classification

Visualize sigmoid function:



Loss function (logistic loss or cross-entropy loss) $y \in \{1, 0\}$

$$\boxed{\ell(h, (x_i, y_i)) = -\left[y_i \log(h(w^T x_i)) + (1-y_i) \log(1 - h(w^T x_i)) \right]}$$

What is the connection between loss function and likelihood function?

Total loss function is the negative of log-likelihood function:

$$\ell(h, (X, y)) = -\log L(w \mid data)$$
$$= -\sum_{i=1}^m \left[(1-y_i)(-w^T x_i) + \log(1 + \exp(-w^T x_i)) \right]$$
$$= -\left[y \log(h(w^T x)) + (1-y) \log(1 - h(w^T x)) \right]$$

Opposite to likelihood, we minimize loss function to find best fit w :

Goal:

$$\boxed{\min_w \sum_{i=1}^{m} \left[(1-y_i)(w^T x_i) + \log(1 + \exp(-w^T x_i)) \right]}$$

With $y \in \{+1, -1\}$ instead of $\{1, 0\}$

So far we have discuss the case of label $y \in \{1, 0\}$, what if label $y \in \{+1, -1\}$?

Loss function, $y \in \{+1, -1\}$

$$\boxed{\ell(h, (x_i, y_i)) = \log(1 + \exp(-y_i w^T x_i))}$$

Why? Converting between $y \in \{0, 1\}$ and $y \in \{\pm 1\}$

- We know that loss function for $y \in \{1, 0\}$ is:

$$Loss_{\{0,1\}} = -\left[y_i \log(h(w^T x_i)) + (1-y_i) \log(1 - h(w^T x_i)) \right] \quad (1)$$

- We also know that:

$$Loss_{\{+1,-1\}} = \log(1 + \exp(-y_i' w^T x_i)) \quad (2)$$

Let $y_i' = 2y_i - 1$, this maps : $\begin{cases} y_i = 1 \iff y_i' = +1 \\ y_i = 0 \iff y_i' = -1 \end{cases}$

- Now consider how the loss function transformed when $y \in \{+1, -1\}$

Recall that $h(w^T x) = \dfrac{1}{1 + \exp(-w^T x_i)}$

  - Case 1: $y_i = 0$, $y_i' = -1$

  (1) becomes : $Loss_{\{0,1\}} = -\log(1 - h(w^T x_i))$

  (2) becomes : $Loss_{\{+1,-1\}} = \log(1 + \exp(w^T x_i))$

  With some arithmetic, we can prove that :

$$-\log(1 - h(w^T x_i)) = \log(1 + \exp(w^T x_i))$$

  - Case 2: $y_i = 1$, $y_i' = +1$

  (1) becomes : $Loss_{\{0,1\}} = -\log(h(w^T x_i))$

  (2) becomes : $Loss_{\{+1,-1\}} = \log(1 + \exp(-w^T x_i))$

  With some arithmetic, we can prove that:

$$-\log(h(w^T x_i)) = \log(1 + \exp(-w^T x_i))$$

So conclude that $Loss_{\{0,1\}}$ and $Loss_{\{+1,-1\}}$ are equivalent