Logistic regression for binary classification

Given train set $T = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$,

where m is the number of data, n is the number of features.

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \qquad W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix},$$

$$X \in \mathbb{R}^{m \times (n+1)} \qquad W \in \mathbb{R}^{n+1}$$

then we can write the model as

$$h_w(X) = \frac{1}{1 + \exp^{-XW}} \quad \text{where } h_w(X) = P(Y=1 \mid X; W)$$

Naturally, we use MLE to estimate the parameters W

(What is written above is the vectorized version, if we look it in sample-wise perspective):

$$L(W) = P(Y \mid X; W) = \prod_{i=1}^{m} P(y_i \mid \vec{x}^{(i)}; w)$$

$$= \prod_{i=1}^{m} P(y_i = 1 \mid \vec{x}^{(i)}; w)^{y^{(i)}} \cdot P(y_i = 0 \mid \vec{x}^{(i)}; w)^{(1-y^{(i)})}$$

$$= \prod_{i=1}^{m} h_w(X^{(i)})^{y^{(i)}} \left[1 - h_w(X^{(i)})\right]^{(1-y^{(i)})}$$

as log is a strictly monotonically increasing, we take log at both sides

$$\log L(w) = \sum_{i=1}^{m}\left[y^{(i)}\log h_w(X^{(i)}) + (1-y^{(i)})\log[1-h_w(X^{(i)})]\right]$$

Then we just need to find W to maximize the above equation

Gradient ascend method:

$$\frac{\partial \log L(w)}{\partial w_j} = \sum_{i=1}^{m}(y^{(i)} - h_w(X^{(i)}))X_j$$

$$W_j := W_j + \frac{\alpha}{m}\sum_{i=1}^{m}(y^{(i)} - h_w(X^{(i)}))X_j$$

However, this implementation is higly inefficent. we are going to adapt vectorized version.

$h_w(X)$ is of dimension $(m, 1)$, one sample per row,

we denote $Y \log h_w(X)$ as a element-wise vector product.

then $\log L(w) = \text{np.sum}\left[Y \log h_w(X) + (1-Y)\log(1-h_w(X))\right]$

$$W = W + \frac{\alpha}{m}X^T(Y - h_w(X))$$

We can further prove that the $\log L(w)$ is concex by show its Hessian is negative semi-definite

Or equivalent $J(w) = -\log L(w)$ is convex, its Hessian is positive semi-definite.

In this case, the problem does not have local optimal other than the global one

$$\nabla J(w) = \frac{1}{m} X^T (h_w(X) - Y)$$

$$H = \frac{1}{m} X^T [h_w(X)(1 - h_w(X))] X$$

Also, we can use Newton method in this secnario.