Formulation

Logistic regression is a supervised, discriminative, binary classification, linear model. In discriminative model, $p(y|x)$ is learned.

$$p(y=+1|x) = \frac{1}{1+e^{-(w^Tx+b)}} = \sigma(f(x))$$
$$\sigma(u) = \frac{1}{1+e^{-u}}, f(x) = w^Tx + b$$

The derivation of the above formula is like:

$$\begin{aligned}
p(y=+1|x) &= \frac{p(x|y=+1)p(y=+1)}{p(x)} \\
&= \frac{p(x|y=+1)p(y=+1)}{p(x|y=+1)p(y=+1)+p(x|y=-1)p(y=-1)} \\
&= \frac{1}{1+\frac{p(x|y=-1)p(y=-1)}{p(x|y=+1)p(y=+1)}} \\
&= \frac{1}{1+g(x)} = \sigma(f(x))
\end{aligned}$$

$$\begin{aligned}
g(x) &= \frac{p(x|y=-1)p(y=-1)}{p(x|y=+1)p(y=+1)} \\
&= e^{f(x)}
\end{aligned}$$

When $f(x) \geq 0, y=+1$; when $f(x) < 0, y=-1$. When x^1, x^2 both have the same labels, i.e., $f(x^1), f(x^2) \geq 0$, their convex combination $\alpha x^1 + (1-\alpha)x^2, 0 \leq \alpha \leq 1$ also has the same label as x^1, x^2. (LR is a linear model in that its decision boundary is a linear hyperplane).

Training

LR maximizes the following likelihood term:

$$\begin{aligned}
L(w,b) &= \prod_i p(y=y^i|x^i;w,b) \\
&= \prod_i \sigma(f(x^i;w,b))
\end{aligned}$$

Its log-likelihood is:

$$l(w,b) = \sum_i \log \sigma(f(x^i; w, b))$$

Maximizing the above is equivalent to minimizing the foll

$$J(w,b) = -\sum_i \log \sigma(f(x^i; w, b))$$

$$\arg\max_{w,b} l(w,b) = \arg\min_{w,b} J(w,b)$$

To take derivative, note that

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

$$\sigma'(u) = \frac{(e^{-u})}{(1+e^{-u})^2} = \sigma(u)(1-\sigma(u))$$

$$\text{Gradient descent:}$$

$$\frac{\partial J}{\partial w} = -\sum_i (1 - \frac{1}{1+e^{-(w^T x^i + b)}}) x^i$$

$$\frac{\partial J}{\partial b} = -\sum_i (1 - \frac{1}{1+e^{-(w^T x^i + b)}})$$

$$\text{Stochastic gradient descent:}$$

$$\frac{\partial J}{\partial w} = -\sum_{i \in I} (1 - \frac{1}{1+e^{-(w^T x^i + b)}}) x^i$$

$$\frac{\partial J}{\partial b} = -\sum_{i \in I} (1 - \frac{1}{1+e^{-(w^T x^i + b)}})$$

In practice, people add a regularization term to the loss:

$$J'(w,b) = J(w,b) + c||w^T, b||_2^2$$