Chapter 4

Logistic Regression

4.1 Definition

We know that regression is for predicting real-valued output Y, while classification is for predicting (finite) discrete-valued Y. But is there a way to connect regression to classification? Can we predict the "probability" of a class label? The answer is generally yes, but we have to keep in mind the constraint that the probability value should lie in $[0, 1]$.

Definition 4: (Logistic Regression)
Assume the following functional form for $P(Y \mid X)$:

$$P(Y = 1 \mid X) = \frac{1}{1 + \exp(-(w_0 + \sum_i w_i X_i))}, \tag{4.1}$$

$$P(Y = 0 \mid X) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)}. \tag{4.2}$$

In essence, logistic regression means applying the logistic function $\sigma(z) = \frac{1}{1+\exp(-z)}$ to a linear function of the data. However, note that it is still a linear classifier.

Diving in the Math 6 - Logistic Regression as linear classifier
Note that $P(Y = 1 \mid X)$ can be rewritten as

$$P(Y = 1 \mid X) = \frac{\exp(w_0 + \sum_i w_i X_i)}{1 + \exp(w_0 + \sum_i w_i X_i)}.$$

We would assign label 1 if $P(Y = 1 \mid X) > P(Y = 0 \mid X)$, which is equivalent to

$$\exp(w_0 + \sum_i w_i X_i) > 1 \Leftrightarrow w_0 + \sum_i w_i X_i > 0.$$

Similarly, we would assign label 0 if $P(Y = 1 \mid X) < P(Y = 0 \mid X)$, which is equivalent to

$$\exp(w_0 + \sum_i w_i X_i) < 1 \Leftrightarrow w_0 + \sum_i w_i X_i < 0.$$

In other words, the decision boundary is the line $w_0 + \sum_i w_i X_i$, which is linear.

4.2 Training logistic regression

Given training data $\{(x_i, y_i)\}_{i=1}^n$ where the input has d features, we want to learn the parameters $w_0, w_1, \ldots, w_d$. We can do so by MCLE:

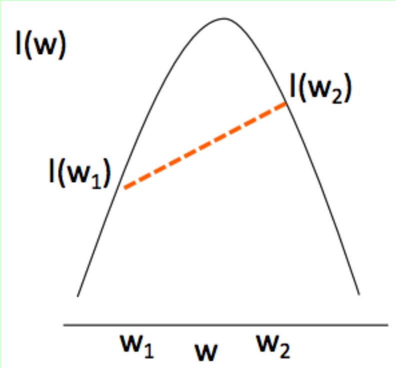
$$\hat{w}_{MCLE} = \arg\max_w \prod_{i=1}^n P(y^{(i)} \mid x^{(i)}, w). \tag{4.3}$$

Note the Discriminative philosophy: *don't waste effort learning $P(X)$, focus on $P(Y \mid X)$ - that's all that matters for classification!* Using (4.1) and (4.2), we can then compute the log-likelihood:

$$\begin{aligned} l(w) &= \ln\left(\prod_{i=1}^n P(y^{(i)} \mid x^{(i)}, w)\right) \\ &= \sum_{i=1}^n \left[y^{(i)}(w_0 + \sum_{j=1}^d w_i x_j^{(i)}) - \ln(1 + \exp(w_0 + \sum_{j=1}^d w_i x_j^{(i)}))\right]. \end{aligned} \tag{4.4}$$

There is no closed-form solution to maximize $l(w)$, but we note that it is a concave function.

Definition 5: (Concave function)
A function $l(w)$ is called *concave* if the line joining two points $l(w_1), l(w_2)$ on the function does not lie above the function on the interval $[w_1, w_2]$.



Equivalently, a function $l(w)$ is *concave* on $[w_1, w_2]$ if

$$l(tx_1 + (1-t)x_2) \geq tl(x_1) + (1-t)l(x_2)$$

for all $x_1, x_2 \in [w_1, w_2]$ and $t \in [0, 1]$. If the sign is reversed, l is a *convex* function.

Diving in the Math 7 - Log likelihood of logistic regression is concave
For convenience we denote $x_0^{(i)} = 1$, so that $w_0 + \sum_{i=j}^{d} w_i x_j^{(i)} = w^T x^{(i)}$.
We first note the following lemmas:

1. If f is convex then $-f$ is concave and vice versa.
2. A linear combination of n convex (concave) functions $f_1, f_2, \ldots, f_n$ with nonnegative coefficients is convex (concave).
3. Another property of twice differentiable convex function is that the second derivative is nonnegative. Using this property, we can see that $f(x) = \log(1 + \exp x)$ is convex.
4. If f and g are both convex, twice differentiable and g is non-decreasing, then $g \circ f$ is convex.

Now we rewrite $l(w)$ as follows:

$$\begin{aligned} l(w) &= \sum_{i=1}^{n} y^{(i)} w^T x^{(i)} - \log(1 + \exp(w^T x^{(i)})) \\ &= \sum_{i=1}^{n} y^{(i)} w^T x^{(i)} - \sum_{i=1}^{n} \log(1 + \exp(w^T x^{(i)})) \\ &= \sum_{i=1}^{n} y^{(i)} f_i(w) - \sum_{i=1}^{n} g(f_i(w)), \end{aligned}$$

where $f_i(w) = w^T x^{(i)}$ and $g(z) = \log(1 + \exp z)$.
$f_i(w)$ is of the form $Ax + b$ where $A = x^{(i)}$ and $b = 0$, which means it's affine (i.e., both concave and convex). We also know that $g(z)$ is convex, and it's easy to see g is non-decreasing. This means $g(f_i(w))$ is convex, or equivalently, $-g(f_i(w))$ is concave.
To sum up, we can express $l(w)$ as

$$l(w) = \underbrace{\sum_{i=1}^{n} y^{(i)} f_i(w)}_{\text{concave}} + \underbrace{\sum_{i=1}^{n} -g(f_i(w))}_{\text{concave}},$$

hence $l(w)$ is concave.

As such, it can be optimized by the gradient ascent algorthim.

Algorithm 7: (Gradient ascent algorithm)
Initialize: Pick w at random.
Gradient:

$$\nabla_w E(w) = \left(\frac{\partial E(w)}{\partial w_0}, \frac{\partial E(w)}{\partial w_1}, \ldots, \frac{\partial E(w)}{\partial w_d} \right).$$

Update:

$$\begin{aligned} \Delta w &= \eta \nabla_w E(w) \\ w_t^{(t+1)} &\leftarrow w_i^{(t)} + \eta \frac{\partial E(w)}{\partial w_i}, \end{aligned}$$

where $\eta > 0$ is the learning rate.

In this case our likelihood function is specified in (4.4), so we have the following steps for training logistic regression:

Algorithm 8: (Gradient ascent algorithm for logistic regression)
Initialize: Pick w at random and a learning rate η.
Update:

- Set an $\epsilon > 0$ and denote

$$\hat{P}(y^{(i)} = 1 \mid x^{(i)}, w^{(t)}) = \frac{\exp(w_0^{(t)} + \sum_{j=1}^{d} w_j^{(t)} x_j^{(i)})}{1 + \exp(w_0^{(t)} + \sum_{j=1}^{d} w_j^{(t)} x_j^{(i)})}.$$

- Iterate until $|w_0^{(t+1)} - w_0^{(t)}| < \epsilon$:

$$w_0^{(t+1)} \leftarrow w_0^{(t)} + \eta \sum_{i=1}^{n} \left[y^{(i)} - \hat{P}(y^{(i)} = 1 \mid x^{(i)}, w^{(t)}) \right].$$

- For $k = 1, \ldots, d$, iterate until $|w_k^{(t+1)} - w_k^{(t)}| < \epsilon$:

$$w_k^{(t+1)} \leftarrow w_k^{(t)} + \eta \sum_{i=1}^{n} x_j^{(i)} \left[y^{(i)} - \hat{P}(y^{(i)} = 1 \mid x^{(i)}, w^{(t)}) \right].$$