Logistic Regression

compiled by Alvin Wan from Professor Jitendra Malik's lecture

1 Univariate Derivation for Logistic Regression

Let us consider the single-variable case, when

$$\Pr(x|C_1) \sim \mathcal{N}(\mu_1, \sigma^2), P(C_1) = \pi_1$$
$$\Pr(x|C_2) \sim \mathcal{N}(\mu_2, \sigma^2), P(C_2) = \pi_2$$

We find that

$$\Pr(C_1|x) = \frac{1}{1 + \exp(-z)}$$

where $z = \beta x + \gamma$, with parameters β and γ. We won't prove the univariate case for $\Pr(C_1|x)$ here but instead, we'll prove it for the more general multivariate case below.

2 Multivariate Derivation for the Logistic Regression

2.1 Logit

Before we begin the derivation, here is the reason for the so-called "logit" expression. Begin with some probability p.

$$p$$

However, a linear model for probability doesn't make sense, since we only have $0 \leq p \leq 1$ for any probability distribution. Now, take the odds, which are the probability of success over probability of failure.

$$\frac{p}{1-p}$$

This is better, because our range of values is now lower bounded by 0. However, this isn't symmetric. So we take the log of this value, to get the **log-odds** or the **logit**.

$$\text{logit}(p) = \log \frac{p}{1-p}$$

This is the reason why we consider the logit expression for logistic regression. We will now begin deriving our result for $\Pr(C_1|x)$.

2.2 Computing Logit

We'll begin by applying Bayes' rule.

$$\Pr(C_1|X) = \frac{\Pr(X|C_1)\Pr(C_1)}{\Pr(X)}$$

Then, take the logit of this probability, so that we have the following.

$$\text{logit}(\Pr(C_1|X)) = \log \frac{\Pr(C_1|X)}{1-\Pr(C_1|X)}$$

Note that we're only considering two classes C_1 and C_2, so $\Pr(C_1|X) + \Pr(C_2|X) = 1$. We can thus substitute $1 - \Pr(C_1|X)$ for $\Pr(C_2|X)$

$$\begin{aligned}
&= \log \frac{\Pr(C_1|X)}{\Pr(C_2|X)} \\
&= \log \frac{\Pr(X|C_1)\Pr(C_1)/\Pr(X)}{\Pr(X|C_2)\Pr(C_2)/\Pr(X)} \\
&= \log \frac{\Pr(X|C_1)\Pr(C_1)}{\Pr(X|C_2)\Pr(C_2)} \\
&= \log \frac{\Pr(X|C_1)}{\Pr(X|C_2)} + \log \frac{\Pr(C_1)}{\Pr(C_2)}
\end{aligned}$$

Let us now plug in the gaussian pdf. We find that the preceding coefficients cancel, since both conditional probability densities share the same variance. Since we take the log of these quantities, we are left with only the terms in the exponent.

$$
\begin{aligned}
&= -\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) + \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + \log\frac{\pi_2}{\pi_1} \\
&= -\frac{1}{2}x^T\Sigma^{-1}x - \mu_2^T\Sigma^{-1}x - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \frac{1}{2}x^T\Sigma^{-1}x - \mu_1^T\Sigma^{-1}x + \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log\frac{\pi_2}{\pi_1}
\end{aligned}
$$

We note that the $\frac{1}{2}x^T\Sigma^{-1}x$ terms cancel out.

$$
= -\mu_2^T\Sigma^{-1}x - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \mu_1^T\Sigma^{-1}x - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log\frac{\pi_2}{\pi_1}
$$

We can rearrange the terms to get the following expression.

$$
\begin{aligned}
&= -\mu_2^T\Sigma^{-1}x - \mu_1^T\Sigma^{-1}x - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log\frac{\pi_2}{\pi_1} \\
&= (-\mu_2^T\Sigma^{-1} - \mu_1^T\Sigma^{-1})x + (-\frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log\frac{\pi_2}{\pi_1}) \\
&= \beta x + \alpha
\end{aligned}
$$

2.3 Deriving $\Pr(C_1|X)$

To assess the probability of some output, we will take the inverse of the logit.

$$
\begin{aligned}
\operatorname{logit}(\Pr(C_1|X)) &= \beta x + \alpha \\
\log \frac{\Pr(C_1|X)}{1 - \Pr(C_1|X)} &= \beta x + \alpha \\
\log \frac{1 - \Pr(C_1|X)}{\Pr(C_1|X)} &= -(\beta x + \alpha) \\
\frac{1 - \Pr(C_1|X)}{\Pr(C_1|X)} &= e^{-(\beta x + \alpha)} \\
1 - \Pr(C_1|X) &= \Pr(C_1|X) e^{-(\beta x + \alpha)} \\
1 &= (1 + e^{-(\beta x + \alpha)}) \Pr(C_1|X) \\
\Pr(C_1|X) &= \frac{1}{1 + e^{-(\beta x + \alpha)}}
\end{aligned}
$$

We find that many different class conditional distributions (e.g., Poisson distributions) will give the same result. We have now found two different forms. We will call the first form $\mu(x)$.

$$
\begin{aligned}
\mu(x) = \Pr(x) &= \frac{1}{1 + \exp(-\beta^T x)} \\
\log \frac{\Pr(x)}{1 - \Pr(x)} &= \beta^T x
\end{aligned}
$$

3 Multivariate Derivation for Logistic Classification

However, we're currently interested in a *classifier* and not regression. So, we need to convert this real-valued continuous output into a label. Let us define a new random variable $Y = 1$ to represent class 1 and $Y = 0$ for class 0. We consider the following.

$$
\begin{aligned}
\Pr(Y = 1|x) &= \mu(x) \\
\Pr(Y = 0|x) &= 1 - \mu(x)
\end{aligned}
$$

We can combine both into a single expression; you can convince yourself that the two are equivalent by plugging in $Y = 0$ and $Y = 1$.

$$
\Pr(y|x) = \mu(x)^y (1 - \mu(x))^{1-y}
$$

3.1 Maximum Likelihood Estimate

We will now compute the likelihood. Before we begin, we will compute $\frac{\partial \mu_i}{\partial \beta}$.

$$\begin{aligned}
\frac{\partial \mu_i}{\partial \beta} &= \frac{\partial}{\partial \beta} \frac{1}{1 + e^{-\beta^T x_i}} \\
&= -(1 + e^{-\beta^T x_i})^{-2}(-x_i e^{-\beta^T x_i}) \\
&= \frac{x_i e^{-\beta^T x_i}}{1 + e^{-\beta^T x_i}} \frac{1}{1 + e^{-\beta^T x_i}} \\
&= x_i^T (1 - \mu_i)\mu_i
\end{aligned}$$

Assume all x_i are independent, and take the product of their probabilities. We will let $D = \{x_1, x_2 ... x_n\}$ and assume that all priors are identical.

$$\Pr(D|\theta) = L(\theta|D) = \Pi_{i=1}^{n} \mu_i^{y_i} (1 - \mu_i)^{(1-y_i)}$$

As usual, the log-likelihood is more informative.

$$l(\theta|D) = \log L(\theta|D) = \sum_i y_i \log \mu_i + (1 - y_i) \log 1 - \mu_i$$

Now, compute the gradient with respect to β, keeping in mind that μ_i are functions of β.

$$\nabla_\beta l = \sum_i \frac{y_i}{\mu_i} \frac{\partial \mu_i}{\partial \beta} - \frac{(1 - y_i)}{(1 - \mu_i)} \frac{\partial \mu_i}{\partial \beta}$$

Plugging in what we have above, we have the following:

$$
\begin{aligned}
&= \sum_i (\frac{y_i}{\mu_i} - \frac{(1-y_i)}{(1-\mu_i)}) x_i^T (1-\mu_i)\mu_i \\
&= \sum_i (\frac{y_i(1-\mu_i) - \mu_i(1-y_i)}{\mu_i(1-\mu_i)}) x_i^T (1-\mu_i)\mu_i \\
&= \sum_i (y_i(1-\mu_i) - \mu_i(1-y_i)) x_i^T \\
&= \sum_i (y_i - y_i\mu_i - \mu_i + \mu_i y_i) x_i^T \\
&= \sum_i (y_i - \mu_i) x_i^T
\end{aligned}
$$

We can take the general stochastic gradient descent update equation $\beta^{(t+1)} = \beta^{(t)} + \alpha \nabla_\beta l_i$ and plug in our gradient.

$$\beta^{(t+1)} = \beta^{(t)} + \alpha(y_i - \mu_i) x_i^T$$

With online updates in the perceptron algorithm, we take $\hat{y}_i = \text{sgn}(\beta^T x_i)$, then we have the following:

$$\beta^{(t+1)} = \beta^{(t)} + \alpha(y_i - \hat{y}_i^{(t)}) x_i$$

For least squares linear regression, we then have the following.

$$\beta^{(t+1)} = \beta^{(t)} + \alpha(y_i - \beta^{(t)T} x_i) x_i$$

We know that this iterative algorithm converges to the correct solution: It converges to the local optimum, and for a strictly convex function, the local optimum is the global optimum.