Logistic Regression

1 Introduction

[TBW]

2 Logistic regression

Logistic Regression is so similar to linear regression methods introduced, though its goal is to model categorical data. It combines logistic function, \( \pi(x) \) with linear regression. By defining logistic function:

\[
\frac{e^x}{1 + e^x},
\]

we define the structure of the Logistic Regression:

\[
\pi(x) = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}},
\]

where the exponent term \( \beta.x \) is the linear combination of variables, with an intercept, the same as what we had in linear regression:

\[
\beta^T.x = \beta_0^T + \sum_{i=1}^n \beta_i^T .x_i.
\]

Or equivalently

\[
\beta^T.x = \beta_0^T + \sum_{i=1}^n \beta_i^T .x_i = \ln \frac{\pi(x)}{1 + \pi(x)},
\]

By observing that \( 0 \leq \pi(x) \leq 1 \), we can assume that \( \Pr(G = 1|X = x) = \pi(x) \), which is two-category classification. We can similarly generalize the model to \( K \)-category class by defining needed equations:

\[
\beta_k^T .x = \beta_{k,0}^T + \sum_{i=1}^n \beta_{k,i}^T .x_{k,i} = \ln \frac{\Pr(G = k|X = x)}{1 + \Pr(G = k|X = x)}, \ 1 \leq k \leq K.
\]
Now we classify the data $X = x$ using $G^* = \arg\max_k \Pr(G = k \mid X = x)$. Note that in the $K$-category case, we have $\sum_{i=1}^{n} \Pr(G = i \mid X = x) = 1$; thus we can write the following:

$$
\Pr(G = k \mid X = x) = \frac{\exp(\beta^\top_k \cdot x)}{1 + \sum_{k=1}^{K-1} \exp(\beta^\top_k \cdot x)}, \quad 1 \leq k \leq K - 1,
$$

and

$$
\Pr(G = K \mid X = x) = \frac{1}{1 + \sum_{k=1}^{K-1} \exp(\beta^\top_k \cdot x)}.
$$

Fitting the multinomial regression is straightforward using maximum-likelihood. By a little abuse of notation, we define $p_g(x; \Theta) = \Pr(G = g \mid X = x)$. We define the likelihood on $N$ data points, as following:

$$
l(\theta) = \sum_{i=1}^{N} \log p_{g_i}(x_i; \Theta),
$$

where $\Theta$ is the parameter matrix of the model. By having the training data-set, we can find the optimal parameters of the problem, by an interative maximization of the likelihood(ML),

$$
\Theta^* = \arg\max_{\Theta} l(\theta) = \arg\max_{\Theta} \sum_{i=1}^{N} \log p_{g_i}(x_i; \Theta),
$$

using Newton-method,

$$
\Theta_{n+1} = \Theta_n - [Hl(x_n)]^{-1}\nabla l(x_n), \quad n \geq 0.
$$

One can derive the above gradient and Hessian matrix for the parameters of the problem, and train the system using the resulting *Iterative Reweighting Least Squares.*