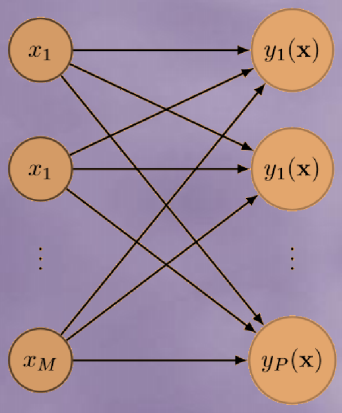

Logistic regression

Daniel Khashabi[1]
KHASHAB2@ILLINOIS.EDU

0.1 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*:

$$\pi(\mathbf{x}) = \frac{e^{\beta^{\mathbf{T}}.\mathbf{x}}}{1+e^{\beta^{\mathbf{T}}.\mathbf{x}}},$$

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

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

Or equivalently

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

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:

$$\beta_{\mathbf{k}}^{\mathbf{T}}.\mathbf{x} = \beta_{k,0}^T + \sum_{i=1}^{n} \beta_{k,i}^T.x_{k,i} = \ln \frac{\Pr(G=k|X=x)}{1+\Pr(G=k|X=x)},\ 1 \leq k \leq K.$$

[1]This is part of my notes; to see the complete list of notes check web.engr.illinois.edu/ khashab2/learn.html.

Now we classify the data $X = x$ using $G^* = \arg\max_k \Pr(G = k|X = x)$. Note that in the K-categoty case, we have $\sum_{i=1}^{n} \Pr(G = i|X = x) = 1$; thus we can write the following:

$$\Pr(G = k|X = x) = \frac{\exp\left(\beta_{\mathbf{k}}^{\mathbf{T}}.\mathbf{x}\right)}{1 + \sum_{k=1}^{K-1} \exp\left(\beta_{\mathbf{k}}^{\mathbf{T}}.\mathbf{x}\right)},\ 1 \leq k \leq K-1,$$

and

$$\Pr(G = K|X = x) = \frac{1}{1 + \sum_{k=1}^{K-1} \exp\left(\beta_{\mathbf{k}}^{\mathbf{T}}.\mathbf{x}\right)}.$$

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

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

where $\mathbf{\Theta}$ 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),

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

using Newton-method,

$$\mathbf{\Theta}_{n+1} = \mathbf{\Theta}_n - [Hl(\mathbf{x}_n)]^{-1} \nabla l(\mathbf{x}_n),\ n \geq 0.$$

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*.