CSE 6363: Machine Learning **Spring 2020**

Lecture 1: Logistic Regression

Lecturer: Dr. Di Ming *TA: Qicheng Wang*

1.1 Two-Class Logistic Regression

$Prob(event\ happens) = p.$

$Prob(event\ not\ happens) = 1 - p = q.$

The **Odds** of winning event happens is defined as

$$Odds \triangleq \frac{p}{q}. \tag{1.1}$$

In logistic regression, we model the log odds as a linear equation

$$log(\frac{p}{q}) = \beta^T x + b \tag{1.2}$$

From this definition, we obtain

$$\begin{aligned}
log(\tfrac{p}{q}) =& \quad \beta^T x + b \\
log(\tfrac{p}{1-p}) =& \quad \beta^T x + b \\
\tfrac{p}{1-p} =& \quad e^{\beta^T x + b} \\
p =& \quad (1-p)e^{\beta^T x + b} \\
p(1 + e^{\beta^T x + b}) =& \quad e^{\beta^T x + b} \\
p =& \quad \tfrac{e^{\beta^T x + b}}{1 + e^{\beta^T x + b}} = \tfrac{1}{1 + e^{-(\beta^T x + b)}} \\
q =& \quad 1 - p = 1 - \tfrac{1}{1 + e^{-(\beta^T x + b)}} = \tfrac{e^{-(\beta^T x + b)}}{1 + e^{-(\beta^T x + b)}} = \tfrac{1}{1 + e^{\beta^T x + b}}.
\end{aligned} \tag{1.3}$$

Now, consider 2-class classification. If we assign class label as $y_i = \pm 1$, we can express the probability as

$$p(y_i) = \frac{1}{1 + e^{-y_i(\beta^T x + b)}}. \tag{1.4}$$

Maximum Likelihood Estimation (MLE) In the following, we set $y_i = 1$ if the event happens, otherwise $y_i = 0$. We obtain the model parameters from the maximization of the likelihood of the events $x_1, \cdots, x_n$, which is defined as

$$L = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}, \tag{1.5}$$

where $p = P(y_i = +1)$. This is **binomial distribution**.

Rather than maximizing the likelihood function, equivalently, we maximizing the **log-likelihood** loss function

$$l = logL = \sum_{i=1}^{n} \left[y_i \cdot log(p) + (1 - y_i) \cdot log(1 - p) \right], \tag{1.6}$$

to obtain the model parameters:

$$\underset{\theta}{max} \;\; l(\theta). \tag{1.7}$$

Model 1. Binomial distribution

For binomial distribution, the model parameter is $\theta = p$. Then we obtain the optimal model parameter by maximizing the log-likelihood loss. The derivative of l with respect to p is computed as

$$\begin{aligned} \frac{\partial l}{\partial p} &= \textstyle\sum_{i=1}^{n} \left(\frac{y_i}{p} - \frac{1-y_i}{1-p} \right) \\ &= \textstyle\sum_{i=1}^{n} \frac{y_i(1-p) - p(1-y_i)}{p(1-p)} \\ &= \textstyle\sum_{i=1}^{n} \frac{y_i - p}{p(1-p)} \\ &= \frac{n_+ - np}{p(1-p)}. \end{aligned} \tag{1.8}$$

Set $\frac{\partial l}{\partial p} = 0$, we have

$$\frac{n_+ - np}{p(1-p)} = 0. \tag{1.9}$$

With the assumption that $p \neq 0$ and $1 - p \neq 0$, optimal parameter is obtained as

$$p = \frac{n_+}{n}. \tag{1.10}$$

Model 2. Logistic regression

For logistic regression, the model parameters are $\theta = (\beta, b)$ in the $p_i = \frac{1}{1+e^{-(\beta^T x_i + b)}}$. After padding, $p_i = \frac{1}{1+e^{-\tilde{\beta}^T \tilde{x}_i}}$, where $\tilde{\beta} = [\beta, b]$ and $\tilde{x}_i = [x_i, 1]$. The log-likelihood is now

$$l = logL = \sum_{i=1}^{n} \left[y_i \cdot log(p_i) + (1 - y_i) \cdot log(1 - p_i) \right], \tag{1.11}$$

The derivative of l with respect to $\tilde{\beta}$ is computed as

$$\begin{aligned} \frac{\partial l}{\partial \tilde{\beta}} &= \frac{\partial l}{\partial p_i} \cdot \frac{\partial p_i}{\partial \tilde{\beta}} \\ &= \textstyle\sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i} \right) \cdot \frac{\partial p_i}{\partial \tilde{\beta}}. \end{aligned} \tag{1.12}$$

In logistic regression, we use the activation function and its derivative

$$\sigma(z) = \frac{1}{1 + e^{(-z)}}, \;\; \frac{d\sigma(z)}{dz} = \sigma \cdot (1 - \sigma), \tag{1.13}$$

where $z_i = \tilde{\beta}^T \tilde{x}_i$ and $p_i = \sigma(z_i)$. Therefore, we have

$$\begin{aligned} \frac{\partial p_i}{\partial \tilde{\beta}} &= \frac{\partial p_i}{\partial z} \cdot \frac{\partial z}{\partial \tilde{\beta}} \\ &= \sigma \cdot (1 - \sigma) \cdot \tilde{x}_i \\ &= p_i \cdot (1 - p_i) \cdot \tilde{x}_i, \end{aligned} \tag{1.14}$$

Finally, we have

$$\begin{aligned}\frac{\partial l}{\partial \tilde{\beta}} &= \sum_{i=1}^{n}\left(\frac{y_i}{p_i}-\frac{1-y_i}{1-p_i}\right)\cdot p_i\cdot(1-p_i)\cdot\tilde{x}_i\\ &= \sum_{i=1}^{n}\left(y_i(1-p_i)-(1-y_i)p_i\right)\cdot\tilde{x}_i\\ &= \sum_{i=1}^{n}\left(y_i-p_i\right)\cdot\tilde{x}_i.\end{aligned} \tag{1.15}$$

Here, because of the complicated expression, we cannot find an expression for the parameter $\tilde{\beta} = \cdots$ as in Equ. 1.10. $\tilde{\beta}$ is computed numerically using an optimization algorithm such as the gradient descent algorithm.

1.2 Multi-Class Logistic Regression

For K class, we have probabilities $\{p_1, p_2, \cdots, p_K\}$ with $\sum_{i=1}^{K} p_i = 1$. Then we define

$$\begin{cases} log\frac{p_1}{p_K} = \beta_1^T x + b_1\\ log\frac{p_2}{p_K} = \beta_2^T x + b_2\\ \cdots\\ log\frac{p_{K-1}}{p_K} = \beta_{K-1}^T x + b_{K-1}\end{cases} \tag{1.16}$$

$$\begin{cases} \frac{p_1}{p_K} = e^{\beta_1^T x + b_1}\\ \frac{p_2}{p_K} = e^{\beta_2^T x + b_2}\\ \cdots\\ \frac{p_{K-1}}{p_K} = e^{\beta_{K-1}^T x + b_{K-1}}\end{cases} \tag{1.17}$$

$$\begin{cases} p_1 = p_K\cdot e^{\beta_1^T x + b_1} = \left[1-(p_1+\cdots+p_{K-1})\right]e^{\beta_1^T x + b_1}\\ p_2 = p_K\cdot e^{\beta_2^T x + b_2} = \left[1-(p_1+\cdots+p_{K-1})\right]e^{\beta_2^T x + b_2}\\ \cdots\\ p_{K-1} = p_K\cdot e^{\beta_{K-1}^T x + b_{K-1}} = \left[1-(p_1+\cdots+p_{K-1})\right]e^{\beta_{K-1}^T x + b_{K-1}}\end{cases} \tag{1.18}$$

Add all equations in Equ 1.18 together

$$\begin{aligned} p_1+\cdots+p_{K-1} &= \left[1-(p_1+\cdots+p_{K-1})\right]\left(e^{\beta_1^T x + b_1}+e^{\beta_2^T x + b_2}+\cdots+e^{\beta_{K-1}^T x + b_{K-1}}\right)\\ 1-p_K &= p_K\cdot\left(\sum_{i=1}^{K-1} e^{\beta_i^T x + b_i}\right)\\ 1 &= p_K\cdot\left(1+\sum_{i=1}^{K-1} e^{\beta_i^T x + b_i}\right)\\ p_K &= \frac{1}{1+\sum_{i=1}^{K-1} e^{\beta_i^T x + b_i}}.\end{aligned} \tag{1.19}$$

Put p_K in Equ 1.19 back into Equ 1.18

$$p_k = p_K\cdot e^{\beta_k^T x + b_k} = \frac{e^{\beta_k^T x + b_k}}{1+\sum_{i=1}^{K-1} e^{\beta_i^T x + b_i}}, \tag{1.20}$$

where k varies in $\{1, 2, \cdots, K-1\}$.

Using padding $\tilde{\beta_k} = [\beta_k, b_k]$ and $\tilde{x} = [x, 1]$, Equ 1.20 is transformed as

$$\begin{cases} p_1 = \frac{e^{\tilde{\beta}_1^T \tilde{x}}}{1+\sum_{i=1}^{K-1} e^{\tilde{\beta}_i^T \tilde{x}}} \\ p_2 = \frac{e^{\tilde{\beta}_2^T \tilde{x}}}{1+\sum_{i=1}^{K-1} e^{\tilde{\beta}_i^T \tilde{x}}} \\ \cdots \\ p_{K-1} = \frac{e^{\tilde{\beta}_{K-1}^T \tilde{x}}}{1+\sum_{i=1}^{K-1} e^{\tilde{\beta}_i^T \tilde{x}}} \\ p_K = \frac{1}{1+\sum_{i=1}^{K-1} e^{\tilde{\beta}_i^T \tilde{x}}} \end{cases} \tag{1.21}$$

Maximum Likelihood Estimation (MLE)

$$L \propto p_1^{n_1} p_2^{n_2} \cdots p_K^{n_K} \tag{1.22}$$

where $\sum_{k=1}^{K} n_k = n$ and $\sum_{k=1}^{K} p_k = 1$. This is **multinomial distribution**.

The log-likelihood loss function

$$l = logL \propto \sum_{i=1}^{K} n_i \cdot logp_i. \tag{1.23}$$

Model 3. Multinomial distribution

If we do multinomial distribution, model parameters are $\{\theta_1 = p_1, \cdots, \theta_K = p_K\}$. Then we obtain optimal model parameters by maximizing the log-likelihood loss

$$\begin{array}{c} \max\limits_{\{p_1, \cdots, p_K\}} \sum_{i=1}^{K} n_i \cdot logp_i, \\ subject\ to\ \sum_{i=1}^{K} p_i = 1. \end{array} \tag{1.24}$$

Using **Augmented Lagrangian Multiplier (ALM)**, we have the following equivalent optimization form

$$\max_{\{p_1, \cdots, p_K\}} \mathcal{L}(\lambda) = \left(\sum_{i=1}^{K} n_i \cdot logp_i\right) - \lambda \cdot \left(\sum_{i=1}^{K} p_i - 1\right) \tag{1.25}$$

The derivative of $\mathcal{L}$ with respect to p_i and λ is computed respectively as

$$\begin{array}{c} \frac{\partial \mathcal{L}}{\partial p_i} = \frac{n_i}{p_i} - \lambda, \\ \frac{\partial \mathcal{L}}{\partial \lambda} = \sum\limits_{i=1}^{K} p_i - 1. \end{array} \tag{1.26}$$

Set $\frac{\partial \mathcal{L}}{\partial p_i} = 0$, we have

$$n_i = \lambda p_i \ (i = 1, \cdots, K). \tag{1.27}$$

Add all the K equations in Equ 1.27 together,

$$\begin{aligned} n_1 + n_2 + \cdots + n_K &= \lambda \cdot (p_1 + p_2 + \cdots + p_K) \\ n &= \lambda \cdot 1 \\ \lambda &= n \end{aligned} \tag{1.28}$$

Put $\lambda = n$ back in Equ 1.27,

$$n_i = np_i \ (i = 1, \cdots, K). \tag{1.29}$$

Finally, optimal parameters are obtained

$$p_i = \frac{n_i}{n} \ (i=1,\cdots,K). \tag{1.30}$$

Model 4. Multinomial logistic regression

Now we do multinomial logistic regression, model parameters $\theta = \tilde{\beta}_k$ are in $p_{k,i} = \frac{e^{\tilde{\beta}_k^T \tilde{x}_i}}{1+\sum_{h=1}^{K-1} e^{\tilde{\beta}_h^T \tilde{x}_i}}$ $(k = 1,\cdots,K-1)$ and $p_{K,i} = \frac{1}{1+\sum_{h=1}^{K-1} e^{\tilde{\beta}_h^T \tilde{x}_i}}$.

For generality, set $\tilde{\beta}_K = 0$. Then we have

$$p_{k,i} = \frac{e^{\tilde{\beta}_k^T \tilde{x}_i}}{\sum_{h=1}^{K} e^{\tilde{\beta}_h^T \tilde{x}_i}}, \quad k=\{1,\cdots,K\}. \tag{1.31}$$

Let $Y_i^k = I(Y_i = k)$ be one-of-K encoding of Y_i , also known as indicator vector with $\sum_{i=1}^{K} Y_i^k = 1$. Define the log-likelihood loss function as

$$\begin{aligned} l = \ & logL \\ = \ & log\left[\prod_{i=1}^{n}\left(p_{1,i}^{Y_i^1}\cdots p_{K,i}^{Y_i^K}\right)\right] \\ = \ & \sum_{i=1}^{n} log\left(p_{1,i}^{Y_i^1}\cdots p_{K,i}^{Y_i^K}\right) \\ = \ & \sum_{i=1}^{n}\sum_{k=1}^{K} Y_i^k log(p_{k,i}) \\ = \ & \sum_{i=1}^{n}\left\{\sum_{k=1}^{K} Y_i^k log\frac{e^{\tilde{\beta}_k^T \tilde{x}_i}}{\sum_{h=1}^{K} e^{\tilde{\beta}_h^T \tilde{x}_i}}\right\} \\ = \ & \sum_{i=1}^{n}\left\{\sum_{k=1}^{K} Y_i^k(\beta_k^T \tilde{x}_i) - log\left(\sum_{h=1}^{K} e^{\tilde{\beta}_h^T \tilde{x}_i}\right)\right\} \end{aligned} \tag{1.32}$$

The derivative of l with respect to $\tilde{\beta}_k$ $(k = 1,\cdots,K-1)$ is computed as

$$\begin{aligned} \frac{\partial l}{\partial \tilde{\beta}_k} = \ & \sum_{i=1}^{n} Y_i^k \cdot \tilde{x}_i - \left(\frac{e^{\tilde{\beta}_k^T \tilde{x}_i}}{\sum_{h=1}^{K} e^{\tilde{\beta}_h^T \tilde{x}_i}}\right)\cdot \tilde{x}_i \\ = \ & \sum_{i=1}^{n}\left(Y_i^k - p_{k,i}\right)\cdot \tilde{x}_i, \end{aligned} \tag{1.33}$$

which is same as the derivative result of binomial logistic regression in Equ 1.15 .

1.3 Softmax Regression

Probability is defined as

$$\begin{cases} p_1 & = \frac{e^{\tilde{\beta}_1^T \tilde{x}}}{\sum_{k=1}^{K} e^{\tilde{\beta}_k^T \tilde{x}}}, \\ & \cdots \\ p_K & = \frac{e^{\tilde{\beta}_K^T \tilde{x}}}{\sum_{k=1}^{K} e^{\tilde{\beta}_k^T \tilde{x}}}. \end{cases} \tag{1.34}$$

Actually, **Softmax Regression** is the general form of **Multi-class Logistic Regression** without specifically setting $\tilde{\beta}_K$ to 0.

Therefore, the derivative of log-likelihood loss with respect to $\tilde{\beta}_k$ is same as the results in Equ 1.33.