Logistic Regression

Statistical Machine Learning (ENGG*6600*08)

School of Engineering,
University of Guelph, ON, Canada

Course Instructor: Benyamin Ghojogh

Fall 2023

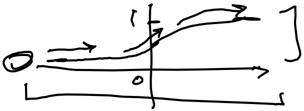
Logistic Regression

- Logistic regression is popular in bio-statistics and bio-informatics.
- Let $\boldsymbol{x} \in \mathbb{R}^d$ be data and $y \in \mathbb{R}$ be class label. Baye's rule:

$$\boxed{\mathbb{P}(y|\boldsymbol{x}) = \frac{\mathbb{P}(\boldsymbol{x}|y)\mathbb{P}(y)}{\mathbb{P}(\boldsymbol{x})},} \tag{1}$$

  where $\mathbb{P}(y|\boldsymbol{x})$ and $\mathbb{P}(\boldsymbol{x}|y)$ are the posterior and likelihood, respectively, and $\mathbb{P}(\boldsymbol{x})$ and $\mathbb{P}(y)$ are the priors.
- In contrast to Linear Discriminant Analysis (LDA), logistic regression works on the posterior $\mathbb{P}(y|\boldsymbol{x})$ directly rather than working on likelihood $\mathbb{P}(\boldsymbol{x}|y)$ and prior $\mathbb{P}(y)$.

Logistic Regression



- Logistic regression is a binary classifier where it assigns probability between zero and one for belonging to one of the classes.
- The logistic function, used in logistic regression, was initially proposed in 1845 for modeling the population growth [1]. It was further improved in the 20th century [2]. See [3] for the history of logistic regression.
- It considers the classification problem as a regression problem where it regresses (predicts) the probability of belonging to a class. It first considers a linear regression $\boldsymbol{\beta}^\top \boldsymbol{x} + \boldsymbol{\beta}_0$. However, in order to not have the bias, it assumes that $\boldsymbol{x}$ is $d+1$ dimensional with an additional element of 1 for bias, i.e., $\boldsymbol{x} = [x_1, \dots, x_d, 1]^\top$. The $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$ is the learnable parameter of the logistic regression model. As a result, the linear regression becomes $\boldsymbol{\beta}^\top \boldsymbol{x}$.
- However, there is no bound on this regression while logistic regression desires the output to be in the range $[0, 1]$ to behave like a probability. Therefore, Logistic regression models the posterior using a **logistic function**, also called the **sigmoid function**, to make this regression between zero and one.

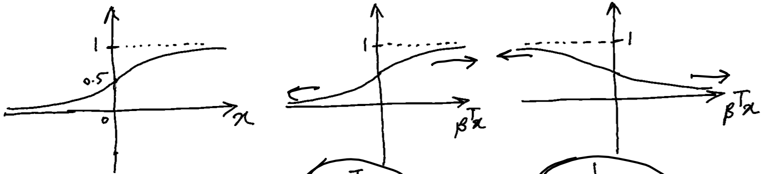
Logistic Regression

- Assume we have two classes $y \in \{0, 1\}$.
- Logistic regression models the posterior using a **logistic function**, also called the **sigmoid function**:

$$\mathbb{P}(y = 1 | X = x) = \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}^\top \boldsymbol{x}}}, \quad (2)$$

$$\mathbb{P}(y = 0 | X = x) = 1 - \mathbb{P}(y = 1 | X = x) = \frac{1}{1 + e^{\boldsymbol{\beta}^\top \boldsymbol{x}}}, \quad (3)$$

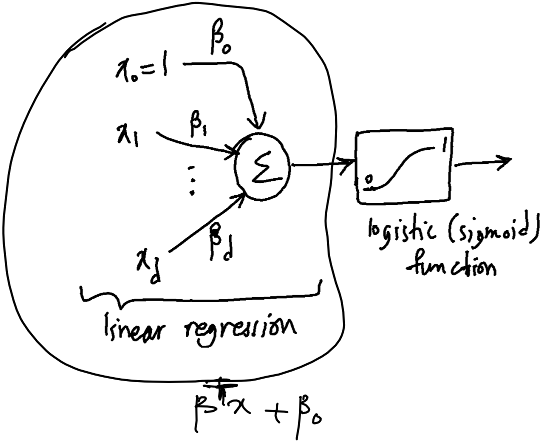
where $\boldsymbol{\beta} \in \mathbb{R}^d$ is the learnable parameter of the logistic regression model.



$\frac{1}{1+e^{-x}}$ (sigmoid) $\qquad \frac{e^{\beta^T x}}{1+e^{\beta^T x}} \qquad \frac{1}{1+e^{\beta^T x}}$

Logistic Regression as a Neural Network

- Logistic regression can be seen as a neural network with one neuron where the activation function is the nonlinear sigmoid (logistic) function.



Logistic Regression

- Consider n data points $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ in the dataset. Assuming that they are independent and identically distributed (i.i.d), the posterior over all data points is:

$$\mathbb{P}(y|X) = \prod_{i=1}^{n} \Big(\mathbb{P}(y_i = 1|X = x_i)\mathbb{I}(y_i = 1) + \mathbb{P}(y_i = 0|X = x_i)\mathbb{I}(y_i = 0)\Big), \tag{4}$$

where $\mathbb{I}(.)$ is the indicator function which is one if its condition is satisfied and is zero otherwise.

- As the labels are either zero or one, i.e., $y_i \in \{0, 1\}$, this equation can be restated as:

$$\mathbb{P}(y|X) = \prod_{i=1}^{n} \big(\mathbb{P}(y_i = 1|X = x_i)\big)^{y_i}\big(\mathbb{P}(y_i = 0|X = x_i)\big)^{1-y_i}. \tag{5}$$

- Substituting Eqs. (2) and (3) in this equation gives:

$$\mathbb{P}(y|X) = \prod_{i=1}^{n} \Big(\frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}{1 + e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}\Big)^{y_i}\Big(\frac{1}{1 + e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}\Big)^{1-y_i}. \tag{6}$$

Logistic Regression

- The log posterior is:

$$
\begin{aligned}
\ell(\boldsymbol{\beta}) := \mathbb{P}(y|X=x) &= \log \prod_{i=1}^{n} \Big(\frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}{1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}\Big)^{y_i} \Big(\frac{1}{1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}\Big)^{1-y_i} \\
&= \sum_{i=1}^{n} \Big(\log \Big(\frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}{1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}\Big)^{y_i} + \log \Big(\frac{1}{1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}\Big)^{1-y_i} \Big) \\
&= \sum_{i=1}^{n} \Big(y_i \log(e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}) - y_i \log(1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}) - (1-y_i)\log(1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}) \Big) \\
&= \sum_{i=1}^{n} \Big(y_i \boldsymbol{\beta}^\top \boldsymbol{x}_i - \cancel{y_i \log(1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i})} - \log(1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}) + \cancel{y_i \log(1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i})} \Big) \\
&= \sum_{i=1}^{n} \Big(y_i \boldsymbol{\beta}^\top \boldsymbol{x}_i - \log(1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}) \Big).
\end{aligned}
$$

Logistic Regression

- The log posterior is:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i \boldsymbol{\beta}^\top \mathbf{x}_i - \log(1 + e^{\boldsymbol{\beta}^\top \mathbf{x}_i}) \right).$$

- Newton's method can be used to find the optimum $\boldsymbol{\beta}$. The first derivative, or the gradient, it:

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left(y_i \mathbf{x}_i - \frac{1}{1 + e^{\boldsymbol{\beta}^\top \mathbf{x}_i}} e^{\boldsymbol{\beta}^\top \mathbf{x}_i} \mathbf{x}_i \right) = \sum_{i=1}^{n} (y_i - \frac{e^{\boldsymbol{\beta}^\top \mathbf{x}_i}}{1 + e^{\boldsymbol{\beta}^\top \mathbf{x}_i}}) \mathbf{x}_i. \tag{7}$$

Its transpose is:

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} = \sum_{i=1}^{n} (y_i - \frac{e^{\boldsymbol{\beta}^\top \mathbf{x}_i}}{1 + e^{\boldsymbol{\beta}^\top \mathbf{x}_i}}) \mathbf{x}_i^\top.$$

Logistic Regression

- The second derivative is:

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = \frac{\partial}{\partial \boldsymbol{\beta}} \Big(\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} \Big) = \frac{\partial}{\partial \boldsymbol{\beta}} \Big(\sum_{i=1}^{n} \big(y_i - \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}{1 + e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}} \big) \boldsymbol{x}_i^\top \Big)$$

$$= \sum_{i=1}^{n} \Big(- \frac{\partial}{\partial \boldsymbol{\beta}} \big(\frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}{1 + e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}} \big) \Big) \boldsymbol{x}_i^\top.$$

- We define:

$$\mathbb{P}(\boldsymbol{x}_i | \boldsymbol{\beta}) := \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}{1 + e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}. \tag{8}$$

Therefore:

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = - \sum_{i=1}^{n} \Big(\frac{\partial}{\partial \boldsymbol{\beta}} \big(\mathbb{P}(\boldsymbol{x}_i | \boldsymbol{\beta}) \big) \Big) \boldsymbol{x}_i^\top. \tag{9}$$

Logistic Regression

- We have:

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}}\big(\mathbb{P}(\boldsymbol{x}_i|\boldsymbol{\beta})\big) &= \frac{\partial}{\partial \boldsymbol{\beta}}\Big(\frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}{1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}\Big) \\
&= \frac{1}{(1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i})^2}\Big(e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}\boldsymbol{x}_i(1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}) - e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}(e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}\boldsymbol{x}_i)\Big) \\
&= \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}{(1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i})^2}\Big(1+\cancel{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}} - \cancel{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}\Big)\boldsymbol{x}_i = \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}{(1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i})^2}\boldsymbol{x}_i \\
&= \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}{(1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i})}\frac{1}{(1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i})}\boldsymbol{x}_i \overset{(8)}{=} \mathbb{P}(\boldsymbol{x}_i|\boldsymbol{\beta})\big(1-\mathbb{P}(\boldsymbol{x}_i|\boldsymbol{\beta})\big)\boldsymbol{x}_i
\end{aligned}
$$

- Substituting it in Eq. (9) gives the second derivative, i.e., the Hessian matrix:

$$
\star \quad \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} = -\sum_{i=1}^{n}\Big(\mathbb{P}(\boldsymbol{x}_i|\boldsymbol{\beta})\big(1-\mathbb{P}(\boldsymbol{x}_i|\boldsymbol{\beta})\big)\boldsymbol{x}_i\Big)\boldsymbol{x}_i^\top. \quad \to 1\times(d+1) \tag{10}
$$

$(d+1)\times 1$

Logistic Regression

- It is possible to write the Newton's method in matrix form. We define:

$$
\begin{aligned}
\underbrace{\mathbb{R}^{(d+1)\times n}} \ni \underbrace{\boldsymbol{X}} &:= \begin{bmatrix} 1 & 1 & \dots & 1 \\ \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_n \end{bmatrix}, \\
\underbrace{\mathbb{R}^{n\times n}} \ni \boldsymbol{W} &:= \underbrace{\mathbf{diag}}\Big(\underbrace{\mathbb{P}(\boldsymbol{x}_i|\boldsymbol{\beta})\big(1-\mathbb{P}(\boldsymbol{x}_i|\boldsymbol{\beta})\big)}\Big), \\
\mathbb{R}^{n} \ni \boldsymbol{y} &:= [y_1, \dots, y_n]^\top, \\
\mathbb{R}^{n} \ni \boldsymbol{p} &:= \Big[\frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_1}}{\underbrace{1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_1}}}, \dots, \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_n}}{\underbrace{1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}_n}}}\Big]^\top.
\end{aligned}
$$

- The Eqs. (7) and (10) can be restated as:

$$
\mathbb{R}^{(d+1)} \ni \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{y} - \boldsymbol{p}), \tag{11}
$$

$$
\mathbb{R}^{(d+1)\times(d+1)} \ni \underbrace{\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}} = -\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^\top. \tag{12}
$$

- Using Newton's method for maximization of the log posterior is:

$$
\boxed{\boldsymbol{\beta}^{(\tau+1)} := \boldsymbol{\beta}^{(\tau)} + \Big(\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\Big)^{-1} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}} \implies
$$

$$
\boldsymbol{\beta}^{(\tau+1)} := \boldsymbol{\beta}^{(\tau)} - (\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^\top)^{-1} \boldsymbol{X}(\boldsymbol{y} - \boldsymbol{p}), \tag{13}
$$

where τ is the iteration index. It is repeated until convergence of $\boldsymbol{\beta}$.

Logistic Regression

- In the test phase, the class of a point $\boldsymbol{x}$ is determined as:

$$y = \begin{cases} 1 & \text{if } \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}}}{1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}}} \geq 0.5, \\ 0 & \text{Otherwise.} \end{cases} \tag{14}$$

- Comparison to LDA:
  - Logistic regression estimates $(d+1)$ parameters in $\boldsymbol{\beta}$, but LDA estimates many more parameters:
    - ★ prior of each class: 1. We have two classes: $2 \times 1 = 2$.
    - ★ mean of each class: d. We have two classes: $2 \times d = 2d$.
    - ★ covariance matrix of each class: $d(d+1)/2$. We have two classes: $2 \times (d(d+1)/2) = d(d+1)$.
    - ★ so, in total: $2 + 2d + d(d+1) = d^2 + 2d + 2$.
  - LDA assumes the distribution of each class is Gaussian which may not be true. However, logistic regression does not assume anything about the distribution of data.

Acknowledgment

- Some slides of this slide deck were inspired by teachings of Prof. Ali Ghodsi (at University of Waterloo, Department of Statistics).

References

[1] P. F. Verhulst, “Resherches mathematiques sur la loi d’accroissement de la population,” *Nouveaux memoires de l’academie royale des sciences*, vol. 18, pp. 1–41, 1845.

[2] S. H. Walker and D. B. Duncan, “Estimation of the probability of an event as a function of several independent variables,” *Biometrika*, vol. 54, no. 1-2, pp. 167–179, 1967.

[3] J. S. Cramer, “The origins of logistic regression,” 2002.