Logistic Regression

Nathanaël Carraz Rakotonirina

Mathématiques Informatique et Statistique Appliquées (MISA)
Université d'Antananarivo

Model

It is a classification model $p(y|\boldsymbol{x}; \boldsymbol{\theta})$.

- $\boldsymbol{x} \in \mathbb{R}^D$: input
- $y \in \{1, ..., C\}$: class label
- $\boldsymbol{\theta}$: parameters

If $C = 2$, it is called **binary logistic regression** and if $C > 2$, it is known as **multiclass logistic regression**.

Binary logistic regression

Since we want to predict $y \in 0, 1$ given some inputs $\boldsymbol{x}$, the model is of the form

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathrm{Ber}(y; f(\boldsymbol{x};\boldsymbol{\theta}))$$

where $f(\boldsymbol{x};\boldsymbol{\theta})$ is a function giving the parameter of the distribution hence must satisfy $0 \leq f(\boldsymbol{x};\boldsymbol{\theta}) \leq 1$. To allow f to be any function, we use:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathrm{Ber}(y; \sigma(f(\boldsymbol{x};\boldsymbol{\theta})))$$

where σ is the **sigmoid** (S-shaped) or **logistic** function:

$$\begin{aligned} \sigma :& \mathbb{R} \to [0, 1] \\ & z \mapsto \sigma(z) = \frac{1}{1 + e^{-z}} \end{aligned}$$

z is called the logit or the pre-activation.

For logistic regression, we choose a linear function $f(\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{w}^\top \boldsymbol{x} + b$. The model has the form

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathrm{Ber}(y;\sigma(\boldsymbol{w}^\top \boldsymbol{x} + b))$$

This means

$$p(y=1|\boldsymbol{x};\boldsymbol{\theta}) = \sigma(\boldsymbol{w}^\top \boldsymbol{x} + b)) = \frac{1}{1+e^{-(\boldsymbol{w}^\top \boldsymbol{x}+b)}}$$

Decision boundary

During prediction, we have

$$\hat{y} = \begin{cases} 1 & \text{if } p(y = 1|\boldsymbol{x}; \boldsymbol{\theta}) > 0.5 \\ 0 & \text{if } p(y = 1|\boldsymbol{x}; \boldsymbol{\theta}) < 0.5 \end{cases}$$

which is the same as

$$\hat{y} = \begin{cases} 1 & \text{if } \boldsymbol{w}^\top \boldsymbol{x} + b > 0 \\ 0 & \text{if } \boldsymbol{w}^\top \boldsymbol{x} + b < 0 \end{cases}$$

The **decision boundary** is $\boldsymbol{w}^\top \boldsymbol{x} + b = 0$. It is a linear hyperplane with normal vector $\boldsymbol{w}$ and an offset b from the origin. It separates the spase into 2 half-spaces.
The data is said to be **linearly separable** when the examples can be perfectly separated by the linear hyperplane.

Maximum likelihood estimation

We note $\mu_i = \sigma(z_i) = \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)$. The negative log likelihood is

$$\begin{aligned}
\mathrm{NLL}(\boldsymbol{w}) &= -\sum_{i=1}^{N} \log p(y_i | \boldsymbol{x}_i; \boldsymbol{\theta}) = -\sum_{i=1}^{N} \log Ber(y; \mu_i) \\
&= -\sum_{i=1}^{N} \log[\mu_i^{y_i} + (1-\mu_i)^{1-y_i}] \\
&= -\sum_{i=1}^{N} [y_i \log(\mu_i) + (1-y_i)\log(1-\mu_i)] \\
&= \sum_{i=1}^{N} \mathbb{H}(y_i, \mu_i)
\end{aligned}$$

where $\mathbb{H}$ is the **binary cross entropy**. This objective is convex and can be minimized using gradient-based methods.

Multinomial logistic regression

It is a classification model of the form

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathrm{Cat}(y; f(\boldsymbol{x};\boldsymbol{\theta}))$$

We note $\mu = f(\boldsymbol{x};\boldsymbol{\theta})$ (here $f : \mathbb{R}^D \to \mathbb{R}^C$) . It must satisfy $0 \leq \mu_i \leq 1$ and $\sum_{i=1}^{C} \mu_i = 1$. To allow f to be any function, we pass it to the **softmax** function

$$\begin{aligned} \mathcal{S} : \mathbb{R}^C &\to [0,1]^C \\ z &\mapsto \mathcal{S}(z) = \left[\frac{e^{z_1}}{\sum_{i=1}^{C} e^{z_i}}, \ldots, \frac{e^{z_C}}{\sum_{i=1}^{C} e^{z_i}}\right] \end{aligned}$$

You might want to use the log-sum-exp trick to avoid numerical overflow when computing the softmax.

We use a linear function $f(\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$ where $\boldsymbol{W}$ is a $C \times D$ matrix and $\boldsymbol{b}$ is a C dimensional vector. The model is of the form

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathrm{Cat}(y;\mathcal{S}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}))$$

If we note $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$ the C dimensional vector of logits, we have

$$p(y = c|\boldsymbol{x};\boldsymbol{\theta}) = \frac{e^{z_c}}{\sum_{i=1}^{C} e^{z_i}}$$

Maximum likelihood estimation

We keep $\boldsymbol{\mu} =$. The negative log likelihood is

$$
\begin{aligned}
\mathrm{NLL}(\boldsymbol{w}) &= -\log \prod_{i=1}^{N} \prod_{c=1}^{C} \mu_{ij}^{y_{ic}} \\
&= -\sum_{i=1}^{N} \sum_{i=c}^{C} y_{ic} \log \mu_{ic} \\
&= \sum_{i=1}^{N} \mathbb{H}(y_i, \mu_i)
\end{aligned}
$$

where $\mu_{ic} = p(y_i = c | x_i; \boldsymbol{\theta}) = (\mathcal{S}(\boldsymbol{W}\boldsymbol{x}_i + \boldsymbol{b}))_c$ and $y_{ic} = \mathbb{I}(y_i = c)$. This objective is also convex and can be minimized using gradient descent.

Do you want more ?

Of course you do !

Explore further

- Robust logistic regression
- Bayesian logistic regression
- Multilabel classification
- Hierarchical classification