CM50264 Machine Learning 1, Lecture 7

Logistic Regression

Xi Chen

UNIVERSITY OF BATH

Review - Binomial distribution

Definition

Binomial distribution $f(k, n, \pi)$: discrete probability distribution that describes the probability of k successes out of a sequence of n independent experiments, with success probability π for the True cases and $1 - \pi$ for the False cases.

The formula:

$$f(k, n, \pi) = \left(\begin{array}{c} n \\ k \end{array} \right) \pi^k (1 - \pi)^{N-k},$$

where

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

Review - Example

Coin toss

A biased coin, with 60% chance of landing heads up when tossed. What is the probability of achieving 3 heads after 6 tosses?

$$f(k, n, \pi) = \left(\begin{array}{c} n \\ k \end{array} \right) \pi^k (1 - \pi)^{n-k},$$

$$f(3, 6, 0.6) = \left(\begin{array}{c} 6 \\ 3 \end{array} \right) 0.6^3 0.4^3,$$

Review - Bernoulli distribution

Definition

Bernoulli distribution $f(k, \pi)$: special case of Binomial distribution when $n = 1$, it is a discrete probability distribution of a random variable which takes the value True with probability π and the value False with probability $1 - \pi$.

$$f(k, \pi) = \begin{cases} \pi, & \text{if } \; k = 1; \\ 1 - \pi, & \text{if } \; k = 0 \end{cases}$$

Alternatively, we have:

$$f(k, \pi) = \pi^k (1 - \pi)^{(1-k)},$$

for $k \in \{0, 1\}$

Binary classification problem

Given a set of data points (pairs of input and output)

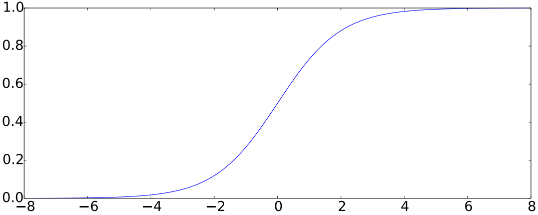
$$D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\} \subset \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^n \times \{0, 1\},$$

in which y can take on only two values, 0(False) and 1(True), and $\mathbf{x}_i$ is a vector of features/properties. The corresponding y_i is called label for the training example.

Goal: train a model to decide the Boolean-valued outcome y_{N+1} when we observe new data $\mathbf{x}_{N+1}$.

Logistic function

Also called *sigmoid* function.



$$\sigma(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z}, \tag{1}$$

which maps $[-\infty, +\infty]$ to $[0, 1]$.

Logistic function - the inverse

It is called *Logit function*, or *log odds* function.

$$logit(\pi) = \sigma^{-1}(z) = \log \frac{\pi}{1-\pi},$$

which maps $[0, 1]$ to $[-\infty, +\infty]$.

Logistic function - special property

A special property of the sigmoid function as follows:

$$\begin{aligned}\sigma^{'}(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}}\\ &= \sigma(z)(1-\sigma(z)).\end{aligned}$$

Logistic regression

Recall the form in linear regression in the previous lecture:

$$f_{\mathbf{w}}(\mathbf{x}) = w_0 x_0 + w_1 x_1 + \cdots + w_M x_M = \mathbf{w}^\top \mathbf{x}.$$

M is the number of input dimensions.

It is now modified as:

$$f_{\mathbf{w}}(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}}},$$

where $\sigma(\cdot)$ is the sigmoid function, we have $z = \mathbf{w}^\top \mathbf{x}$ in function $\sigma(z)$.

Notice that both $\sigma(\cdot)$ and $f_{\mathbf{w}}(\cdot)$ are bounded between 0 and 1.

Sigmoid function form

So why in the form of sigmoid function?

$$f_{\mathbf{w}}(\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}}},$$

Mainly depends on the distribution and form of samples.

It is derived from the Generalised Linear Model (GLM) with samples follow a member in the exponential family of probability distributions.

Including Bernoulli, Normal, Poisson, Dirichlet distributions etc..

Connecting to classification problem

- Binary classification problem: output y, is 0 or 1.
- For now assume single input feature $\mathbf{x} = x$ and we have:

$$z = \mathbf{w}^\top \mathbf{x} = w_0 + w_1 x.$$

- We assume

$$P(y = 1|x; \mathbf{w}) = f_{\mathbf{w}}(\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}}},$$

$$P(y = 0|x; \mathbf{w}) = 1 - f_{\mathbf{w}}(\mathbf{x}) = 1 - \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}}} = \frac{e^{-\mathbf{w}^\top \mathbf{x}}}{1 + e^{-\mathbf{w}^\top \mathbf{x}}},$$

- w_0 and w_1 are parameters which we fit to data.

Likelihood and log-likelihood

The probability can now be expressed in a compact form similar to that in the Bernoulli distribution:

$$p(y|x;\mathbf{w}) = (f_{\mathbf{w}}(\mathbf{x}))^{y}(1 - f_{\mathbf{w}}(\mathbf{x}))^{1-y}.$$

Our data set with N data points:

$$D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\} \subset \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^n \times \{0, 1\},$$

We now have the likelihood of parameters $\mathbf{w}$ as:

$$L(\mathbf{w}) = \prod_{i=1}^{N} p(y_i|x_i;\mathbf{w})$$

Likelihood and log-likelihood

The likelihood equation

$$L(\mathbf{w}) = \prod_{i=1}^{N} p(y_i | x_i; \mathbf{w})$$

can be written in the form of log-likelihood as:

$$\log L(\mathbf{w}) = \sum_{i=1}^{N} [y_i \log(f_{\mathbf{w}}(x_i)) + (1 - y_i) \log(1 - f_{\mathbf{w}}(x_i))]$$

Similar to the linear regression case, the optimum solution can be written as:

$$\mathbf{w}_* = \arg\max_{\mathbf{w}} \log L(\mathbf{w}).$$

How do we maximise this?

Maximum likelihood estimation

Recall the gradient descent way to update **w** in our previous lecture:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t \mathbf{p}_t.$$

The direction $\mathbf{p}_t$ can now be obtained by gradient descent:

$$\mathbf{p}_t = \nabla \log L(\mathbf{w}).$$

Derive.
To update each $w_{t,j}$, we have:

$$\mathbf{p}_{t,j} = (y - f_{\mathbf{w}}(x))x_j$$

to update the *j*th entry of **w** at the *i*th iteration step.

Model fitting - gradient descent

Input: step size parameter $\alpha_t > 0$;

1. $t = 0$; Make an initial guess, $w_{0,0} = w_{0,1} = 0$;
2. Iterate i, the samples, until termination conditions are met, $\log L(\mathbf{w})$ stops increasing.
    1. $w_{t+1,j} = w_{t,j} + \alpha_t \frac{\partial \log L(\mathbf{w})}{\partial w_j}$ for $j = 0, 1$;
    2. $t = t + 1$;

Model fitting - multivariate case

- Replace z with:

$$z = \mathbf{w}^\top \mathbf{x},$$

where

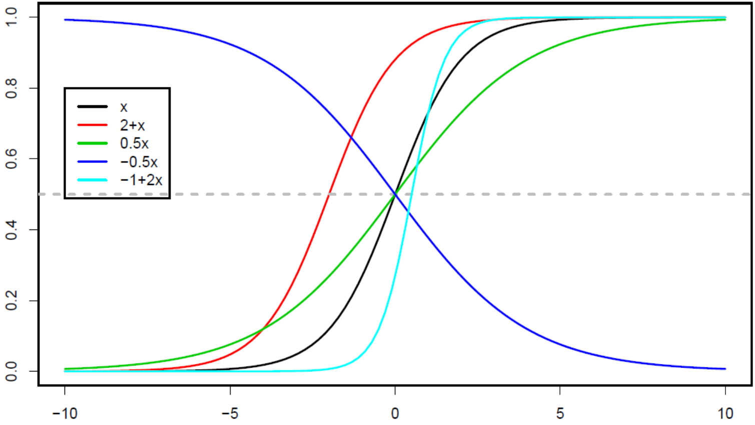
$$\mathbf{w} = (w_0, w_1, w_2, \dots)^\top, \qquad \mathbf{x} = (1, x_1, x_2, \dots)^\top.$$

- Input: step size parameter $\alpha_t > 0$;

```
I   $t = 0$; Make an initial guess, $\mathbf{w}_0 = 0$;
II  Iterate until termination conditions are met,
    $\log L(\mathbf{w})$ stops increasing.
        i   $\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t \nabla_{\mathbf{w}} \log L(\mathbf{w})$;
        ii  $t = t + 1$;
```

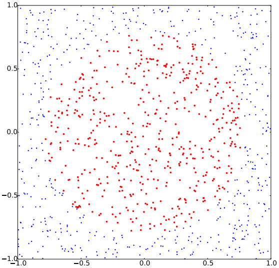
Logistic curves

Logistic curves in the simple univariate case with different **w**.



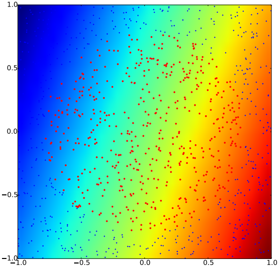
Decision boundary

Separate with a straight line?



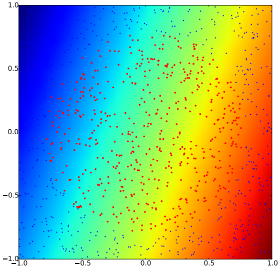
Decision boundary

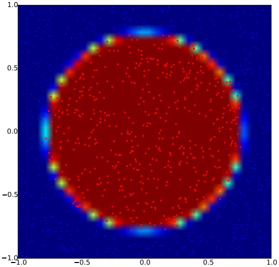
Separate with a straight line?



Decision boundary

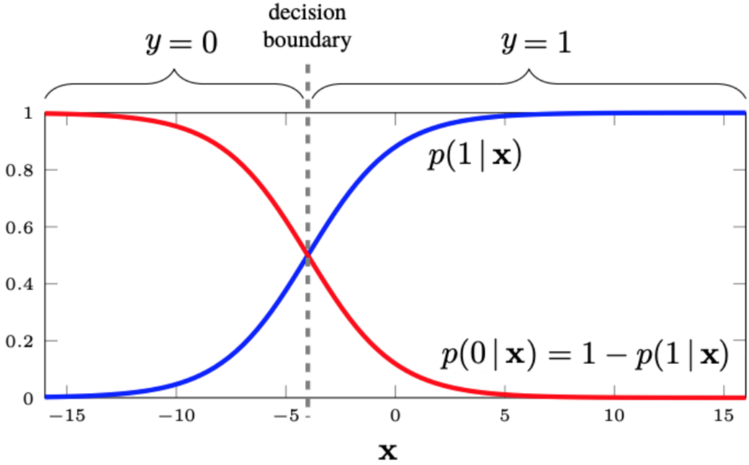
Separate with a straight line?





Decision boundary

- Decision boundary: $w_0 + w_1 x = 0$.
- It is a linear decision boundary.

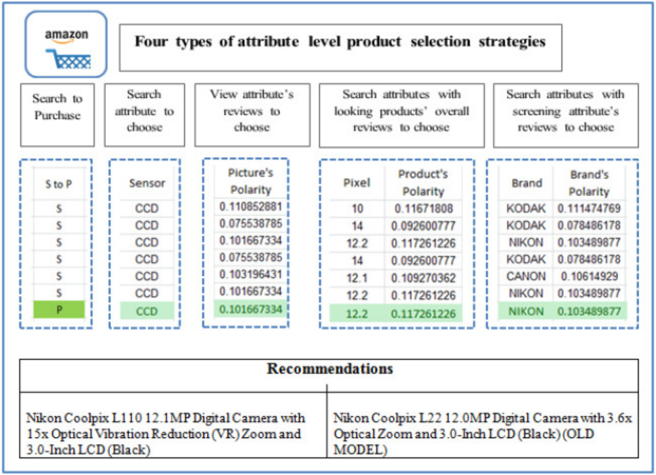


Summary

- Efficient.
- Give probabilistic interpretation of outputs for categorical classification problems.
- Easily regularised to prevent overfitting.
- Linear decision boundary can be insufficient for complicated problems.

Example 1

For customer online purchase intent.



Four types of attribute level product selection strategies

Search to Purchase	Search attribute to choose	View attribute's reviews to choose	Search attributes with looking products' overall reviews to choose	Search attributes with screening attribute's reviews to choose

S to P
S
S
S
S
S
S
P

Sensor
CCD
CCD
CCD
CCD
CCD
CCD
CCD

Picture's Polarity
0.110852881
0.075538785
0.101667334
0.075538785
0.103196431
0.101667334
0.101667334

Pixel	Product's Polarity
10	0.11671808
14	0.092600777
12.2	0.117261226
14	0.092600777
12.1	0.109270362
12.2	0.117261226
12.2	0.117261226

Brand	Brand's Polarity
KODAK	0.111474769
KODAK	0.078486178
NIKON	0.103489877
KODAK	0.078486178
CANON	0.10614929
NIKON	0.103489877
NIKON	0.103489877

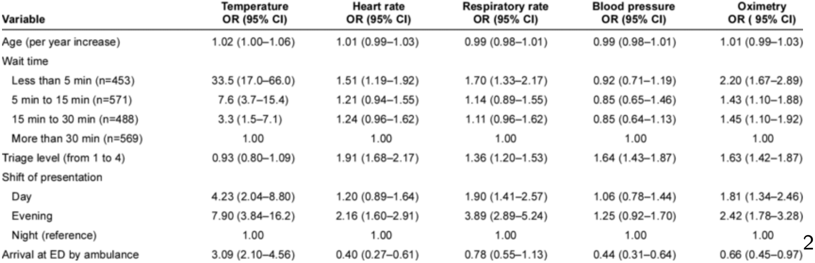
Recommendations	
Nikon Coolpix L110 12.1MP Digital Camera with 15x Optical Vibration Reduction (VR) Zoom and 3.0-Inch LCD (Black)	Nikon Coolpix L22 12.0MP Digital Camera with 3.6x Optical Zoom and 3.0-Inch LCD (Black) (OLD MODEL)

[1]

[1] Bag, S., Tiwari, M. K., Chan, F. T. (2019). Predicting the consumer's purchase intention of durable goods: An attribute-level analysis. Journal of Business Research, 94, 408-419.

Example 2

In disease diagnostic.



Variable	Temperature OR (95% CI)	Heart rate OR (95% CI)	Respiratory rate OR (95% CI)	Blood pressure OR (95% CI)	Oximetry OR (95% CI)
Age (per year increase)	1.02 (1.00–1.06)	1.01 (0.99–1.03)	0.99 (0.98–1.01)	0.99 (0.98–1.01)	1.01 (0.99–1.03)
Wait time					
Less than 5 min (n=453)	33.5 (17.0–66.0)	1.51 (1.19–1.92)	1.70 (1.33–2.17)	0.92 (0.71–1.19)	2.20 (1.67–2.89)
5 min to 15 min (n=571)	7.6 (3.7–15.4)	1.21 (0.94–1.55)	1.14 (0.89–1.55)	0.85 (0.65–1.46)	1.43 (1.10–1.88)
15 min to 30 min (n=488)	3.3 (1.5–7.1)	1.24 (0.96–1.62)	1.11 (0.96–1.62)	0.85 (0.64–1.13)	1.45 (1.10–1.92)
More than 30 min (n=569)	1.00	1.00	1.00	1.00	1.00
Triage level (from 1 to 4)	0.93 (0.80–1.09)	1.91 (1.68–2.17)	1.36 (1.20–1.53)	1.64 (1.43–1.87)	1.63 (1.42–1.87)
Shift of presentation					
Day	4.23 (2.04–8.80)	1.20 (0.89–1.64)	1.90 (1.41–2.57)	1.06 (0.78–1.44)	1.81 (1.34–2.46)
Evening	7.90 (3.84–16.2)	2.16 (1.60–2.91)	3.89 (2.89–5.24)	1.25 (0.92–1.70)	2.42 (1.78–3.28)
Night (reference)	1.00	1.00	1.00	1.00	1.00
Arrival at ED by ambulance	3.09 (2.10–4.56)	0.40 (0.27–0.61)	0.78 (0.55–1.13)	0.44 (0.31–0.64)	0.66 (0.45–0.97)

[2]

[2] Gravel, J., Opatrny, L., Gouin, S. (2006) High rate of missing vital signs data at triage in a paediatric emergency department. Paediatrics child health, 11(4), 211-215.

Example 3

In fraud detection.

The datasets contains transactions made by credit cards in September 2013 by european cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains 28 numerical input variables which are the result of a PCA transformation.

Check Kaggle competition:
`https://www.kaggle.com/mlg-ulb/creditcardfraud`.

Questions

- What is the relationship between logistic regression, linear regression and generalised linear model.
- Which kinds of problem is suitable for applying logistic regression?
- Logistic regression applied to multi-class classification problem?

Reading list

MUR Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*, chapter 8.

BIS Christopher Bishop. *Pattern Recognition and Machine Learning*, section 4.3.