Logistic Regression

Nội dung

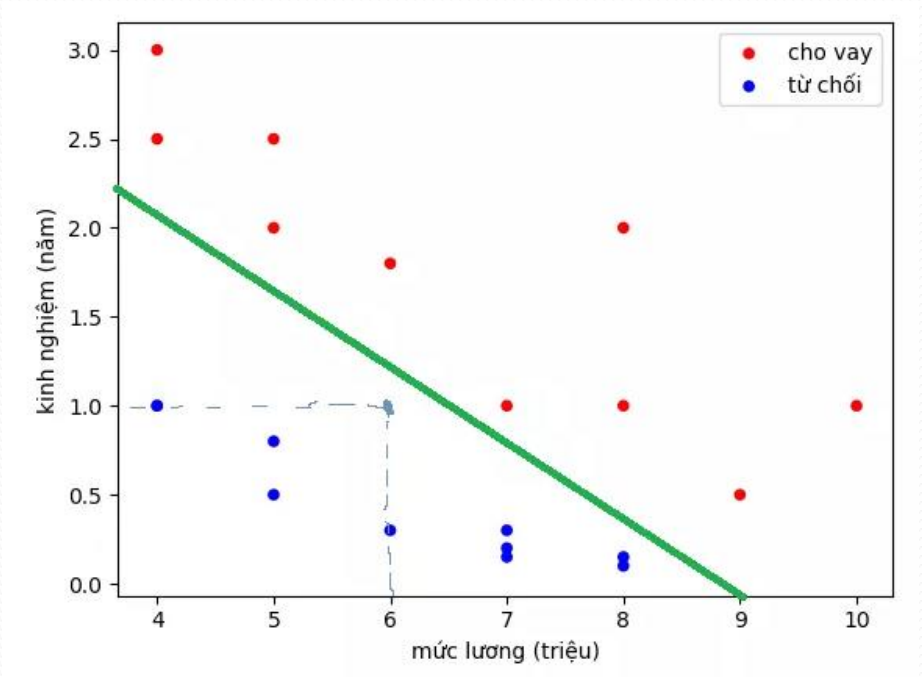
- Khái niệm hồi qui logistic (Logistic Regression)
- Mô hình hóa
- Sigmoid function
- Logistic Regression và bài toán phân loại 2 lớp
  - Logistic Regression dùng SGD
- Mở rộng
- Bài Tập

Logistic Regression

- Phương pháp hồi quy logistic là một mô hình hồi quy nhằm dự đoán output *rời rạc* (*discrete target variable*) **y** ứng với một vector input **x**.
- Việc này tương đương với chuyện phân loại các **x** vào các nhóm **y** tương ứng.
- Thường dùng trong *binary classification*. Có thể mở rộng cho multiclass (softmax regression)

Logistic Regression

- Ví dụ: Ngân hàng có chương trình cho vay ưu đãi cho các đối tượng mua chung cư. Số lượng hồ sơ gửi về 1000-2000 hồ sơ mỗi ngày.
  - Input: mức lương và thời gian công tác
  - Output: cho vay hoặc từ chối



thời kỳ khó khăn nên việc cho vay bị thắt lại, chỉ những hồ sơ nào chắc chắn trên **80%** mới được vay.

cần tìm xác xuất nên cho hồ sơ ấy vay là bao nhiêu

Logistic Regression

- Modeling:

Linear Regression: $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$

Output của logistic regression thường được viết chung dưới dạng:

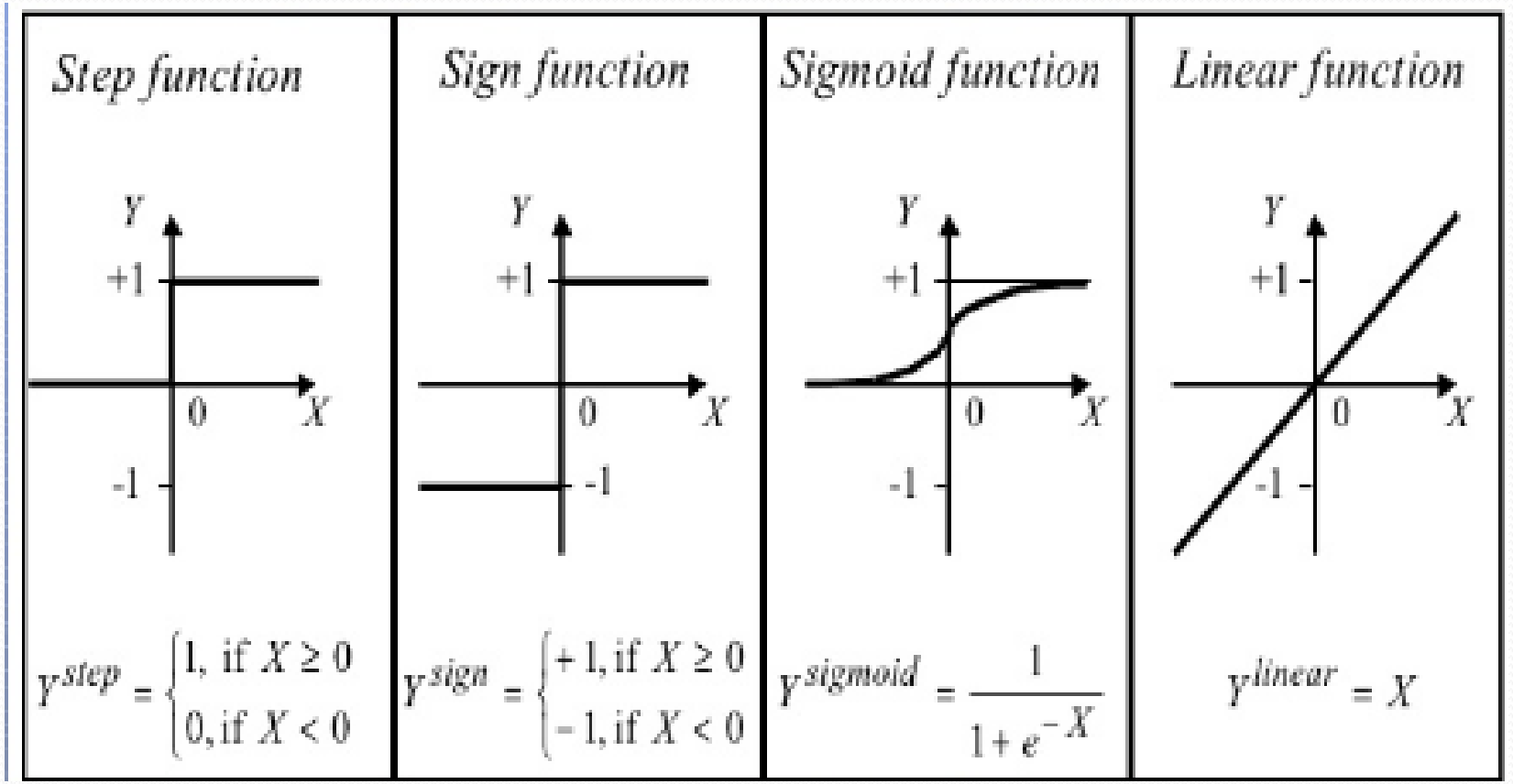
$$f(\mathbf{x}) = \theta(\mathbf{w}^T\mathbf{x})$$

Trong đó θ được gọi là logistic function

Tổng quát $\theta(.)$ được gọi là một **activation function** (hàm kích hoạt)

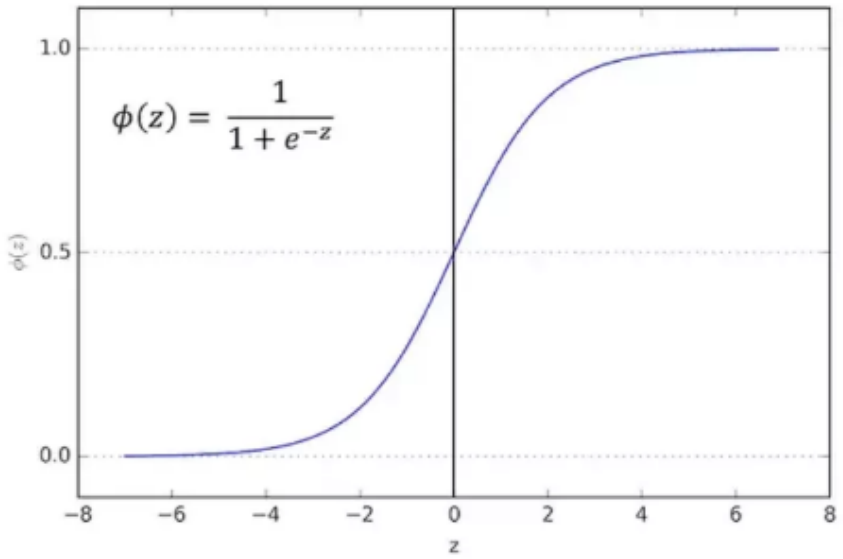
Logistic Regression

- Ví dụ: Một số activation function phổ biến



Logistic Regression

- Sigmoid function
- Ví dụ: cần tìm xác xuất của hồ sơ mới nên cho vay. Hay giá trị của hàm cần trong khoảng [0,1]. Rõ ràng là giá trị của phương trình đường thẳng như bài trước có thể ra ngoài khoảng [0,1] nên cần một hàm mới luôn có giá trị trong khoảng [0,1]



Đồ thị hàm sigmoid

Logistic Regression

- Sigmoid function

$$f(s) = \frac{1}{1+e^{-s}} \triangleq \sigma(s)$$

  - bị chặn trong khoảng (0,1)

$$\lim_{s\to-\infty} \sigma(s) = 0; \quad \lim_{s\to+\infty} \sigma(s) = 1$$

  - có đạo hàm tại mọi điểm

(có thể áp dụng gradient descent)

$$\begin{aligned}\sigma'(s) &= \frac{e^{-s}}{(1+e^{-s})^2} \\ &= \frac{1}{1+e^{-s}}\frac{e^{-s}}{1+e^{-s}} \\ &= \sigma(s)(1-\sigma(s))\end{aligned}$$

Logistic Regression

- Modeling:
- Xem xét bài toán binary classification (phân loại 2 lớp, 0 và 1)
- Giả sử rằng xác suất để một điểm dữ liệu **x** rơi vào
  - class **1** là $f(\mathbf{w}^T\mathbf{x})$
  - class 0 là $1 - f(\mathbf{w}^T\mathbf{x})$
- Dựa vào dữ liệu training (đã biết output **y** và input **x**), ta có thể viết như sau

$$P(y_i = 1|\mathbf{x}_i; \mathbf{w}) = f(\mathbf{w}^T\mathbf{x}_i)$$
$$P(y_i = 0|\mathbf{x}_i; \mathbf{w}) = 1 - f(\mathbf{w}^T\mathbf{x}_i)$$

$P(y_i = 1|\mathbf{x}_i; \mathbf{w})$ được hiểu là xác suất xảy ra sự kiện đầu ra $y_i=1$ khi biết tham số mô hình **w** và dữ liệu đầu vào $\mathbf{x}_i$

Logistic Regression

- Modeling:

- Goal: tìm các hệ số **w** sao cho $f(\boldsymbol{w}^T\boldsymbol{x}_i)$ càng gần với 1 càng tốt với các điểm dữ liệu thuộc class 1 và càng gần với 0 càng tốt với những điểm thuộc class 0.

- Ví dụ : Nếu $f(\boldsymbol{w}^T\boldsymbol{x}_i) \geq \varepsilon$ thì $\boldsymbol{x}_i \in$ class 1

  Nếu $f(\boldsymbol{w}^T\boldsymbol{x}_i) < \varepsilon$ thì $\boldsymbol{x}_i \in$ class 0

Logistic Regression

- Modeling:
- Giả sử $z_i = f(\mathbf{w}^T \mathbf{x}_i)$

$$P(y_i = 1|\mathbf{x}_i; \mathbf{w}) = f(\mathbf{w}^T \mathbf{x}_i)$$
$$P(y_i = 0|\mathbf{x}_i; \mathbf{w}) = 1 - f(\mathbf{w}^T \mathbf{x}_i)$$

$$P(y_i|\mathbf{x}_i; \mathbf{w}) = z_i^{y_i}(1 - z_i)^{1-y_i}$$

- Xem xét toàn bộ mẫu trong tập huấn luyện (training data)

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{d \times N} \text{ và } \mathbf{y} = [y_1, y_2, \ldots, y_N],$$

cần tìm **w** để biểu thức sau đây đạt giá trị lớn nhất: $\mathbf{w} = \arg\max_{\mathbf{w}} P(\mathbf{y}|\mathbf{X}; \mathbf{w})$

Logistic Regression

- Vấn đề trên được gọi là bài toán ***maximum likelihood estimation*** với hàm số phía sau argmax được gọi là ***likelihood function.***
- Giả sử các điểm dữ liệu được sinh ra một cách ngẫu nhiên độc lập với nhau (independent)

$$P(\mathbf{y}|\mathbf{X};\mathbf{w}) = \prod_{i=1}^{N} P(y_i|\mathbf{x}_i;\mathbf{w})$$
$$= \prod_{i=1}^{N} z_i^{y_i}(1-z_i)^{1-y_i}$$

$\prod$ là ký hiệu của tích

Logistic Regression

- Modeling:
- Quan sát:

$$P(\mathbf{y}|\mathbf{X};\mathbf{w}) = \prod_{i=1}^{N} P(y_i|\mathbf{x}_i;\mathbf{w})$$
$$= \prod_{i=1}^{N} z_i^{y_i}(1-z_i)^{1-y_i}$$

  - N >> : tích của NN số nhỏ hơn 1 có thể dẫn tới sai số trong tính toán (numerial error) vì tích là một số quá nhỏ.
  - Dùng *logarit likelihood function* tránh việc số quá nhỏ.

Logistic Regression

- Modeling:
- **Loss function** (hàm chi phí, hàm mất mát) được định nghĩa bởi

$$J(\mathbf{w}) = -\log P(\mathbf{y}|\mathbf{X};\mathbf{w})$$

$$= -\sum_{i=1}^{N}(y_i \log z_i + (1-y_i)\log(1-z_i))$$

$\mathbf{z_i}$ là một hàm số của **w**, $z_i = f(\mathbf{w}^T\mathbf{x}_i)$

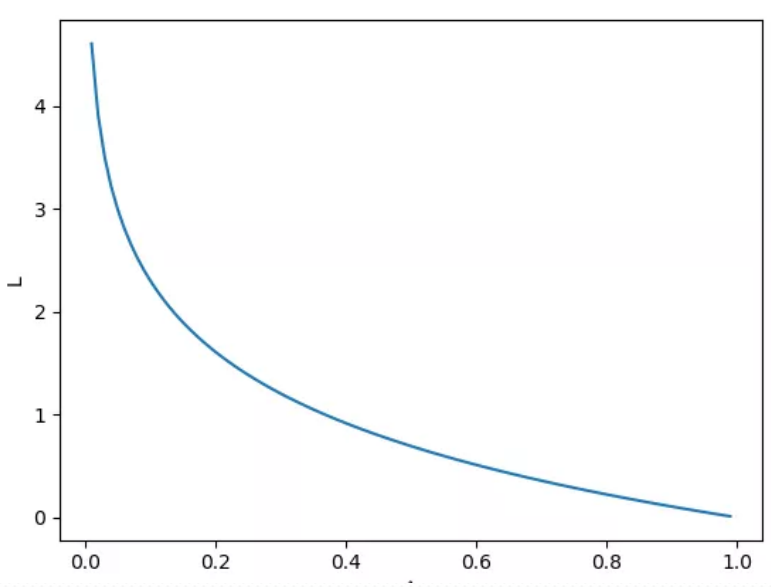
- Dấu " - " để chuyển bài toán ***maximum likelihood estimation*** và dạng **miniminze loss function**

Logistic Regression

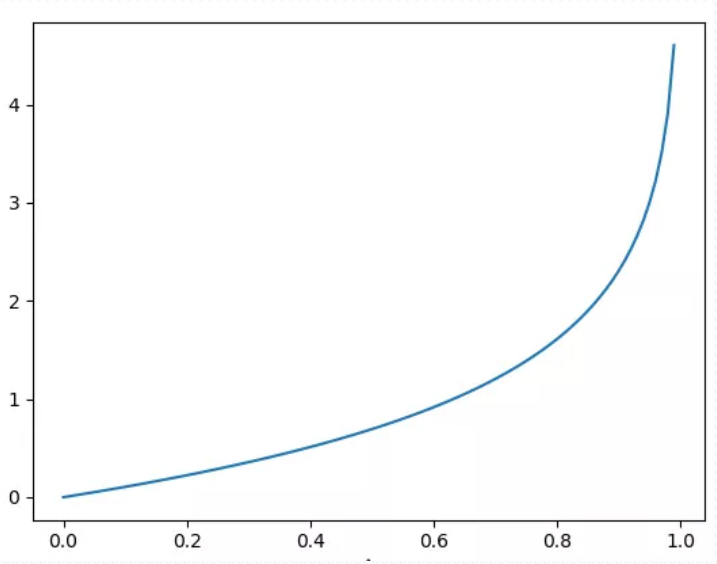
- Ví dụ

$y_i = 1$ thì $J = -\log(z_i)$

$$z_i = f(\mathbf{w}^T \mathbf{x}_i)$$



loss function trong trường hợp $y_i = 1$



loss function trong trường hợp $y_i = o$

Logistic Regression

- Optimize loss function: sử dụng phương pháp Stochastic Gradient Descent (SGD)
- Xem xét : Loss function với chỉ một điểm dữ liệu (x_i, y_i) là

$$J(\mathbf{w}; \mathbf{x}_i, y_i) = -(y_i \log z_i + (1 - y_i) \log(1 - z_i))$$

Đạo hàm theo **w**:

(dựa vào chain rule)

$$\frac{\partial J(\mathbf{w}; \mathbf{x}_i, y_i)}{\partial \mathbf{w}} = -\left(\frac{y_i}{z_i} - \frac{1 - y_i}{1 - z_i}\right) \frac{\partial z_i}{\partial \mathbf{w}}$$

$$= \frac{z_i - y_i}{z_i(1 - z_i)} \frac{\partial z_i}{\partial \mathbf{w}}$$

Logistic Regression

$$\frac{\partial J(\mathbf{w}; \mathbf{x}_i, y_i)}{\partial \mathbf{w}} = -\left(\frac{y_i}{z_i} - \frac{1 - y_i}{1 - z_i}\right)\frac{\partial z_i}{\partial \mathbf{w}}$$

$$= \frac{z_i - y_i}{z_i(1 - z_i)}\frac{\partial z_i}{\partial \mathbf{w}}$$

- Dựa vào sigmoid function

$$z_i = f(\mathbf{w}^T \mathbf{x}_i)$$

$$z_i = \frac{1}{1 + e^{-w^T x_i}} \qquad \frac{\partial z_i}{\partial w} = z_i(1 - z_i)$$

Khi đó:

$$\frac{\partial J(\mathbf{w}; \mathbf{x}_i, y_i)}{\partial \mathbf{w}} = (z_i - y_i)\mathbf{x}_i$$

Logistic Regression

- Công thức cập nhật (theo thuật toán Stochastic Gradient Descent (SGD) cho logistic regression là

$$\mathbf{w} = \mathbf{w} + \eta(y_i - z_i)\mathbf{x}_i$$

Trong đó:

$$z_i = \frac{1}{1+e^{-w^T x_i}}$$

$$\frac{\partial J(\mathbf{w}; \mathbf{x}_i, y_i)}{\partial \mathbf{w}} = (z_i - y_i)\mathbf{x}_i$$

Logistic Regression dùng SGD

- Khởi tạo ngẫu nhiên giá trị w_o:
- Tính loss function

$$J(\mathbf{w}; \mathbf{x}_i, y_i) = -(y_i \log z_i + (1 - y_i) \log(1 - z_i))$$

- Lặp (cho đến khi loss hội tụ hoặc số lượng vòng lặp vượt quá một ngưỡng)

  { Đối với mỗi sample trong training data

  Cập nhật

$$\mathbf{w} = \mathbf{w} + \eta(y_i - z_i)\mathbf{x}_i$$

  }

Logistic Regression

❑ Tính chất:

- Logistic Regression được sử dụng nhiều trong các bài toán Classification.
- Việc xác định class y cho một điểm dữ liệu x được xác định bằng việc so sánh hai biểu thức xác suất

$$P(y = 1|\mathbf{x}; \mathbf{w}); \quad P(y = 0|\mathbf{x}; \mathbf{w})$$

- Nếu biết $\boldsymbol{x_i}$ và $\mathbf{w}$, công thức xác suất được tính dựa vào sigmoid function

$$P(y_i = 1 | w; x_i) = f(w^T x_i) = \frac{1}{1 + e^{-w^T x_i}}$$

Logistic Regression

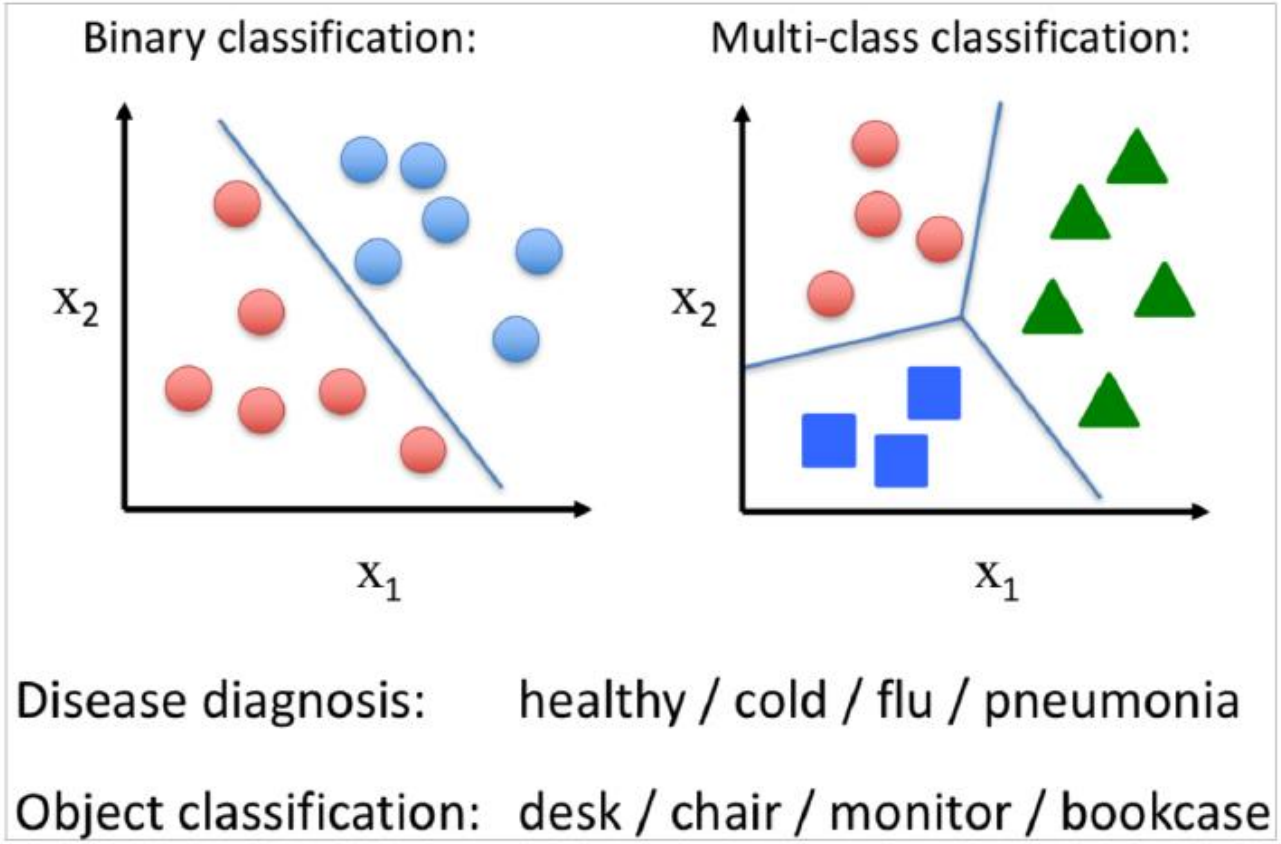
- Tính chất:
- Boundary tạo bởi Logistic Regression có dạng tuyến tính

$$
\begin{aligned}
& P(y=1|\mathbf{x};\mathbf{w}) > 0.5 \\
\Leftrightarrow & \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}} > 0.5 \\
\Leftrightarrow & e^{-\mathbf{w}^T\mathbf{x}} < 1 \\
\Leftrightarrow & \mathbf{w}^T\mathbf{x} > 0
\end{aligned}
$$

boundary giữa hai class là đường có phương trình $\mathbf{w}^T\mathbf{x}$ **(còn gọi là hyperplane)**

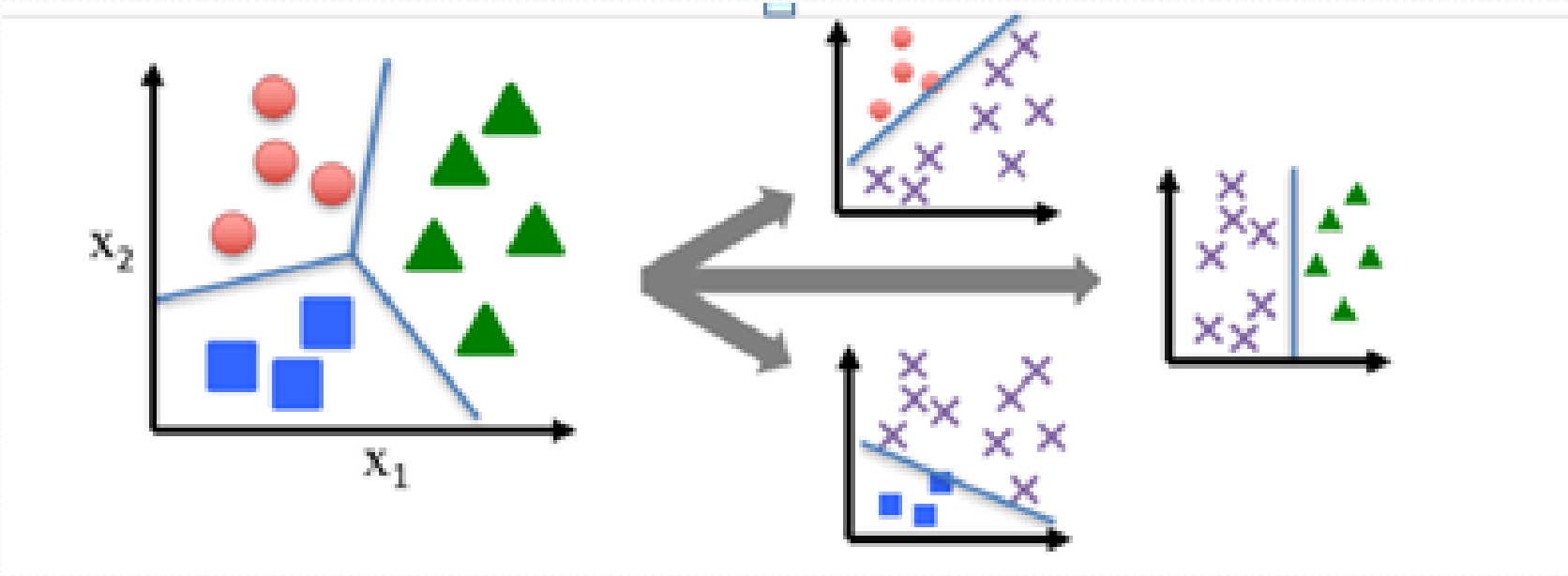
Mở Rộng

- Multi-class classification



Mở Rộng

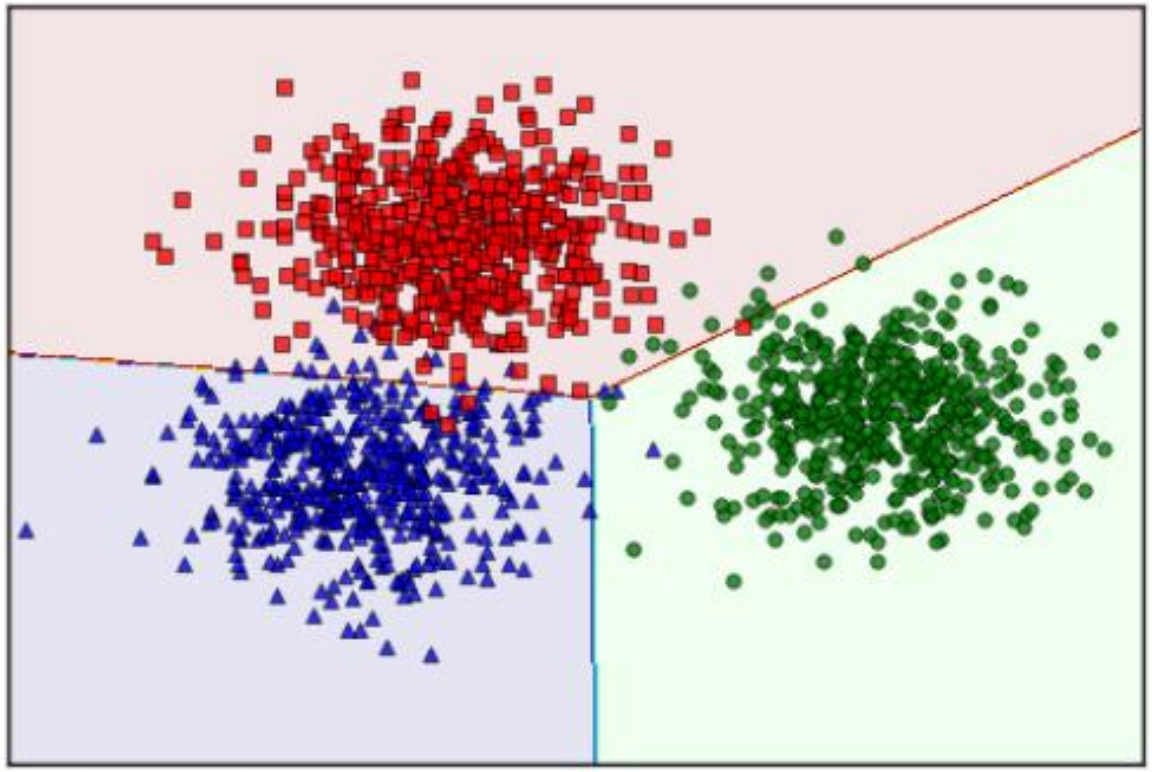
- One-vs-all (one-vs rest)



$$p(y_k|\mathbf{x}) = \max p(y_j|\mathbf{x}) \quad , \forall j = \overline{1, K}$$

Tìm hiểu thêm

- Softmax Regression (Multi-class classification)



Hình 6: Ranh giới giữa các class tìm được bằng Softmax Regression.

Bài Tập

1) Dựa đoán trúng tuyển đại học dựa vào điểm thi. Cài đặt chương trình demo bằng python mô phỏng thuật toán Logistic Regression (dùng thư viện scikit-learn)

Data:

- marks of two exams for 100 applicants
- 1 means the applicant was admitted to the university
- 0 means the applicant didn't get an admission

```
34.62365962451697,78.0246928153624,0
30.2867107682260 7,43.894997524001 01,0
35.84740876993872,72.9021980270836 4,0
60.18259938620976,86.3085520954682 6,1
79.0327360507101,75.3443764369103,1
45.083277476683 39,56.3163717815305,0
61.10666453684766,96.51142588489624,1
75.02474556738889,46.55401354116538,1
76.09878670226257,87.42056971926803,1
84.43281996120035,43.53339331072109,1
```

Bài Tập

2) Phân lại hoa dùng thuật toán Logistic Regression

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html
https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html

- Phân loại hoa

The iris dataset is a classic and very easy multi-class classification dataset.

Classes	3
Samples per class	50
Samples total	150
Dimensionality	4
Features	real, positive