ITF: **Information Technology Faculty**
Thuyloi University

Hồi quy logistic

Trình bày: PGS.TS Nguyễn Hữu Quỳnh

Bài giảng được dựa trên giáo trình machine learning cơ bản

Giới thiệu

Nhắc lại hai mô hình tuyến tính

- Hai mô hình tuyến tính (linear models) Linear Regression và Perceptron Learning Algorithm (PLA) chúng ta đã biết đều có chung một dạng:

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

  - trong đó f() được gọi là *activation function*
  - Với linear regression thì f(s)=s, với PLA thì f(s)=f(s)=sgn(s)

- Trong linear regression, tích vô hướng w^Tx được trực tiếp sử dụng để dự đoán output y, loại này phù hợp nếu chúng ta cần dự đoán một giá trị thực của đầu ra không bị chặn trên và dưới.

- Trong PLA, đầu ra chỉ nhận một trong hai giá trị 1 hoặc −1, phù hợp với các bài toán *binary classification*.

Giới thiệu

- Trong phần này, ta sẽ giới thiệu mô hình có tên là *logistic regression*:
  - Đầu ra có thể được thể hiện dưới dạng xác suất (probability). Ví dụ: xác suất thi đỗ nếu biết thời gian ôn thi, xác suất ngày mai có mưa dựa trên những thông tin đo được trong ngày hôm nay,…
  - L*ogistic regression*:
    - giống với linear regression ở khía cạnh đầu ra là số thực,
    - và giống với PLA ở việc đầu ra bị chặn (trong đoạn [0,1]).
  - Mặc dù trong tên có chứa từ *regression*, logistic regression thường được sử dụng nhiều hơn cho các bài toán classification.

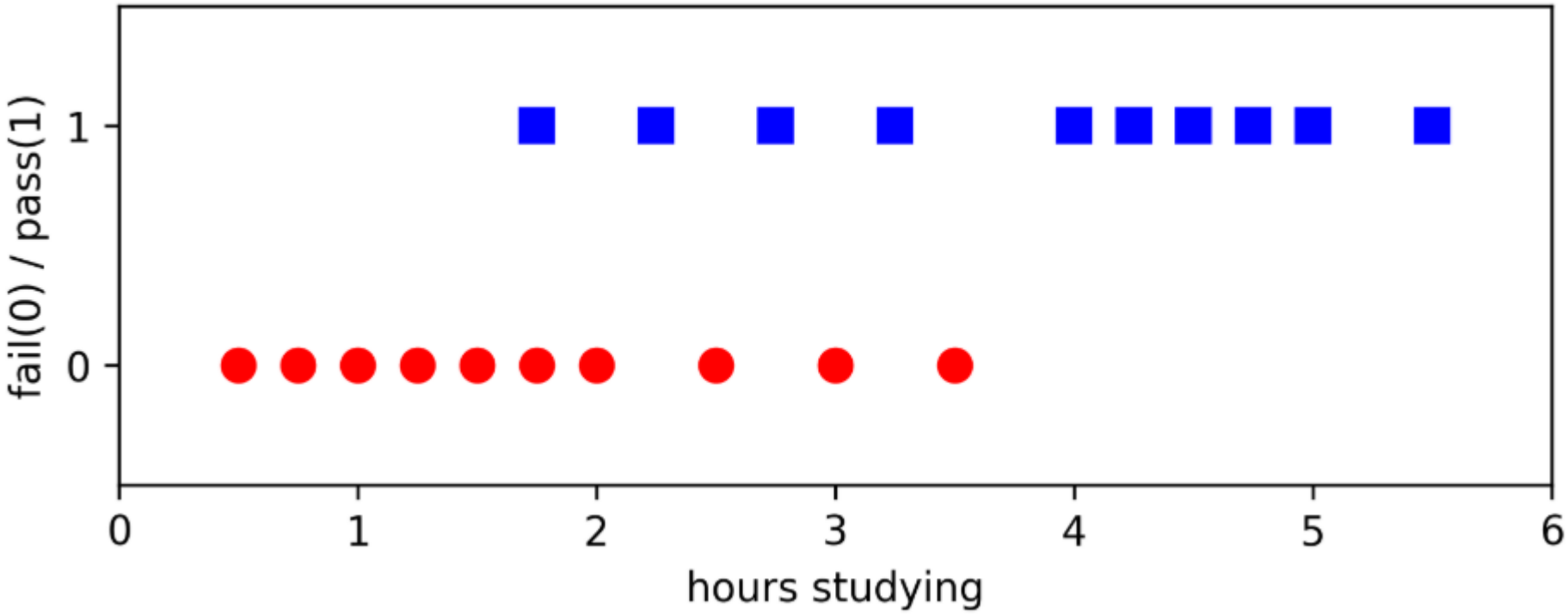
Giới thiệu

Ví dụ: Một nhóm 20 sinh viên dành thời gian trong khoảng từ 0 đến 6 giờ cho việc ôn thi. Thời gian ôn thi này ảnh hưởng đến xác suất sinh viên vượt qua kỳ thi như thế nào?

Kết quả thu được

Hours	Pass	Hours	Pass
.5	0	2.75	1
.75	0	3	0
1	0	3.25	1
1.25	0	3.5	0
1.5	0	4	1
1.75	0	4.25	1
1.75	1	4.5	1
2	0	4.75	1
2.25	1	5	1
2.5	0	5.5	1

Giới thiệu



- Nhận thấy rằng cả linear regression và PLA đều không phù hợp với bài toán này, chúng ta cần một mô hình *flexible* hơn.

Mô hình Logistic Regression

Đầu ra dự đoán của:

- Linear Regression:

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

- PLA:

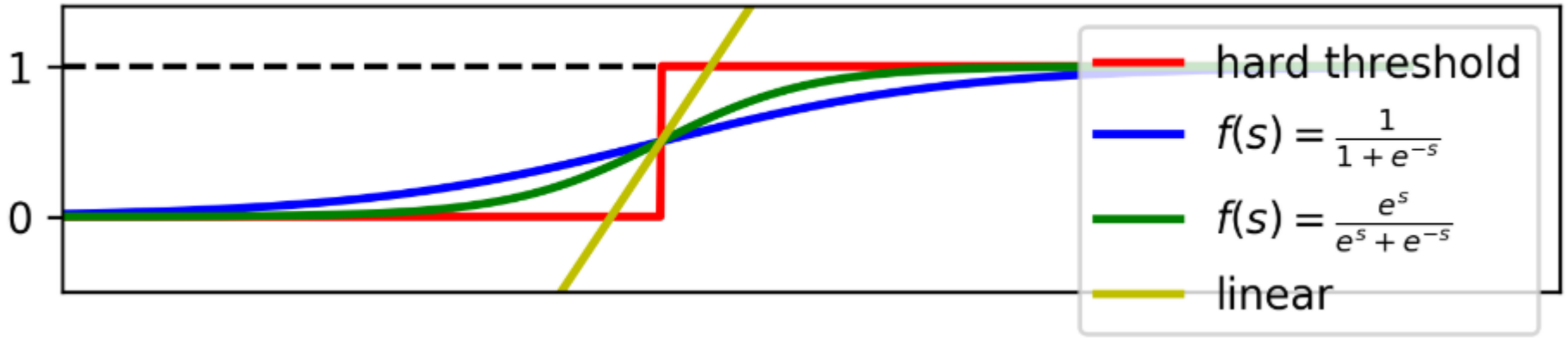
$$f(\mathbf{x}) = \text{sgn}(\mathbf{w}^T\mathbf{x})$$

- Logistic regression

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

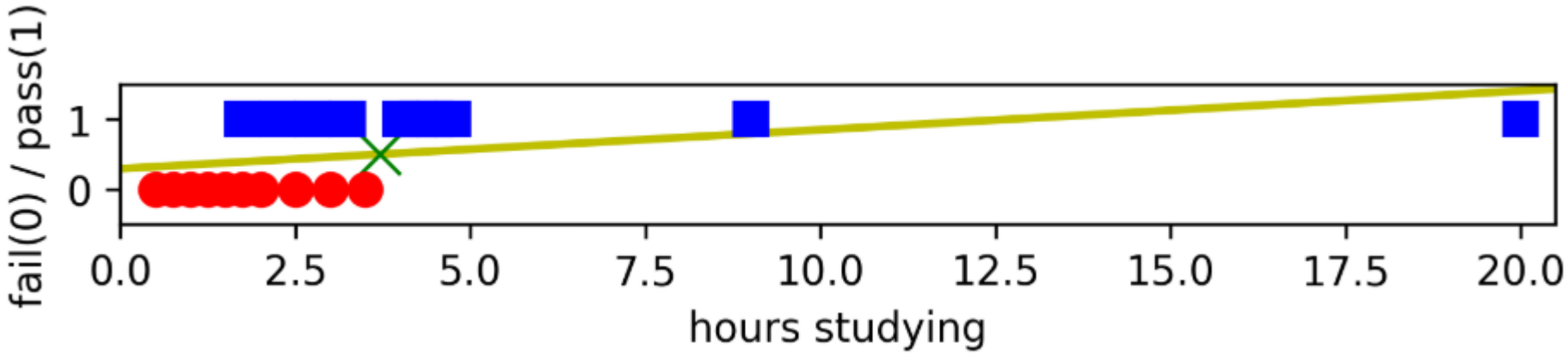
Mô hình Logistic Regression

- Một số activation cho mô hình tuyến tính được cho trong hình dưới đây:



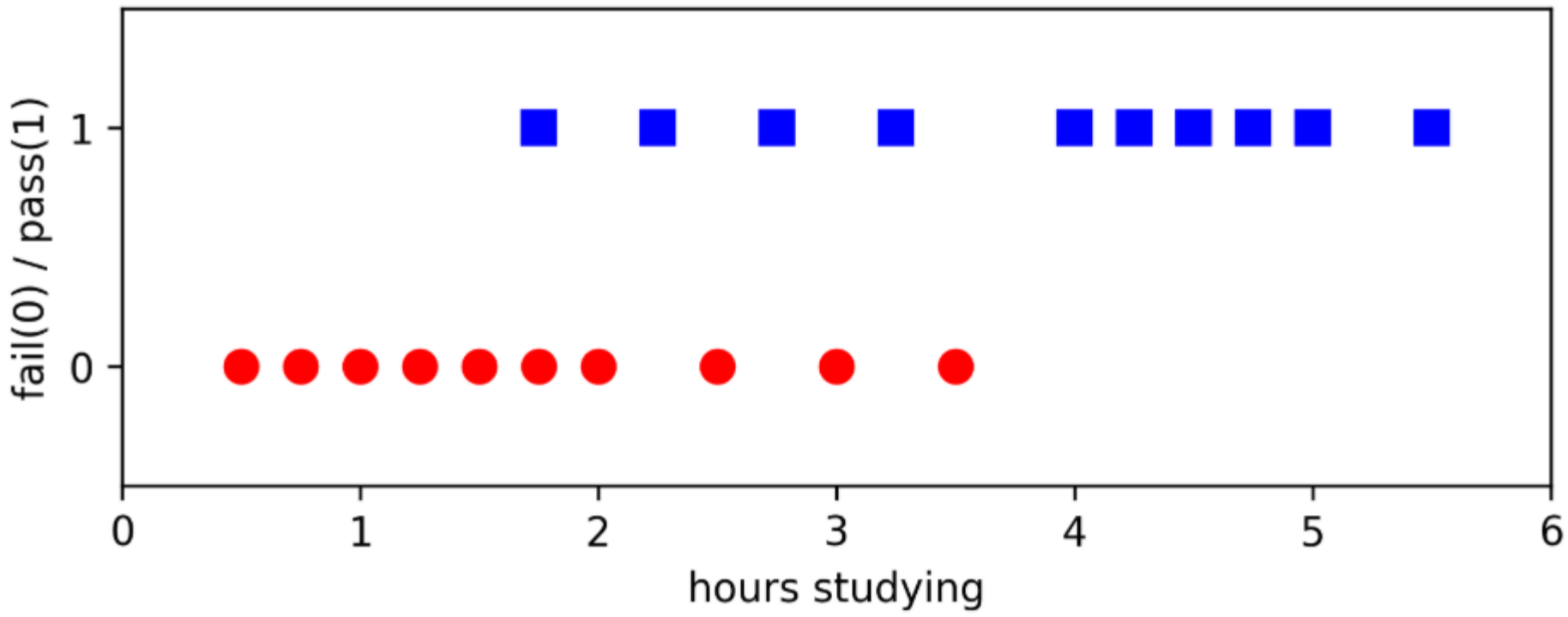
Mô hình Logistic Regression

- Mô hình Logistic Regression không phù hợp cho bài toán này vì:



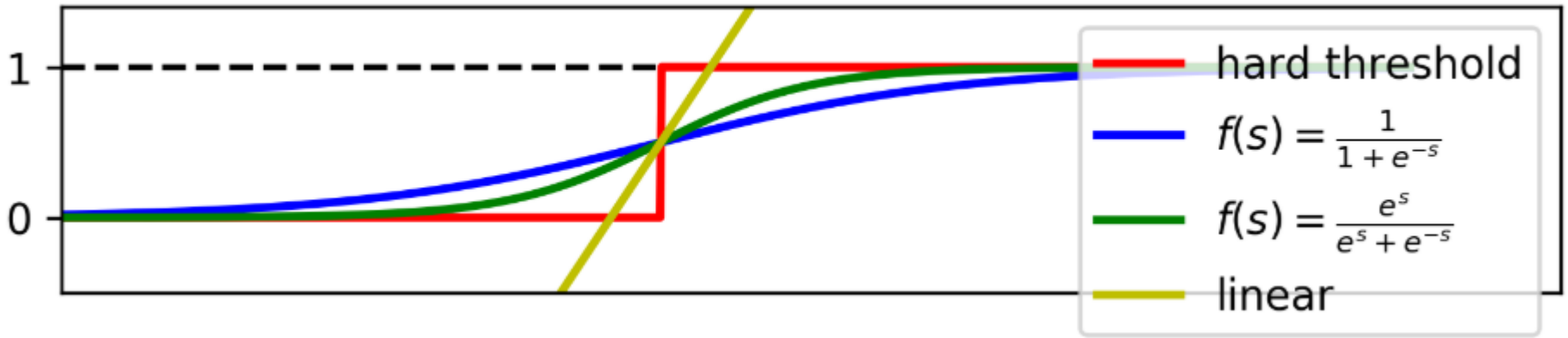
Mô hình Logistic Regression

- PLA không hoạt động trong bài toán này vì dữ liệu đã cho không *linearly separable*.



Mô hình Logistic Regression

- Các đường màu xanh lam và xanh lục phù hợp với bài toán của chúng ta hơn. Chúng có một vài tính chất quan trọng sau:
  - Là hàm số liên tục nhận giá trị thực, bị chặn trong khoảng (0,1).
  - Nếu coi điểm có tung độ là $\frac{1}{2}$ làm điểm phân chia thì các điểm càng xa điểm này về phía bên trái có giá trị càng gần 0. Ngược lại, các điểm càng xa điểm này về phía phải có giá trị càng gần 1. Điều này *khớp* với nhận xét rằng học càng nhiều thì xác suất đỗ càng cao và ngược lại.
  - *Mượt* (smooth) nên có đạo hàm mọi nơi, có thể được lợi trong việc tối ưu.



Mô hình Logistic Regression

- Sigmoid function

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

- Hàm này được sử dụng nhiều nhất vì:
  - nó bị chặn trong khoảng (0,1)
  - Có giới hạn $\lim_{s\to-\infty} \sigma(s) = 0; \quad \lim_{s\to+\infty} \sigma(s) = 1$
  - Có đạo hàm đơn giản

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

Hàm mất mát

Xây dựng hàm mất mát

- Với mô hình (các activation màu xanh lam và lục), ta có thể giả sử rằng xác suất để một điểm dữ liệu x rơi vào class 1 là $f(wTx)$ và rơi vào class 0 là $1 - f(wTx)$.
- Với mô hình được giả sử này, với các điểm dữ liệu training (đã biết đầu ra y), ta có thể viết như sau:

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

- Mục đích của ta là tìm các hệ số w sao cho $f(w^T x_i)$ 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.

Hàm mất mát

- Ký hiệu $z_i = f(w^T x_i)$ và viết gộp lại hai biểu thức

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

ta có:

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

  - Biểu thức ở dưới tương đương với hai biểu thức ở trên vì:
    - khi $y_i = 1$, phần thứ hai của vế phải sẽ triệt tiêu,
    - khi $y_i = 0$, phần thứ nhất sẽ bị triệt tiêu!
  - Chúng ta muốn mô hình gần với dữ liệu đã cho nhất, tức xác suất

$$P(y_i|\mathbf{x}_i; \mathbf{w})$$

đạt giá trị cao nhất.

Hàm mất mát

- Xét toàn bộ training set với $X = [x_1, x_2, \dots, x_N] \in R^{d \times N}$ và $y = [y_1, y_2, \dots, y_N]$, chúng ta cần tìm w để
  - $P(\mathbf{y}|\mathbf{X}; \mathbf{w})$ đạt giá trị lớn nhất hay $\mathbf{w} = \arg\max_{\mathbf{w}} P(\mathbf{y}|\mathbf{X}; \mathbf{w})$
- GS các điểm dữ liệu được sinh ra một cách ngẫu nhiên độc lập với nhau (independent), ta có thể viết:

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

Hàm mất mát

- Để thuận tiện cho việc tính toán, ta tấy log cơ số e ta được

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

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

Tối ưu hàm mất mát

- Hàm mất mát với chỉ một điểm dữ liệu (x_i, yi) là:

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

- Với đạo hàm:

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

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

- Để cho biểu thức này trở nên *gọn* và *đẹp* hơn, chúng ta sẽ tìm hàm $z = f(wTx)$ sao cho mẫu số bị triệt tiêu.

Tối ưu hàm mất mát

- Nếu đặt $s = w^T x$, chúng ta sẽ có:

$$\frac{\partial z_i}{\partial \mathbf{w}} = \frac{\partial z_i}{\partial s}\frac{\partial s}{\partial \mathbf{w}} = \frac{\partial z_i}{\partial s}\mathbf{x}$$

- Một cách trực quan nhất, ta sẽ tìm hàm số $z = f(s)$ sao cho:

$$\frac{\partial z}{\partial s} = z(1-z)$$

Tối ưu hàm mất mát

- Để triệt tiêu mẫu số trong biểu thức

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

- Cho $\frac{\partial z}{\partial s} = z(1-z)$ tương đương với

$$\begin{aligned}
\frac{\partial z}{z(1-z)} &= \partial s \\
\Leftrightarrow (\frac{1}{z} + \frac{1}{1-z})\partial z &= \partial s \\
\Leftrightarrow \log z - \log(1-z) &= s \\
\Leftrightarrow \log \frac{z}{1-z} &= s \\
\Leftrightarrow \frac{z}{1-z} &= e^s \\
\Leftrightarrow z &= e^s(1-z) \\
\Leftrightarrow z = \frac{e^s}{1+e^s} &= \frac{1}{1+e^{-s}} = \sigma(s)
\end{aligned}$$

Công thức cập nhật cho logistic sigmoid regression

- Đến đây, chúng ta có:

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

- Và công thức cập nhật (theo thuật toán SGD) cho logistic regression là:

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