Sigmoid function

Trong số các hàm số có 3 tính chất nói trên thì hàm *sigmoid*:

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

được sử dụng nhiều nhất, vì nó bị chặn trong khoảng $(0,1)$. Thêm nữa:

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

Đặc biệt hơn nữa:

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

Công thức đạo hàm đơn giản thế này giúp hàm số này được sử dụng rộng rãi. Ở phần sau, tôi sẽ lý giải việc *người ta đã tìm ra hàm số đặc biệt này như thế nào*.

Ngoài ra, hàm *tanh* cũng hay được sử dụng:

$$\tanh(s) = \frac{e^{s}-e^{-s}}{e^{s}+e^{-s}} \qquad (2)$$

Hàm số này nhận giá trị trong khoảng $(-1,1)$ nhưng có thể dễ dàng đưa nó về khoảng $(0,1)$. Bạn đọc có thể chứng minh được:

$$\tanh(s) = 2\sigma(2s) - 1$$

$$\tanh(s) = \frac{e^{s}-e^{-s}}{e^{s}+e^{-s}} = \frac{1-e^{-2s}}{1+e^{-2s}} = \frac{2}{1+e^{-2s}} - 1 = 2\sigma(2s) - 1$$

Tối ưu hàm mất mát

Chúng ta lại sử dụng phương pháp Stochastic Gradient Descent (SGD) ở đây (*Bạn đọc được khuyến khích đọc SGD trước khi đọc phần này*) . Hàm mất mát với chỉ một điểm dữ liệu $(\mathbf{x}_i, y_i)$ là:

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

Với đạo hàm:

$$\frac{\partial J(\mathbf{w}; \mathbf{x}_i, y_i)}{\partial \mathbf{w}} = -(\frac{y_i}{z_i} - \frac{1 - y_i}{1 - z_i})\frac{\partial z_i}{\partial \mathbf{w}} = \frac{z_i - y_i}{z_i(1 - z_i)}\frac{\partial z_i}{\partial \mathbf{w}} \quad (3)$$

Để cho biểu thức này trở nên *gọn* và *đẹp* hơn, chúng ta sẽ tìm hàm $z = f(\mathbf{w}^T\mathbf{x})$ sao cho mẫu số bị triệt tiêu. Nếu đặt $s = \mathbf{w}^T\mathbf{x}$, 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:

$$\frac{\partial z}{\partial s} = z(1 - z) \quad (4)$$

để triệt tiêu mẫu số trong biểu thức (3). Chúng ta cùng khởi động một chút với phương trình vi phân đơn giản này. Phương trình (4) tương đương với:

$$\frac{\partial z}{z(1 - z)} = \partial s \Leftrightarrow (\frac{1}{z} + \frac{1}{1 - z})\partial z = \partial s \Leftrightarrow \log z - \log(1 - z) = s \Leftrightarrow \log \frac{z}{1 - z} = s \Leftrightarrow \frac{z}{1 - z} = e^s$$

$$\Leftrightarrow z = e^s(1 - z) \Leftrightarrow z = \frac{e^s}{1 + e^s} = \frac{1}{1 + e^{-s}} = \sigma(s)$$

CONVEX SETS & CONVEX FUNCTIONS

1. Convex sets:

Định nghĩa 1: Một tập hợp được gọi là *tập lồi* (convex set) nếu đoạn thẳng nối hai điểm *bất kỳ* trong tập hợp hợp đó nằm trọn vẹn trong tập hợp đó.

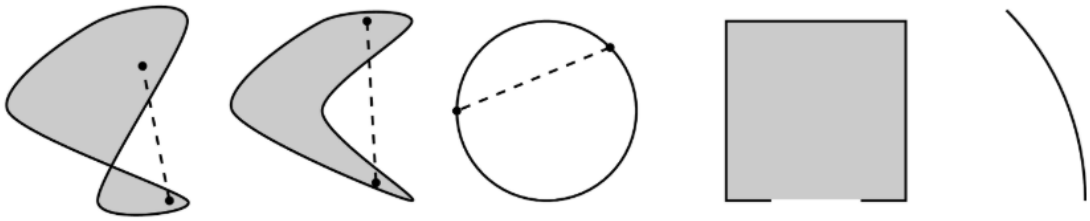
Một vài ví dụ về convex sets:

Example of convex sets

Hình 1: Các ví dụ về convex sets.

Các hình với đường biên màu đen thể hiện việc bao gồm cả biên, biên màu trắng thể hiện việc biên đó không nằm trong tập hợp đang xét. Đường hoặc đoạn thẳng cũng là một tập lồi theo định nghĩa phía trên.

Dưới đây là một vài ví dụ về *nonconvex sets*, tức tập hợp mà không phải là lồi:



Examples of nonconvex sets

Hình 2: Các ví dụ về nonconvex sets.

Định nghĩa 2: Một tập hợp $\mathcal{C}$ được gọi là *convex* nếu với hai điểm bất kỳ $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}$, điểm $\mathbf{x}_\theta = \theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2$ cũng nằm trong $\mathcal{C}$ với bất kỳ $0 \leq \theta \leq 1$.

Có thể thấy rằng, tập hợp các điểm có dạng $(\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2)$ chính là *đoạn thẳng* nối hai điểm $\mathbf{x}_1$ và $\mathbf{x}_2$.

2. Convex functions:

Để trực quan, trước hết ta xem xét các hàm 1 biến, đồ thị của nó là một đường trong một mặt phẳng. Một hàm số được gọi là *lồi* nếu **tập xác định của nó là một tập lồi** và nếu ta nối hai điểm bất kỳ trên đồ thị hàm số đó, ta được một đoạn thẳng nằm về phía trên hoặc nằm trên đồ thị (xem Hình 3)

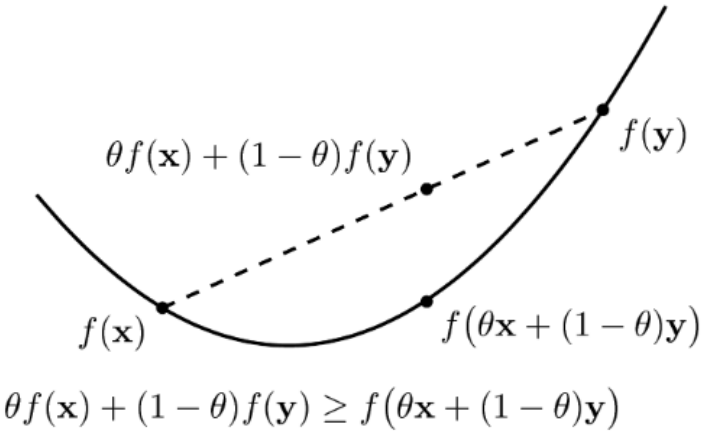
Định nghĩa theo toán học:

Định nghĩa convex function: Một hàm số $f : \mathbb{R}^n \to \mathbb{R}$ được gọi là một *hàm lồi* (convex function) nếu $\mathrm{dom} f$ là một *tập lồi*, và:

$$f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$$

với mọi $\mathbf{x}, \mathbf{y} \in \mathrm{dom} f, 0 \leq \theta \leq 1$.

Điều kiện $\mathrm{dom} f$ là một *tập lồi* là rất quan trọng, vì nếu không có nó, ta không định nghĩa được $f(\theta \mathbf{x} + (1-\theta)\mathbf{y})$.



Hình 3. Convex function.

Một hàm số f được gọi là **concave** (nếu bạn muốn dịch là *lõm* cũng được, tôi không thích cách dịch này) nếu $-f$ là **convex**. Một hàm số có thể không thuộc hai loại trên. Các hàm tuyến tính vừa *convex*, vừa *concave*.

Định nghĩa strictly convex function: (tiếng Việt có một số tài liệu gọi là *hàm lồi mạnh*, *hàm lồi chặt*) Một hàm số $f : \mathbb{R}^n \to \mathbb{R}$ được gọi là *strictly convex* nếu $\mathrm{dom} f$ là một *tập lồi*, và:

$$f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) < \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$$

với mọi $\mathbf{x}, \mathbf{y} \in \mathrm{dom} f, \mathbf{x} \neq \mathbf{y}, 0 < \theta < 1$.

Tương tự với định nghĩa **strictly concave**.

Đây là một điểm quan trọng: **Nếu một hàm số là *strictly convex* và có điểm cực trị, thì điểm cực trị đó là duy nhất và cũng là *global minimum***.

3. Kiểm tra tính chất lồi dựa vào đạo hàm:

a/ First-order condition

Trước hết chúng ta định nghĩa phương trình đường (mặt) tiếp tuyến của một hàm số f khả vi tại một điểm nằm trên đồ thị (mặt) của hàm số đó $(\mathbf{x}_0, f(\mathbf{x}_0))$. Với hàm một biến, bạn đọc đã quen thuộc:

$$y = f'(x_0)(x - x_0) + f(x_0)$$

Với hàm nhiều biến, đặt $\nabla f(\mathbf{x}_0)$ là gradient của hàm số f tại điểm $\mathbf{x}_0$, phương trình mặt tiếp tuyến được cho bởi:

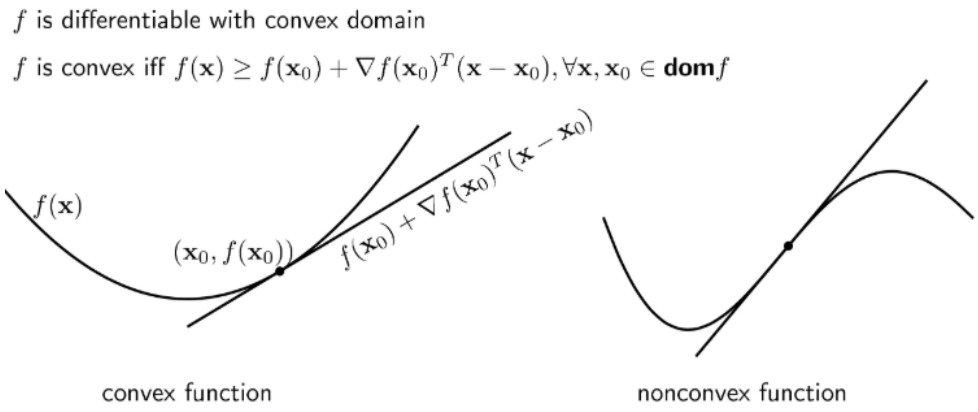
$$y = \nabla f(\mathbf{x}_0)^T (\mathbf{x} - \mathbf{x}_0) + f(\mathbf{x}_0)$$

First-order condition nói rằng: Giả sử hàm số f có tập xác định là một tập lồi, có đạo hàm tại mọi điểm trên tập xác định đó. Khi đó, hàm số f là *lồi* **nếu và chỉ nếu** với mọi $\mathbf{x}, \mathbf{x}_0$ trên tập xác định của hàm số đó, ta có:

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T(\mathbf{x} - \mathbf{x}_0) \quad (*)$$

Tương tự như thế, một hàm số là *stricly convex* nếu dấu bằng trong $(*)$ xảy ra khi và chỉ khi $\mathbf{x} = \mathbf{x}_0$.

Nói một cách trực quan hơn, một hàm số là lồi nếu đường (mặt) tiếp tuyến tại một điểm bất kỳ trên đồ thị (mặt) của hàm số đó **nằm dưới** đồ thị (mặt) đó. (Đừng quên điều kiện về tập xác định là lồi) Dưới đây là ví dụ về *hàm lồi* và *hàm không lồi*.



Hình 4. Kiểm tra tính convexity dựa vào đạo hàm bậc nhất. (Trái: hàm lồi; Phải: hàm không lồi)

Note:

First-order condition ít được sử dụng để tìm tính chất lồi của một hàm số, thay vào đó, người ta thường dùng *Second-order condition* với các hàm có đạo hàm tới bậc hai.

b/ Second-order condition

Với hàm nhiều biến, tức biến là một vector, giả sử có chiều là d, đạo hàm bậc nhất của nó là một vector cũng có chiều là d. Đạo hàm bậc hai của nó là một ma trận vuông có chiều là $d \times d$. Đạo hàm bậc hai của hàm số $f(\mathbf{x})$ được ký hiệu là $\nabla^2 f(\mathbf{x})$. Đạo hàm bậc hai còn được gọi là *Hessian*.

Second-order condition: Một hàm số có đạo hàm bậc hai là *convex* nếu **dom**f là *convex* và Hessian của nó là một ma trận *nửa xác định dương* với mọi $\mathbf{x}$ trong tập xác định:

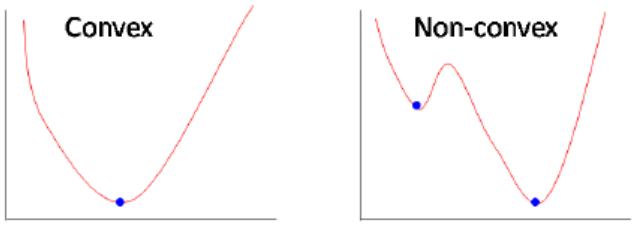
$$\nabla^2 f(\mathbf{x}) \succeq 0.$$

Nếu Hessian là một ma trận *xác định dương* thì hàm số đó *strictly convex*. Tương tự, nếu Hessian là một ma trận *xác định âm* thì hàm số đó là *strictly concave*.

Với hàm số một biến $f(x)$, điều kiện này tương đương với $f''(x) \geq 0$ với mọi x thuộc tập xác định (và tập xác định là *lồi*).

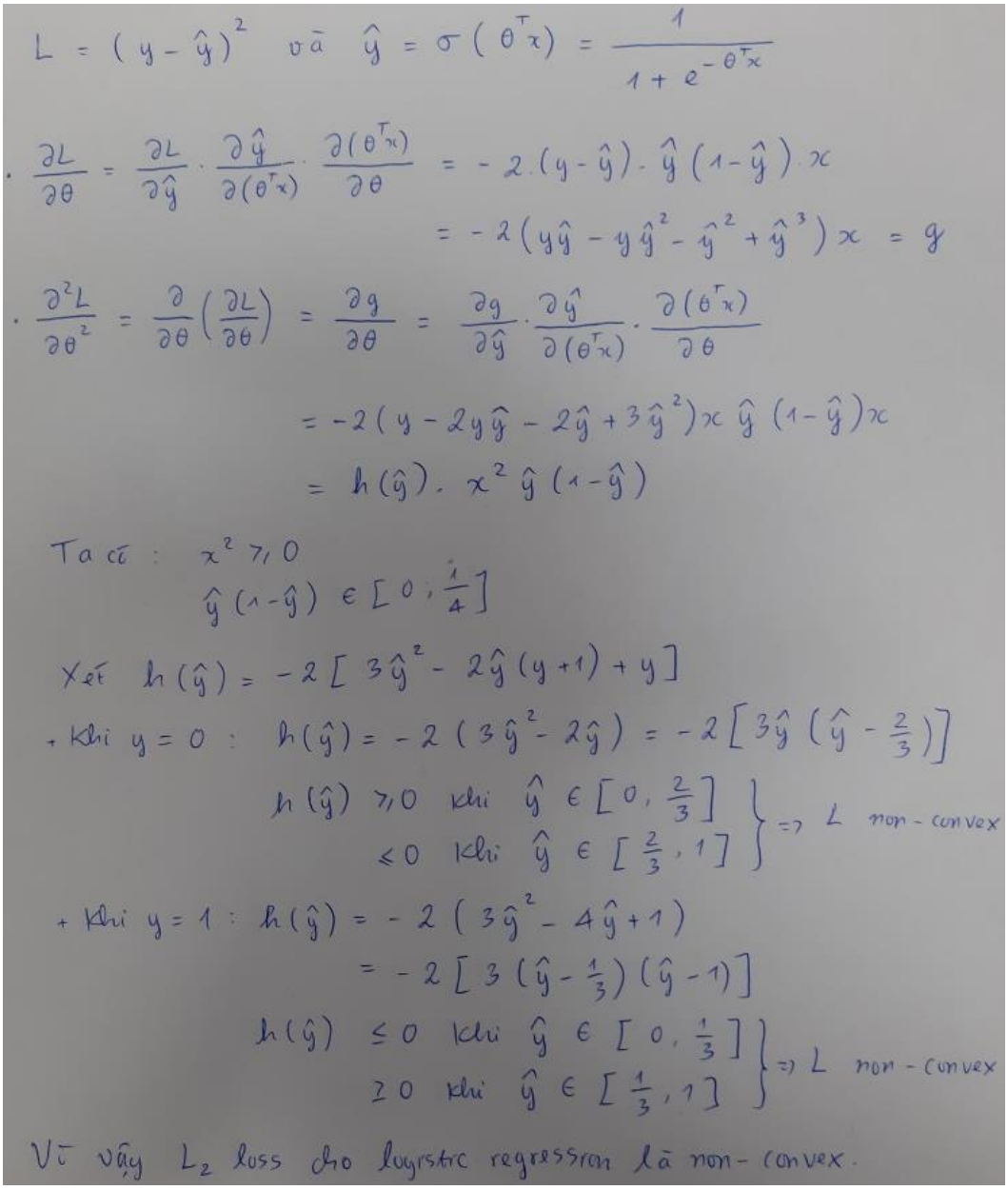
MSE and problem of Non-Convexity in Logistic Regression

Trong các bài toán phân loại, chúng ta thường sử dụng các kỹ thuật dựa trên gradient (Newton Raphson, gradient descent, v.v.) để tìm các giá trị tối ưu cho các hệ số bằng cách giảm thiểu hàm mất mát. Do đó, nếu hàm mất mát không lồi, thì không đảm bảo rằng chúng ta sẽ luôn đạt tới cực tiểu toàn cục, thay vào đó chúng ta có thể bị mắc kẹt ở cực tiểu cục bộ.



Hình 5. Convex and non-Convex functions

1. Chứng minh L2 loss cho logistic regression là non-convex:



2. Chứng minh Binary cross-entropy cho logistic regression là convex: