Convexity of Linear and Logistic Regression

David Zhu | Spring 2017 | Machine Learning Independent Study

Why convexity?

In machine learning, most of the time we would like to minimize a cost function - a function that determines the delta between predicted and actual data. Minimizing the cost function would mean our predictive model would get more accurate at modeling real world behaviors. Using cost functions that are strictly convex may provide handy benefits when training a model, due to the properties of convexity.

When a function is strictly convex, we can be certain that there is one unique global minimum. We can arrive at the absolute minimum regardless of where we begin. For non-convex functions, this is not the case, and therefore optimization techniques may land us in local minima. For strictly convex functions, we could also derive analytical equations to directly calculate the minimum without having to search for it through techniques such as gradient descent. (One caveat, however, is that sometimes these equations become non-performant when the data/features become large due to large matrix calculations.)

Several of the most popular machine learning algorithms, such as linear and logistic regression, are strictly convex. While the Coursera lectures teach the purpose and applications of these techniques, it was beyond their scope to explain the proofs behind why these methods are convex. To expand my learning, I would like to use this write-up to explore these proofs for linear and logistic regression.

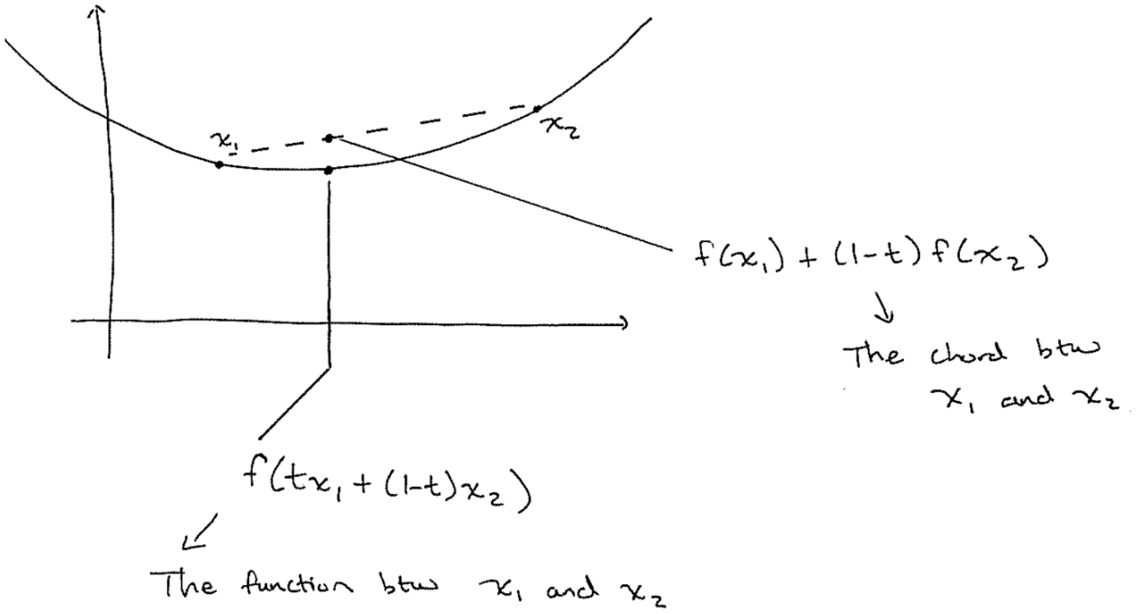
Definition of convexity

First, we need a definition for convexity. A convex function is defined as follows:

if $\forall x_1, x_2 \in \mathcal{X}$ $\forall t \in [0,1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

$f(x)$ is convex.



x_1
x_2
$f(x_1) + (1-t)f(x_2)$
↓
The chord btw x_1 and x_2
$f(tx_1 + (1-t)x_2)$
↙
The function btw x_1 and x_2

This means that for any given range, any point of the function within that range must be below the chord that forms that range.

Using Taylor expansion, we can prove that if the second derivative of a function is always positive, then the function is convex.

If a function f has a 2nd derivative that is positive over an interval, the function is convex over the interval.

First, use Taylor series expansion of function at x_0

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f''(a)}{2}(x-x_0)^2$$

a is btw x_0 and x.

According to the given, $f''(a) \geq 0$.

$\frac{f''(a)}{2}(x-x_0)^2 \geq 0$ since $(x-x_0)^2 \geq 0$. for all x.

$\therefore \quad f(x) \geq f(x_0) + f'(x_0)(x-x_0)$

Let $x_0 = tx_1 + (1-t)x_2$ (placing x_0 between x_1 and x_2.)

①

At $x = x_1$,

$$f(x_1) \geq f(x_0) + f'(x_0)(x_1 - x_0)$$
$$\geq f(x_0) + f'(x_0)(x_1 - x_2)(1-t)$$

where
$$x_1 - x_0 = x_1 - (tx_1 + (1-t)x_2)$$
$$= x_1 - tx_1 - (1-t)x_2$$
$$= x_1(1-t) - x_2(1-t)$$
$$= (x_1 - x_2)(1-t).$$

②

At $x = x_2$,

$$f(x_2) \geq f(x_0) + f'(x_0)(x_2 - x_0)$$
$$\geq f(x_0) + f'(x_0)(t)(x_2 - x_1)$$

where
$$x_2 - x_0 = x_2 - (tx_1 + (1-t)x_2)$$
$$= x_2 - tx_1 - (1-t)x_2$$
$$= \cancel{x_2} - tx_1 - \cancel{x_2} + tx_2$$
$$= t(x_2 - x_1).$$

Now, multiply t to ① and $(1-t)$ to ② and combine.

$$\begin{aligned} t f(x_1) + (1-t) f(x_2) &\geq f(x_0)(t) + f'(x_0)\cancel{(t)(x_1-x_2)}(1-t) \\ &\quad + f(x_0)(1-t) + f'(x_0)(1-t)\cancel{(x_1-x_2)}(t)(-1) \\ &\geq f(x_0) \\ &\geq f(t x_1 + (1-t) x_2) \end{aligned}$$

By definition, $f(x)$ is then convex when $f''(x) \geq 0$

This also makes sense intuitively. If we think about the first derivative as the change in direction of where the function is going, the second derivative would be the change in the rate of change in direction. If the second derivative is always positive, this would mean that the slope is constantly increasing, and there would be no opportunity for the function to turn and form more than one minimum.

Positive semi-definite implies convexity

Positive semi-definite (PSD) is a handy property of a symmetric n x n matrix. You could think of a PSD matrix as analogous to non-negative numbers in the realm of matrices. Following is the definition of a matrix being PSD.

A Hermitian matrix M $(n \times n)$ is PSD if

$x^T M x \geq 0$ for all x in R^n.

A Hessian matrix of a multivariable function *f* organizes all second partial derivatives into a matrix [1]. Hessians are analogous to second derivatives in single variable calculus, and they are useful in

[1] khanacademy.org/math/multivariable-calculus/applications-of-multivariable-derivatives/quadratic-approximations/a/the-hessian

quadratic approximations (analogous to Taylor expansions) and finding max and mins. The Hessian matrix has a special property where if it is PSD, then the function is convex[2].

First, the definition of a convex function:

$$f(tx+(1-t)y) \leq tf(x)+(1-t)f(y)$$

Let's fix x and y to look at the one dimension case.

Let $g(t) = f(tx+(1-t)y)$

Using the chain rule,

$$g'(t) = (x-y)^T \nabla f(tx+(1-t)y)$$

$$g''(t) = (x-y)^T \nabla^2 f(tx+(1-t)y)(x-y)$$

Given that the Hessian $\nabla^2 f$ is positive semi-definite,

or $\nabla^2 f \succeq 0$, then:

$$a^T \nabla^2 f a \geq 0 \qquad \text{(definition)}$$

where $a = x-y$.

Then $g''(t) \geq 0$.

As shown before, we can use the Taylor expansion to show that $g(x)$ is convex.

[2] https://math.stackexchange.com/a/946484/443165

If we can show that the second derivative, or Hessian, of our function to be PSD, then we can demonstrate that the function is convex. In the following, we will prove that linear and logistic regression are convex by demonstrating that their Hessian matrix is positive semi-definite.

Linear regression

Linear regression is a popular algorithm, generally used to for predicting outcomes across a continuous space. Its cost function can be defined as a least squares between outcomes and predictions.

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Basic form

Let’s first consider the convexity of a basic form.

First, let's only consider

$$J(\theta_0, \theta_1) = \frac{1}{2}(\theta_0 + \theta_1 x - y)^2$$

Then,

$$\frac{\partial J}{\partial \theta_0} = \theta_0 \qquad \frac{\partial J}{\partial \theta_1} = x(x\theta_1) = x^2\theta$$

$$H(J) = \begin{pmatrix} 1 & 0 \\ 0 & x^2 \end{pmatrix}$$

Using the determinant test, the determinants from the upper left are:

$1, x^2 \implies$ $1 \geq 0$
$x^2 \geq 0.$ Therefore H is PSD.

Therefore J is convex.

Matrix form

Now let's look at the convexity of lin. reg.
in matrix form.

$$J(\theta) = \frac{1}{2m}\|X\theta - y\|^2 = \frac{1}{2m}(X\theta - y)^T(X\theta - y).$$

$$J(\theta) = \frac{1}{2m}((X\theta)^T - y^T)(X\theta - y) \qquad \text{(transpose identity)}$$

Drop the $\frac{1}{2m}$ since we only care if $\nabla^2 J$ is PSD, for convenience

$$J(\Theta) \propto (X\Theta)^T X\Theta - \underbrace{(X\Theta)^T y}_{} - \underbrace{y^T (X\Theta)}_{} + y^T y$$

These two can be combined
b/c we don't care which order
$X\Theta$ and y multiply through the transpose.

$$J(\Theta) \propto \Theta^T X^T X\Theta - 2(X\Theta)^T y + y^T y$$

Now to calculate the gradient.

First, let $A(\Theta) = 2(X\Theta)^T y$

$$A(\Theta) = 2\left[\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & \ddots & & \\ & & \ddots & \\ x_{m1} & & & x_{mn} \end{pmatrix} \begin{pmatrix} \Theta_1 \\ \Theta_2 \\ \vdots \\ \Theta_n \end{pmatrix} \right]^T \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$A(\Theta) = 2\left((x_{11}\Theta_1 + \ldots x_{1n}\Theta_n)\, y_1 + \ldots (x_{m1}\Theta_1 + \ldots x_{mn}\Theta_n)\, y_m \right)$$

$$= 2\sum_{r=1}^{m} y_r (x_{r1}\Theta_1 + \ldots x_{rn}\Theta_n) = 2\sum_{r=1}^{m} y_r \sum_{c=1}^{n} x_{rc}\Theta_c$$

$$\nabla_\theta A = \begin{pmatrix} 2\sum_{r=1}^{m} y_r x_{r1} \\ \vdots \\ 2\sum_{r=1}^{m} y_r x_{rn} \end{pmatrix} = 2X^T y.$$

Now let $B(\Theta) = \Theta^T x^T x \Theta$

Dimensions of x are $n \times m$

of Θ are $n \times 1$

of $x^T x$ are $n \times n$

of $x^T x \Theta$ are $n \times 1$

of $\Theta^T x^T x \Theta$ are 1×1

$$B(\Theta) = \Theta_1 (x^2_{11} \Theta_1 + x^2_{1n} \Theta_n) + \dots + \Theta_n (x^2_{n1} \Theta_1 + \dots + x^2_{nn} \Theta_n)$$

$$= \sum_{i=1}^{n} \Theta_i \sum_{j=1}^{n} x^2_{ij} \Theta_j = \sum_{i=1}^{n} \sum_{j=1}^{n} x^2_{ij} \Theta_i \Theta_j$$

$$\nabla_\Theta B = \left(\begin{array}{c} 2\Theta_1 x^2_{11} + \Theta_2 x^2_{12} + \dots + \Theta_n x^2_{1n} + \Theta_2 x^2_{21} + \dots \Theta_n x^2_{n1} \\ \vdots \\ 2\Theta_n x^2_{n1} + \dots\dots \end{array} \right.$$

But $x^2_{12} = x^2_{21}$ since x^2 is symmetric.

$$\nabla_\Theta B = \begin{pmatrix} 2\Theta_1 x^2_{11} + \dots \; 2\Theta_n x^2_{1n} \\ \vdots \end{pmatrix} = 2x^2 \Theta = 2x^T x \Theta$$

Now,

$$\nabla_\Theta J = \nabla_\Theta B - \nabla_\Theta A = 2x^T x \Theta - 2x^T y$$

Now, we are ready to derive the Hessian.

$$\nabla_\theta^2 J = \nabla_\theta^2 (2X^T X \Theta)$$

$$X^T X \Theta = \begin{pmatrix} x_{11}^2 \Theta_1 \ldots + x_{1n}^2 \Theta_n \\ \vdots \\ x_{n1}^2 \Theta_1 + \ldots x_{nn}^2 \Theta_n \end{pmatrix}$$

$$\nabla_\theta^2 (X^T X \Theta) = \begin{pmatrix} x_{11}^2 & x_{12}^2 & \cdots & x_{1n}^2 \\ & \vdots & & \\ x_{n1}^2 & & \cdots & x_{nn}^2 \end{pmatrix} = X^2 = X^T X$$

For nxn symmetric matrix M,

$M^T M$ is positive semi definite.

Proof:

A is PSD if $v^T A v \geq 0$

Let $A = M^T M$.

$$v^T M^T M v = (Mv)^T (Mv) = (Mv) \cdot (Mv) \geq 0$$

$\therefore$ $M^T M$ is psd.

Therefore $\nabla_\theta^2 J = X^T X$ is psd.

Therefore J is convex.

Logistic regression

Logistic regression is another popular algorithm, mostly used in classification problems. Logistic regression has a two-part cost function, depending on whether y is 0 or 1.[3]

$$E(\theta, y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

when combined, where $y \in \{0, 1\}$

$$E(\theta, y) = -y \log(h_\theta(x)) + (1-y)(-\log(1 - h_\theta(x))))$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} E(h_\theta(x^{(i)}), y^{(i)})$$

Show that for logistic regression, $J(\theta)$ is convex.

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(-y_i \log(h_\theta(x_i)) + (1-y_i)(-\log(1 - h_\theta(x_i)))\right)$$

where $y_i \in \{0, 1\}$

To prove convexity, we can do so by parts.

This is b/c linear comb. of nonnegative convex functions are convex.

First, let $A(\theta) = -\log h_\theta x^{(i)}$

First, we expand:

$$A(\theta) = -\log h_\theta x^i = -\log\left(\frac{1}{1+e^{-\theta^T x}}\right) = -\left(\overset{0}{\cancel{\log(1)}} - \log(1+e^{-\theta^T x})\right)$$

$$= \log(1+e^{-\theta^T x})$$

[3] http://stackoverflow.com/a/32987352/2204868

Before proceeding, for convenience, observe that:

$$g(z) = \frac{1}{1+e^{-z}} \qquad 1-g(z) = 1 - \frac{1}{1+e^{-z}}$$

$$= \frac{1+e^{-z}-1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}}$$

$$\frac{\partial}{\partial z}(g(z)) = \frac{\partial}{\partial z}\left((1+e^{-z})^{-1}\right)$$

$$= \cancel{(-1)}(1+e^{-z})^{-2}(\cancel{-}e^{-z}) = \frac{e^{-z}}{(1+e^{-z})^{2}} = -g(z)(1-g(z))$$

To summarize,

$$g(z) = \frac{1}{1+e^{-z}}, \quad 1-g(z) = \frac{e^{-z}}{1+e^{-z}}, \quad \frac{\partial g}{\partial z} = -g(z)(1-g(z))$$

$$\nabla_\theta A = \nabla_\theta \left(\log(1+e^{-\theta^T x})\right)$$

$$= \frac{(1+e^{-\theta^T x})'}{1+e^{-\theta^T x}} = \frac{(-x)e^{-\theta^T x}}{1+e^{-\theta^T x}} = x\underbrace{(h_\theta(x)-1)}_{\hookrightarrow \text{ scalar.}}$$

$$\nabla_\theta^2 A = \nabla_\theta \left(x(h_\theta(x)-1)\right)$$

$$= x(h_\theta(x)-1)' = xx^T(\cancel{-}h_\theta(x)(1-h_\theta(x))\cancel{(-1)}$$

To show that $\nabla_\theta^2 A$ is PSD,

$$v^T\left(\nabla_\theta^2 A\right)v \geq 0$$

$$\Downarrow$$

$$v^T x x^T h_\theta(x)\left(1-h_\theta(x)\right) v$$

$$= \underbrace{h_\theta(x)\left(1-h_\theta(x)\right)}_{\nearrow \text{ Scalar}} v^T x x^T v$$

$$= \underbrace{h_\theta(x)\left(1-h_\theta(x)\right)}_{\nearrow \text{ logistic from 0 to 1, so it has to be positive.}} \underbrace{(v^T x)^2}_{\uparrow \text{ positive}} \geq 0$$

Therefore $\nabla_\theta^2 A$ is PSD.

Now let $B(\theta) = -\log\left(1-h_\theta(x)\right)$

$$B(\theta) = -\log\left(\frac{e^{-\theta^T x}}{1+e^{-\theta^T x}}\right) = \theta^T x + \log\left(1+e^{-\theta^T x}\right)$$

$$= \theta^T x - \log\left(\frac{1}{1+e^{-\theta^T x}}\right)$$

$$B(\theta) = \theta^T x + A(\theta)$$

But we proved that $A(\theta)$ is PSD.

So then, let's consider $C(\theta) = \theta^T x$

$$\nabla_\theta C = x$$

$$\nabla_\theta^2 C = 0$$

Since $v^T (\nabla_\theta^2 C) v = v^T (0) v = 0 \geq 0$

C is also PSD.

And since $J(\theta) = \frac{1}{m} \sum_{i=1}^{m} y_i A(\theta) + (1-y_i)\left(C(\theta) + B(\theta)\right)$

$J(\theta)$ is also PSD.

$\therefore$ $J(\theta)$ is convex.

Summary

In this report, we've explored the convexity of linear and logistic regression. We derived the connection between a matrix being positive semi-definite and being convex, and we explained the importance of convexity. Convex analysis is a large and complex field, especially important in the realm of machine learning for evaluating modeling techniques. By demonstrating the convexity of linear and logistic regression, I gained a deeper understanding of the properties of these methods, and I hope to take this learning and apply it to other techniques moving forward.

Additional Sources

- Elements of Information Theory. https://liqiangguo.files.wordpress.com/2011/04/the-proof-of-why-the-kl-divergence-is-not-smaller-than-zero6.pdf
- https://spin.atomicobject.com/2014/06/24/gradient-descent-linear-regression/
- http://mathgotchas.blogspot.com/2011/10/why-is-error-function-minimized-in.html
- http://qwone.com/~jason/writing/convexLR.pdf
- https://math.stackexchange.com/q/513887/443165
- https://math.stackexchange.com/a/322743/443165
- https://math.stackexchange.com/a/133355/443165
- https://www.youtube.com/watch?v=u8JrE9JlZPM
- https://www.youtube.com/watch?v=tccVVUnLdbc
- https://www.quora.com/Why-is-Convex-Optimization-such-a-big-deal-in-Machine-Learning/answer/Ferenc-Husz%C3%A1r?srid=2uYn