Convex Functions

Sujay Sanghavi

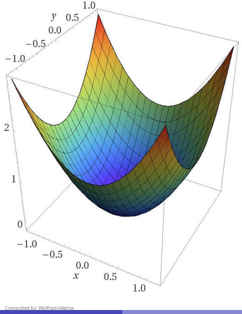
Machine Learning and Opimization

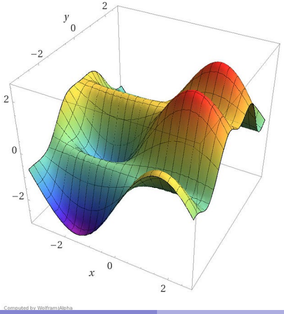
Machine learning: find the values of the parameters of a model so that error is minimized on a training dataset

Optimization: methods to efficiently find the minimum of a given function

Modern trend: do machine learning using methods from optimization

e.g.: Linear regression (and Ridge and LASSO), Logistic Regression



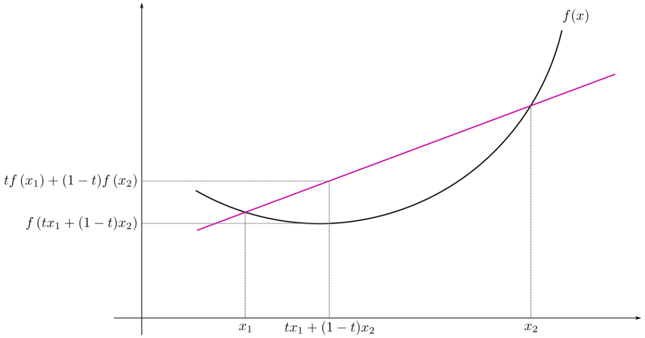


Convex Functions

A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is said to be **convex** if

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

for any pair of points x_1, x_2 in $\mathbb{R}^d$ and any $0 \leq t \leq 1$.



How do I know if my function is convex ?

Scalar case i.e. $f : \mathbb{R} \to \mathbb{R}$

$$\boxed{f \text{ is convex} \Leftrightarrow f''(x) \geq 0 \text{ for all } x}$$

Recall: Logistic Regression in scalar case, i.e. $\beta \in \mathbb{R}$

$$f(\beta) \;=\; -\log\left(\frac{1}{1+e^{-\beta x}}\right) \;=\; \log\left(1+e^{-\beta x}\right)$$

How do I know if my function is convex ?

Scalar case i.e. $f : \mathbb{R} \to \mathbb{R}$

$$\boxed{f \text{ is convex} \Leftrightarrow f''(x) \geq 0 \text{ for all } x}$$

Recall: Logistic Regression in scalar case, i.e. $\beta \in \mathbb{R}$

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

$$f'(\beta) = \frac{-xe^{-\beta x}}{1+e^{-\beta x}} = \frac{-x}{e^{\beta x}+1}$$

How do I know if my function is convex ?

Scalar case i.e. $f : \mathbb{R} \to \mathbb{R}$

$$\boxed{f \text{ is convex} \Leftrightarrow f''(x) \geq 0 \text{ for all } x}$$

Recall: Logistic Regression in scalar case, i.e. $\beta \in \mathbb{R}$

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

$$f'(\beta) = \frac{-xe^{-\beta x}}{1+e^{-\beta x}} = \frac{-x}{e^{\beta x}+1}$$

$$f''(\beta) = (-x) \times \left(-\frac{xe^{\beta x}}{(e^{\beta x}+1)^2}\right) = \frac{x^2 e^{\beta x}}{(e^{\beta x}+1)^2}$$

Recall: Vector Calculus

Vector case i.e. $f : \mathbb{R}^d \to \mathbb{R}$

The **gradient** $\nabla f(\cdot)$ is its derivative.

For any point $x \in \mathbf{R}^d$, $\nabla f(x)$ is a d-length vector.

$$[\nabla f(x)]_i = \frac{\partial f(x)}{\partial x_i} \quad \text{for all coordinates } i \text{ in } 1, \ldots, d$$

For any vector a and any point x,

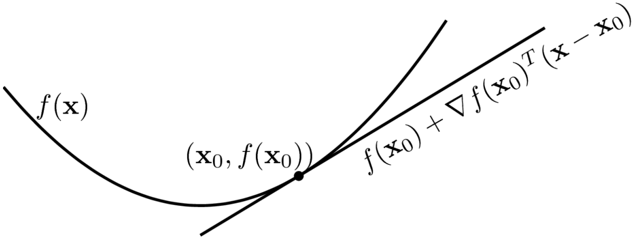
$$\lim_{\delta \to 0} \frac{f(x + \delta a) - f(x)}{\delta} = a^\top \nabla f(x)$$

That is, $a^\top \nabla f(x)$ represents the rate of change of f in the direction fo a.

Convex Functions

A function f is convex if and only if it is always "above its gradient", i.e. for any points x and x_0

$$f(x) \;\geq\; f(x_0) + \nabla f(x_0)^\top (x - x_0)$$



Recall: Vector Calculus

Vector case i.e. $f : \mathbb{R}^d \to \mathbb{R}$

Recall: the **Hessian** is the "second derivative" of f. It is a **matrix** and is denoted by $\nabla^2 f(\cdot)$

$$\nabla^2 f(x) \quad = \quad \begin{bmatrix} \frac{\partial^2 f}{(\partial x_1)^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{(\partial x_2)^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots & & & \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f}{(\partial x_d)^2} \end{bmatrix}$$

It's $(i,j)^{th}$ element is $\frac{\partial^2 f}{\partial x_i \partial x_j}$ for i and j in $(1, \ldots, d)$

Note: $\nabla^2 f(x)$ is a **symmetric** matrix.

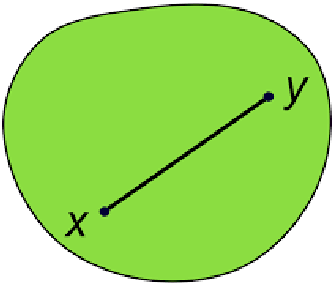
How do I know if my function is convex ?

- f is convex $\Leftrightarrow f''(x) \geq 0$ for all x
  - For vector case, this means the Hessian $\nabla^2 f(x)$ is **positive definite** for all x
- f is convex $\Leftrightarrow f(y) \geq f(x) + (y-x)^\top \nabla f(x)$ for all points x and y
- If f is convex and $g(x) = f(Ax+b)$ for all x and some matrix A and vector b, then g is convex
- If f_1 and f_2 are convex, and $g = f_1 + f_2$, then g is convex

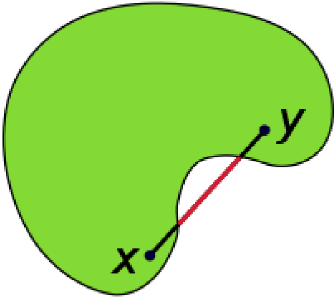
many other such properties can be used to check for convexity ...

Convex Sets

A set $\mathcal{C}$ is.convex if for every pair of points x and y in $\mathcal{C}$, the line joining x and y is also in $\mathcal{C}$.



Convex set



Not convex set

Convex Optimization Problem and Gradient Descent

Convex Optimization is of the form

$$\begin{array}{ll} \min_{x} & f(x) \\ s.t. & x \in \mathcal{C} \end{array}$$

where f is a convex function and $\mathcal{C}$ is a convex set.

Gradient Descent:

- Start from some x_0
- Update by moving along the direction of fastest decrease:

$$x_{t+1} = \mathcal{P}_{\mathcal{C}}\left(x_t - \eta_t \nabla f(x_t) \right)$$

where η_t is called the step size at time t, and $\mathcal{P}_{\mathcal{C}}(\cdot)$ is the projection back onto the set $\mathcal{C}$

$\star$ If $f(\cdot)$ is convex, then for *well chosen* step sizes, $x_t \to x^\star$ as $t \to \infty$

Gradient Descent for Linear Regression

Convex Loss function:

$$\min_{\beta} \ \|y - X\beta\|_2^2$$

Gradient of the loss function: $-2X^\top(y - X\beta)$

Gradient descent:

$$\beta_{t+1} \ = \ \beta_t \ - \ \eta_t \left[-2X^\top(y - X\beta_t) \right]$$

Then for certain choices of the step sizes η_t, we have that $\beta_t \to \beta^*$ as $t \to \infty$

What the choice of η_t's should be, and how quickly it will converge, depends on what the X is.

Gradient Descent for Ridge Regression

Convex Loss function:

$$\min_{\beta} \; \|y - X\beta\|_2^2 \; + \; \lambda \, \|\beta\|_2^2$$

Gradient of the loss function: $-2X^\top(y - X\beta) + 2\lambda\beta$

Gradient descent:

$$\beta_{t+1} \; = \; \beta_t \; - \; \eta_t \left[-2X^\top(y - X\beta_t) + 2\lambda\beta \right]$$

again, $\beta_t \to \beta^*$ as $t \to \infty$ for well-chosen η_ts ...

Gradient Descent for LASSO

Convex Loss function:

$$\min_{\beta} \; \|y - X\beta\|_2^2 \; + \; \lambda \, \|\beta\|_1$$

Gradient of the loss function: does not exist
because $\|\beta\|_1 = \sum_j |\beta_j|$ is not differentiable

In the case of LASSO, we are lucky because this specific non-differentiable part $\|\beta\|_1$ can be "dealt with" in a different way:

$$\beta_{t+1} \; = \; \mathcal{S}_\lambda \left\{ \beta_t \; - \; \eta \left[-2X^\top (y - X\beta_t) \right] \right\}$$

Where $\mathcal{S}_\lambda$ is the shrinkage operator:

$$[\mathcal{S}_\lambda(\beta)]_i \; = \; \begin{cases} \beta_i - \lambda & \text{if } \beta_i > \lambda \\ 0 & \text{if } -\lambda \le \beta_i \le \lambda \\ \beta_i + \lambda & \text{if } \beta_i < -\lambda \end{cases}$$

"shrinks each β_i by λ"