Convex optimization

February 9, 2021

Abstract

As part of my machine learning study, I summerize what I learnd about basic optimization problem. Among the optimization problems, I write about the convex optimization. The convex otimization is the problem of minimizing convex functions over convex sets. Due to the nature of the convex function, the convex optimization is relatively easy to solve the optimal solution of the objective function, and the solution method has been established to some extent. This time, I show about a classifier that uses convex optimization such as a log-linear classifier.

Convex optimization

Before writing about the convex optimization, brifly write what the optimization problem is. An optimization problem is a problem of finding valuable values and function value that optimizes (minimize or maximize) a function under certain constraints. And, the function to be optimized is called the objective function. For example, suppose there is an objective function $f(x_1, x_2) = x_1 x_2$ and a constraint $x_1 - x_2 - a = 0$ (a is a constant). Also, assume that the optimization problem minimizes $f(x_1, x_2)$. The formal writing of these is as follows,

$$\begin{cases} \mathrm{min.} \quad \mathrm{f(x_1, x_2) = x_1 x_2} \\ \mathrm{s.t.} \quad \mathrm{x_1 - x_2 - a = 0} \end{cases}$$

and s.t. means subject to. Solving this gives $x_1 x_2 = (x_1 - \frac{a}{2})^2 - \frac{a^2}{4}$, and is obtained optimal solutions $-\frac{a^2}{4}$ and $(x_1, x_2) = (\frac{a}{2}, -\frac{a}{2})$ with the smallest $f(x_1, x_2)$. Next, I describe what a convex set and convex function is and those properties.

0.1 Convex set and convex function

First, I write the defnition of the convexity of the set on $\mathbb{R}^n$.

Definition 0.1 (Convex set)**.**
$\mathrm{X} \subset \mathbb{R}^n$ is a convex set. $\overset{def}{\iff}$ $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathrm{X}$ and $\forall \lambda \in [0, 1] \to \lambda \boldsymbol{x}_1 + (1 - \lambda)\boldsymbol{x}_2 \in \mathrm{X}$

As you can see by drawing the convex set in the figure, the convex set means a set in which the line segment connecting arbitrary points in the convex set does not protrude from the set. Next, although I write about the definition of a convex function, there are two types of convex functions, before that, I state the dediniton of epigraph, which is a concept that connects covex sets and convex downward functions.

Definition 0.2 (Epigraph)**.**
Given the real-valued function f, the following set is called the epigraph of f

$$\mathrm{epi}\, f := \{(\boldsymbol{x}, y) \in \mathbb{R}^n \times \mathbb{R} : y \geq f(\boldsymbol{x})\}$$

Definition 0.3 (Convex function)**.**
$f : \mathbb{R}^n \to \mathbb{R}$ is a convex function $\overset{def}{\Longleftrightarrow}$ epi f is a convex set.[*1]

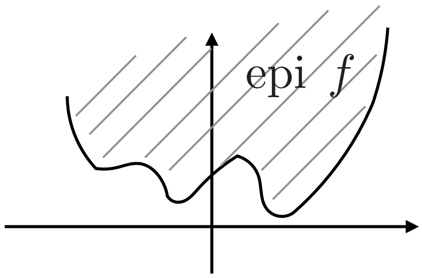

Figure1 epigraph

The epigraph in Figure1 is an epigraph that is not a convex set. Using these definitions (0.1 $\sim$ 0.3), the following theorem holds.

Theorem 0.1.
The fact that the function f on $\mathbb{R}^n$ is a convex function is equal to the following cinditon : for $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$ and $\forall \lambda \in [0, 1]$,

$$\lambda f(\boldsymbol{x}_1) + (1 - \lambda) f(\boldsymbol{x}_2) \geq f(\lambda \boldsymbol{x}_1 + (1 - \lambda) \boldsymbol{x}_2)$$

Proof. For $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$ and $\forall \lambda \in [0, 1]$, if there is a there is a $\tilde{\boldsymbol{x}}$ such that $\tilde{\boldsymbol{x}} = \lambda \boldsymbol{x}_1 + (1 - \lambda) \boldsymbol{x}_2$, then think of y that satisfies $y \geq f(\tilde{\boldsymbol{x}})$. Recalling the definion of a convex set (0.1), when Y is a set of y, of course, $f(\boldsymbol{x}_1)$ and $f(\boldsymbol{x}_2) \in$ Y, so $\lambda f(\boldsymbol{x}_1) + (1 - \lambda) f(\boldsymbol{x}_2) \in$Y. That is, $\lambda f(\boldsymbol{x}_1) + (1 - \lambda) f(\boldsymbol{x}_2)$ is also included in the value of y, and $\lambda f(\boldsymbol{x}_1) + (1 - \lambda) f(\boldsymbol{x}_2) \geq f(\lambda \boldsymbol{x}_1 + (1 - \lambda) \boldsymbol{x}_2)$ holds. □

Next, I explain the characteristics of the convex function. If the convex fuction is differentiable, the following necessary and sufficient conditions hold.

[*1]

Definition 0.4 (Concave function)**.** $f : \mathbb{R}^n \to \mathbb{R}$ is a concave function $\overset{def}{\Longleftrightarrow}$ $-f$ is a convex function.

Theorem 0.2 (First-order convexity condition)**.**
Suppose the real-valued function $f : \mathbb{R}^n \to [-\infty, \infty]$ is differentiable. At this time, the necessary and sufficient conditons for f to be a convex function are as follows:
for $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$,

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \boldsymbol{\nabla} f(\boldsymbol{x})^{\mathrm{T}} \mid_{\boldsymbol{x}=\boldsymbol{x}_1} (\boldsymbol{x}_2 - \boldsymbol{x}_1)$$

Proof. Since the equal sign holds when $\lambda = 0$ and 1, prove it for $\lambda \in (0, 1)$. In addition, assuming there are $\forall \boldsymbol{x}_1$ and $\boldsymbol{x}_2 \in \mathbb{R}^n$ $(\boldsymbol{x}_1 \neq \boldsymbol{x}_2)$.
$(\Longrightarrow)$
Since f is a convex function, $\lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2) \geq f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2)$ holds. Therefore,

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \frac{f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) - f(\boldsymbol{x}_1)}{1-\lambda}$$

Let $\tilde{\lambda} = 1 - \lambda$, the above formula becomes,

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \frac{f(\boldsymbol{x}_1 + \tilde{\lambda}(\boldsymbol{x}_2 - \boldsymbol{x}_1))}{\tilde{\lambda}(\boldsymbol{x}_2 - \boldsymbol{x}_1)} (\boldsymbol{x}_2 - \boldsymbol{x}_1)$$

Since f is differentiable, perform taylor expansion around $\Delta \boldsymbol{x} = \tilde{\lambda}(\boldsymbol{x}_2 - \boldsymbol{x}_1)$ on f, the RHS (right-hand side) is,

$$\begin{aligned} &\mathrm{RHS} = \\ &f(\boldsymbol{x}_1) + \boldsymbol{\nabla} f(\boldsymbol{x})^{\mathrm{T}} \mid_{\boldsymbol{x}=\boldsymbol{x}_1} (\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{1}{2!}(\boldsymbol{x}_2 - \boldsymbol{x}_1)^{\mathrm{T}} \boldsymbol{\nabla}^2 f(\boldsymbol{x}) \mid_{\boldsymbol{x}=\boldsymbol{x}_1} (\boldsymbol{x}_2 - \boldsymbol{x}_1) + \mathcal{O}(\Delta^3) \end{aligned} \quad (1)$$

Of course, even if the $\Delta \boldsymbol{x}$ is as close to $\mathbf{0}$ as possible, the above inequality holds, so

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \boldsymbol{\nabla} f(\boldsymbol{x})^{\mathrm{T}} \mid_{\boldsymbol{x}=\boldsymbol{x}_1} (\boldsymbol{x}_2 - \boldsymbol{x}_1)$$

$(\Longleftarrow)$
If the inequality (0.2) holds, then the following relational expression can be created.

$$\begin{cases} f(\boldsymbol{x}_2) \geq f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) + \boldsymbol{\nabla} f(\boldsymbol{x})^{\mathrm{T}} \mid_{\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2} (\boldsymbol{x}_2 - (\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2)) \\ f(\boldsymbol{x}_1) \geq f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) + \boldsymbol{\nabla} f(\boldsymbol{x})^{\mathrm{T}} \mid_{\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2} (\boldsymbol{x}_1 - (\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2)) \end{cases}$$

Multiply these relational expressions by $1 - \lambda$ and λ, and take the sum,

$$\lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2) \geq f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2)$$

Therefore, f is a convex function. □

For example, if there is only one variable, for $\forall x$ and x_0, the theorem (0.2) becomes,

$$f(x) \geq f(x_0) + \frac{df(x)}{dx} \mid_{x=x_0} (x - x_0)$$

Especially when $\frac{df(x_0)}{dx} = 0$, $f(x)$ is the minimum value at $x = x_0$, as is clear from the formula. So far, I have explained the first-order convexity condition, but the condition of the convex has second, and now I describe the second-order convexity condition.

Theorem 0.3 (Second-order convexity condition)**.**
Let $f : \mathbb{R}^n \to [-\infty, \infty]$ be a real valued function that can be differented twice. At this time, the necessary and sufficient condition for f to be a convex function is to satisfy the following inequality about Hessian matrix H for $\boldsymbol{x} = \{x_i\}_{i=1,2,\cdots,n} \in \mathbb{R}^n$:
for $\forall \boldsymbol{d} = \{d_i\}_{i=1,2,\cdots,n} \in \mathbb{R}^n$,

$$\boldsymbol{d}^{\mathrm{T}} H(\boldsymbol{x})\boldsymbol{d} \geq 0$$

this ineuality is called **positive semi-difine.**

Here, the Hessian matrix is defined as follows:

Definition 0.5 (Hessian matrix)**.**
At $\boldsymbol{x} = \{x_i\}_{i=1,2,\cdots,n}$ and $f(\boldsymbol{x})$ is differentiable twice, the Hessian matrix defined as

$$H(\boldsymbol{x})_{ij} := \boldsymbol{\nabla}^2 f(\boldsymbol{x})_{ij} \quad (i, j = 1, 2, \cdots, n)$$

Proof. For $\forall \lambda \in (0, 1)$, utilize the formula used in the proof (0.2).
$(\Longrightarrow)$
From equation (1), for $\boldsymbol{d} := (\boldsymbol{x_2} - \boldsymbol{x_1})$, perform the taylor expansion around $\boldsymbol{d}$ of $f(\boldsymbol{x})$,

$$f(\boldsymbol{x} + \boldsymbol{d}) = f(\boldsymbol{x}) + \frac{1}{\lambda}\boldsymbol{\nabla} f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d} + \frac{1}{2!}\frac{1}{\lambda^2}\boldsymbol{d}^{\mathrm{T}} H(\boldsymbol{x})\boldsymbol{d} + \mathcal{O}(\Delta^3)$$

Also, the following relational expression holds from the theorem (0.2),

$$f(\boldsymbol{x} + \boldsymbol{d}) \geq f(\boldsymbol{x}) + \frac{1}{\lambda}\boldsymbol{\nabla} f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}$$

Compare the two formulas above,

$$\frac{1}{2!}\frac{1}{\lambda^2}\boldsymbol{d}^{\mathrm{T}} H(\boldsymbol{x})\boldsymbol{d} + \mathcal{O}(\Delta^3) \geq 0$$

This formula holds no matter how small $O(\Delta^3)$ is, so

$$\boldsymbol{d}^{\mathrm{T}} H(\boldsymbol{x})\boldsymbol{d} \geq 0$$

$(\Longleftarrow)$
Suppose the Hessian matrix is positive semi-difine. Perform the taylor expansion around $\boldsymbol{d}$ of $f(\boldsymbol{x})$ and move the contents of the formula,

$$f(\boldsymbol{x} + \boldsymbol{d}) - \left(f(\boldsymbol{x}) + \frac{1}{\lambda}\boldsymbol{\nabla} f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}\right) = \frac{1}{2!}\frac{1}{\lambda^2}\boldsymbol{d}^{\mathrm{T}} H(\boldsymbol{x})\boldsymbol{d} + \mathcal{O}(\Delta^3)$$

Since the Hessian matrix is positive semi-difine, that is,

$$f(\boldsymbol{x} + \boldsymbol{d}) - \left(f(\boldsymbol{x}) + \frac{1}{\lambda}\boldsymbol{\nabla} f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}\right) \geq \mathcal{O}(\Delta^3)$$

Of course, it holds even if $\mathcal{O}(\Delta^3)$ is small enough,

$$f(\boldsymbol{x}+\boldsymbol{d}) - \left(f(\boldsymbol{x}) + \frac{1}{\lambda}\boldsymbol{\nabla} f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}\right) \geq 0$$

Therefore, accoding to the theorem (0.2), $f(\boldsymbol{x})$ is a convex function. □

By utilizing the fact that it is differentiable, it is possible to know whether a function is a convex function without using inequalities. For example, if $f(x) = exp(x)$, then $f''(x) \geq 0$, so $f(x)$ is a convex function.

0.2 Convex optimization

lagrange multiplier

0.3 Log-linear classifier

logistic regression

References

[1] 奥村 学, 高村 大也 (2010), 言語処理のための機械学習入門 (自然言語処理シリーズ), コロナ社.