Tutorial 3 Basic Optimization

Yifan WANG

School of Data Science

2022.9.27

Self Introduction

- name: Yifan WANG
- email: 119010317@link.cuhk.edu.cn
- major: statistics
- office hour: Thu. 10:00-11:00, by appointment

Introduction

- Definition of convex optimization
- How to solve convex optimization problem
- Exercise: convex optimization in machine learning

Goal

Solve some basic optimization problem in machine learning. If you are interested in optimization, you can take MAT3007 for more advanced knowledge.

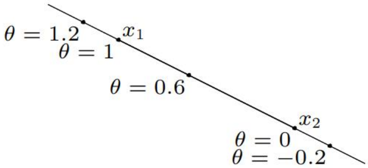
Recap

line: through x_1, x_2: all points satisfy:

$$x = \theta x_1 + (1 - \theta) x_2 \qquad \theta \in \mathbb{R}$$

line segment: between x_1, x_2: all points satisfy:

$$x = \theta x_1 + (1 - \theta) x_2 \qquad \theta \in [0, 1]$$



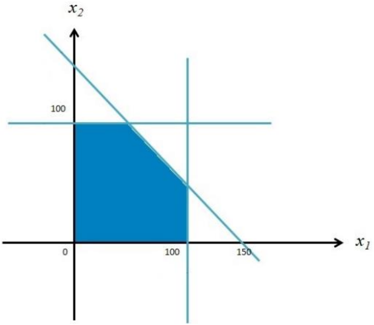
Recap

affine set: contains the line through any two distinct points in set

- Example: solution set of linear equations $\{x | Ax = b\}$

convex set: contains line segment between any two points in set

- Example:

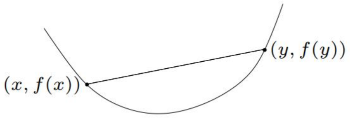


Recap

convex function: f: $R^n \to R$ is convex if **dom** f is a convex set and

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

for all $x, y \in$ **dom** f, $0 \leq \theta \leq 1$



- f is concave if $-f$ is convex
- f is strictly convex if **dom** f is convex and

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

FOC and SOC

FOC: differentiable f with convex domain is convex iff

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y \in \textbf{dom } f$$

SOC: for twice differentiable f with convex domain

- f is convex if and only if

$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x \in \textbf{dom} \text{ f}$$

- if $\nabla^2 f(x) \succ 0$ for all $x \in \textbf{dom } f$, then f is strictly convex

Positive definite: for $A \in R^{M \times M}$, if $A = A^T$, and for any $X \in R^{M \times 1} \neq \mathbf{0}$, $X^T AX > 0$, then A is called a positive definite matrix

Exercise 1

Logistic Regression(l_2 loss)

The l_2 loss for logistic regression is

$$J(\mathrm{w}) = \frac{1}{2m}\sum_i^m (g(\mathrm{w}^T x_i) - y_i)^2$$

show that it's not convex, where $g(w^T x) = \frac{1}{1+exp(-w^T x)}$. For convenience, you can take $J(\mathrm{w}) = (g(\mathrm{w}^T x) - y)^2$.

Hint: consider the special case

Standard Form

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{aligned}$$

Convex Optimization

- f_0, f_1, ..., f_m are convex functions
- $h_i(x) = 0$ are affine constrains

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & Ax = b \end{aligned}$$

- For convex optimization, any local optimum is also a global optimum

Unconstrained problem

$$\text{minimize} \quad f_0(x)$$

optimal condition: $\nabla f(x^*) = 0$

- 1. take the derivative
- 2. gradient descent method
  $x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)}$ with $f(x^{(k+1)}) < f(x^{(k)})$

Exercise 2

Ridge Regression(Shrinkage estimator)

The ridge regression uses the penalty form $R(\beta) = \lambda ||\beta||_2^2$ and to minimize

$$min_{\beta}(Y - X\beta)^T(Y - X\beta) + \lambda ||\beta||_2^2$$

show that $\beta = (X^TX + \lambda I)^{-1}X^TY$ with the fact that $||\beta||_2^2 = \beta^T\beta$
Hint: $\frac{\nabla X^TX}{\nabla X} = 2X$, $\frac{\nabla X^Tw}{\nabla w} = X$

Constrained problem

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & h_i(x) \leq 0, \quad i = 1, \dots, m \\ & l_j(x) = 0, \quad j = 1, \dots, r \end{aligned}$$

- Lagrange function

  $$L(w, u, v) = f(w) + \sum_{i=1}^{m} u_i h_i(w) + \sum_{j=1}^{r} v_j l_j(w)$$

- KKT conditions

  stationarity: $0 \in \partial f(w) + \sum_{i=1}^{m} u_i \partial h_i(w) + \sum_{j=1}^{r} v_j \partial l_j(w)$
  complementary slackness: $u_i \cdot h_i(w) = 0$ for all i
  primal feasibility: $h_i(w) \leq 0$, $l_j(w) = 0$ for all i, j
  dual feasibility: $u_i \geq 0$ for all i

Exercise 3

Support vector machine(hard-margin)

The objective function of support vector machine is

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & 1 - y_i(w^T x_i + b) \leq 0, \forall i \end{aligned}$$

write the Lagrange function and KKT conditions of it.

Solution 1-1

$$
\begin{aligned}
J(w) =& (g(w^Tx_i) - y_i)^2 \\
\frac{\nabla g(w^Tx)}{\nabla w} =& \frac{x \cdot exp(-w^Tx))}{(1 + exp(-w^Tx))^2} = x \cdot g(w^Tx)(1 - g(w^Tx)) \\
\frac{\nabla J(w)}{\nabla w} =& \frac{\nabla J(w)}{\nabla g(w^Tx)} \cdot \frac{\nabla g(w^Tx)}{\nabla w} \\
=& 2x(g(w^Tx) - y)g(w^Tx)(1 - g(w^Tx)) \\
=& 2x(-g(w^Tx)^3 + (y+1)g(w^Tx)^2 - yg(w^Tx)) \\
\frac{\nabla^2 J(w)}{\nabla w^2} =& \frac{\nabla^2 J(w)}{\nabla g(w^Tx)} \cdot \frac{\nabla g(w^Tx)}{\nabla w}
\end{aligned}
$$

$$= 2xx^Tg(w^Tx)(1 - g(w^Tx)(-3g(w^Tx)^2 + 2(y+1)g(w^Tx) - y)$$

Solution 1-2

Special case: take y=-1

$$
\begin{aligned}
&2xx^Tg(w^Tx)(1-g(w^Tx)(-3g(w^Tx)^2+2(y+1)g(w^Tx)-y)\\
=&2xx^Tg(w^Tx)(1-g(w^Tx)(-3g(w^Tx)^2+1)
\end{aligned}
$$

When $g(w^Tx)\in(-\frac{\sqrt{3}}{3},0)\cup(\frac{\sqrt{3}}{3},1)$, this expression is smaller than 0, so it's not convex

Solution 2

$$min_{\beta}(Y - X\beta)^T(Y - X\beta) + \lambda||\beta||_2^2$$

$$\frac{\nabla f(x)}{\nabla \beta} = 2X^T(X\beta - Y) + 2\lambda\beta = 0$$

$$2(X^TX + \lambda I)\beta = 2X^TY$$

$$\beta = (X^TX + \lambda I)^{-1}X^TY$$

Solution 3

Lagrange function: $L(w, b, \alpha) = \frac{1}{2}||w||^2 + \sum \alpha_i(1 - y_i(w^T x_i + b))$

KKT conditions:

$$
\begin{aligned}
\textit{stationary} :& \frac{\partial L}{\partial w} = 0 \rightarrow w = \sum \alpha_i y_i x_i \\
& \frac{\partial L}{\partial b} = 0 \rightarrow \sum \alpha_i y_i = 0 \\
\textit{primal} :& \alpha_i \geq 0, \quad 1 - y_i(w^T x_i + b) \leq 0, \forall i \\
\textit{complementary} :& \alpha_i(1 - y_i(w^T x_i + b)) = 0, \forall i \\
\textit{dualfeasibility} :& \alpha_i \geq 0, \forall i
\end{aligned}
$$