ML OPT

GeoComput & ML

2021-05-06 Thur

ML OPT

ML often cast as an <u>optimisation problem</u>,
the problem of determining an argument for which
a given function has extreme values on a given domain

Optimisation

Given a function $f : \mathbb{R}^n \to \mathbb{R}$ and a set $S \subseteq \mathbb{R}^n$, we seek $\boldsymbol{x}^*$ such that f attains minimum at $\boldsymbol{x}^*$, $\forall \boldsymbol{x} \in S$

We call f the objective function, S the feasible set and any vector $\boldsymbol{x} \in S$ a feasible point.

Generally, a continuous optimisation problem takes the form

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad s.t. \quad \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{o} \text{ and } \boldsymbol{h}(\boldsymbol{x}) \leq \boldsymbol{o}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$, and $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^p$.

Existence and Uniqueness

If f is continuous on a close and bounded set $S \subseteq R^n$, then f has a global minimum on S

Coercive functions

A continuous function f on an unbounded set $S \subseteq R^n$ is said to coercive if

$$\lim_{\|x\| \to \infty} f(\boldsymbol{x}) = +\infty$$

If f is coercive on a close, unbounded set $S \subseteq R^n$, then f has a global minimum on S

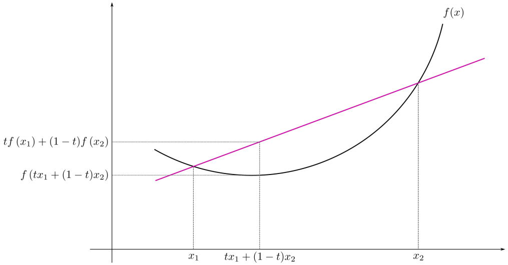
Convexity

Definition

If a function f satisfies the condition

$$f(\alpha \boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2) \leq \alpha f(\boldsymbol{x}_1) + (1-\alpha) f(\boldsymbol{x}_2)$$

$\forall \alpha \in [0,1] \cup \forall \boldsymbol{x}_i \in S$



Any local minimum of a convex function f on convex set $S \subseteq \mathbb{R}^n$ is a global minimum of f on S

Unconstrained Optimality Conditions

1^{st} order necessary condition

$$\nabla f(\boldsymbol{x}^*) = 0$$

where gradient $\nabla f(\boldsymbol{x}) = \left(\frac{\partial \boldsymbol{x}}{\partial x_1} \frac{\partial \boldsymbol{x}}{\partial x_1} \dots \frac{\partial \boldsymbol{x}}{\partial x_n} \right)^T$ and $\boldsymbol{x}^*$ is called a critical point of f.

Unconstrained Optimality Conditions

2^{nd} order sufficient condition

At a critical point $\boldsymbol{x}^*$, if $\boldsymbol{H}_f(\boldsymbol{x}^*)$ is

- positive definite, then $\boldsymbol{x}^*$ is a minimum.
- negative definite, then $\boldsymbol{x}^*$ is a maximum.
- indefinite, then $\boldsymbol{x}^*$ is a saddle point.

where

$$\boldsymbol{H}_f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial^2 \boldsymbol{x}}{\partial x_1^2} & \frac{\partial^2 \boldsymbol{x}}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 \boldsymbol{x}}{\partial x_1 \partial x_n} \\ \frac{\partial^2 \boldsymbol{x}}{\partial x_2 \partial x_1} & \frac{\partial^2 \boldsymbol{x}}{\partial x_2^2} & \cdots & \frac{\partial^2 \boldsymbol{x}}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \boldsymbol{x}}{\partial x_n \partial x_1} & \frac{\partial^2 \boldsymbol{x}}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 \boldsymbol{x}}{\partial x_n^2} \end{pmatrix}$$

Constrained Optimality Conditions

Lagrange multiplier

Consider problem

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad s.t. \quad \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{o}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$, and $m \leq n$.

Define the *Lagrangian function*, $\boldsymbol{\mathcal{L}} : \mathbb{R}^{n+m} \to \mathbb{R}$.

$$\boldsymbol{\mathcal{L}}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{x})$$

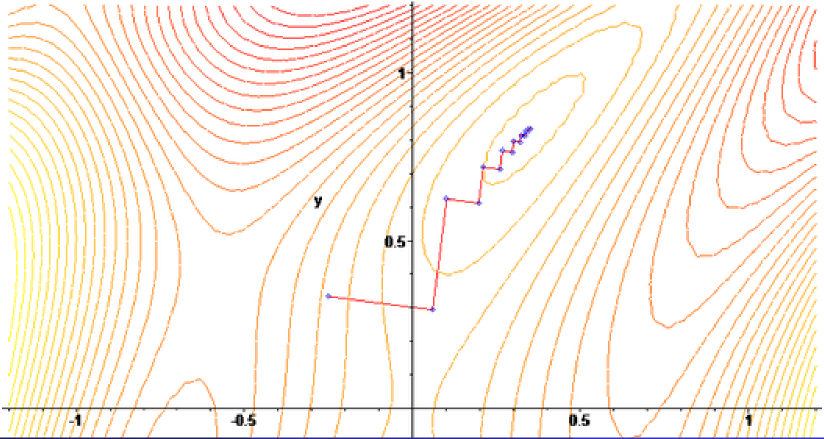
then the necessary condition for a critical point is

$$\nabla \boldsymbol{\mathcal{L}} = \begin{pmatrix} \nabla f(\boldsymbol{x}) + \boldsymbol{J}_{\boldsymbol{g}}^T(\boldsymbol{x})\boldsymbol{\lambda} \\ \boldsymbol{g}(\boldsymbol{x}) \end{pmatrix} = \boldsymbol{o}$$

Unconstrained Optimisation

Gradient Descent

$$\boldsymbol{x}_{k+1} = f(\boldsymbol{x}_k) + \alpha \boldsymbol{s}_k$$
$$\boldsymbol{s}_k = -\nabla f(\boldsymbol{x}_k)$$

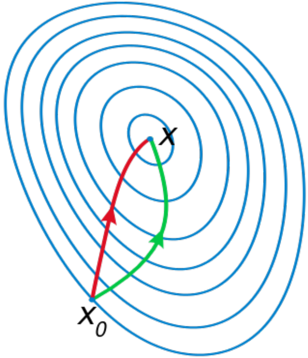


Unconstrained Optimisation

Newton Method

$$f(\boldsymbol{x}+\boldsymbol{s}) \approx f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T \boldsymbol{s} + \boldsymbol{s}^T \boldsymbol{H}_f(\boldsymbol{x})\boldsymbol{s}/2$$

$$\boldsymbol{H}_f(\boldsymbol{x})\boldsymbol{s} = -\nabla f(\boldsymbol{s})$$



OLR : LS

learn a decision function $h : \boldsymbol{U} \to \boldsymbol{V}$

$$\begin{cases} h(\boldsymbol{u}) = \sum_{i=0}^{n} \theta_i u_i = \boldsymbol{\theta}^T \boldsymbol{u} \\ u_0 = 1 \end{cases}$$

search for $\boldsymbol{\theta} \in \mathbb{R}^{n+1}$

$$\min_{\boldsymbol{\theta}} \{ f(\boldsymbol{\theta}) := \sum_i (h_{\boldsymbol{\theta}}(u_i) - v_i)^2 \}$$

$$\begin{aligned} f(\boldsymbol{\theta}) &= (\boldsymbol{U\theta} - \boldsymbol{v})^T (\boldsymbol{U\theta} - \boldsymbol{v}) \\ &= \boldsymbol{\theta}^T \boldsymbol{U}^T \boldsymbol{U\theta} - 2\boldsymbol{\theta}^T \boldsymbol{U}^T \boldsymbol{v} + \boldsymbol{v}^T \boldsymbol{v} \\ f'(\boldsymbol{\theta}) &= \boldsymbol{U}^T \boldsymbol{U\theta}^T - \boldsymbol{U}^T \boldsymbol{v} = 0 \\ \boldsymbol{\theta}^* &= (\boldsymbol{U}^T \boldsymbol{U})^{-1} \boldsymbol{U}^T \boldsymbol{v} \end{aligned}$$

OLR : MLE

Let $\epsilon_i = v_i - \boldsymbol{\theta}^T u_i$ and $\epsilon_i \sim \mathcal{N}(\mu, \sigma^2)$

$$p(v_i|u_i;\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(v_i - \boldsymbol{\theta} u_i)^2}{2\sigma^2}\right)$$

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \sum_i log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(v_i - \boldsymbol{\theta} u_i)^2}{2\sigma^2}\right)\right) \\
&= N \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2}\sum_i (v_i - \boldsymbol{\theta}^T u_i)^2
\end{aligned}$$

Logistic Model

Given $\boldsymbol{U} = (u_1, \ldots, u_n)^T \cup \boldsymbol{v} \in \{0, 1\}$

decision function $h_{\boldsymbol{\theta}} = g(\boldsymbol{\theta}^T u) = \dfrac{1}{1 + \exp(-\boldsymbol{\theta}^T u)}$

$v_i \sim \mathcal{B}(p)$

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_i \log(h(u_i)^{v_i}(1 - h(u_i))^{1-v_i})$$

$$\max_{\boldsymbol{\theta}} \sum_i (-\log(1 + \exp(-\boldsymbol{\theta}^T u)) - (1 - v_i)\boldsymbol{\theta}^T u)$$

MAP

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{X}) \\
&= \max_{\boldsymbol{\theta}} p(\mathcal{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})
\end{aligned}
$$

Philosophical Discussions

- ML and optimisation
- Intelligence

Project Discussions

- time frame
- my expectations

References

- M. Heath. Scientific Computing An Introductory Survey (2018)
- C. Aggarwal. Linear Algebra and Optimization for Machine Learning (2020)
- S. Theodoridis. Machine Learning : A Bayesian and Optimization Perspective (2020)
- S. Sar *et. al.* Optimization for Machine Learning (2012)
- J. Gallier. Fundamentals of Optimization Theory With Applications to Machine Learning (2019)
- https://en.wikipedia.org/wiki/Convex_function
- https://en.wikipedia.org/wiki/Gradient_descent
- https://en.wikipedia.org/wiki/Newton%27s_method_in_optimization
- https://en.wikipedia.org/wiki/Bernoulli_distribution