Chapter 7. continous optimization

Machine learning model을 training = objective function에 좋은 parameter set을 발견하는것.

↳ optimization algorithm ⇒ objective function을 minimize하는 최적의 parameter를 발견.

Optimization using Gradient Descent.

∘ unconstrained optimization problem : $\min_x f(x)$ ⇒ objective function $f(x)$를 최소화하는 x찾기.

여기서 f는 미분가능해야함., $\mathbb{R}^d$ dimension인 vector를 Scalar로 mapping.

gradient descent: $f(x)$가 단순하게 미분을 하여 찾기 어려울 때 사용하는 방법.

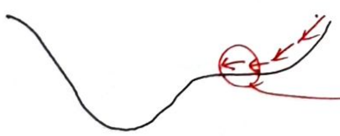
function의 gradient의 반대방향으로 움직여 local minimum을 향해 가는 것.

$$x_{i+1} = x_i - \gamma_i((\nabla f)(x_i))^T \quad (\gamma_i : \text{step size})$$

gradient descent with momentum

$$x_{i+1} = x_i - \gamma_i((\nabla f)(x_i))^T - \alpha \Delta x_i \quad (\Delta x_i = x_i - x_{i-1})$$

이전에 얼만큼 x가 움직였는지를 memory에 추가해서 계산 ⇒ 더 멀리 움직이도록.



→ 이 부분에서 매우 천천히 진행 ⇒ momentum을 활용해서 가속.

∘ stochastic gradient descent.

gradient descent가 converge 하기 위해서는 ~~true~~ gradient의 unbiased estimate(평균이 같은) 추정 gradient 필요

⇒ gradient의 간단한 approximation을 하여 추정 gradient를 활용.

◦ SGD in typical machine learning.

- standard gradient descent: 전체 data를 가지고 optimization.
- SGD(with minibatch) : N개의 data 포인트에서 S개만 가지고 계산.

$S\uparrow \Rightarrow$ Variance $\downarrow$, convergence 안정 $\uparrow$, 계산량 $\uparrow$.

Constrained optimization problem.

◦ Constrained optimization problem. :

$$\min_{x} f(x)$$

$$\text{subject to. } g_i(x) \leq 0 \quad i = 1 \sim M.$$

◦ Lagrange multipliers.

$$\text{Lagrangian} = \mathcal{E}(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) = f(x) + \lambda^T g(x).$$

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_M \end{bmatrix} \quad g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_m(x) \end{bmatrix}$$

↳ Lagrange multiplier for i-th constraint.

· Duality.

- original optimization problem – x (primal variable)
- another optimization problem – λ (dual variable)

⇒ x에 관한 optimization 문제를 λ에 대한 optimization 문제로 변환한다.

· Lagragian duality.

$$\left.\begin{aligned} &\min_{x} f(x) \\ &\text{subject to } g_i(x) \leq 0. \end{aligned}\right\} \Rightarrow \text{primal problem} \Longrightarrow \begin{aligned} &\max_{\lambda \in R^D} D(\lambda) \\ &\text{subject to } \lambda \geq 0. \end{aligned}$$

λ: dual variable, $D(\lambda) = \min_{x \in R^d} \mathcal{D}(x, \lambda)$

x에 대한 dependency를 없애고 λ에 대한 식이 됨.

- Weak duality.

minimax inequality : two argument function $\varphi(x, y)$

$$\max_{y} \min_{x} \varphi(x, y) \le \min_{x} \max_{y} \varphi(x, y).$$

weak duality: primal value $\ge$ dual value.

$$\min_{x \in \mathbb{R}^d} \max_{\lambda \ge 0} \mathcal{E}(x, \lambda) \ge \max_{\lambda \ge 0} \underbrace{\boxed{\min_{x \in \mathbb{R}^d} \mathcal{E}(x, \lambda)}}_{D(\lambda)} \Rightarrow \min_{x \in \mathbb{R}^d} J(x) \ge \max_{\lambda \ge 0} D(\lambda)$$

- Dual problem을 푸는 법.

① Inner problem : $\min_{x \in \mathbb{R}^d} \mathcal{E}(x, \lambda)$ $\quad D(\lambda)$

② outer problem : $\max_{\lambda \in \mathbb{R}^m} D(\lambda)$

subject to $\lambda \ge 0$.

- 항등식 constraint 를 다루는 법.

$$\min_{x} f(x)$$

subject to $g_i(x) \le 0$.

$h_j(x) = 0. \iff h_j(x) \le 0$ and $h_j(x) \ge 0. \iff h_j(x) \le 0$ and $-h_j(x) \le 0.$

$$\mathcal{E}(x, \lambda, v) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{n} v_j h_j(x)$$

$$= f(x) + \lambda^T g(x) + v^T h(x)$$

$$\sum_{j=1}^{n} \alpha_j h_j(x) + \sum_{j=1}^{n} \beta_j(-h_j(x))$$

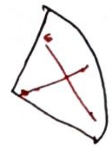
$$= \sum_{j=1}^{n} (\alpha_j - \beta_j) h_j(x) \quad (\alpha_j, \beta_j \ge 0.)$$

$$= \sum_{j=1}^{n} v_j h_j(x)$$

Lagrangian multiplier.

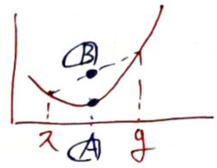
* convex optimization.

$C \rightarrow$ convex set $\quad a, b \in C. \Rightarrow \cancel{\theta(x)}\ \theta a + (1-\theta) b \in C.$

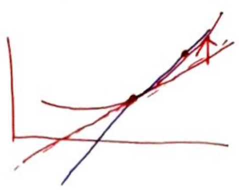


∘ convex function.

$$f(\underbrace{\theta x + (1-\theta) y}_{A}) \le \underbrace{\theta f(x) + (1-\theta) f(y)}_{B}$$



f: 미분가능, f: convex. $\Rightarrow f(y) \ge f(x) + \nabla_x f(x)^T (y - x)$



∘ convex optimization problem.

$$\min_x f(x)$$

subject to $g_i(x) \le 0$

$h_j(x) = 0.$

f: convex function, g, h를 만족하는 x가 convex set

$\Rightarrow$ convex optimization problem.

Dual problem의 solution $\xrightleftharpoons{\text{strong duality.}}$ primal problem의 solution

∘ Linear programming.

$$\min_{x \in \mathbb{R}^d} c^T x \rightarrow \text{linear}$$

subject to $Ax \le b. \rightarrow$ affine.

Dual optimization problem.

$$\max_{\lambda \in \mathbb{R}^m} -b^T \lambda$$

subject to $c + A^T \lambda = 0$

$\lambda \ge 0.$

Lagrangian $\mathcal{E}(x, \lambda) = c^T x + \lambda^T (Ax - b)$

$= (c + A^T \lambda)^T x - \lambda^T b \xRightarrow{\text{미분}} c + A^T \lambda = 0.$

$\Rightarrow$ Dual Lagrangian : $D(\lambda) = \min_{x \in \mathbb{R}^d} \mathcal{E}(x, \lambda) = -\lambda^T b.$

- Quadratic programming. (2차 함수 같은 형태).

$$\min_{x \in \mathbb{R}^d} \frac{1}{2} x^T Q x + c^T x$$

Q : positive definite $\mathbb{R}^{d \times d}$ → symmetric.

Subject to $Ax \le b$.

Lagrangian : $\mathcal{E}(x, \lambda) = \frac{1}{2} x^T Q x + c^T x + \lambda^T (Ax - b)$

$= \frac{1}{2} x^T Q x + (c + A^T \lambda)^T x - \lambda^T b.$ $\overset{미분}{\Rightarrow}$ $Qx + (c + A^T \lambda) = 0.$

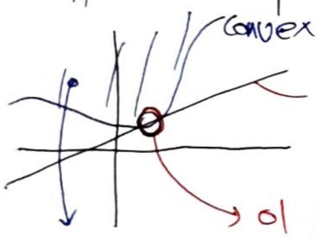
$$x = -Q^{-1}(c + A^T \lambda)$$

Dual optimization problem.

$$\max_{\lambda \in \mathbb{R}^m} -\frac{1}{2}(c + A^T \lambda)^T Q^{-1} (c + A^T \lambda) - \lambda^T b$$

Subject to $\lambda \ge 0$.

- Supporting hyperplane and gradient.



⇒ convex function ⟷ supporting hyperplane ⟷ gradient

Legendre transform : convex function을 gradient로 표현한것.

- Legendre transform. (convex conjugate)

$f : \mathbb{R}^p \to \mathbb{R}$ 의 convex conjugate : $f^*(s) = \sup(\langle s, x \rangle - f(x))$

f의 gradient.

Convex conjugate를 정의하면 원래 함수 f에 대한 optimization 문제를

f^*에 대한 optimization 문제로 변환가능 ⇒ 요컨대 dual problem 정의.